Volume 14 Number 5 1986 Nucleic Acids Research

The isolation and nucleotide sequence of the complex *AROM* locus of *Aspergillus nidulans*

Ian G.Charles[1], John W.Keyte[1], William J.Brammar[1], Melanie Smith[2] and Alastair R.Hawkins[1,2*]

[1]Department of Biochemistry, University of Leicester, Leicester LE1 7RH and [2]Department of Genetics, Ridley Building, University of Newcastle upon Tyne, Newcastle upon Tyne NE1 7RU, UK

Received 30 December 1985; Revised and Accepted 10 February 1986

ABSTRACT.

The AROM locus of A.nidulans, which governs five consecutive steps in pre-chorismate aromatic amino acid biosynthesis, has been cloned in a bacteriophage vector. The nucleotide sequence of the locus reveals a single, open reading-frame of 4,812 base-pairs, apparently without introns. An internal segment of the A.nidulans AROM sequence has extensive homology with the E.coli aroA gene that encodes the 5-enolpyruvylshikimate 3-phosphate synthase.

INTRODUCTION

Pre-chorismate aromatic amino acid biosynthesis has been the subject of intensive genetical and biochemical study in a variety of microorganisms. The enzymes catalysing steps two to six in the pathway (see Fig 1, ref. 1) have been shown to co-sediment within a number of genera (2, 3, 4). Genetic analysis of the AROM locus in Neurospora crassa has indicated the presence of a "cluster gene", containing five, distinct genetic units corresponding to the five enzymatic functions (5). Biochemical analysis of the N.crassa AROM-specified protein complex reveals a single, pentafunctional polypeptide of 165 KDa which is active as a homodimer (6,7,8). A similar conclusion has arisen from genetic studies in Schizosaccharomyces pombe, where a 4.5 kb mRNA species has been shown to be specified by the aro-3 locus (9, 10). These findings contrast with the situation in E.coli, where the corresponding biosynthetic genes are unlinked (11).

The genes governing aromatic amino acid biosynthesis in several microorganisms are now being subjected to molecular analysis. A DNA fragment containing the biosynthetic dehydroquinase function of the A.nidulans AROM locus has been cloned (12) and sequenced (13). The entire aro loci from S.cerevisiae (14) and S.pombe (10) and part of the N.crassa AROM locus (15) have also been isolated by molecular cloning.

These continuing studies should provide answers to a number of important questions, including:-

(i) Is the AROM locus a single large gene, with or without introns, or is it a cluster of five distinct genes?

(ii) What is the nature of the AROM-specified messenger RNA?

(iii) What are the evolutionary relationships between the unlinked aro genes of E.coli and the eukaryotic AROM loci, and between the biosynthetic and catabolic dehydroquinase iso- enzymes in A.nidulans?

(iv) How many functional domains within the AROM-specified polypeptide can be detected by molecular analysis and how do they relate to the sequence of the AROM locus?

We have previously (13) addressed questions (ii) and (iii) for A.nidulans. This communication describes the isolation and presents the coding sequence of the complete AROM locus of that organism.

MATERIALS AND METHODS

Strains

A white spored, pyridoxine-requiring strain R153 (wA3; pyroA4) of Aspergillus nidulans was used for the preparation of genomic DNA. The genotypes of the bacterial strains used are shown in Table 1 of reference 16.

Media

Defined minimal medium and provision of nutritional supplements for auxotrophic strains of A.nidulans were those previously described (16). Media for liquid culture contained the wetting agent Tween 80 (diluted 10^{-5} v/v) and $MgSO_4$ and carbon source (glucose) added aseptically after sterilization. Bacterial minimal medium and provision of nutritional supplements were as previously described (17,18).

Growth of mycelium

Conditions for the growth of mycelium were those previously described (19).

DNA preparation

A.nidulans chromosomal DNA was prepared from freshly grown mycelium rapidly frozen in liquid nitrogen. Frozen mycelium was powdered in a Waring Blender, resuspended in 10 mM Tris-HCl pH 8.1, 1 mM EDTA, 4% (w/v) SDS, 25% sucrose at 20 ml g^{-1}, shaken with an equal volume of phenol, chloroform, isoamylalcohol (PCA; 48:48:4 v/v) and centrifuged to remove debris. The resulting supernatant was made 1 M in potassium acetate, the precipitate

removed by centrifugation, the procedure repeated twice, and the nucleic acids in the supernatant precipitated by the addition of 2 volumes of cold ethanol. The precipitated nucleic acids were resuspended in TE (10 mM Tris-HCl pH 8.1, 1 mM EDTA) and further purified by buoyant density centrifugation in CsCl. DNA prepared in this manner was free of RNA contamination and was greater than 50 Kb in length. Plasmid DNA was prepared by a scaled-up version of the method of Birnboim and Doly (20).

The vector λDB286 (21) was propagated by heat-induction of a lysogenic bacterial strain, precipitated by polyethylene glycol and purified on a CsCl block-gradient as previously described (22). Vector DNA was prepared from the phage particles by treatment with RNAse and pronase followed by phenol-extraction. Recombinant phages derived from λDB286 were grown by the method of Blattner et al (23) using E.coli strain C600 as the host.

Construction of a gene-bank

The 'freeze thaw' and 'sonicated' extracts for in vitro λ packaging were prepared by the method of Scalenghe et al. (24) and found to have an efficiency of packaging of $1.5\text{-}3.0 \times 10^8$ pfu μg^{-1} with λcI857 control DNA. Gene-banks were prepared following standard protocols (22), using 0.5 μg of λDB286 'arms' and a 3-fold molar excess of fungal DNA in the packaging reaction. The λDB286 vector 'arms' were purified from potassium acetate gradients following digestion with endonucleases BamHI and SalI. A.nidulans DNA fragments in the size range 8-16 Kb were prepared from sucrose gradients as previously described (22) following partial digestion with endonuclease Sau3A.

DNA probes.

DNA fragments and plasmid DNAs were labelled with α-^{32}P-dATP as previously described (25).

Purification of DNA fragments

DNA fragments produced by endonuclease action were purified using DE-81 paper as previously described (26), following separation by gel electrophoresis.

Southern blotting and plaque-hybridisation

DNA fragments separated by gel electrophoresis and DNA from λ phage particles were transferred to nitrocellulose, denatured and fixed by standard methods (27, 28). Filters were screened with specific radioactive DNA probes using the method of Jeffreys et al. (29).

E.coli transformation

E.coli strains were transformed using the method of Kushner (30) with modifications as previously described (18).

Construction of recombinant plasmids

Plasmids were subjected to a ten-fold over-digestion with the appropriate endonuclease in the presence of calf intestinal phosphatase (0.5 units μg^{-1}). Digested plasmid DNA (20 ng) was mixed with a three-fold molar excess of fungal DNA in a buffer containing 50 mM Tris- HCl pH 7.5, 10 mM $MgCl_2$, 1 mM rATP, 1 mM DTT in a volume of 10-30 µl. Parallel reactions containing half and double the proposed amount of fungal DNA (to allow for errors in DNA estimations from ethidium bromide-stained fragments in agarose), and no fungal DNA (to check the effectiveness of the phosphatase treatment) were set up and included in the bacterial transformations.

Plasmid and M13 RF DNA purification

The methods used for preparing M13 RF DNA and plasmid DNA have been detailed previously (13).

DNA sequencing

The DNA sequencing methods using ^{35}S-dATP, dideoxynucleoside triphosphates, and buffer-gradient gels have been previously described (31,32).

Synthesis of specific oligonucleotides

Oligodeoxyribonucleotides were made by the method of Matthes et al. (33) with modifications to the wash-cycle as described by Sproat and Gait (34).

Sources of material

Restriction endonucleases, DNA polymerase I, nuclease-free sucrose, ultra pure urea and SDS were from BRL Ltd. (Gibco) or NBL Ltd; T4 DNA ligase, calf intestinal phosphatase, large fragment DNA polymerase were from BCL Ltd. or Pharmacia Ltd.; Fujimax X-Ray film was from Hannimex U.K. Ltd.; Nitrocellulose was from Sartorius; DE81 paper was from Whatman, BBL-trypticase powder was from BBL Microbiology Systems; Agar was from Difco Laboratories; Agarose was from Miles Ltd.; Tryptone, peptone, and yeast extract powder were from Oxoid Ltd.; Salmon sperm DNA; polyvinyl pyrollidone, bovine serum albumin, ficoll 400, polyethylene glycol 6000, dithiothreitol, ampicillin, tetracycline, chloramphenicol, nucleoside triphosphates and dideoxynucleoside triphosphates were from Sigma Ltd., ^{35}S-dATP, and α^{32}PdCTP were from Amersham, chemicals for oligonucleotide synthesis were from Cruachem Chemical Co.; all other reagents were of Analar or greater purity.

RESULTS AND DISCUSSION

Cloning the AROM locus

The biosynthetic dehydroquinase function of the AROM locus of A.nidulans is located on a 1.9 Kb HindIII fragment, previously cloned on plasmid pHK 29 (12). The isolated HindIII fragment was used as a DNA probe to locate AROM-containing recombinant phages in a gene-bank of A.nidulans DNA cloned in the vector λDB286.

Approximately 10^5 recombinant phage, identified by their Spi^- phenotype on P2 lysogenic strain Q359, were plated on the recB, C^- E.coli strain ED8910 on ten 9 cm plates. The phages were transferred to nitrocellulose, the DNAs were denatured and screened for the presence of AROM sequences using the ^{32}P-labelled 1.9 Kb HindIII fragment as probe. Positively hybridising phages were plaque-purified through several cycles, checked for purity, propagated in bulk and their DNAs subjected to restriction enzyme mapping. Ten AROM-positive isolates analysed contained approximately 19 Kb of unique A.nidulans DNA. The physical maps of two such isolates, λARH 3.1 and λARH 8.1, which cover the entire region, are shown in Figure 1 together with the position of the 1.9 Kb HindIII fragment used as probe.

DNA sequence of the AROM locus.

The nucleotide sequence of the 1.9 Kb HindIII fragment that contains the coding sequence for the biosynthetic dehydroquinase function has been determined (13), allowing us to deduce its reading-frame and the orientation of the AROM locus. Assuming similarity with the structure of the AROM locus of N.crassa, where the sequence encoding the dehydroquinase function has been localised to the 3'-end of the locus (8), we concentrated our search for the AROM locus of A.nidulans to the region extending from 1 Kb to the right to 5 Kb to the left of the 1.9 Kb HindIII fragment. To facilitate the determination of the entire AROM DNA sequence, appropriate restriction endonuclease-generated fragments were subcloned into the RF form of the DNA sequencing vectors M13mp8 and 9 (35). Fragments were inserted in both orientations and sequences determined from both ends using universal primer. Gaps in the sequence were filled by using synthetic oligonucleotides as specific primers to extend and overlap the sequences on both strands. The overall sequencing strategy is summarized in Fig. 1.

Analysis of 6.5 Kb of unique sequence revealed a single, open reading-frame of 4,812 bp in the same orientation and phase as previously determined for that encoding the dehydroquinase function (13). This open reading-frame

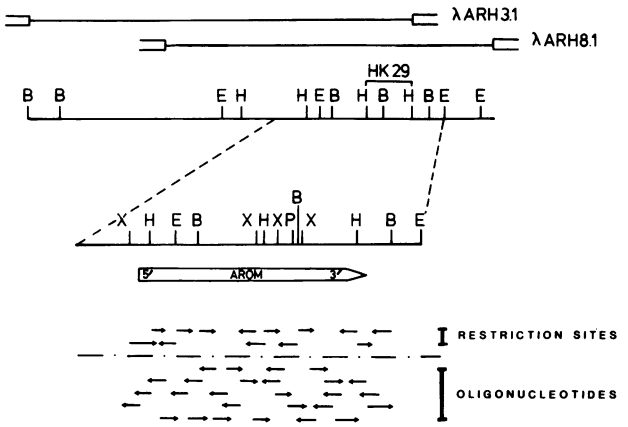

Figure 1. Restriction enzyme map of the cloned AROM DNA. The positions of the cloned DNAs within two recombinant phages, λARH3.1 and λARH8.1, are shown above the restriction map derived using the endonucleases BamHI (B), EcoRI (E) and HindIII (H). The 1.9 Kb HindIII fragment used as probe, previously cloned in plasmid pHK29, is indicated. A section of 7 Kb illustrated at a relative scale expansion of two is shown with the recognition sites for XhoI (X) and PstI (P) included. The arrows beneath this section show the direction and extent of sequencing reactions using universal primer (restriction sites) or specific oligonucleotide primers (oligonucleotides).

begins 217 b.p. downstream from the leftward XhoI site, incorporates the previously determined sequence (13) and extends it rightwards for a further 388 b.p. The sequence of the open reading- frame is shown in Fig. 2 with the deduced amino acid sequence. The inferred molecular weight of the AROM-polypeptide encoded by this open reading-frame is 175,101, which is in close agreement (6%) with that of 165,000 for the arom polypeptide of N.crassa measured by its mobility in SDS-polyacrylamide gels (8).

The observation that the AROM locus has a single open reading-frame of a size that is in close agreement with that required to specify the multi-functional polypeptide product suggests that the AROM locus contains a single structural gene lacking introns. This conclusion is supported by the finding that segments of the AROM locus can be expressed in E.coli to provide the biosynthetic activities missing in aroD or aroA mutants of the bacterium (12 and Hawkins, unpublished). Other structural genes of A.nidulans have been shown to contain introns (36,37), which are flanked by the eukaryotic consensus boundary sequences. Inspection of the A.nidulans AROM sequence does not reveal any likely intron sequences. Together these findings strongly suggest the absence of introns in the A.nidulans AROM

```
      10        20        30        40        50        60        70        80        90       100       110       120
ATGTCGAACCCTACAAAAATCAGCATCCTTGGCCGGGAAAGCATCATCGCTGATTTCGGGCTTTGGCGAAACTATGTGGCTAAAGACCTGATCAGTGACTGCTCCTCTACCACCTACGTC
MetSerAsnProThrLysIleSerIleLeuGlyArgGluSerIleIleAlaAspPheGlyLeuTrpArgAsnTyrValAlaLysAspLeuIleSerAspCysSerSerThrThrTyrVal

      130       140       150       160       170       180       190       200       210       220       230       240
CTTGTCACTGATACGAATATCGGATCGATCTACACCCCCAGCTTTGAAGAAGCTTTTCGAAAACGTGCCGCCGAGATTACTCCCTCCCCACGCCTCCTTATTTATAATCGTCCCCCCGGA
LeuValThrAspThrAsnIleGlySerIleTyrThrProSerPheGluGluAlaPheArgLysArgAlaAlaGluIleThrProSerProArgLeuLeuIleTyrAsnArgProProGly

      250       260       270       280       290       300       310       320       330       340       350       360
GAAGTTTCCAAGTCCCGACAGACGAAGGCCGATATTGAGGATTGGATGTTGAGTCAGAACCCTCCATGTGGCCGCGATACCGTTGTAATTGCATTGGGTGGAGGAGTCATCGGAGATCTG
GluValSerLysSerArgGlnThrLysAlaAspIleGluAspTrpMetLeuSerGlnAsnProProCysGlyArgAspThrValValIleAlaLeuGlyGlyGlyValIleGlyAspLeu

      370       380       390       400       410       420       430       440       450       460       470       480
ACAGGATTCGTCGCTTCCACCTACATGCGCGGTGTCCGTTATGTGCAGGTTCCCACTACTCTTCTGGCCATGGTAGATTCATCGATCGGCGGGAAAACTGCCATCGACACTCCGCTGGGT
ThrGlyPheValAlaSerThrTyrMetArgGlyValArgTyrValGlnValProThrThrLeuLeuAlaMetValAspSerSerIleGlyGlyLysThrAlaIleAspThrProLeuGly

      490       500       510       520       530       540       550       560       570       580       590       600
AAGAACCTGATCGGCGCAATCTGGCAACCGACGAAAATCTACATTGACCTTGAGTTCTTGGAGACGCTGCCGGTGAGAGAGTTCATCAATGGTATGGCAGAGGTCATCAAGACGGCAGCA
LysAsnLeuIleGlyAlaIleTrpGlnProThrLysIleTyrIleAspLeuGluPheLeuGluThrLeuProValArgGluPheIleAsnGlyMetAlaGluValIleLysThrAlaAla

      610       620       630       640       650       660       670       680       690       700       710       720
ATCTCTAGCGAAGAAGAGTTTACAGCTTTAGAAGAGAACGCGGAGACAATCTTGAAGGCGGTCCGTCGCGAGGTCACGCCAGGAGAACATCGGTTTGAGGGCACAGAAGAGATTCTGAAG
IleSerSerGluGluGluPheThrAlaLeuGluGluAsnAlaGluThrIleLeuLysAlaValArgArgGluValThrProGlyGluHisArgPheGluGlyThrGluGluIleLeuLys

      730       740       750       760       770       780       790       800       810       820       830       840
GCCCGAATTCTGGCATCTGCGCGGCACAAGGCGTATGTTGTCTCAGCAGACGAACGTGAGGGGGGTCTCCGGAACCTTTTGAACTGGGGTCACTCTATTGGCCATGCCATTGAAGCCATT
AlaArgIleLeuAlaSerAlaArgHisLysAlaTyrValValSerAlaAspGluArgGluGlyGlyLeuArgAsnLeuLeuAsnTrpGlyHisSerIleGlyHisAlaIleGluAlaIle

      850       860       870       880       890       900       910       920       930       940       950       960
CTGACGCCGCAAATTCTCCACGGAGAGTGTGTCGCAATCGGCATGGTGAAGGAAGCAGAACTGGCCCGTCATCTCGGTATCCTGAAGGGCGTTGCAGTCTCCCGCATTGTCAAATGTCTC
LeuThrProGlnIleLeuHisGlyGluCysValAlaIleGlyMetValLysGluAlaGluLeuAlaArgHisLeuGlyIleLeuLysGlyValAlaValSerArgIleValLysCysLeu

      970       980       990      1000      1010      1020      1030      1040      1050      1060      1070      1080
GCTGCTTACGGACTGCCCACCTCGCTGAAGGACGCTCGCATCCGGAAATTGACCGCCGGCAAGCACTGCTCGGTCGACCAGTTGATGTTTAACATGGCTCTTGACAAGAAGAACGATGGT
AlaAlaTyrGlyLeuProThrSerLeuLysAspAlaArgIleArgLysLeuThrAlaGlyLysHisCysSerValAspGlnLeuMetPheAsnMetAlaLeuAspLysLysAsnAspGly

     1090      1100      1110      1120      1130      1140      1150      1160      1170      1180      1190      1200
CCCAAGAAGAAAATTGTCCTCCTTTCCGCCATTGGAACTCCATATGAGACCCGCGCCAGCGTTGTTGCCAATGAGGATATTCGCGTGGTCCTTGCTCCCAGCATTGAGGTACACCCAGGT
ProLysLysLysIleValLeuLeuSerAlaIleGlyThrProTyrGluThrArgAlaSerValValAlaAsnGluAspIleArgValValLeuAlaProSerIleGluValHisProGly

     1210      1220      1230      1240      1250      1260      1270      1280      1290      1300      1310      1320
GTTGCACACTCCTCGAATGTCATCTGTGCACCTCCTGGATCCAAAAGTATTTCGAACCGGGCCTTGGTCTTGGCCGCCCTCGGCTCAGGTACTTGCCGCATCAAAAACCTGCTGCACTCC
ValAlaHisSerSerAsnValIleCysAlaProProGlySerLysSerIleSerAsnArgAlaLeuValLeuAlaAlaLeuGlySerGlyThrCysArgIleLysAsnLeuLeuHisSer

     1330      1340      1350      1360      1370      1380      1390      1400      1410      1420      1430      1440
GATGATACAGAAGTGATGCTAAATGCCCTGGAAAGGCTTGGTGCTGCGACCTTTTCCTGGGAAGAGGAAGGTGAAGTTCTGGTGGTGAACGGGAAAGGGGGAAATCTACAAGCATCTTCC
AspAspThrGluValMetLeuAsnAlaLeuGluArgLeuGlyAlaAlaThrPheSerTrpGluGluGluGlyGluValLeuValValAsnGlyLysGlyGlyAsnLeuGlnAlaSerSer

     1450      1460      1470      1480      1490      1500      1510      1520      1530      1540      1550      1560
TCGCCATTGTACCTTGGAAACGCTGGTACAGCTTCCCGATTCCTCACAACCGTCGCTACCCTTGCCAACTCAAGCACTGTTGATTCCAGCGTCCTTACTGGTAATAACCGCATGAAGCAA
SerProLeuTyrLeuGlyAsnAlaGlyThrAlaSerArgPheLeuThrThrValAlaThrLeuAlaAsnSerSerThrValAspSerSerValLeuThrGlyAsnAsnArgMetLysGln

     1570      1580      1590      1600      1610      1620      1630      1640      1650      1660      1670      1680
CGTCCTATTGGGGACTTGGTGGATGCATTGACCGCCAACGTCCTTCCATTGAATACGTCGAAAGGACGGGCCAGTTTGCCTCTCAAGATCGCTGCTTCGGGCGGATTTGCCGGAGGGAAC
ArgProIleGlyAspLeuValAspAlaLeuThrAlaAsnValLeuProLeuAsnThrSerLysGlyArgAlaSerLeuProLeuLysIleAlaAlaSerGlyGlyPheAlaGlyGlyAsn

     1690      1700      1710      1720      1730      1740      1750      1760      1770      1780      1790      1800
ATCAACCTGGCCGCCAAAGTGTCGTCGCAGTACGTCTCTTCACTGCTCATGTGTGCTCCTTACGCCAAGGAACCTGTCACTCTGAGGTTGGTCGGCGGGAAGCCTATCTCCCAACCGTAC
IleAsnLeuAlaAlaLysValSerSerGlnTyrValSerSerLeuLeuMetCysAlaProTyrAlaLysGluProValThrLeuArgLeuValGlyGlyLysProIleSerGlnProTyr

     1810      1820      1830      1840      1850      1860      1870      1880      1890      1900      1910      1920
ATTGATATGACAACGGCAATGATGCGTCGTTTGGCATTGATGTACAAAAGTCGACAACCGAAGAGCACACCTACCACATTCCCCAGGGCCGTTACGTGAACCCTGCGGAATACGTTATC
IleAspMetThrThrAlaMetMetArgSerPheGlyIleAspValGlnLysSerThrThrGluGluHisThrTyrHisIleProGlnGlyArgTyrValAsnProAlaGluTyrValIle

     1930      1940      1950      1960      1970      1980      1990      2000      2010      2020      2030      2040
GAAAGTGACGCCAGCTGTGCCACTTACCCACTCGCCGTTGCAGCCGTCACTGGCACTACCTGCACCGTTCCCAACATTGGCTCGGCTTCGCTGCAAGGAGATGCTCGTTTCGCTGTCGAG
GluSerAspAlaSerCysAlaThrTyrProLeuAlaValAlaAlaValThrGlyThrThrCysThrValProAsnIleGlySerAlaSerLeuGlnGlyAspAlaArgPheAlaValGlu
```

```
2050       2060       2070       2080       2090       2100       2110       2120       2130       2140       2150       2160
GTTTTGAGGCCTATGGGGTGTACTGTGGAGCAGACTGAAACTTCTACTACTGTTACTGGACCGTCGGACGGCATCTTACGGGCCACTTCCAAACGTGGATATGGAACCAATGACCGATGC
ValLeuArgProMetGlyCysThrValGluGlnThrGluThrSerThrThrValThrGlyProSerAspGlyIleLeuArgAlaThrSerLysArgGlyTyrGlyThrAsnAspArgCys

2170       2180       2190       2200       2210       2220       2230       2240       2250       2260       2270       2280
GTTCCTCGGTGCTTCCGTACTGGCAGCCATCGCCCGATGGAAAAGAGTCAAACCACACCACCACCGTATCTAGCGGTATTGCCAACCAGCGTGTCAAAGAATGCAACCGCATTAAGGCCATG
ValProArgCysPheArgThrGlySerHisArgProMetGluLysSerGlnThrThrProProValSerSerGlyIleAlaAsnGlnArgValLysGluCysAsnArgIleLysAlaMet

2290       2300       2310       2320       2330       2340       2350       2360       2370       2380       2390       2400
AAGGATGAGCTTGCCAAGTTTGGAGTTATCTGCCGCGAGCACGACGACGGTCTCGAGATTGACGGTATTGACCGCTCAAACCTGCGCCAGCCCGTTGGCGGTGTTTTCTGCTATGATGAT
LysAspGluLeuAlaLysPheGlyValIleCysArgGluHisAspAspGlyLeuGluIleAspGlyIleAspArgSerAsnLeuArgGlnProValGlyGlyValPheCysTyrAspAsp

2410       2420       2430       2440       2450       2460       2470       2480       2490       2500       2510       2520
CACCGAGTTGCATTCAGCTTTAGTGTACTCTCTCTCGTGACACCCCAGCCGACACTTATTCTCGAAAAGGAATGTGTTGGGAAGACATGGCCCGGCTGGTGGGATACCCTCCGTCAGCTA
HisArgValAlaPheSerPheSerValLeuSerLeuValThrProGlnProThrLeuIleLeuGluLysGluCysValGlyLysThrTrpProGlyTrpTrpAspThrLeuArgGlnLeu

2530       2540       2550       2560       2570       2580       2590       2600       2610       2620       2630       2640
TTCAAGGTGAAGCTTGAGGGCAAGGAATTAGAAGAAGAGCCTGTGGCTGCCAGCGGGCCCGATCGCGGTAACGCTTCAATTTACATTATTGGCATGCGTGGTGCTGGAAAGTCGACAGCT
PheLysValLysLeuGluGlyLysGluLeuGluGluProValAlaAlaSerGlyProAspArgGlyAsnAlaSerIleTyrIleIleGlyMetArgGlyAlaGlyLysSerThrAla

2650       2660       2670       2680       2690       2700       2710       2720       2730       2740       2750       2760
GGCAACTGGGTTTCTAAGGCTCTCAACCGGCCATTCGTTGATTTGGATACAGAGCTTGAGACTGTTGAGGGCATGACTATTCCGGACATCATCAAGACCCGCGGTYGGCAAGGCTTTAGA
GlyAsnTrpValSerLysAlaLeuAsnArgProPheValAspLeuAspThrGluLeuGluThrValGluGlyMetThrIleProAspIleIleLysThrArgGlyTrpGlnGlyPheArg

2770       2780       2790       2800       2810       2820       2830       2840       2850       2860       2870       2880
AATGCGGAGCTGGAAATCCTGAAACGTACCTTGAAGGAACGCTCGAGAGGATACGTTTTTGCTTGTGGTGGCGGTGTAGTTGAGATGCCTGAAGCTCGGAAATTGCTGACTGACTACCAT
AsnAlaGluLeuGluIleLeuLysArgThrLeuLysGluArgSerArgGlyTyrValPheAlaCysGlyGlyGlyValValGluMetProGluAlaArgLysLeuLeuThrAspTyrHis

2890       2900       2910       2920       2930       2940       2950       2960       2970       2980       2990       3000
AAGACCAAGGGCAACGTTCTGCTCCTCATGCGTGACATCAAGAAGATTATGGACTTTTTGTCAATCGACAAGTCCCGTCCTGCTTACGTCGAGGATATGATGGGCGTGTGGCTTCGACGG
LysThrLysGlyAsnValLeuLeuMetArgAspIleLysLysIleMetAspPheLeuSerIleAspLysSerArgProAlaTyrValGluAspMetMetGlyValTrpLeuArgArg

3010       3020       3030       3040       3050       3060       3070       3080       3090       3100       3110       3120
AAGCCATGGTTCCAGGAATGCAGTAATATTCAATATTACAGCCGAGATGCTTCTCCCAGTGGGCTGGCTCGCGCGTCCGAGGACTTTAACCGGTTCCTGCAGGTAGCTACCGGGCAGATT
LysProTrpPheGlnGluCysSerAsnIleGlnTyrTyrSerArgAspAlaSerProSerGlyLeuAlaArgAlaSerGluAspPheAsnArgPheLeuGlnValAlaThrGlyGlnIle

3130       3140       3150       3160       3170       3180       3190       3200       3210       3220       3230       3240
GACAGCCTCAGTATTATCAAGGAGAAAGAGCACTCTTTCTTTGCGTCGCTGACTCTACCGGATCTGCGTGAAGCTGGCGACATCCTCGAGGAAGTGTGTGTCGGATCTGACGCTGTGGAG
AspSerLeuSerIleIleLysGluLysGluHisSerPhePheAlaSerLeuThrLeuProAspLeuArgGluAlaGlyAspIleLeuGluGluValCysValGlySerAspAlaValGlu

3250       3260       3270       3280       3290       3300       3310       3320       3330       3340       3350       3360
CTACGGGTCGATCTTCTTAAGGATCCCGCCTCGAACAACGACATTCCGTCTGTTGACTATGTCGTAGAGCAGCTCTCTTTTCTGCGAAGTCGTGTCACACTCCCTATCATCTTCACCATC
LeuArgValAspLeuLeuLysAspProAlaSerAsnAsnAspIleProSerValAspTyrValValGluGlnLeuSerPheLeuArgSerArgValThrLeuProIleIlePheThrIle

3370       3380       3390       3400       3410       3420       3430       3440       3450       3460       3470       3480
CGTACCCAGAGCCAGGGCGGTCGTTTCCCCGATAACGCCCACGACGCTGCGTTGGAGCTGTACCGTCTTGCATTCAGGTCCGGGTGTGAATTTGTCGATCTCGATATCGCTTTCCCTGAA
ArgThrGlnSerGlnGlyGlyArgPheProAspAsnAlaHisAspAlaAlaLeuGluLeuTyrArgLeuAlaPheArgSerGlyCysGluPheValAspLeuAspIleAlaPheProGlu

3490       3500       3510       3520       3530       3540       3550       3560       3570       3580       3590       3600
GACATGCTGCGGGCCGTTACAGAGATGAAGGGCTTCTCTAAGATCATTGCGTCGCACCACGACCCCAAGGGCGAGCTTTCTTGGGCCAACATGTCCTGGATCAAGTTCTACAATAAAGCG
AspMetLeuArgAlaValThrGluMetLysGlyPheSerLysIleIleAlaSerHisHisAspProLysGlyGluLeuSerTrpAlaAsnMetSerTrpIleLysPheTyrAsnLysAla

3610       3620       3630       3640       3650       3660       3670       3680       3690       3700       3710       3720
CTTGAATATGGTGACATTATCAAACTTGTGGGTGTTGCACGAAACATCGACGACAACACAGCCCTGCGGAAGTTCAAGAACTGGGCTGCTGAAGCTCACGATGTCCCTCTCATCGCCATA
LeuGluTyrGlyAspIleIleLysLeuValGlyValAlaArgAsnIleAspAspAsnThrAlaLeuArgLysPheLysAsnTrpAlaAlaGluAlaHisAspValProLeuIleAlaIle

3730       3740       3750       3760       3770       3780       3790       3800       3810       3820       3830       3840
AACATGGGTGACCAGGGGCAGCTGAGCCGCATCCTCAACGGCTTCATGACACCCGTGTCTCACCCTAGCTTGCCATTCAAGGCAGCTCCCGGCCAACTCTCCGCAACCGAGATCCGCAAG
AsnMetGlyAspGlnGlyGlnLeuSerArgIleLeuAsnGlyPheMetThrProValSerHisProSerLeuProPheLysAlaAlaProGlyGlnLeuSerAlaThrGluIleArgLys

3850       3860       3870       3880       3890       3900       3910       3920       3930       3940       3950       3960
GGCCTGTCTCTCATGGGTGAAATCAAGCCCAAGAAGTTCGCCATCTTCGGCAGTCCCATATCCCAATCCGCTCCCCAGCTCTCCACAACACCCTATTTGCCCAGGTCGGCCTCCCCCATA
GlyLeuSerLeuMetGlyGluIleLysProLysLysPheAlaIlePheGlySerProIleSerGlnSerAlaProGlnLeuSerThrThrProTyrLeuProArgSerAlaSerProIle

3970       3980       3990       4000       4010       4020       4030       4040       4050       4060       4070       4080
ACTACACCCGCCTGGAGACTACGAACGCCCAAGATGTGCAGGAGTTCATCCGCTCTCCTGACTTCGGCGGCGCCTTCCGTAACAATTCGCTCAAGCTCGACATCATGCCCCTTCTCGACG
ThrThrProAlaTrpArgLeuArgThrProLysMetCysArgSerSerSerAlaLeuLeuThrSerAlaAlaProSerValThrIleArgSerSerSerThrSerCysProPheSerThr
```

```
     4090      4100      4110      4120      4130      4140      4150      4160      4170      4180      4190      4200
AAGTTGCCGCGGAAGCCGAGATCATCGGAGCTGTTAACACAATCATTCCCGTGTCGACTGGCAAGAACACTCCATCACGCCTACGTCGGCCGCAACACCGACTGGCAGGGAATGATTCTG
LysLeuProArgLysProArgSerSerGluLeuLeuThrGlnSerPheProCysArgLeuAlaArgThrLeuHisHisAlaTyrValGlyArgAsnThrAspTrpGlnGlyMetIleLeu

     4210      4220      4230      4240      4250      4260      4270      4280      4290      4300      4310      4320
TCCCTCCGCAAAGCGGGAGTCTACGGACCCAAGAGAAAGGATCAAGAGCAGTCTGCCCTCGTCGTCGGCGGCGGCGGCACGGCCCGTGCAGCCATCTACGCCCTGCACAACATGGGCTAC
SerLeuArgLysAlaGlyValTyrGlyProLysArgLysAspGlnGluGlnSerAlaLeuValValGlyGlyGlyGlyThrAlaArgAlaAlaIleTyrAlaLeuHisAsnMetGlyTyr

     4330      4340      4350      4360      4370      4380      4390      4400      4410      4420      4430      4440
TCTCCCATCTACATCGTTGGCCGCACCCCGTCTAAGCTGGAAAACATGGTCTCTTCTTTCCCCAGCAGCTACAACATCCGCATCGTTGAGAGCCCTTCAAGCTTCGAGTCCGTTCCGCAC
SerProIleTyrIleValGlyArgThrProSerLysLeuGluAsnMetValSerSerPheProSerSerTyrAsnIleArgIleValGluSerProSerSerPheGluSerValProHis

     4450      4460      4470      4480      4490      4500      4510      4520      4530      4540      4550      4560
GTCGCGATTGGTACAATCCCCGCCGATCAACCAATTGACCCGACTATGCGTGAGACACTGTGCCACATGTTCGAGCGCGCGCAGGAGGCAGACGCTGAAGCTGTGAAGGCCATTGAGCAT
ValAlaIleGlyThrIleProAlaAspGlnProIleAspProThrMetArgGluThrLeuCysHisMetPheGluArgAlaGlnGluAlaAspAlaGluAlaValLysAlaIleGluHis

     4570      4580      4590      4600      4610      4620      4630      4640      4650      4660      4670      4680
GCGCCGCGTATCCTGCTTGAGATGGCGTACAAGCCTCAGGTGACCGCACTGATGAGGCTGGCGTCTGATTCAGGCTGGAAGACTATTCCTGGTTTGGAGGTGCTAGTTGGTCAAGGGTGG
AlaProArgIleLeuLeuGluMetAlaTyrLysProGlnValThrAlaLeuMetArgLeuAlaSerAspSerGlyTrpLysThrIleProGlyLeuGluValLeuValGlyGlnGlyTrp

     4690      4700      4710      4720      4730      4740      4750      4760      4770      4780      4790      4800
TATCAGGTTTGTTTCCTTGCTTCAATCATACTCATTGCATGCGAGCTAACTGAACGTAGTTTAAATACTGGACTGGGATCTCGCCGCTATATGAGAGTGCCAGGGCATGTAGCTCCCCCC
TyrGlnValCysPheLeuAlaSerIleIleLeuIleAlaCysGluLeuThrGluArgSerLeuAsnThrGlyLeuGlySerArgArgTyrMetArgValProGlyHisValAlaProPro

     4810
TCATTTAACTAA
SerPheAsn***
```

Figure 2. The nucleotide sequence of the protein-coding region of the AROM locus. The DNA strand shown has the same orientation as the messenger RNA, and the deduced amino acid sequence is translated from a single open reading-frame.

locus, although the possibility of small, in-phase introns containing no translation-termination codons cannot be excluded.

Codon-usage in the AROM sequence.

The amino acid composition of the AROM-specified protein, together with the corresponding codon-usage, are shown in Table 1. The marked bias in codon-utilisation is evident, with strong preference for G or C and the diminution of A in the third position. The overall G + C content of the AROM coding sequence is 54%, compared with a value of 53% for A.nidulans DNA (38). A similar bias in codon-usage has been observed for other genes in A.nidulans, including PGK (40) and QutE (41).

Evolutionary relationships with E.coli aro genes and implications for domain structure within the AROM polypeptide.

It has previously been reported that the sequences encoding the two dehydroquinase iso-enzymes of A.nidulans, the biosynthetic enzyme specified within the AROM locus and the catabolic dehydroquinase determined by the quinic acid-utilisation (Qut) gene-cluster, show no sequence homology, strongly suggesting convergent evolutionary pathways (13). It is therefore of interest to determine whether there is any sequence homology between the

Table 1. Codon-usage within the AROM messenger RNA of A.nidulans.

	U	C	A	G	
U	PHE 18	SER 26	TYR 13	CYS 13	U
	PHE 31	SER 29	TYR 27	CYS 14	C
	LEU 5	SER 17	TER 1	TER 0	A
	LEU 25	SER 27	TER 0	TRP 19	G
C	LEU 28	PRO 23	HIS 8	ARG 27	U
	LEU 34	PRO 36	HIS 21	ARG 29	C
	LEU 8	PRO 14	GLN 16	ARG 13	A
	LEU 47	PRO 18	GLN 25	ARG 17	G
A	ILE 45	THR 32	ASN 14	SER 14	U
	ILE 56	THR 29	ASN 42	SER 23	C
	ILE 4	THR 29	LYS 21	ARG 8	A
	MET 39	THR 12	LYS 61	ARG 7	G
G	VAL 35	ALA 45	ASP 35	GLY 31	U
	VAL 39	ALA 49	ASP 37	GLY 44	C
	VAL 11	ALA 26	GLU 44	GLY 29	A
	VAL 24	ALA 20	GLU 53	GLY 17	G

A.nidulans AROM locus and the corresponding aro genes of E.coli. The sequence of the E.coli aroA gene, encoding the 5-enolpyruvylshikimate 3-phosphate (EPSP) synthase, is available (39) and this was compared with the A.nidulans AROM sequence using a conventional dot-matrix programme at 70% stringency. This matrix showed distinct homology between the two sequences from nucleotide 1005 to 2361 in the AROM sequence. The inferred amino acid sequence from the appropriate section of AROM DNA is shown in Fig. 3, aligned with the amino acid sequence of the E.coli aroA gene. The relatedness of the two sequences is evident: overall there is 36% homology, though this is considerably higher in discrete regions (v. residues 511 to 530). In addition, there are many examples of conservative amino acid substitutions.

The clear homology between the A.nidulans AROM sequence and the E.coli aroA gene has implications concerning the evolution of the complex locus and the domain structure of the AROM polypeptide. The data suggest that an ancestral EPSP synthase gene has been fused into a complex gene encoding a multifunctional polypeptide in A.nidulans. Alternatively, though less likely, an ancestral, multifunctional locus or group of clustered genes has become dispersed in E.coli. This observation suggests the likelihood that homology will be found between other E.coli aro genes and corresponding regions of the A.nidulans AROM locus.

The homology between AROM and the E.coli aroA gene locates the sequence encoding the EPSP synthase function within the overall AROM locus sequence. Overlapping sections of A.nidulans DNA containing this region of

```
392 APSIEVHPGVAHSSNVICAPPGSKSISNRALVLAALGSGTCRIKNLLHSDDTEVMLNALE
    *   *  *    *  **** ***** **** *     *** ***  *****
  1 MESLTLQP IARVDGTI NLPGSKTVSNRALLLAALAHGKTVLTNLLDSDDVRHMLNALT

452 RLGAATFSWEEEGEVLVVNGKGGNLQASSS PLYLGNAGTASRFLTTVATLANSSTVDSS
    **          * ** * *   * ******* * *     *
 59 ALGVS YTLSADRTRCEIIGNGGPLHAEGALELFLGNAGTAMRPLAAALCLGSDD I

511 VLTGNNRMKQRPIGDLVDALTANVLPLNTSKGRASLP LKIAASGGFAGGNINLAAKVSS
    ****  *** **** *****       *     * *    *** ***       ***
114 VLTGEPRMKERPIGHLVDALRLGGAKI TYLEQENYPPLRLQ  GGFTGGNVDVDGSVSS

570 QYVSSLLMCAPYAKEPVTLRVLGGKPISQPYIDMTTAMMRSFGIDVQKSTTEEHTYHIPQ
    *    *** ** * *    *  *    * **** *   *  **
171 QFLTALLMTAPLAPEDTVIRIKGDL VSKPYIDITLNLMKTFGVEIENQHYQQ  FVVKG

630 GR YVNPAEYVIESDASCATYPLAVAAVTGTTCTVPNIGSASLQGDARFAVEVLRPMGCT
    *  *  *  *  * *** * * ** **  * *  *  **  * *** ***  **  ** *
228 GQSYQSPGTYLVEGDASSASYFLAAAAIKGGTVKVTGIGRNSMQGDIRFA NVLEKMGAT

689 VE QTETSTTVTGPSDGILRATSKRGYGTNDRCVPRCFRTGSHRPMEKSQTTPPVSSGIA
              *   *                   *     *  **     *
287 ICWGDDYISCTHGELNAIDMDMN H IPDAAMTIA    TAAL  FAKG TTR LRN IY

748 NQRVKECNRIKAMKDELAKFGVICREHDDGLEIDGIDRSNLRQPVGGVFCYDDHRVAFSF
    * **** *  **  ** * *   *  *   *     *          * *** *  *
336 NWRVKETDRLFAMATELRKVGAEVEEGHDYIRITPPEKLNF AEMAT   YNDHRMAMCF

808 SVLSLVTPQPTLILEKECVGKTWPGWWDTL R QLFKV
    *  *    *  **   *  ** *     * *
392 SLVAL SDTPVTILDPKCTAKTFPDYFEQLARISQAA

MATCHES/LENGTH     =   36    percent
```

Figure 3. Sequence alignment between the E.coli EPSP synthase and the A.nidulans AROM-product. Both amino acid sequences are shown in the single letter code; the upper sequence is that deduced for the A.nidulans AROM-product from residues 392 to 843 and the lower the E.coli EPSP synthase sequence. The alignment is by computer analysis using the Beckman Micro-Genie suite of programmes; stars indicate positions of identical residues.

AROM have been expressed in E.coli to suppress the auxotrophy of aroA mutants (A.R.H., unpublished), suggesting that the relevant segment of the AROM Polypeptide is independently capable of folding into a catalytically active structure.

The 1.9 Kb HindIII fragment of the A.nidulans AROM locus can also be expressed to give an enzymatically active polypeptide, the biosynthetic dehydroquinase (12,13). The HindIII site at the 5' end of this DNA fragment lies exactly at the 3'end of the sequence encoding the EPSP synthase activity and may be fortuitously located in a nucleotide sequence encoding a peptide link between adjacent domains. The ability of the 1.9 Kb HindIII fragment from the 3' end of the AROM gene to encode an active dehydroquinase demonstrates that the folding of this polypeptide into a catalytically active configuration can occur quite independently of the rest of the protein structure.

In many fungi quinate catabolism and aromatic amino acid synthesis

share two intermediates, dehydroquinate and dehydroshikimate, and have two distinct dehydroquinase iso-enzymes (12, 42). Enzymes encoded by the AROM locus have been implicated in the channelling of intermediates in aromatic biosynthesis, keeping the levels of the common intermediates low enough to prevent induction of the catabolic pathway in wild-type strains (1). The existence of adjacent catalytic domains on a single, multifunctional enzyme might be well suited to achieve that end.

With the complete sequence of the AROM locus now available, we are in a stronger position to probe the relationship between domain-structure and function in the complex enzyme, using molecular genetic dissection and *in vitro* mutagenesis.

ACKNOWLEDGEMENT

We are grateful to the SERC for financial support through Project Grants GR/C/21144 and GR/C/94711, to the Society for General Microbiology for a grant partially funding computing facilities in Newcastle, and to Anne Bates for typing the manuscript.

*To whom correspondence should be sent at the University of Newcastle upon Tyne

REFERENCES

1. Giles, N.H. (1978). The American Naturalist. 162, 641-657.
2. Ahmed, S.I. and Giles, N.H. (1969). J. Bacteriol. 99, 231-237.
3. Burgoyne, L.M., Case, M.E. and Giles, N.H. (1969). Biochim. Biophys. Acta. 191, 452-462.
4. Patel, V.B. and Giles, N.H. (1979). Biochim. Biophys. Acta. 576, 24-34.
5. Case, M.E. and Giles, N.H. (1968). Genetics 60, 49-58.
6. Gaertner, F.H. and Cole, K.W. (1977). Biochem. Biophys. Res. Comm. 75, 259-264.
7. Lumsden, J. and Coggins, J.R. (1977). Biochem. J. 161, 599-607.
8. Smith, D.D.S. and Coggins, J.R. (1983). Biochem. J. 213, 405-415.
9. Strauss, A. (1979). Molec. Gen. Genet. 172, 233-241.
10. Nakanishi, N. and Yamamoto, M. (1984). Molec. Gen. Genet. 195, 164-169.
11. Bachman, B.J. (1983). Microbiol. Revs. 47, 180-230.
12. Kinghorn, J.R. and Hawkins, A.R. (1982) Molec. Gen. Genet. 186, 145-152.
13. Charles, I.G., Keyte, J.W., Brammar, W.J. and Hawkins, A.R. (1985). Nucleic Acids Res. 13, 8119-8128.
14. Larimer, F.W., Morse, C.C., Beck, A.K., Cole, K.W. and Gaertner, F.H. (1983). Mol. Cell. Biol. 3, 1609-1614.
15. Catcheside, D.E.A., Storer, P.J. and Klein, B. (1985). Molec. Gen. Genet. 199, 446-451.
16. Hawkins, A.R., Da Silva, A.J.F. and Roberts, C.F. (1985). Current Genet. 9, 305-311.

17. Armitt, S., McCollough, W. and Roberts, C.F. (1976). J. Gen. Microbiol. 92, 263-282.
18. Willetts, N.S., Clark, A.J. and Low, B. (1969). J. Bacteriol 97, 244-259.
19. Hawkins, A.R., Da Silva, A.J.F. and Roberts, C.R. (1984). J. Gen. Microbiol. 130, 567-574.
20. Birnboim, H.C. and Doly, J. (1979). Nucleic Acid Res. 7, 1513-1523.
21. Brammar, W.J. (1982). In Genetic Engineering, vol 3, 53-82 Ed. R. Williamson, Acad. Press, London.
22. Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982). In Molecular Cloning, a laboratory manual. Cold Spring Harbor laboratory.
23. Blattner, F.R., Williams, B.G., Blechl, A.E., Denniston-Thomson, K., Faber, H.E., Furlong, L., Grunwald, D.J., Kiefer, D.O., Moore, D.D., Schumm, J.W., Sheldon, E.L. and Smithies, O. (1977). Science 194, 161-169.
24. Scalenghe, F., Turco, E., Edstrom, J.E., Pirotta, V., and Melli, M. (1981). Chromosoma, 82, 205-216.
25. Rigby, P.W.J., Diekmann, M., Rhodes, C., and Berg, P. (1977). J. Mol. Biol. 113, 237-251.
26. Dretzen, G., Bellard, M., Sarsonne-Corri, P. and Chambon, P. (1981). Anal. Biochem. 112, 295-298.
27. Southern, E.M. (1975). J. Mol. Biol. 98, 503-517.
28. Benton, W.D. and Davies, R.W. (1977). Science, 196, 180-182.
29. Jeffreys, A.J., Wilson, V., Wood, D., Simons, J.P., Kay, R.M. and Williams, J.G. (1980). Cell 21, 555-564.
30. Kushner, S.R. (1978). In Genetic Engineering eds. Boyer, H.W. and Nicosia, S. North Holland Biomedical Press, Amsterdam, 17-23.
31. Sanger, F., Nicklen, S. and Coulson, A.R. (1979). Proc. Natl. Acad. Sci. USA 74, 5463-5467.
32. Biggin, M.D., Gibson, T.S. and Hong, C.F. (1983). Proc. Natl. Acad. Sci. USA 80, 3963-3965.
33. Matthes, M.W.D., Zenke, W.M., Grundstrom, T., Staub, A., Wintzerith, M. and Chambon, P. (1984). EMBO. J. 3, 801-805.
34. Sproat, B.S. and Gait, M.J. (1985). Nucleic Acid Res. 13, 2959-2978.
35. Messing, J. (1983). Methods in Enzymol. 101, 20-78.
36. McKnight, G.L., Kato, H., Upshall, A., Parker, M.D., Saari, G. and O'Hara, P.J. (1985). EMBO J. 4, 2093-2099.
37. Clements, J.M. and Roberts, C.F. (1985). Current Genet. 9, 293-298.
38. Storck, R. and Alexopolous, C.J. (1970). Bact. Revs. 34, 2, 126-154.
39. Duncan, K., Lewendon, A. and Coggins, J.R. (1984). FEBS Lett. 170, 59-63.
40. Clements, J.M. (1985). Ph.D. Thesis, University of Leicester, U.K.
41. Da Silva, A.J.F. (1985). Ph.D. Thesis, University of Leicester, U.K.
42. Hawkins, A.R., Giles, N.H. and Kinghorn, J.R. (1982). Biochem. Genet. 20, 271-286.