The pentafunctional *arom* enzyme of *Saccharomyces cerevisiae* is a mosaic of monofunctional domains

Kenneth DUNCAN,* R. Mark EDWARDS† and John R. COGGINS*‡
*Department of Biochemistry, University of Glasgow, Glasgow G12 8QQ, and †Searle Research and Development, Lane End Road, High Wycombe, Bucks. HP12 4HL, U.K.

The nucleotide sequence of the *Saccharomyces cerevisiae ARO1* gene which encodes the *arom* multifunctional enzyme has been determined. The protein sequence deduced for the pentafunctional *arom* polypeptide is 1588 amino acids in length and has a calculated M_r of 174555. Functional regions within the polypeptide chain have been identified by comparison with the sequences of the five monofunctional *Escherichia coli* enzymes whose activities correspond with those of the *arom* multifunctional enzyme. The observed homologies demonstrate that the *arom* polypeptide is a mosaic of functional domains and are consistent with the hypothesis that the *ARO1* gene evolved by the linking of ancestral *E. coli*-like genes.

INTRODUCTION

Multifunctional enzymes are found in all classes of organisms, including bacteria, higher plants and mammals, but they appear to be particularly common on the biosynthetic pathways of the lower eukaryotes (Kirschner & Bisswanger, 1976; Schmincke-Ott & Bisswanger, 1980; Hardie & Coggins, 1986). Most multifunctional enzymes catalyse two or more consecutive reactions on a biosynthetic pathway. In some cases the pathway intermediates are covalently bound to the enzyme, as in the case of the fatty acid synthases (Schweizer, 1986; Hardie & McCarthy, 1986), but more frequently the products of the individual reactions are free to diffuse away from the enzyme surface. The *arom* multifunctional enzyme (Lambert *et al.*, 1985; Coggins *et al.*, 1985; Coggins & Boocock, 1986) which catalyses the five central steps of the shikimate pathway (see Fig. 1) is an example of this latter type of multifunctional enzyme.

One remarkable feature of the *arom* system is the diversity in the patterns of gene and enzyme organization found in different species. Genetic and biochemical studies have revealed the presence of an '*arom* gene cluster' in *Neurospora crassa* (Giles *et al.*, 1967*a*; Catcheside *et al.*, 1985), *Aspergillus nidulans* (Ahmed & Giles, 1969; Charles *et al.*, 1986), *Saccharomyces cerevisiae* (de Leeuw, 1967; Larimer *et al.*, 1983), *Schizosaccharomyces pombe* (Strauss, 1979; Nakanishi & Yamamoto, 1984), a number of other fungal and yeast species (Ahmed & Giles, 1969; Böde & Birnbaum, 1981), and in *Euglena gracilis* (Berlyn *et al.*, 1970). In contrast, the corresponding structural genes for the five central enzymes of the shikimate pathway in *Escherichia coli*, *Salmonella typhimurium* and *Bacillus subtilis* are widely scattered about the genome (Bachmann, 1983; Sanderson & Roth, 1983; Henner & Hoch, 1980) and in the case of *E. coli* the five enzymes have also been shown to be separable (Berlyn & Giles, 1969; Chaudhuri & Coggins, 1985; Coggins *et al.*, 1985). In plants three of the enzymes of the pathway are separable but two, 3-dehydroquinase and shikimate dehydrogenase, copurify (Polley, 1978; Koshiba, 1979; Fiedler & Schultz, 1985; Coggins, 1986; Mousdale *et al.*, 1987) and have been shown to occur on a single bifunctional polypeptide chain (Polley, 1978; Fiedler & Schultz, 1985; Mousdale *et al.*, 1987).

These observations raise the question of what relationship there is between the prokaryotic monofunctional shikimate pathway enzymes and the multifunctional eukaryotic enzymes. On the basis of limited proteolysis experiments on the *N. crassa arom* multifunctional enzyme (Smith & Coggins, 1983; Coggins *et al.*, 1985; Coggins & Boocock, 1986) and from knowledge of its subunit molecular mass and the subunit molecular mass of the five corresponding *E. coli* enzymes we have proposed that the *arom* protein has a mosaic structure consisting of five autonomous, monofunctional domains each one homologous to the appropriate *E. coli* enzyme (Coggins *et al.*, 1985; Chaudhuri & Coggins, 1985; Coggins & Boocock, 1986; Hardie & Coggins, 1986).

To confirm this hypothesis we set out to determine the complete sequence of the *S. cerevisiae arom* protein and the five corresponding monofunctional. *E. coli* enzymes. While this work was in progress Hawkins and his co-workers reported the partial (Charles *et al.*, 1985) and later the complete (Charles *et al.*, 1986) sequence of the *A. nidulans arom* multifunctional enzyme and pointed out that it contained a region homologous to *E. coli* EPSP synthase (Charles *et al.*, 1986; Duncan *et al.*, 1984). Here we report the complete sequence of the *S. cerevisiae arom* multifunctional enzyme and compare it with the sequences, determined in our laboratory, of all five of the corresponding monofunctional *E. coli* enzymes (Duncan *et al.*, 1984, 1986; Millar *et al.*, 1986; Millar & Coggins, 1986; Anton & Coggins, 1987). The results confirm that the *arom* polypeptide is a 'mosaic' of five functional domains, each of which is homologous to a monofunctional *E. coli* polypeptide.

Abbreviation used: EPSP, 5-enoylpyruvylshikimate 3-phosphate,

‡ To whom correspondence and reprint requests should be addressed.

These sequence data have been submitted to the EMBL/GenBank Data Libraries under the accession number Y00313.

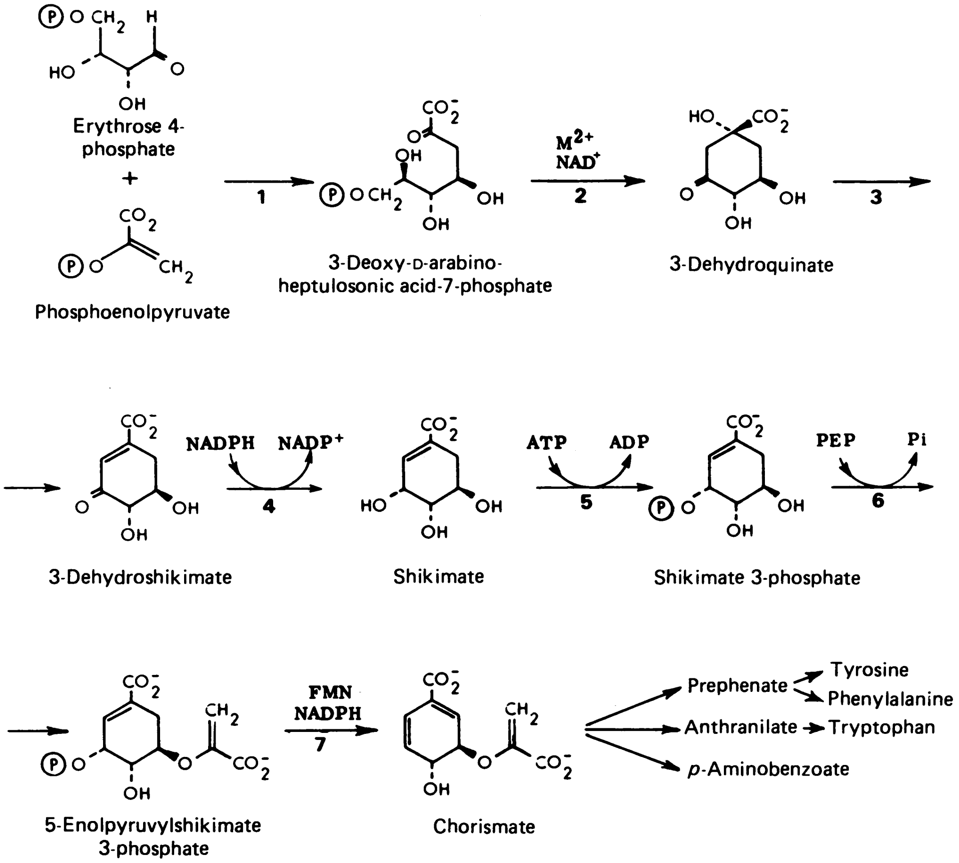

Fig. 1. Reactions of the early common pathway of aromatic amino acid biosynthesis

The numbers on the Figure refer to the enzymes of the pathway: (1) 3-deoxy-D-*arabino*-heptulosonic acid 7-phosphate synthase (EC 4.1.2.15), (2) 3-dehydroquinate synthase (EC 4.6.1.3), (3) 3-dehydroquinase (EC 4.2.1.10), (4) shikimate dehydrogenase (EC 1.1.1.25), (5) shikimate kinase (EC 2.7.1.71), (6) 5-enoylpyruvylshikimate 3-phosphate (EPSP) synthase (EC 2.5.1.19, alternative name 3-phosphoshikimate 1-carboxyvinyltransferase), (7) chorismate synthase (EC 4.6.1.4). The *arom* multifunctional enzyme catalyses reactions 2–6.

MATERIALS AND METHODS

Materials

Restriction enzymes were purchased from a number of commercial suppliers and were used in accordance with the manufacturers' instructions. T4 DNA ligase and *E. coli* DNA polymerase I were from Bethesda Research Laboratories, Paisley, U.K. All reagents for DNA sequencing, including α-[^{35}S]thioATP, were purchased from Amersham International, Amersham, Bucks., U.K.

Cloning and DNA sequence analysis

Plasmid preparations and manipulations were as described in Maniatis *et al.* (1982). DNA sequencing methods have been described previously (Duncan *et al.*, 1984). The paired vectors M13mp8 and M13mp9 (Messing & Vieira, 1982) or M13mp18 and M13mp19 (Norrander *et al.*, 1983) were used.

RESULTS AND DISCUSSION

Nucleotide sequence determination of the *ARO1* gene

The characterization of two independently isolated *S. cerevisiae ARO1* clones, pFL6 [a derivative of YpAR1 (Larimer *et al.*, 1983)] and pME173, has been described (Duncan *et al.*, 1987). These plasmids, which by restriction analysis have almost identical genomic inserts, are capable of complementing the auxotrophic lesions in a number of *E. coli* aromatic pathway mutant strains, namely *aroA*, *aroB*, *aroD* and *aroE* strains. The ability of a series of deletion derivatives of pME173 to complement the various aromatic pathway mutants lead directly to the location on the genomic insert of the 'sub-regions' within *ARO1*, and suggested the likely direction of transcription by comparison with the known order of the activities on the analogous *N. crassa* polypeptide (Duncan *et al.*, 1987).

The nucleotide sequence of pFL6 was determined, using the M13/dideoxy method (Sanger *et al.*, 1977; Messing & Vieira, 1982; Biggin *et al.*, 1983; Norrander *et al.*, 1983) and the sequence confirmed by sequence analysis of the *Sau*3A fragments between the *Kpn*I and *Hin*dIII sites of pME173. The sequencing strategy is outlined in Fig. 2. Both strands were sequenced in their entirety and all the restriction sites used to generate fragments for sequencing were overlapped by the sequenced fragments.

Translation of the DNA sequence revealed a single open reading frame which is sufficiently long to encode the *arom* polypeptide. This 1588 amino acid sequence (shown in Fig. 3) has a calculated M_r of 174555, which compares with an estimate by SDS/polyacrylamide-gel

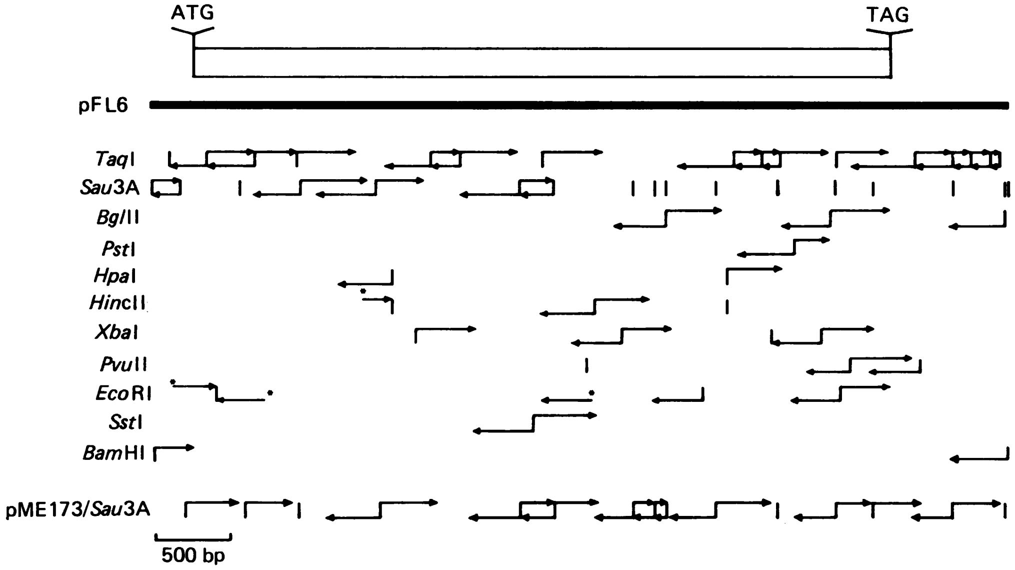

Fig. 2. Sequencing strategy for the yeast *ARO1* gene

Restriction sites used for sub-cloning into M13 prior to DNA sequence analysis are shown. Arrows indicate the direction and length of sequence obtained; *indicates clones resulting from digestion at *Eco*RI* sites, or blunt-end cloning of sheared DNA.

electrophoresis of 165000 for the corresponding enzyme isolated from *Neurospora crassa* (Lumsden & Coggins, 1977). Although direct evidence in support of the assignment of the *N*-terminal methionine is lacking, this particular ATG codon is favoured for two reasons. The next methionine in the open reading frame, Met-104, is within the region of the *arom* polypeptide which is homologous to the product of the *E. coli aroB* gene (see below). Also, it has been shown that the first AUG codon in eukaryotic mRNA usually serves as the translation initiation codon (Kozak, 1984); evidence has been obtained that this codon is the first AUG in the *ARO1* transcript (Duncan *et al.*, 1987).

Overall comparison of *S. cerevisiae ARO1* gene with the five corresponding *aro* genes of *E. coli* and the *arom* gene of *Aspergillus nidulans*

The subunit M_r of each of the individual *E. coli* enzyme is listed in Table 1. The yeast *arom* subunit M_r of 174555 corresponds closely to the M_r of 159698 for the combined *E. coli* enzymes. In order to confirm that the *arom* protein contained all five *E. coli* domains the predicted *ARO1* coding sequence was compared, in turn, with each of the individual *E. coli* protein sequences corresponding to the five activities of the complex. This was carried out using the programs 'BESTFIT' and 'GAP' (University of Wisconsin Genetics Computer Group Package of DNA sequence analysis programs; Devereux *et al.*, 1984). BESTFIT uses the 'local homology' algorithm of Smith & Waterman (1981) to find the best segments of similarity between two sequences. The alignments obtained are illustrated in Fig. 4. There are very clear homologies between the sequence of each *E. coli* enzyme and a region of corresponding length in the *S. cerevisiae* multifunctional enzyme. The order of the activities on the *arom* polypeptide chain is the same as that predicted for *N. crassa* (Giles *et al.*, 1967*a*), and for *S. pombe* (Strauss, 1979; Nakanishi & Yamamoto, 1984), and not as originally deduced for *S. cerevisiae* by de Leeuw (1967). The computer programs were also used to align the sequence of the *A. nidulans arom* polypeptide (Charles *et al.*, 1985, 1986) with the *S. cerevisiae* sequence (Fig. 5). The order of functional regions is the same for these two fungal multifunctional enzymes, which are more homologous to each other than they are to the five individual *E. coli* enzymes. The number of position identities and the percentage homologies for pair-wise combinations of the bacterial and fungal enzymes is shown in Table 2. The *S. cerevisiae*, *A. nidulans* and *E. coli* sequences are all clearly homologous although the exact degree of homology varies with the different domains as discussed below.

Specific homologies between the functional domains

The first 392 amino acid residues of the *S. cerevisiae arom* polypeptide are homologous with the *E. coli aroB* gene product, 3-dehydroquinate synthase (Millar & Coggins, 1986). In the alignment shown in Fig. 4 there is 36% identity between the two sequences. The distribution of the homology shows two very highly conserved sub-domains, consisting of residues 100–213 and 258–387 in the *S. cerevisiae* sequence. One of these sub-domains includes the $\beta\alpha\beta$ nucleotide-binding fold previously identified between residues 96 and 126 in the *E. coli* 3-dehydroquinate synthase sequence (Millar & Coggins, 1986). Linking these two conserved sub-domains there is a region of very low homology (residues 214–257 in the *S. cerevisiae* sesquence) which contains only one conserved residue and where, in the *S. cerevisiae* sequence, there is a 27 amino acid insertion. The *A. nidulans* polypeptide contains a shorter insertion (13 amino acids compared with the *E. coli* sequence) in this region (Fig. 5) which cannot be essential for enzyme activity.

A sequence of 11 amino acids (residues 393–403) connects the *aroB* region to a region homologous with the *E. coli aroA* gene product, EPSP synthase. The

```
                 .         .         .         .         .         .         .         .         .
    1  ATGGTGCAGTTAGCCAAAGTCCCAATTCTAGGAAATGATATTATCCACGTTGGGTATAACATTCATGACCATTTGGTTGAAACCATAATT    90
       MetValGlnLeuAlaLysValProIleLeuGlyAsnAspIleIleHisValGlyTyrAsnIleHisAspHisLeuValGluThrIleIle
                 .         .         .         .         .         .         .         .         .
   91  AAACATTGTCCTTCTTCGACATACGTTATTTGCAATGATACGAACTTGAGTAAAGTTCCATACTACCAGCAATTAGTCCTGGAATTCAAG   180
       LysHisCysProSerSerThrTyrValIleCysAsnAspThrAsnLeuSerLysValProTyrTyrGlnGlnLeuValLeuGluPheLys
                 .         .         .         .         .         .         .         .         .
  181  GCTTCTTTGCCAGAAGGCTCTCGTTTACTTACTTATGTTGTTAAACCAGGTGAGACAAGTAAAAGTAGAGAAACCAAAGCGCAGCTAGAA   270
       AlaSerLeuProGluGlySerArgLeuLeuThrTyrValValLysProGlyGluThrSerLysSerArgGluThrLysAlaGlnLeuGlu
                 .         .         .         .         .         .         .         .         .
  271  GATTATCTTTTAGTGGAAGGATGTACTCGTGATACGGTTATGGTAGCGATCGGTGGTGGTGTTATTGGTGACATGATTGGGTTCGTTGCA   360
       AspTyrLeuLeuValGluGlyCysThrArgAspThrValMetValAlaIleGlyGlyGlyValIleGlyAspMetIleGlyPheValAla
                 .         .         .         .         .         .         .         .         .
  361  TCTACATTTATGAGAGGTGTTCGTGTTGTCCAAGTACCAACATCCTTATTGGCAATGGTCGATTCCTCCATTGGTGGTAAAACTGCTATT   450
       SerThrPheMetArgGlyValArgValValGlnValProThrSerLeuLeuAlaMetValAspSerSerIleGlyGlyLysThrAlaIle
                 .         .         .         .         .         .         .         .         .
  451  GACACTCCTCTAGGTAAAAACTTTATTGGTGCATTTTGGCAACCAAAATTTGTCCTTGTAGATATTAAATGGCTAGAAACGTTAGCCAAG   540
       AspThrProLeuGlyLysAsnPheIleGlyAlaPheTrpGlnProLysPheValLeuValAspIleLysTrpLeuGluThrLeuAlaLys
                 .         .         .         .         .         .         .         .         .
  541  AGAGAGTTTATCAATGGGATGGCAGAAGTTATCAAGACTGCTTGTATTTGGAACGCTGACGAATTTACTAGATTAGAATCAAACGCTTCG   630
       ArgGluPheIleAsnGlyMetAlaGluValIleLysThrAlaCysIleTrpAsnAlaAspGluPheThrArgLeuGluSerAsnAlaSer
                 .         .         .         .         .         .         .         .         .
  631  TTGTTCTTAAATGTTGTTAATGGGGCAAAAAATGTCAAGGTTACCAATCAATTGACAAACGAGATTGACGAGATATCGAATACAGATATT   720
       LeuPheLeuAsnValValAsnGlyAlaLysAsnValLysValThrAsnGlnLeuThrAsnGluIleAspGluIleSerAsnThrAspIle
                 .         .         .         .         .         .         .         .         .
  721  GAAGCTATGTTGGATCATACATATAAGTTAGTTCTTGAGAGTATTAAGGTCAAAGCGGAAGTTGTCTCTTCGGATGAACGTGAATCCAGT   810
       GluAlaMetLeuAspHisThrTyrLysLeuValLeuGluSerIleLysValLysAlaGluValValSerSerAspGluArgGluSerSer
                 .         .         .         .         .         .         .         .         .
  811  CTAAGAAACCTTTTGAACTTCGGACATTCTATTGGTCATGCTTATGAAGCTATACTAACCCCACAAGCATTACATGGTGAATGTGTGTCC   900
       LeuArgAsnLeuLeuAsnPheGlyHisSerIleGlyHisAlaTyrGluAlaIleLeuThrProGlnAlaLeuHisGlyGluCysValSer
                 .         .         .         .         .         .         .         .         .
  901  ATTGGTATGGTTAAAGAGGCGGAATTATCCCGTTATTTCGGTATTCTCTCCCCTACCCAAGTTGCACGTCTATCCAAGATTTTGGTTGCC   990
       IleGlyMetValLysGluAlaGluLeuSerArgTyrPheGlyIleLeuSerProThrGlnValAlaArgLeuSerLysIleLeuValAla
                 .         .         .         .         .         .         .         .         .
  991  TACGGGTTGCCTGTTTCGCCTGATGAGAAATGGTTTAAAGAGCTAACCTTACATAAGAAAACACCATTGGATATCTTATTGAAGAAAATG  1080
       TyrGlyLeuProValSerProAspGluLysTrpPheLysGluLeuThrLeuHisLysLysThrProLeuAspIleLeuLeuLysLysMet
                 .         .         .         .         .         .         .         .         .
 1081  AGTATTGACAAGAAAAACGAGGGTTCCAAAAAGAAGGTGGTCATTTTAGAAAGTATTGGTAAGTGCTATGGTGACTCCGCTCAATTTGTT  1170
       SerIleAspLysLysAsnGluGlySerLysLysLysValValIleLeuGluSerIleGlyLysCysTyrGlyAspSerAlaGlnPheVal
                 .         .         .         .         .         .         .         .         .
 1171  AGCGATGAAGACCTGAGATTTATTCTAACAGATGAAACCCTCGTTTACCCCTTCAAGGACATCCCTGCTGATCAACAGAAAGTTGTTATC  1260
       SerAspGluAspLeuArgPheIleLeuThrAspGluThrLeuValTyrProPheLysAspIleProAlaAspGlnGlnLysValValIle
                 .         .         .         .         .         .         .         .         .
 1261  CCCCCTGGTTCTAAGTCCATCTCCAATCGTGCTTTAATTCTTGCTGCCCTCGGTGAAGGTCAATGTAAAATCAAGAACTTATTACATTCT  1350
       ProProGlySerLysSerIleSerAsnArgAlaLeuIleLeuAlaAlaLeuGlyGluGlyGlnCysLysIleLysAsnLeuLeuHisSer
                 .         .         .         .         .         .         .         .         .
 1351  GATGATACTAAACATATGTTAACCGCTGTTCATGAATTGAAAGGTGCTACGATATCATGGGAAGATAATGGTGAGACGGTAGTGGTGGAA  1440
       AspAspThrLysHisMetLeuThrAlaValHisGluLeuLysGlyAlaThrIleSerTrpGluAspAsnGlyGluThrValValValGlu
                 .         .         .         .         .         .         .         .         .
 1441  GGACATGGTGGTTCCACATTGTCAGCTTGTGCTGACCCCTTATATCTAGGTAATGCAGGTACTGCATCTAGATTTTTGACTTCCTTGGCT  1530
       GlyHisGlyGlySerThrLeuSerAlaCysAlaAspProLeuTyrLeuGlyAsnAlaGlyThrAlaSerArgPheLeuThrSerLeuAla
                 .         .         .         .         .         .         .         .         .
 1531  GCCTTGGTCAATTCTACTTCAAGCCAAAAGTATATCGTTTTAACTGGTAACGCAAGAATGCAACAAAGACCAATTGCTCCTTTGGTCGAT  1620
       AlaLeuValAsnSerThrSerSerGlnLysTyrIleValLeuThrGlyAsnAlaArgMetGlnGlnArgProIleAlaProLeuValAsp
                 .         .         .         .         .         .         .         .         .
 1621  TCTTTGCGTGCTAATGGTACTAAAATTGAGTACTTGAATAATGAAGGTTCCCTGCCAATCAAAGTTTATACTGATTCGGTATTCAAAGGT  1710
       SerLeuArgAlaAsnGlyThrLysIleGluTyrLeuAsnAsnGluGlySerLeuProIleLysValTyrThrAspSerValPheLysGly
                 .         .         .         .         .         .         .         .         .
 1711  GGTAGAATTGAATTAGCTGCTACAGTTTCTTCTCAGTACGTATCCTCTATCTTGATGTGTGCCCCATACGCTGAAGAACCTGTAACTTTG  1800
       GlyArgIleGluLeuAlaAlaThrValSerSerGlnTyrValSerSerIleLeuMetCysAlaProTyrAlaGluGluProValThrLeu
                 .         .         .         .         .         .         .         .         .
 1801  GCTCTTGTTGGTGGTAAGCCAATCTCTAAATTGTACGTCGATATGACAATAAAAATGATGGAAAAATTCGGTATCAATGTTGAAACTTCT  1890
       AlaLeuValGlyGlyLysProIleSerLysLeuTyrValAspMetThrIleLysMetMetGluLysPheGlyIleAsnValGluThrSer
                 .         .         .         .         .         .         .         .         .
 1891  ACTACAGAACCTTACACTTATTATATTCCAAAGGGACATTATATTAACCCATCAGAATACGTCATTGAAAGTGATGCCTCAAGTGCTACA  1980
       ThrThrGluProTyrThrTyrTyrIleProLysGlyHisTyrIleAsnProSerGluTyrValIleGluSerAspAlaSerSerAlaThr
                 .         .         .         .         .         .         .         .         .
 1981  TACCCATTGGCCTTCGCCGCAATGACTGGTACTACCGTAACGGTTCCAAACATTGGTTTTGAGTCGTTACAAGGTGATGCCAGATTTGCA  2070
       TyrProLeuAlaPheAlaAlaMetThrGlyThrThrValThrValProAsnIleGlyPheGluSerLeuGlnGlyAspAlaArgPheAla
                 .         .         .         .         .         .         .         .         .
 2071  AGAGATGTCTTGAAACCTATGGGTTGTAAAATAACTCAAACGGCAACTTCAACTACTGTTTCGGGTCCTCCTGTAGGTACTTTAAAGCCA  2160
       ArgAspValLeuLysProMetGlyCysLysIleThrGlnThrAlaThrSerThrThrValSerGlyProProValGlyThrLeuLysPro
                 .         .         .         .         .         .         .         .         .
 2161  TTAAAACATGTTGATATGGAGCCAATGACTGATGCGTTCTTAACTGCATGTGTTGTTGCCGCTATTTCGCACGACAGTGATCCAAATTCT  2250
       LeuLysHisValAspMetGluProMetThrAspAlaPheLeuThrAlaCysValValAlaAlaIleSerHisAspSerAspProAsnSer
                 .         .         .         .         .         .         .         .         .
 2251  GCAAATACAACCACCATTGAAGGTATTGCAAACCAGCGTGTCAAAGAGTGTAACAGAATTTTGGCCATGGCTACAGAGCTCGCCAAATTT  2340
       AlaAsnThrThrThrIleGluGlyIleAlaAsnGlnArgValLysGluCysAsnArgIleLeuAlaMetAlaThrGluLeuAlaLysPhe
                 .         .         .         .         .         .         .         .         .
 2341  GGCGTCAAAACTACAGAATTACCAGATGGTATTCAAGTCCATGGTTTAAACTCGATAAAAGATTTGAAGGTTCCTTCCGACTCTTCTGGA  2430
       GlyValLysThrThrGluLeuProAspGlyIleGlnValHisGlyLeuAsnSerIleLysAspLeuLysValProSerAspSerSerGly
```

```
                  .         .         .         .         .         .         .         .         .
2431     CCTGTCGGTGTATGCACATATGATGATCATCGTGTGGCCATGAGTTTCTCGCTTCTTGCAGGAATGGTAAATTCTCAAAATGAACGTGAC     2520
         ProValGlyValCysThrTyrAspAspHisArgValAlaMetSerPheSerLeuLeuAlaGlyMetValAsnSerGlnAsnGluArgAsp
                  .         .         .         .         .         .         .         .         .
2521     GAAGTTGCTAATCCTGTAAGAATACTTGAAAGACATTGTACTGGTAAAACCTGGCCTGGCTGGTGGGATGTGTTACATTCCGAACTAGGT     2610
         GluValAlaAsnProValArgIleLeuGluArgHisCysThrGlyLysThrTrpProGlyTrpTrpAspValLeuHisSerGluLeuGly
                  .         .         .         .         .         .         .         .         .
2611     GCCAAATTAGATGGTGCAGAACCTTTAGAGTGCACATCCAAAAAGAACTCAAAGAAAAGCGTTGTCATTATTGGCATGAGAGCAGCTGGC     2700
         AlaLysLeuAspGlyAlaGluProLeuGluCysThrSerLysLysAsnSerLysLysSerValValIleIleGlyMetArgAlaAlaGly
                  .         .         .         .         .         .         .         .         .
2701     AAAACTACTATAAGTAAATGGTGCGCATCCGCTCTGGGTTACAAATTAGTTGACCTAGACGAGCTGTTTGAGCAACAGCATAACAATCAA     2790
         LysThrThrIleSerLysTrpCysAlaSerAlaLeuGlyTyrLysLeuValAspLeuAspGluLeuPheGluGlnGlnHisAsnAsnGln
                  .         .         .         .         .         .         .         .         .
2791     AGTGTTAAACAATTTGTTGTGGAGAACGGTTGGGAGAAGTTCCGTGAGGAAGAAACAAGAATTTTCAAGGAAGTTATTCAAAATTACGGC     2880
         SerValLysGlnPheValValGluAsnGlyTrpGluLysPheArgGluGluGluThrArgIlePheLysGluValIleGlnAsnTyrGly
                  .         .         .         .         .         .         .         .         .
2881     GATGATGGATATGTTTTCTCAACAGGTGGCGGTATTGTTGAAAGCGCTGAGTCTAGAAAAGCCTTAAAAGATTTTGCCTCATCAGGTGGA     2970
         AspAspGlyTyrValPheSerThrGlyGlyGlyIleValGluSerAlaGluSerArgLysAlaLeuLysAspPheAlaSerSerGlyGly
                  .         .         .         .         .         .         .         .         .
2971     TACGTTTTACACTTACATAGGGATATTGAGGAGACAATTGTCTTTTTACAAAGTGATCCTTCAAGACCTGCCTATGTGGAAGAAATTCGT     3060
         TyrValLeuHisLeuHisArgAspIleGluGluThrIleValPheLeuGlnSerAspProSerArgProAlaTyrValGluGluIleArg
                  .         .         .         .         .         .         .         .         .
3061     GAAGTTTGGAACAGAAGGGAGGGGTGGTATAAAGAATGCTCAAATTTCTCTTTCTTTGCTCCTCATTGCTCCGCAGAAGCTGAGTTCCAA     3150
         GluValTrpAsnArgArgGluGlyTrpTyrLysGluCysSerAsnPheSerPhePheAlaProHisCysSerAlaGluAlaGluPheGln
                  .         .         .         .         .         .         .         .         .
3151     GCTCTAAGAAGATCGTTTAGTAAGTACATTGCAACCATTACAGGTGTCAGAGAAATAGAAATTCCAAGCGGAAGATCTGCCTTTGTGTGT     3240
         AlaLeuArgArgSerPheSerLysTyrIleAlaThrIleThrGlyValArgGluIleGluIleProSerGlyArgSerAlaPheValCys
                  .         .         .         .         .         .         .         .         .
3241     TTAACCTTTGATGACTTAACTGAACAAACTGAGAATTTGACTCCAATCTGTTATGGTTGTGAGGCTGTAGAGGTCAGAGTAGACCATTTG     3330
         LeuThrPheAspAspLeuThrGluGlnThrGluAsnLeuThrProIleCysTyrGlyCysGluAlaValGluValArgValAspHisLeu
                  .         .         .         .         .         .         .         .         .
3331     GCTAATTACTCTGCTGATTTCGTGAGTAAACAGTTATCTATATTGCGTAAAGCCACTGACAGTATTCCTATCATTTTTACTGTGCGAACC     3420
         AlaAsnTyrSerAlaAspPheValSerLysGlnLeuSerIleLeuArgLysAlaThrAspSerIleProIleIlePheThrValArgThr
                  .         .         .         .         .         .         .         .         .
3421     ATGAAGCAAGGTGGCAACTTTCCTGATGAAGAGTTCAAAACCTTGAGAGAGCTATACGATATTGCCTTGAAGAATGGTGTTGAATTCCTT     3510
         MetLysGlnGlyGlyAsnPheProAspGluGluPheLysThrLeuArgGluLeuTyrAspIleAlaLeuLysAsnGlyValGluPheLeu
                  .         .         .         .         .         .         .         .         .
3511     GACTTAGAACTAACTTTACCTACTGATATCCAATATGAGGTTATTAACAAAAGGGGCAACACCAAGATCATTGGTTCCCATCATGACTTC     3600
         AspLeuGluLeuThrLeuProThrAspIleGlnTyrGluValIleAsnLysArgGlyAsnThrLysIleIleGlySerHisHisAspPhe
                  .         .         .         .         .         .         .         .         .
3601     CAAGGATTATACTCCTGGGACGACGCTGAATGGGAAAACAGATTCAATCAAGCGTTAACTCTTGATGTGGATGTTGTAAAATTTGTGGGT     3690
         GlnGlyLeuTyrSerTrpAspAspAlaGluTrpGluAsnArgPheAsnGlnAlaLeuThrLeuAspValAspValValLysPheValGly
                  .         .         .         .         .         .         .         .         .
3691     ACGGCTGTTAATTTCGAAGATAATTTGAGACTGGAACACTTTAGGGATACACACAAGAATAAGCCTTTAATTGCAGTTAATATGACTTCT     3780
         ThrAlaValAsnPheGluAspAsnLeuArgLeuGluHisPheArgAspThrHisLysAsnLysProLeuIleAlaValAsnMetThrSer
                  .         .         .         .         .         .         .         .         .
3781     AAAGGTAGCATTTCTCGTGTTTTGAATAATGTTTTAACACCTGTGACATCAGATTTATTGCCTAACTCCGCTGCCCCTGGCCAATTGACA     3870
         LysGlySerIleSerArgValLeuAsnAsnValLeuThrProValThrSerAspLeuLeuProAsnSerAlaAlaProGlyGlnLeuThr
                  .         .         .         .         .         .         .         .         .
3871     GTAGCACAAATTAACAAGATGTATACATCTATGGGAGGTATCGAGCCTAAGGAACTGTTTGTTGTTGGAAAGCCAATTGGCCACTCTAGA     3960
         ValAlaGlnIleAsnLysMetTyrThrSerMetGlyGlyIleGluProLysGluLeuPheValValGlyLysProIleGlyHisSerArg
                  .         .         .         .         .         .         .         .         .
3961     TCGCCAATTTTACATAACACTGGCTATGAAATTTTAGGTTTACCTCACAAGTTCGATAAATTTGAAACTGAATCCGCACAATTGGTGAAA     4050
         SerProIleLeuHisAsnThrGlyTyrGluIleLeuGlyLeuProHisLysPheAspLysPheGluThrGluSerAlaGlnLeuValLys
                  .         .         .         .         .         .         .         .         .
4051     GAAAAACTTTTGGACGGAAACAAGAACTTTGGCGGTGCTGCAGTCACAATTCCTCTGAAATTAGATATAATGCAGTACATGGATGAATTG     4140
         GluLysLeuLeuAspGlyAsnLysAsnPheGlyGlyAlaAlaValThrIleProLeuLysLeuAspIleMetGlnTyrMetAspGluLeu
                  .         .         .         .         .         .         .         .         .
4141     ACTGATGCTGCTAAAGTTATTGGTGCTGTAAACACAGTTATACCATTGGGTAACAAGAAGTTTAAGGGTGATAATACCGACTGGTTAGGT     4230
         ThrAspAlaAlaLysValIleGlyAlaValAsnThrValIleProLeuGlyAsnLysLysPheLysGlyAspAsnThrAspTrpLeuGly
                  .         .         .         .         .         .         .         .         .
4231     ATCCGTAATGCCTTAATTAACAATGGCGTTCCCGAATATGTTGGTCATACCGCTGGTTTGGTTATCGGTGCAGGTGGCACTTCTAGAGCC     4320
         IleArgAsnAlaLeuIleAsnAsnGlyValProGluTyrValGlyHisThrAlaGlyLeuValIleGlyAlaGlyGlyThrSerArgAla
                  .         .         .         .         .         .         .         .         .
4321     GCCCTTTACGCCTTGCACAGTTTAGGTTGCAAAAAGATCTTCATAATCAACAGGACAACTTCGAAATTGAAGCCATTAATAGAGTCACTT     4410
         AlaLeuTyrAlaLeuHisSerLeuGlyCysLysLysIlePheIleIleAsnArgThrThrSerLysLeuLysProLeuIleGluSerLeu
                  .         .         .         .         .         .         .         .         .
4411     CCATCTGAATTCAACATTATTGGAATAGAGTCCACTAAATCTATAGAAGAGATTAAGGAACACGTTGGCGTTGCTGTCAGCTGTGTACCA     4500
         ProSerGluPheAsnIleIleGlyIleGluSerThrLysSerIleGluGluIleLysGluHisValGlyValAlaValSerCysValPro
                  .         .         .         .         .         .         .         .         .
4501     GCCGACAAACCATTAGATGACGAACTTTTAAGTAAGCTGGAGAGATTCCTTGTGAAAGGTGCCCATGCTGCTTTTGTACCAACCTTATTG     4590
         AlaAspLysProLeuAspAspGluLeuLeuSerLysLeuGluArgPheLeuValLysGlyAlaHisAlaAlaPheValProThrLeuLeu
                  .         .         .         .         .         .         .         .         .
4591     GAAGCCGCATACAAACCAAGCGTTACTCCCGTTATGACAATTTCACAAGACAAATATCAATGGCACGTTGTCCCTGGATCACAAATGTTA     4680
         GluAlaAlaTyrLysProSerValThrProValMetThrIleSerGlnAspLysTyrGlnTrpHisValValProGlySerGlnMetLeu
                  .         .         .         .         .         .         .         .
4681     GTACACCAAGGTGTAGCTCAGTTTGAAAAGTGGACAGGATTCAAGGGCCCTTTCAAGGCCATTTTTGATGCCGTTACGAAAGAGTAG   4767
         ValHisGlnGlyValAlaGlnPheGluLysTrpThrGlyPheLysGlyProPheLysAlaIlePheAspAlaValThrLysGluEnd
```

Fig. 3. DNA sequence of the *ARO1* coding region, and the corresponding *arom* protein sequence

Table 1. Structure of the five *E. coli* enzymes which correspond to the *S. cerevisiae arom* activities

Note that the length of shikimate kinase is reported here as 173 amino acids and is 174 amino acids in the text. The *N*-terminal methionine is cleaved post-translationally (Millar *et al.*, 1986).

Pathway step	Enzyme activity	*E. coli* gene	Calculated M_r	Length (amino acids)	Quaternary structure
2	3-Dehydroquinate synthase	*aroB*	38880	362	Monomer
3	3-Dehydroquinase	*aroD*	26377	240	Dimer
4	Shikimate dehydrogenase	*aroE*	29380	272	Monomer
5	Shikimate kinase	*aroL*	18937	173	Monomer
6	EPSP synthase	*aroA*	46112	427	Monomer
	Total		159689	1475	

S. cerevisiae EPSP synthase domain is located between amino acids 404 and 866. Of the five functional domains of the *arom* multifunctional enzyme this is the best conserved, with 38% identity between the *E. coli* and *S. cerevisiae* sequences and 55% identity between the fungal sequences (Table 2). As with the 3-dehydroquinate synthase domain there are two very well conserved sub-domains separated by a region with no homology. In the *S. cerevisiae* sequence this unconserved region of 51 residues (Ile-701 to Thr-753), like the unconserved region separating the two 3-dehydroquinate synthase subdomains, contains an 11-residue insertion compared with the *E. coli* sequence. This two sub-domain pattern is illustrated in Fig. 6, which shows the alignment of two bacterial and two fungal EPSP synthase sequences.

Much attention has been focused recently on EPSP synthase since the discovery that the commercially important herbicide glyphosate (*N*-phosphonomethylglycine) acts on plants by inhibiting this enzyme (Amrhein *et al.*, 1980; Mousdale & Coggins, 1984). Glyphosate is also a potent inhibitor of the *N. crassa* and *E. coli* enzymes (Boocock & Coggins, 1983; Lewendon & Coggins, 1983). A glyphosate-insensitive form of EPSP synthase has been isolated from a *Salmonella typhimurium* mutant resistant to glyphosate (Comai *et al.*, 1983) and it has been shown that the only alteration in the enzyme structure is a Pro to Ser change at position 101 in the enzyme sequence (Stalker *et al.*, 1985). Pro-101 is conserved between *E. coli* and *S. typhimurium* (Fig. 6), but in both the fungal sequences it is replaced by Phe (position 505 in the *S. cerevisiae* sequence). This position follows a highly conserved region in the bacterial and fungal sequences and precedes a less well conserved sequence which includes a 5-amino-acid insertion in both fungal sequences (Fig. 6). The absence of a conserved Pro at position 505 in the fungal sequences indicates that this residue cannot be an essential feature of glyphosate-sensitive forms of the enzyme.

It has been proposed that the mechanism of EPSP synthase involves a cysteine residue at the active site (Ganem, 1978). The greater than 98% inactivation of the multifunctional *N. crassa* EPSP synthase by *N*-ethylmaleimide and the protection against inactivation by this reagent observed in the presence of shikimate 3-phosphate and glyphosate are consistent with this suggestion (M. R. Boocock & J. R. Coggins, unpublished work), but it should be noted that cysteine-directed reagents do not completely inactivate the monofunctional *E. coli* (Lewendon, 1984) and *Aerobacter aerogenes* (Steinrucken & Amrhein, 1984) enzymes. This implies that an important cysteine residue is near to but not necessarily at the active site of the enzyme. There is a single cysteine which is conserved in all four EPSP synthase (Cys-853 in the *S. cerevisiae* sequence, see Fig. 6); further experiments are required to establish the precise functional role of this residue.

The EPSP synthase region is linked to a region homologous to the *E. coli aroL* gene product, shikimate kinase II, by a 20-amino-acid sequence (residues 867–886 in the *S. cerevisiae* sequence). Homology with *E. coli* shikimate kinase II (Millar *et al.*, 1986) extends to residue 1059. The homology found in this region, which is 23% for the *E. coli* versus fungal sequences and 40% for the two fungal species, is lower than that found in the EPSP synthase region. Although the overall degree of homology between the yeast and *E. coli* shikimate kinase sequences is rather low, there is one well conserved region which has sequence homology with the 'A' sequence of the ATP-binding site of phosphofructokinase and adenylate kinase (Walker *et al.*, 1982). This 'A' sequence, G-X_4-G-K-(T)-X_6-I/V, occurs between residues 895 and 909 in the *S. cerevisiae* sequence (corresponding to *E. coli* shikimate kinase residues 9–23); the final residue, conserved between these two species, is an alanine rather than the usual isoleucine or valine. Comparing the *S. cerevisiae* and *A. nidulans* shikimate kinase domains, the sequences around this 'A' region of the ATP-binding site are very homologous (9/11 matches). However, the *A. nidulans* enzyme does not have the GKT motif; instead it was GKS. Although not unique in this respect (Midgeley & Murray, 1985), this feature is unusual in that the GKT is highly conserved over a wide species range and over a wide range of different ATP-utilizing enzymes, and it is conserved between the *S. cerevisiae* and *E. coli* shikimate kinases (Millar *et al.*, 1986).

Following the shikimate kinase region is a region which shows homology to the *E. coli aroD* gene product, 3-dehydroquinase (Duncan *et al.*, 1986). In the alignment shown in Fig. 4 the *N*-terminal amino acid of *E. coli* 3-dehydroquinase overlaps with the *C*-terminal amino acid of *E. coli* shikimate kinase. The percentage

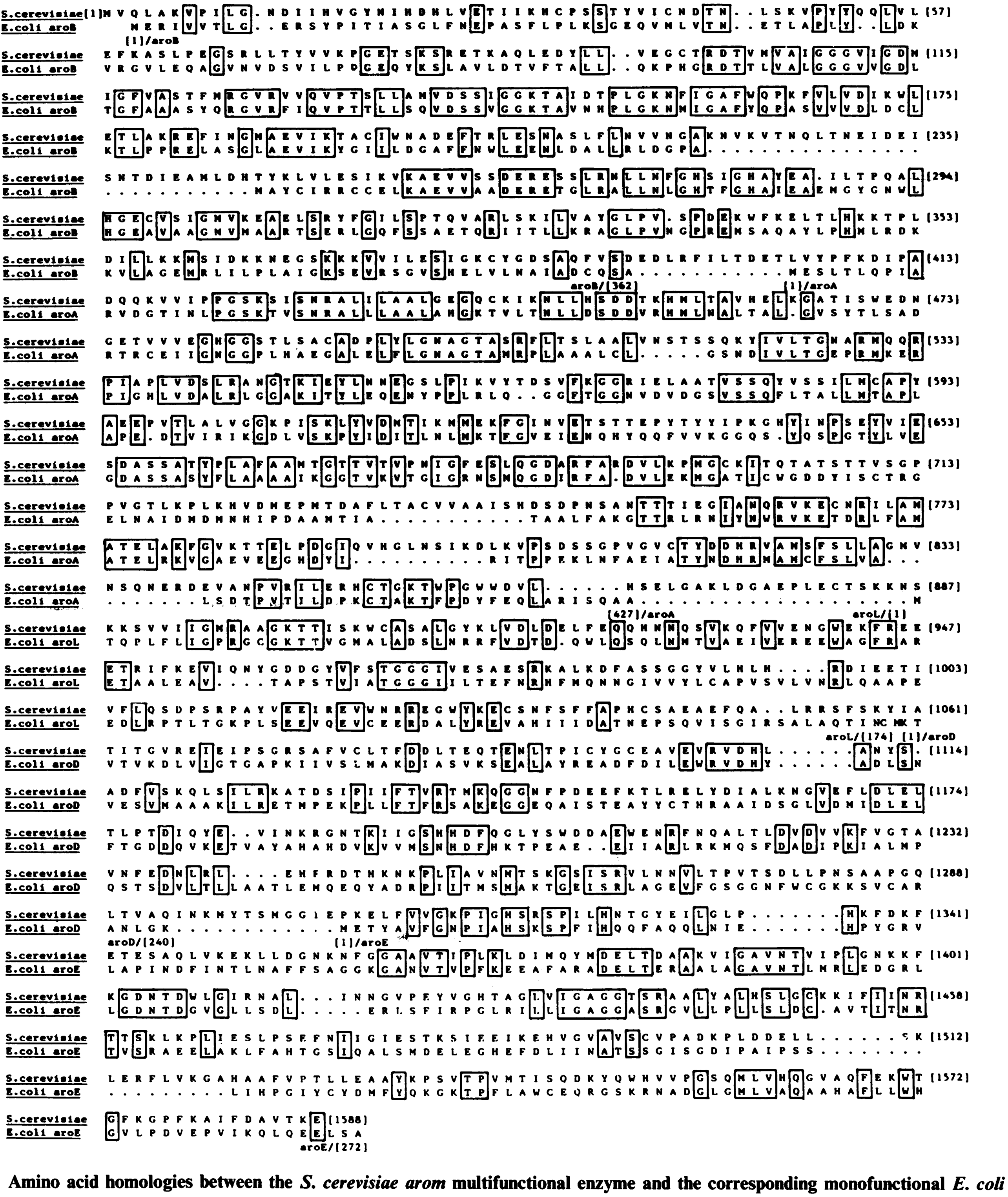

Fig. 4. Amino acid homologies between the *S. cerevisiae arom* multifunctional enzyme and the corresponding monofunctional *E. coli* enzymes

Key: *aroB*, dehydroquinate synthase; *aroA*, EPSP synthase; *aroL*, shikimate kinase; *aroD*, 3-dehydroquinase; *aroE*, shikimate dehydrogenase. Numbers above and below the sequences indicate amino acid positions in the *S. cerevisiae* enzyme and in the individual *E. coli* enzymes, respectively. Gaps in both sequences maintain the format with Fig. 5.

homology among the three species is similar to that found in the shikimate kinase domain (Table 2). Confirmation that this region of the *S. cerevisiae* sequence truly encodes the 3-dehydroquinase activity was provided by the observation that there is homology with a pentadecapeptide isolated from the 3-dehydroquinase active of the *N. crassa arom* polypeptide (S. Chaudhuri & J. R. Coggins, unpublished work). This peptide had been radiolabelled, by treatment with 3-dehydroquinate and NaB^3H_4, on the lysine

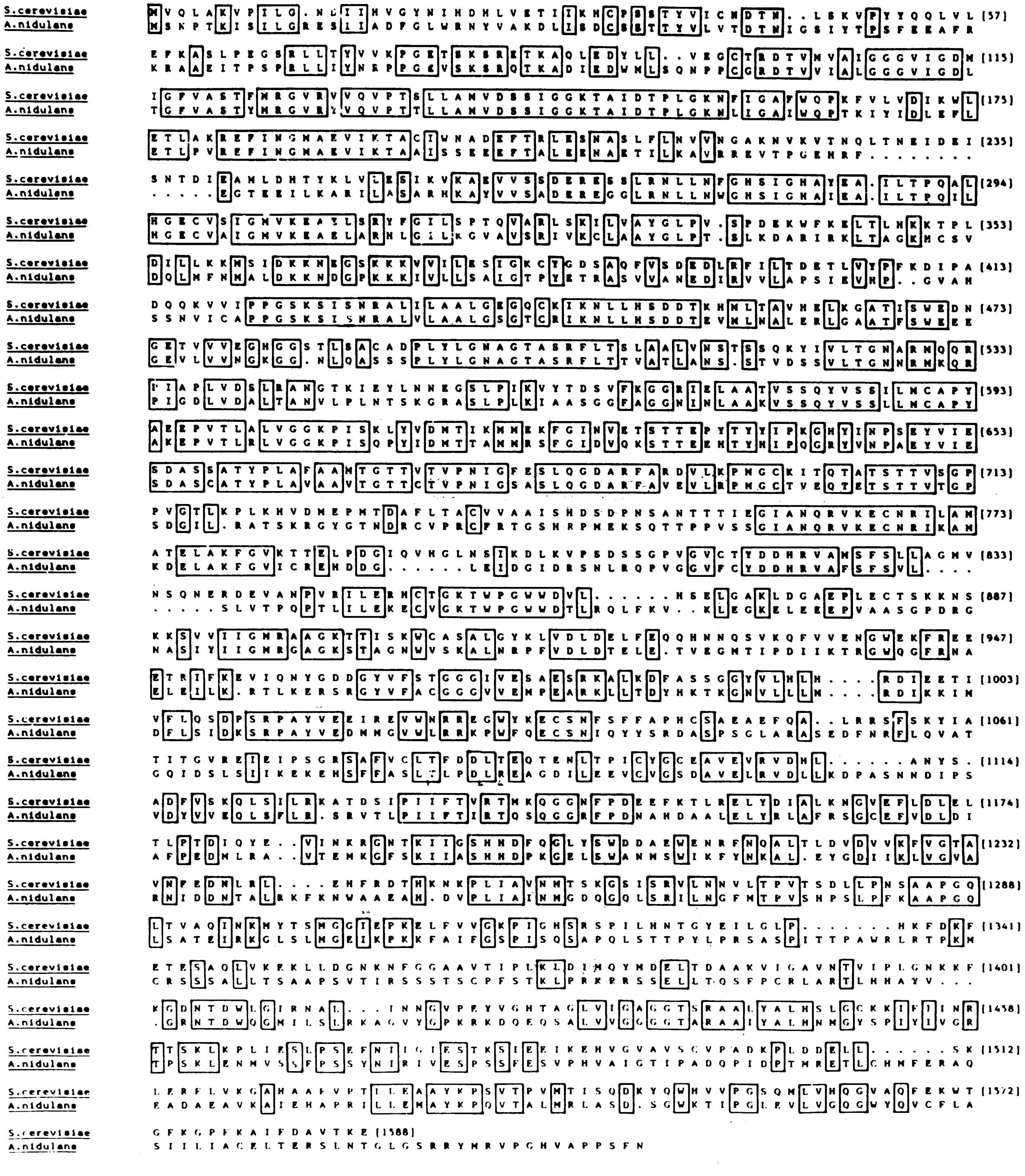

Fig. 5. Amino acid homologies between the *S. cerevisiae* and the *A. nidulans arom* multifunctional enzymes

Numbers indicate amino acid positions in the sequences. Again, gaps in both sequences maintain the format from Fig. 4.

residue which is known to form a imine intermediate during the enzyme-catalysed reaction (Chaudhuri *et al.*, 1986). This homology predicts that Lys-1227 in the *S. cerevisiae arom* polypeptide is the residue involved in imine formation in the 3-dehydroquinase reaction. The alignment places Lys-170 at the active site of the *E. coli* enzyme; this prediction has also been confirmed by the isolation of an active site peptide (K. Duncan & J. R. Coggins, unpublished work). The mechanism for the action of 3-dehydroquinase proposed by Walsh (1979) requires a basic group for proton abstraction. Chaudhuri *et al.* (1986) have provided evidence that, for both the *E. coli* and *N. crassa* enzymes, this group is the imidazole side chain of a histidine residue. Two histidine residues are conserved between the *E. coli* and *S. cerevisiae* sequences, but only one of these (His-1198) is also conserved in the 3-dehydroquinase domain of the *A. nidulans arom* polypeptide (Charles *et al.*, 1985, 1986) (Fig. 5). It is therefore reasonable to propose that this is the active site histidine residue.

Table 2. Summary of the homologies found between the five *E. coli* monofunctional enzymes and the *S. cerevisiae* and *A. nidulans* multifunctional enzymes

E. coli enzyme	Polypeptide chain length (amino acid) residues)	Number of residues conserved between the *E. coli* monofunctional enzymes and each *arom* multifunctional enzyme		Number of residues conserved in the two multifunctional enzymes and in the corresponding *E. coli* monofunctional enzyme	Number of residues conserved in the functional domains of the two *arom* polypeptide chains
		S. cerevisiae	*A. nidulans*		
3-Dehydroquinate synthase	362	130 (36%)	128 (36%)	96 (27%)	201 (51%)
EPSP synthase	427	162 (38%)	146 (34%)	127 (30%)	254 (55%)
Shikimate kinase	174	39 (23%)	35 (20%)	22 (13%)	68 (40%)
3-Dehydroquinase	240	50 (21%)	41 (17%)	30 (13%)	97 (42%)
Shikimate dehydrogenase	272	68 (25%)	41 (15%)	29 (11%)	75 (27%)

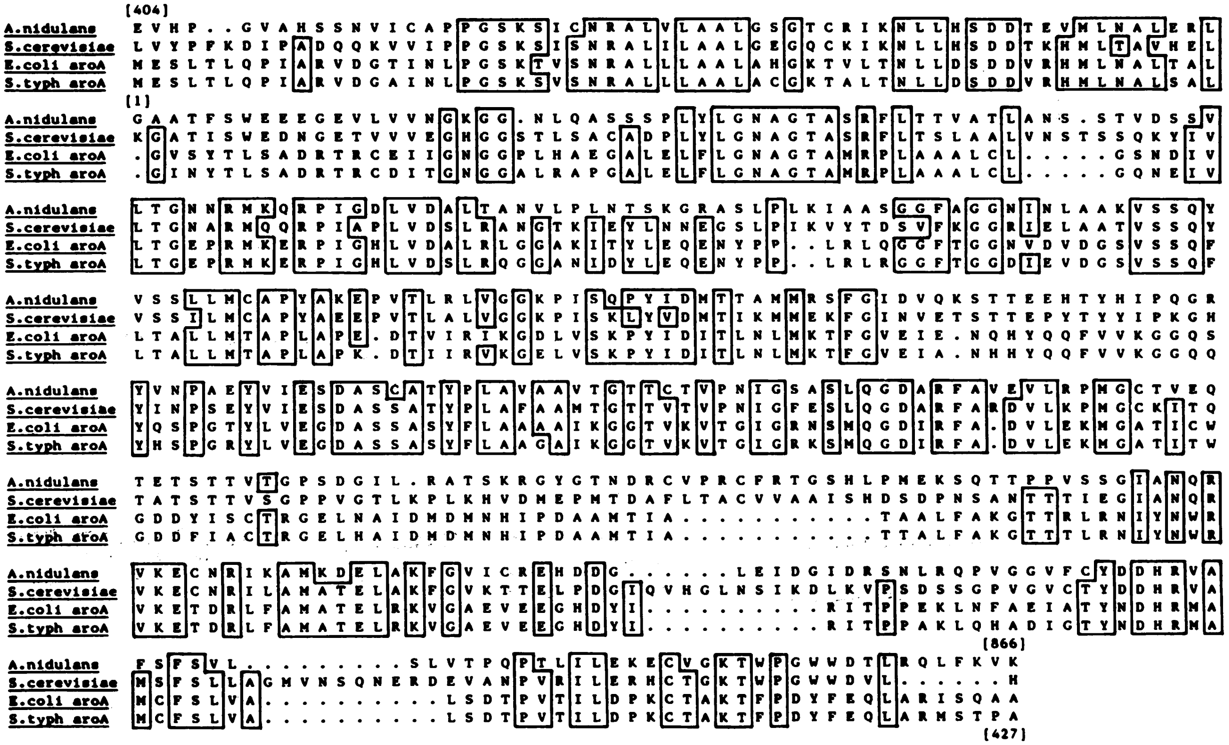
[404]
A.nidulans EVHP..GVAHSSNVICAPPGSKSICNRALVLAALGSGTCRIKNLLHSDDTEVMLNALERL
S.cerevisiae LVYPFKDIPADQQKVVIPPGSKSISNRALILAALGEGQCKIKNLLHSDDTKHMLTAVHEL
E.coli aroA MESLTLQPIARVDGTINLPGSKTVSNRALLLAALAHGKTVLTNLLDSDDVRHMLNALTAL
S.typh aroA MESLTLQPIARVDGAINLPGSKSVSNRALLLAALACGKTALTNLLDSDDVRHMLNALSAL
[1]
A.nidulans GAATFSWEEEGEVLVVNGKGG.NLQASSSPLYLGNAGTASRFLTTVATLANS.STVDSSV
S.cerevisiae KGATISWEDNGETVVVEGHGGSTLSACADPLYLGNAGTASRFLTSLAALVNSTSSQKYIV
E.coli aroA .GVSYTLSADRTRCEIIGNGGPLHAEGALELFLGNAGTAMRPLAAALCL.....GSNDIV
S.typh aroA .GINYTLSADRTRCDITGNGGALRAPGALELFLGNAGTAMRPLAAALCL.....GQNEIV
A.nidulans LTGNNRMKQRPIGDLVDALTANVLPLNTSKGRASLPLKIAASGGFAGGNINLAAKVSSQY
S.cerevisiae LTGNARMQQRPIAPLVDSLRANGTKIEYLNNEGSLPIKVYTDSVFKGGRIELAATVSSQY
E.coli aroA LTGEPRMKERPIGHLVDALRLGGAKITYLEQENYPP..LRLQGGFTGGNVDVDGSVSSQF
S.typh aroA LTGEPRMKERPIGHLVDSLRQGGANIDYLEQENYPP..LRLRGGFTGGDIEVDGSVSSQF
A.nidulans VSSLLMCAPYAKEPVTLRLVGGKPISQPYIDMTTAMMRSFGIDVQKSTTEEHTYHIPQGR
S.cerevisiae VSSILMCAPYAEEPVTLALVGGKPISKLYVDMTIKMMEKFGINVETSTTEPYTYYIPKGH
E.coli aroA LTALLMTAPLAPE.DTVIRIKGDLVSKPYIDITLNLMKTFGVEIE.NQHYQQFVVKGGQS
S.typh aroA LTALLMTAPLAPK.DTIIRVKGELVSKPYIDITLNLMKTFGVEIA.NHHYQQFVVKGGQQ
A.nidulans YVNPAEYVIESDASCATYPLAVAAVTGTTCTVPNIGSASLQGDARFAVEVLRPMGCTVEQ
S.cerevisiae YINPSEYVIESDASSATYPLAFAAMTGTTVTVPNIGFESLQGDARFARDVLKPMGCKITQ
E.coli aroA YQSPGTYLVEGDASSASYFLAAAAIKGGTVKVTGIGRNSMQGDIRFA.DVLEKMGATICW
S.typh aroA YHSPGRYLVEGDASSASYFLAAGAIKGGTVKVTGIGRKSMQGDIRFA.DVLEKMGATITW
A.nidulans TETSTTVTGPSDGIL.RATSKRGYGTNDRCVPRCFRTGSHLPMEKSQTTPPVSSGIANQR
S.cerevisiae TATSTTVSGPPVGTLKPLKHVDMEPMTDAFLTACVVAAISHDSDPNSANTTTIEGIANQR
E.coli aroA GDDYISCTRGELNAIDMDMNHIPDAAMTIA...........TAALFAKGTTRLRNIYNWR
S.typh aroA GDDFIACTRGELHAIDMDMNHIPDAAMTIA...........TTALFAKGTTTLRNIYNWR
A.nidulans VKECNRIKAMKDELAKFGVICREHDDG......LEIDGIDRSNLRQPVGGVFCYDDHRVA
S.cerevisiae VKECNRILAMATELAKFGVKTTELPDGIQVHGLNSIKDLKVPSDSSGPVGVCTYDDHRVA
E.coli aroA VKETDRLFAMATELRKVGAEVEEGHDYI..........RITPPEKLNFAEIATYNDHRMA
S.typh aroA VKETDRLFAMATELRKVGAEVEEGHDYI..........RITPPAKLQHADIGTYNDHRMA
[866]
A.nidulans FSFSVL..........SLVTPQPTLILEKECVGKTWPGWWDTLRQLFKVK
S.cerevisiae MSFSLLAGMVNSQNERDEVANPVRILERHCTGKTWPGWWDVL......H
E.coli aroA MCFSLVA..........LSDTPVTILDPKCTAKTFPDYFEQLARISQAA
S.typh aroA MCFSLVA..........LSDTPVTILDPKCTAKTFPDYFEQLARMSTPA
[427]

Fig. 6. Amino acid homologies in the EPSP synthase domain in four species, *S. cerevisiae*, *A. nidulans*, *E. coli* and *S. typhimurium*

Only positions which are conserved in at least three of the sequences are boxed.

In some species of micro-organism there is an inducible catabolic pathway that allows the utilization of the plant metabolite quinic acid as a carbon source (Giles *et al.*, 1967*b*; Giles, 1978). One of these catabolic enzymes is a 3-dehydroquinase, and it has been reported that there is no discernible homology between the biosynthetic 3-dehydroquinase domain of the *A. nidulans arom* polypeptide and the inducible catabolic 3-dehydroquinases of *N. crassa* and *A. nidulans* (Da Silva *et al.*, 1986). Neither the 3-dehydroquinase domain of the *S. cerevisiae arom* protein nor the *E. coli* biosynthetic 3-dehydroquinase show any homology with the inducible fungal 3-dehydroquinases, which supports the proposal that the biosynthetic and degradative 3-dehydroquinase functions have arisen independently (Da Silva *et al.*, 1986).

The *C*-terminal region of the *arom* polypeptide (residues 1306–1588) is homologous to the *E. coli aroE* gene product, shikimate dehydrogenase (Anton & Coggins, 1987). In this case, the homology between *E. coli* and *S. cerevisiae* (25%) is higher than that for the shikimate kinase and 3-dehydroquinase domains; the *A. nidulans* shikimate dehydrogenase domain however has diverged substantially, being only 15% homologous with the *E. coli* and 27% homologous with the *S. cerevisiae* sequences (Table 2). This final domain of the *arom*

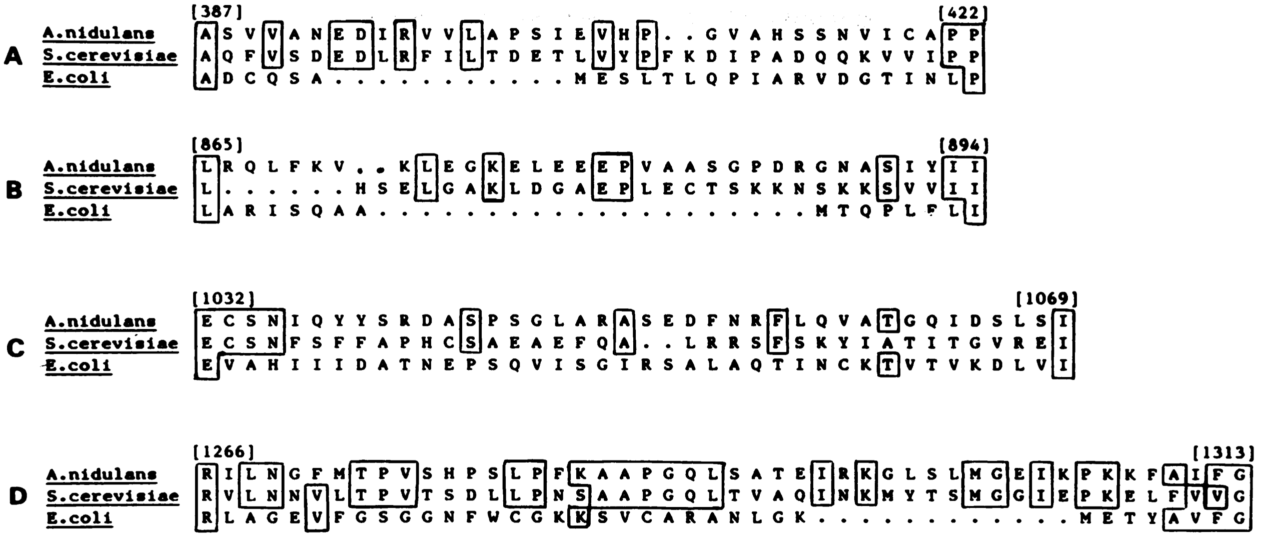
A
[387]
[422]
A.nidulans A S V V A N E D I R V V L A P S I E V H P . . G V A H S S N V I C A P P
S.cerevisiae A Q F V S D E D L R F I L T D E T L V Y P F K D I P A D Q Q K V V I P P
E.coli A D C Q S A . . . . M E S L T L Q P I A R V D G T I N L P
B
[865]
[894]
A.nidulans L R Q L F K V . . K L E G K E L E E E P V A A S G P D R G N A S I Y I I
S.cerevisiae L . . . . . . H S E L G A K L D G A E P L E C T S K K N S K K S V V I I
E.coli L A R I S Q A A . . . . M T Q P L F L I
C
[1032]
[1069]
A.nidulans E C S N I Q Y Y S R D A S P S G L A R A S E D F N R F L Q V A T G Q I D S L S I
S.cerevisiae E C S N F S F F A P H C S A E A E F Q A . . L R R S F S K Y I A T I T G V R E I
E.coli E V A H I I I D A T N E P S Q V I S G I R S A L A Q T I N C K T V T V K D L V I
D
[1266]
[1313]
A.nidulans R I L N G F M T P V S H P S L P F K A A P G Q L S A T E I R K G L S L M G E I K P K K F A I F G
S.cerevisiae R V L N N V L T P V T S D L L P N S A A P G Q L T V A Q I N K M Y T S M G G I E P K E L F V V G
E.coli R L A G E V F G S G G N F W C G K K S V C A R A N L G K . . . . M E T Y A V F G

Fig. 7. Amino acid homologies in the regions which link the functional domains

The linker sequences shown are between: (A) 3-dehydroquinate synthase and EPSP synthase; (B) EPSP synthase and shikimate kinase; (C) shikimate kinase and 3-dehydroquinase; (D) 3-dehydroquinase and shikimate dehydrogenase.

polypeptide chain is connected to the 3-dehydroquinase region by a 12-amino-acid peptide (residues 1294–1305 in the *S. cerevisiae arom* sequence).

Linkage of the domains

The *S. cerevisiae arom* polypeptide chain contains 1588 amino acid residues, which is 113 more than the total number of amino acid residues found in the five corresponding *E. coli* polypeptide chains (Table 1). Many of these extra amino acids occur in the regions linking the various domains (Fig. 7). Zalkin *et al.* (1984) have postulated that connector regions are probably essential for the structural integrity of multifunctional proteins, but that their sequence is not important. The *S. cerevisiae–E. coli* homologies break down towards the end of the *E. coli* sequences, making it impossible to say precisely where one domain ends and another begins in the *arom* sequence and making it difficult to define where the connectors begin and end. There are nonetheless four obvious connector regions linking the five domains of the *S. cerevisiae arom* polypeptide chian. These connector regions are characterized by a lack of homology between the *E. coli* and *S. cerevisiae* sequences that extends over some 30–40 residues and in three of the four cases by an insertion of from 11 to 20 amino acids in the *S. cerevisiae* sequence (Fig. 7). Secondary structure predictions following the method of Chou and Fasman indicates that these non-homologous connector regions are essentially devoid of secondary structure. In the three cases where there are insertions the additional amino acids are mainly hydrophilic.

Codon usage

The codon usage of the *ARO1* gene is shown in Table 3. The pattern resembles that for other *S. cerevisiae* genes involved in amino acid biosynthesis, for example *TRP5* (Zalkin & Yanofsky, 1982), *HIS1* (Hinnesbusch & Fink, 1983) and *HIS4* (Donahue *et al.*, 1982) suggesting that *ARO1* is expressed at about the same level as these other amino acid biosynthetic enzymes. Studies on two highly expressed *S. cerevisiae* genes, alcohol dehydrogenase I and glyceraldehyde-3-phosphate dehydrogenase, allowed Bennetzen & Hall (1982) to identify the 25 preferred codons which correspond to the most abundant isoacceptor tRNA species of *S. cerevisiae*. They have derived a codon bias index which quantifies the degree of the bias towards these selected codons and which correlates well with the extent of expression of a gene. The codon bias index was calculated for *ARO1*; the value obtained (0.25) indicates that is is a moderately expressed gene. This is consistent with the report that the *AROM* gene of *A. nidulans* is also expressed at a low level compared with the highly expressed phosphoglycerate kinase gene (Clements & Roberts, 1986).

Table 3. Codon utilization in the *ARO1* gene

'Term' indicates translation termination codons.

TTT	Phe	32	TCT	Ser	32	TAT	Tyr	23	TGT	Cys	15
TTC	Phe	29	TCC	Ser	26	TAC	Tyr	22	TGC	Cys	8
TTA	Leu	57	TCA	Ser	17	TAA	Term	-	TGA	Term	-
TTG	Leu	44	TCG	Ser	14	TAG	Term	*	TGG	Trp	17
CTT	Leu	17	CCT	Pro	5	CAT	His	27	CGT	Arg	16
CTC	Leu	4	CCC	Pro	33	CAC	His	11	CGC	Arg	0
CTA	Leu	15	CCA	Pro	31	CAA	Glu	34	CGA	Arg	1
CTG	Leu	9	CCG	Pro	0	CAG	Glu	10	CGG	Arg	0
ATT	Ile	66	ACT	Thr	46	AAT	Asn	37	AGT	Ser	19
ATC	Ile	23	ACC	Thr	20	AAC	Asn	34	AGC	Ser	8
ATA	Ile	17	ACA	Thr	33	AAA	Lys	66	AGA	Arg	31
ATG	Met	31	ACG	Thr	9	AAG	Lys	46	AGG	Arg	5
GTT	Val	67	GCT	Ala	48	GAT	Asp	54	GGT	Gly	73
GTC	Val	25	GCC	Ala	31	GAC	Asp	27	GGG	Gly	17
GTA	Val	21	GCA	Ala	28	GAA	Glu	72	GGA	Gly	17
GTG	Val	18	GCG	Ala	6	GAG	Glu	38	GGG	Gly	6

DISCUSSION

There is an increasing body of evidence that long polypeptide chains have evolved by the fusion of smaller pre-existing functional modules (Hardie & Coggins, 1986). In some cases, for example the immunoglobulins, the fusions have involved the repetition and diversification of a single structural element, presumably through gene duplication followed by divergence (Cushley, 1986). In other cases there is evidence that functions which in some species are present as separate monofunctional proteins occur in other species as fused multifunctional proteins. The amino acid sequences of these multifunctional proteins have mosaic structures with recognizable regions that are closely related to their monofunctional counterparts (Hardie & Coggins, 1986). The results presented here for the *S. cerevisiae arom* multifunctional enzyme demonstrate that its pentafunctional polypeptide chain has such a mosaic structure and in this respect is very similar to the *A. nidulans arom* multifunctional enzyme (Charles *et al.*, 1986). The most likely explanation for the origin of the pentafunctional fungal *arom* polypeptides is that they have arisen by the fusion of ancestral *E. coli*-like genes (Hardie & Coggins, 1986; Charles *et al.*, 1986). The alternative explanation, that the multifunctional enzymes are more ancient and that the monofunctional bacterial enzymes arose from them by mutational insertion of stop and start codons, cannot however be totally excluded (Hardie & Coggins, 1986).

Assuming that the gene fusion hypothesis is correct it would be expected that at least some of the functional regions of the *arom* polypeptide chain would maintain a degree of structural autonomy and that functional domains might be isolatable, for example by limited proteolysis. Although no such studies of the *S. cerevisiae arom* polypeptide have been reported the domain structure of the closely related *N. crassa arom* polypeptide has been studied directly by limited proteolysis (Smith & Coggins, 1983; Coggins *et al.*, 1985; Coggins & Boocock, 1986). A very stable *C*-terminal tryptic fragment of M_r 68000 which carries both the 3-dehydroquinase and shikimate dehydrogenase activities has been isolated (Smith & Coggins, 1983; Coggins & Boocock, 1986). One particularly interesting property of this bifunctional fragment of the *arom* polypeptide is that even after denaturation with 8 M-urea or sodium dodecyl sulphate it can refold and regain some of its shikimate dehydrogenase activity (Smith & Coggins, 1983; Coggins & Boocock, 1986). This implies that the *C*-terminal region of the *arom* polypeptide is a truly autonomous functional region. Evidence has also been presented that expression of the *C*-terminal region of the *A. nidulans AROM* gene gives an independently folding polypeptide chain carrying 3-dehydroquinase activity (Kinghorn & Hawkins, 1982) and a truncated bifunctional *A. nidulans AROM* polypeptide carrying EPSP synthase and 3-dehydroquinase activity has been reported (Charles *et al.*, 1986). The early genetic data for the *N. crassa arom* locus, which included the description of many point mutations lacking a single enzyme activity (Giles *et al.*, 1967*a*; Rines *et al.*, 1969; Case & Giles, 1971) is also consistent with the mosaic model for the *arom* polypeptide.

Forty years ago Horowitz proposed that biosynthetic pathways, as they occur today, are the result of retroevolution; that is, they have been progressively built backwards from the final metabolite of the pathway (Horowitz, 1945). The mechanism of this process presumably involved gene duplication followed by divergence (Horowitz, 1945, 1965) and one would therefore expect that the evolved proteins would retain some homology with the ancestral protein at the end of the metabolic sequence. Evidence in support of this hypothesis has recently been provided by the demonstration that two enzymes catalysing successive steps in methionine biosynthesis in *E. coli* are homologous (Belfaiza *et al.*, 1986). While the sequence homologies presented here between the five monofunctional *E. coli* shikimate pathway enzymes and the multifunctional *arom* polypeptides imply that the tertiary structures of the functional domains are conserved, we have so far been unable to identify any homologies, at the primary structure level, between the five shikimate pathway enzymes. The question of whether these five enzymes have common structural features at the tertiary level will have to await detailed three-dimensional structural analysis.

Gaertner and his co-workers have attributed some very interesting catalytic properties to the *N. crassa arom* system (Gaertner *et al.*, 1970; Welch & Gaertner, 1975, 1976). These included 'catalytic facilitation' (Gaertner *et al.*, 1970), 'channelling' (Welch & Gaertner, 1975) and 'co-ordinate regulation' (Welch & Gaertner, 1976). It is now clear that all these experiments were carried out with *arom* that was not only proteolytically degraded (Gaertner, 1978) but was also seriously deficient in 3-dehydroquinate synthase activity (Lambert *et al.*, 1985). Also the kinetic parameters used in the calculations were very different from those determined more recently with well defined preparations of homogeneous enzyme (Lambert *et al.*, 1985; Coggins & Boocock, 1986). At the present time we are not aware of any conclusive evidence of catalytic interactions between the component enzymes, nor have we obtained any evidence of co-ordinate activation (G. A. Nimmo, M. R. Boocock, J. M. Lambert & J. R. Coggins, unpublished work). This lack of evidence for any special catalytic properties for the *arom* system has lead us to consider an alternative adaptive advantage for the occurrence of the pentafunctional *arom* polypeptide chain. By having five enzymic functions involved in catalysing five sequential steps on a biosynthetic pathway on a single multifunctional polypeptide chain the problem of coordinating the expression of the five separate enzyme activities is avoided. In this connection it is interesting to note that the turnover numbers of the five *arom* enzyme activities for the *N. crassa* multifunctional enzyme are very similar (Lambert *et al.*, 1985).

We wish to thank Dr. Frank Larimer (Oak Ridge National Laboratory, U.S.A.) for a gift of plasmid pFL6. The work carried out in Glasgow was supported by a research grant from Searle Research and Development.

REFERENCES

Ahmed, S. I. & Giles, N. H. (1969) J. Bacteriol. **99**, 231–237

Alton, N. K., Buxton, F. Patel, V., Giles, N. H. & Vapnek, D. (1982) Proc. Natl. Acad. Sci. U.S.A. **79**, 1955–1959

Amrhein, N., Schab, J. & Steinrucken, H. C. (1980) Naturwissenschaften **67**, 356–357

Anton, I. A. & Coggins, J. R. (1987) Biochem. J., in the press

Bachmann, B. (1983) Microbiol. Rev. **44**, 180–230

Belfaiza, J., Parsot, C., Martel, A., Bouthier de la Tour, C., Margarita, D., Cohen, G. N. & Saint-Girons, I. (1986) Proc. Natl. Acad. Sci. U.S.A. **83**, 867–871

Bennetzen, J. L. & Hall, B. D. (1982) J. Biol. Chem. **257**, 3026–3031

Berlyn, M. B. & Giles, N. H. (1969) J. Bacteriol. **99**, 222–230

Berlyn, M. B., Ahmed, S. I. & Giles, N. H. (1970) J. Bacteriol. **104**, 768–774

Biggin, M. D., Gibson, T. J. & Hong, G. F. (1983) Proc. Natl. Acad. Sci. U.S.A. **80**, 3963–3965

Böde, R. & Birnbaum, D. (1981) Z. Allg. Mikrobiol. **21**, 417–422

Boocock, M. R. & Coggins, J. R. (1983) FEBS Lett. **154**, 127–133

Case, M. E. & Giles, N. H. (1971) Proc. Natl. Acad. Sci. U.S.A. **68**, 58–62

Catcheside, D. E. A., Storer, P. J. & Klein, B. (1985) Mol. Gen. Genet. **199**, 446–451

Charles, I. J., Keyte, J. W., Brammar, W. J. & Hawkins, A. R. (1985) Nucleic Acids Res. **13**, 8119–8128

Charles, I. J., Keyte, J. W., Brammar, W. J., Smith, M. & Hawkins, A. R. (1986) Nucleic Acids Res. **14**, 2201–2213

Chaudhuri, S. & Coggins, J. R. (1985) Biochem. J. **226**, 217–223

Chaudhuri, S., Lambert, J. M., McColl, L. A. & Coggins, J. R. (1986) Biochem. J. **239**, 699–704

Coggins, J. R. (1986) in Biotechnology and Crop Improvement and Protection (Day, P. R., ed.), pp. 101–110, British Crop Protection Council, Croydon

Coggins, J. R. & Boocock, M. R. (1986) in Multifunctional Proteins: Structure and Evolution (Hardie, D. G. & Coggins, J. R., eds.), pp. 259–281, Elsevier, Amsterdam

Coggins, J. R., Boocock, M. R. Campbell, M. S., Chaudhuri, S., Lambert, J. M., Lewendon, A., Mousdale, D. M. & Smith, D. D. S. (1985) Biochem. Soc. Trans. **13**, 299–303

Comai, L., Sen, L. C. & Stalker, D. M. (1983) Science **221**, 370–371

Clements, J. & Roberts, G. F. (1986) Gene **44**, 97–105

Cushley, W. (1986) in Multifunctional Proteins: Structure and Evolution (Hardie, D. G. & Coggins, J. R., eds.), pp. 13–53, Elsevier, Amsterdam

Da Silva, A. J. F., Whittington, H., Clements, J., Roberts, C. F. & Hawkins, A. R. (1986) Biochem. J. **240**, 481–488

de Leeuw, A. (1967) Genetics **56**, 554–555

Devereux, J., Haeberli, P. & Smithies, O. (1984) Nucleic Acids. Res. **12**, 387–395

Donahue, T. F., Farnbaugh, P. J. & Fink, G. R. (1982) Gene **18**, 47–59

Duncan, K., Lewendon, A. & Coggins, J. R. (1984) FEBS Lett. **170**, 59–63

Duncan, K., Chaudhuri, S., Campbell, M. S. & Coggins, J. R. (1986) Biochem. J. **238**, 475–483

Duncan, K., Dacey, S. A., Edwards, R. M. & Coggins, J. R. (1987) FEBS Lett., in the press

Fiedler, E. & Schultz, (1985) Plant Physiol. **79**, 212–218

Gaertner, F. H., Ericson, M. C. & DeMoss, J. A. (1970) J. Biol. Chem. **245**, 595–600

Gaertner, F. H. (1978) in Microenvironments and Metabolic Compartmentation (Srere, P. A. & Estabrook, R. W., eds.), pp. 345–353, Academic Press, New York

Ganem, B. (1978) Tetrahedron **34**, 3353–3383

Giles, N. H. (1978) Am. Naturalist **112**, 641–658

Giles, N. H., Case, M. E., Partridge, C. W. H. & Ahmed, S. I. (1967*a*) Proc. Natl. Acad. Sci. U.S.A. **58**, 1453–1460

Giles, N. H., Partridge, C. W. H., Ahmed, S. I. & Case, M. E. (1967*b*) Proc. Natl. Acad. Sci. U.S.A. **58**, 1930–1937

Hardie, D. G. & Coggins, J. R. (1986) in Multifunctional Proteins: Structure and Evolution (Hardie, D. G. & Coggins, J. R., eds.), pp. 332–334, Elsevier, Amsterdam

Hardie, D. G. & McCarthy, T. (1986) in Multifunctional Proteins: Structure and Evolution (Hardie, D. G. & Coggins, J. R., eds.), pp. 229–258, Elsevier, Amsterdam

Henner, D. J. & Hoch, J. A. (1980) Microbiol. Rev. **44**, 57–82

Hinnebusch, A. G. & Fink, G. R. (1983) J. Biol. Chem. **258**, 5238–5247

Horowitz, N. H. (1945) Proc. Natl. Acad. Sci. U.S.A. **31**, 153–157

Horowitz, N. H. (1965) in Evolving Genes and Proteins (Bryson, V. & Vogel, H. J., eds.), pp. 15–23, Academic Press, New York

Kinghorn, J. R. & Hawkins, A. R. (1982) Mol. Gen. Genet. **186**, 145–152

Kirschner, K. & Bisswanger, H. (1976) Annu. Rev. Biochem. **45**, 143–166

Koshiba, T. (1979) Plant Cell Physiol. **2**, 667–670

Kozak, M. (1984) Nucleic Acids Res. **12**, 857–872

Lambert, J. M., Boocock, M. R. & Coggins, J. R. (1985) Biochem. J. **226**, 817–829

Larimer, F. W., Morse, C. C., Beck, A. K., Cole, K. W. & Gaertner, F. H. (1983) Mol. Cell. Biol. **3**, 1609–1614

Lewendon, A. (1984) Ph.D. Thesis, University of Glasgow

Lewendon, A. & Coggins, J. R. (1983) Biochem. J. **213**, 187–191

Lumsden, J. & Coggins, J. R. (1977) Biochem. J. **161**, 599–607

Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory, New York

Messing, J. & Vieira, J. (1982) Gene **19**, 269–276

Midgley, C. A. & Murray, N. E. (1985) EMBO J. **4**, 2695–2703

Millar, G. & Coggins, J. R. (1986) FEBS Lett. **200**, 11–17

Millar, G., Lewendon, A., Hunter, M. & Coggins, J. R. (1986) Biochem. J. **237**, 427–437

Miozarri, G. F. & Yanofsky, C. (1979) Nature (London) **277**, 486–489

Mousdale, D. M. & Coggins, J. R. (1984) Planta **160**, 78–83

Mousdale, D. M., Campbell, M. S. & Coggins, J. R. (1987) Phytochemistry **26**, in the press

Nakanishi, N. & Yamamoto, M. (1984) Mol. Gen. Genet. **195**, 164–169

Norrander, J., Kempe, T. & Messing, J. (1983) Gene **26**, 101–106

Polley, L. D. (1978) Biochim. Biophys. Acta **526**, 259–266

Rines, H. W., Case, M. E. & Giles, N. H. (1969) Genetics **61**, 789–800

Sanderson, K. E. & Roth, J. R. (1983) Microbiol. Rev. **47**, 410–553

Sanger, F., Nicklen, S. & Coulson, A. R. (1977) Proc. Natl. Acad. Sci. U.S.A. **74**, 5463–5467

Schmincke-Ott, E. & Bisswanger, H. (1980) in Multifunctional Proteins (Bisswanger, H. & Schmincke-Ott, E., eds.), pp. 1–29, Wiley, New York

Schweizer, M. (1986) in Multifunctional Proteins: Structure and Evolution (Hardie, D. G. & Coggins, J. R., eds.), pp. 195–227, Elsevier, Amsterdam

Smith, D. D. S. & Coggins, J. R. (1983) Biochem. J. **213**, 405–415

Smith, T. F. & Waterman, M. S. (1981) Adv. Appl. Math. **2**, 482–489

Stalker, D. M., Hiatt, W. R. & Comai, L. (1985) J. Biol. Chem. **260**, 4724–4728

Steinrucken, H. C. & Amrhein, N. (1984) Eur. J. Biochem. **143**, 341–349

Strauss, A. (1979) Mol. Gen. Genet. **172**, 233–241

Walker, J. E., Saraste, M., Runswick, M. J. & Gray, M. J. (1982) EMBO J. **1**, 945–951

Walsh, C. (1979) Enzyme Reaction Mechanisms, pp. 553–556, Freeman, San Francisco

Welch, G. R. & Gaertner, F. H. (1975) Proc. Natl. Acad. Sci. U.S.A. **72**, 4218–4222

Welch, G. R. & Gaertner, F. H. (1976) Arch. Biochem. Biophys. **172**, 476–489

Zalkin, H. & Yanofsky, C. (1982) J. Biol. Chem. **257**, 1491–1500

Zalkin, H., Puluh, J. L., van Cleemput, M., Maoye, W. S. & Yanofsky, C. (1984) J. Biol. Chem. **259**, 3985–3992

Received 5 December 1986/2 April 1987; accepted 13 May 1987