Volume 14 Number 4 1986 Nucleic Acids Research

Nucleotide sequence of the gene coding for yeast cytoplasmic aspartyl-tRNA synthetase (APS); mapping of the 5' and 3' termini of AspRS mRNA

Mustafa Sellami, Franco Fasiolo, Guy Dirheimer, Jean-Pierre Ebel and Jean Gangloff*

Institut de Biologie Moléculaire et Cellulaire du CNRS, Laboratoire de Biochimie, 15 rue René Descartes, 67084 Strasbourg, France

Received 26 November 1985; Revised and Accepted 24 January 1986

ABSTRACT

A 3.8 Kb DNA fragment, which contains the structural gene of aspartyl-tRNA synthetase (AspRS) and its flanking regions, has been fully sequenced by the combined M13/dideoxy chain terminator method. From the single open reading frame of correct length (1671 bp) we deduced an amino acid sequence consistent with that of several peptides of AspRS. No significant internal sequence repeats were observed in the primary structure of the protein. The AspRS gene (APS) has a codon usage pattern typical of non abundant proteins. S1 nuclease analysis of APS mRNA showed a major start 17 bases downstream from a "TATA box" and stops near an RNA polymerase terminator sequence.

INTRODUCTION

Aminoacyl-tRNA synthetases specifically attach amino acids to their corresponding tRNAs and are thus of central importance in the mechanism of protein biosynthesis. Studying their structure leads to the understanding of the nature and specificity of the interactions between these enzymes and their substrates : ATP, amino acid and tRNA (1). We decided to focus on yeast aspartyl-tRNA synthetase (AspRS) which we had purified (2). It is an α_2 dimer (2 x 64000) which binds two molecules of tRNA (3) and, as compared to other known aminoacyl-tRNA synthetases, has a very high Km value for aspartic acid. The enzyme has been crystallized (4) and high resolution electron density maps derived from X-ray analysis are actually established in our laboratory. The amino acid sequence of AspRS peptides has been under investigation in our laboratory by classical protein chemistry techniques (5) but sequencing and overlapping of all the peptides necessary to unambigously establish the entire sequence proved laborious. Therefore, we cloned the AspRS structural gene and localised it to a 3.8 kb cloned DNA fragment (6). We have determined its entire nucleotide sequence as well as those of its 3' and 5' flanking regions ; this allowed to speed up and achieve sequence determination of AspRS by chemical way (7). The amino acid

composition was deduced from the coding sequence and its codon usage analysed. Transcription initiation and termination sites were established by S_1 nuclease mapping.

METHODS

Sequencing strategy

DNA fragments of various lengths were obtained by digestion with different restriction enzymes of the 3.8 kb insert. These fragments, after polyacrylamide gel fractionation, were cloned into phage M13mp9 for DNA sequencing by the dideoxy method developped by Sanger et al. (8).

Total Cell RNA isolation

RNA was prepared according to Russel and Hall (9) with minor modifications. Yeast cells grown in one liter YNB to a density of 3.10^7 cells/ml were harvested, washed twice with ice-cold water and ice-cold RNA buffer (1 mM sodium acetate, 5 mM NaCl, 0.1 mM $MgCl_2$ pH 5.1) and resuspended in 1.5 ml RNA buffer, containing 0.5% sodium dodecyl sulphate. Three ml of sterile glass beads (0.45 mm diameter) were added together with 0.5 volume of phenol. Cells were vigorously stirred on a Vortex mixer six times for 30s with 30S cooling in between. The mixture was centrifugated for 20 min at 10,000 rpm and the aqueous phase isolated and extracted with phenol three times. The RNA in the final aqueous phase was precipitated in 0.3 M sodium acetate pH 5.1 with three volumes of alcohol. Polyadenylated-RNA was purified on oligo dT-cellulose columns as described by Maniatis et al. (10).

Mapping of the APS mRNA termini

For 5' end mapping, a 208 bp HinfI-RsaI fragment that spanned the start of APS (-174 to +34) was 5' end labeled with [γ-^{32}P]-ATP using polynucleotide kinase (Boehringer, Mannheim) after dephosphorylation with calf intestinal phosphatase (Boehringer, Mannheim) according to the methods of Maxam and Gilbert (11). For 3' end mapping a 295 bp Sau3a-RsaI fragment hanging 267 bp over the TGA stop codon was 3' end labeled by filling in its recessed 3' Sau3a end with Klenow DNA polymerase (Boehringer, Mannheim) in the presence of [α-^{32}P]-dCTP as described by Smith and Calvo (12).

The 5' and 3' termini of APS mRNA were mapped using S_1 nuclease, essentially according to the method described by Berk and Scharp (13) as modified by Weaver and Weissman (14). A total of 50,000 cpm of strand separated probe and 10 µg of poly[A^+]RNA were ethanol precipitated and resuspended in 10 µl of hybridization buffer (10 mM Pipes pH 6.5, 400 mM NaCl, 1 mM EDTA, 0.1% sodium dodecylsulfate and 50% formamide). The mixture

was treated at 85°C for 3 min and then hybridized at 42°C for 8-14 h. The reaction was chilled in ice-water and then diluted 10 fold into S_1 buffer (final concentration : 0.3 M NaCl, 30 mM sodium acetate pH 4.5, 3 mM $ZnSO_4$). S_1 nuclease (3000 units, Boehringer, Mannheim) were added (this optimal amount was determined previously) and the mixture incubated at 25°C for 2 h. A 2 µg sample of yeast RNA was added and the nucleic acids precipitated with ethanol. The S_1 nuclease resistant DNA together with the fragments obtained by G and T+C chemical sequencing reactions prepared according to Maxam and Gilbert (10) from the same labeled DNA strand were electrophoresed on a 7 M urea - 6% polyacrylamide gel according to Sanger and Coulson (15). The gels were autoradiographed for a few days at -70°C using Kodak XAR-5 film and Dupont Crones Lightning Plus intensifying screens.

RESULTS

The nucleotide sequence of APS

A 3.8 kb fragment of DNA, cloned into a pFL1 plasmid, has been shown to contain the structural gene of AspRS and its promoter sequence (6). This insert was hydrolysed by several restriction enzymes, and overlapping restriction fragments were isolated and sequenced on both strands by the dideoxy chain termination method. All regions were sequenced at least twice. By computer analysis we determined an open reading frame of 1671 bp which length is consistent with AspRS ; it is flanked by upstream and downstream regions of 1548 bp and 544 bp respectively (Fig. 1). The identity of the coding region was confirmed by several AspRS peptides, sequenced by classical methods (5) which are scattered throughout the translated DNA sequence. Their total length corresponded to about half of the total number of the amino acids contained in the whole AspRS protein (7). However, the NH_2 terminal of the protein could not be defined by a corresponding N-terminal peptide since, as previously suggested (5), AspRS, during purification, was probably mildly proteolysed at the N-terminal region during the course of enzyme purification.

Codon usage

Codon usage frequencies and the amino acid composition of APS are given in table I ; it shows that of 61 possible coding triplets 7 are not used at all and there is a strong bias toward codons for Leu, Cys and Gly. In spite of these few exceptions, the overall bias is not strong. The codon bias index defined by Bennetzen and Hall (16) and calculated from data of table I corresponds to 0.6 and is typical of the category of non-abundant to

```
   1                                                                GATCCTTC
   9 TGGGAGGCTA TGGCTTCTCT TTCTTCAAAC AATGTGGAAT GAAACCAAAT ACTTTGCATT TATTGAAAAC
  79 CTAGACAAGG TGTAAATTAG TTTTTTGAGG GTTTTTTTGG AAACCTATCT TCAACCTAAA CGGAGACTTC
 149 CTGAAAAGCA TATAAACATA GGTTATTTGA GCTGTTAATA AATCCAAGGA AGTTAAAAAC TTTTCCCGTT
 219 CATTTGCCAA TAGGTCGCGC TCTGTTAAAT TAACATAATG ACACACTAAG CGAGTGAGGT CATACCAATA
 289 GACGACATCA CAATCATTGA GATAGAACAG ATCGGGGCTT GACGTTATAG GCATCTTTTT ACCAGCTATC
 359 TTCGATATTG CAGGAAAAGA CGAATAGTCG TTCACGGTTT CAACTTGATT TTGTGGTAAA TATGGTAACG
 429 TGCGACAATA AACATCTAAA TCCTGGTTAT AGTACAGCAG CTCACAGCAG CCTTTCCTTT CGTGGACGGG
 499 AGTAGGTGAC CAAGCCAGTT GTTCCTTGAA TTCAATATCG CGACACGAGT CTAAAAAGTC TTCCAATGAT
 569 AAGATCACTG GGTTGCTTGA TAATTCCAGT TTAAGATTAA GACCCGCTTG AAATATTTCC AATTCATTAC
 639 CGATGTCAGG CACCCCGTGT AGTTCGTTTA GAATGGACCG TGTAAAAAGA AGGGGGGAAC CAACCTCTGT
 709 CTAGTATCAA CGAATTATTA TTATTATTAG CGCTGTGTCT GATAAGAACA CCATCCCACC AGCCATTAGA
 779 CCCTTTAGTG AGTACGTAAA TCAAATCGCC TACCCTTAAA GACAGTTTAT TCCGTGATTT TGTAAAATAT
 849 TGATAGGTAC ATTGGACTAC ATCGATAGGT TGTAAGCATG GAATCCCGTG GCCTTCCTTA TCCTTCACCG
 919 GAGTTGTCAT GCCGGCATAT GACGCAGTGC AACTCATTCT GGTGGCAGAT TATATATGTC TTTTAGATAT
 989 GTCTCTCATA TTTTTTTTTT TTTTCACTTT GTTATATGCA AAAAAGCAAT TTTGACTACT ATTGTAGAAG
1059 GGAACCCTTA AATTTTGCGA GCCTGCAGAC ATTATTTTCC AATGATGCGG TACAGCCGAT GTAAATCTAA
1129 TATATCTACA TGACGACATG ACGACATAAC ATAATAACGT ACATTTGGTA CAAAACGTTG CAAATACAGC
1199 TAGAATGGTT TGCTTCTGGT CCTCCAAGCA GGACATATGT TTCCTTGAAG TTCATCTGCA GTAATTGTGC
1269 ACATGCAGTT TTTGTTTCTC AAAGAGACAG GCCAAGCTGA AGATGTTCCT ATGATATTAT GAAAGTATAT
1339 ATTCTATTAC CGTCACATTT TGTACCATTA GGCTACAGAA AACAATAGGA TTGGGACTCA TCCATAAGAA
1409 CTCTTTTACG TTGTGCTGTT GCTTTATTGT CTATGAAATC ATCAGTGACG AGTACCTTAT GACGCATTTT
1479 CGTGAAAAAA GAAGAAATGA AAAATTATTT AATGTCATAT AAGGATGAAG CAAGACAACC GATTTGTTTA
1549 AGTTTTCGAT CATTGGCTGG AACGGCCTCT TAAGGTTAAT TGTAAAAGAA AAAAAGAAAC ATTTTACGTG
1619 ATG TCT CAA GAC GAA AAT ATT GTC AAA GCT GTT GAA GAA TCC GCA GAA CCT GCT CAA GTT
     MET SER GLN ASP GLU ASN ILE VAL LYS ALA VAL GLU GLU SER ALA GLU PRO ALA GLN VAL
1679 ATT CTT GGG GAA GAT GGT AAG CCA TTG TCC AAG AAG GCC TTG AAG AAA TTG CAG AAA GAG
     ILE LEU GLY GLU ASP GLY LYS PRO LEU SER LYS LYS ALA LEU LYS LYS LEU GLN LYS GLU
1739 CAA GAG AAA CAG AGA AAG AAG GAG GAA AGA GCT CTC CAG TTG GAA GCT GAA AGA GAA GCC
     GLN GLU LYS GLN ARG LYS LYS GLU GLU ARG ALA LEU GLN LEU GLU ALA GLU ARG GLU ALA
1799 CGT GAA AAG AAA GCC GCT GCC GAA GAC ACC GCA AAG GAC AAC TAC GGT AAG TTG CCA TTG
     ARG GLU LYS LYS ALA ALA ALA GLU ASP THR ALA LYS ASP ASN TYR GLY LYS LEU PRO LEU
1859 ATC CAG TCT CGT GAC TCT GAC AGA ACT GGT CAG AAG CGT GTC AAG TTT GTT GAC TTG GAT
     ILE GLN SER ARG ASP SER ASP ARG THR GLY GLN LYS ARG VAL LYS PHE VAL ASP LEU ASP
1919 GAG GCT AAG GAT AGC GAC AAA GAA GTC CTC TTC AGG GCA AGA GTC CAC AAC ACC AGA CAA
     GLU ALA LYS ASP SER ASP LYS GLU VAL LEU PHE ARG ALA ARG VAL HIS ASN THR ARG GLN
1979 CAA GGT GCA ACA TTG GCC TTT TTA ACT TTA AGG CAA CAA GCT TCC TTG ATC CAA GGT CTA
     GLN GLY ALA THR LEU ALA PHE LEU THR LEU ARG GLN GLN ALA SER LEU ILE GLN GLY LEU
2039 GTA AAG GCC AAC AAG GAA GGT ACC ATC AGC AAA AAC ATG GTC AAA TGG GCT GGT TCA TTG
     VAL LYS ALA ASN LYS GLU GLY THR ILE SER LYS ASN MET VAL LYS TRP ALA GLY SER LEU
2099 AAT TTG GAG TCC ATT GTC CTT GTC AGA GGT ATT GTC AAG AAG GTA GAT GAG CCA ATC AAG
     ASN LEU GLU SER ILE VAL LEU VAL ARG GLY ILE VAL LYS LYS VAL ASP GLU PRO ILE LYS
2159 TCT GCT ACT GTG CAA AAC CTG GAA ATT CAC ATT ACC AAG ATT TAT ACC ATT TCC GAG ACT
     SER ALA THR VAL GLN ASN LEU GLU ILE HIS ILE THR LYS ILE TYR THR ILE SER GLU THR
2219 CCA GAA GCA TTG CCA ATC CTT TTG GAA GAT GCC TCC CGT TCC GAA GCT GAA GCT GAA GCT
     PRO GLU ALA LEU PRO ILE LEU LEU GLU ASP ALA SER ARG SER GLU ALA GLU ALA GLU ALA
2279 GCA GGT TTG CCC GTG GTC AAC TTG GAC ACC AGA TTA GAC TAC CGT GTC ATT GAC TTG AGA
     ALA GLY LEU PRO VAL VAL ASN LEU ASP THR ARG LEU ASP TYR ARG VAL ILE ASP LEU ARG
2339 ACC GTC ACC AAC CAA GCT ATT TTC AGG ATT CAA GCT GGT GTT TGT GAG TTG TTC AGA GAA
     THR VAL THR ASN GLN ALA ILE PHE ARG ILE GLN ALA GLY VAL CYS GLU LEU PHE ARG GLU
2399 TAT TTG GCC ACA AAG AAA TTT ACC GAA GTA CAC ACA CCA AAA CTG TTG GGT GCA CCA AGT
     TYR LEU ALA THR LYS LYS PHE THR GLU VAL HIS THR PRO LYS LEU LEU GLY ALA PRO SER
```

```
2459 GAA GGT GGT TCC AGT GTG TTT GAG GTG ACA TAC TTC AAA GGG AAG GCC TAC CTA GCT CAA
     GLU GLY GLY SER SER VAL PHE GLU VAL THR TYR PHE LYS GLY LYS ALA TYR LEU ALA GLN
2519 TCT CCA CAA TTT AAC AAG CAA CAA TTG ATT GTG GCC GAC TTT GAA AGA GTT TAC GAA ATC
     SER PRO GLN PHE ASN LYS GLN GLN LEU ILE VAL ALA ASP PHE GLU ARG VAL TYR GLU ILE
2579 GGG CCT GTG TTC AGG GCT GAA AAC TCC AAC ACC CAC CGT CAC ATG ACC GAG TTT ACT GGT
     GLY PRO VAL PHE ARG ALA GLU ASN SER ASN THR HIS ARG HIS MET THR GLU PHE THR GLY
2639 TTG GAC ATG GAA ATG GCT TTC GAA GAA CAT TAC CAC GAA GTT TTG GAC ACG TTG AGT GAG
     LEU ASP MET GLU MET ALA PHE GLU GLU HIS TYR HIS GLU VAL LEU ASP THR LEU SER GLU
2699 TTG TTT GTG TTT ATT TTC AGT GAA TTG CCC AAG AGA TTT GCT CAT GAA ATT GAG TTG GTA
     LEU PHE VAL PHE ILE PHE SER GLU LEU PRO LYS ARG PHE ALA HIS GLU ILE GLU LEU VAL
2759 CGT AAG CAA TAC CCT GTT GAA GAA TTC AAG TTA CCT AAA GAT GGT AAG ATG GTT CGT CTA
     ARG LYS GLN TYR PRO VAL GLU GLU PHE LYS LEU PRO LYS ASP GLY LYS MET VAL ARG LEU
2819 ACA TAC AAA GAA GGT ATT GAA ATG CTA AGA GCT GCC GGT AAG GAA ATT GGT GAT TTT GAA
     THR TYR LYS GLU GLY ILE GLU MET LEU ARG ALA ALA GLY LYS GLU ILE GLY ASP PHE GLU
2879 GAC TTG AGT ACC GAA AAT GAA AAG TTC TTG GGT AAG TTG GTT CGC GAC AAA TAC GAC ACC
     ASP LEU SER THR GLU ASN GLU LYS PHE LEU GLY LYS LEU VAL ARG ASP LYS TYR ASP THR
2939 GAC TTT TAC ATC CTA GAC AAG TTC CCC TTG GAG ATC CGT CCC TTC TAC ACA ATG CCC GAC
     ASP PHE TYR ILE LEU ASP LYS PHE PRO LEU GLU ILE ARG PRO PHE TYR THR MET PRO ASP
2999 CCA GCC AAC CCT AAG TAT TCT AAC TCG TAT GAT TTC TTC ATG AGG GGT GAA GAA ATC TTG
     PRO ALA ASN PRO LYS TYR SER ASN SER TYR ASP PHE PHE MET ARG GLY GLU GLU ILE LEU
3059 TCC GGT GCA CAA CGT ATC CAC GAC CAT GCT CTA TTA CAA GAA AGG ATG AAA GCC CAT GGT
     SER GLY ALA GLN ARG ILE HIS ASP HIS ALA LEU LEU GLN GLU ARG MET LYS ALA HIS GLY
3119 TTG TCT CCT GAG GAC CCA GGT CTA AAG GAC TAC TGT GAC GGC TTC AGC TAT GGG TGT CCT
     LEU SER PRO GLU ASP PRO GLY LEU LYS ASP TYR CYS ASP GLY PHE SER TYR GLY CYS PRO
3179 CCA CAC GCC GGT GGT GGT ATC GGT TTG GAA AGA GTT GTT ATG TTC TAT TTG GAT TTG AAA
     PRO HIS ALA GLY GLY GLY ILE GLY LEU GLU ARG VAL VAL MET PHE TYR LEU ASP LEU LYS
3239 AAT ATC AGA AGA GCT TCA TTC GTT CCA AGA GAT CCA AAG AGA TTA AGA CCA TGA
     ASN ILE ARG ARG ALA SER LEU PHE PRO ARG ASP PRO LYS ARG LEU ARG PRO STOP
3293 AGGAGTCTCC TCCAACGTTT TTATGTAGCG AGTGAAGGTA AAAAAATTTT CAACATTTCT AATTTTTTCC
3363 TGTCACCCAC TTTTATTTTA ACTGAATTAT TATATTTAGC TTTACACATA TAAGTCTCAA CGTCTGACAT
3433 ATGTCACACT TTCTTTACCG CCAGACCACT TCTAATATTC CGAAAGCAAG GGCTATGAGC AGTCCGACAA
3503 TTCGGAAACC ATAGCTCGCC ATAATTTACA CGCTCTATGC ACATATTAAA ATACATTTAA ACAGTACTTA
3573 CAGCGCATCT GGGATGTTCG GTTATGTACC AGGACAATAT TAAGTTGACG CGTCAGTTGA GTAGTAATCG
3643 TCCTAGTATT TTCGCCTACC ATGCCGTCTT TTGCGTTTTT GACCGCCTCT GGTCCCTTAT ACAATCCACA
3713 GGCAAAAAAA AATTACATAA CAAGATTACT CTTCAAGAAA CCTGGAAAAT ATTTTTTTGG GCAGTTTTTT
3783 AAGACGATAA TAAGTTCTAT GCCTCATTTA CGCACCAGCG GTTCTATTAC CAAG
```

Fig. 1. Nucleotide sequence of the APS gene and the primary structure of AspRS. The 5' and 3' ends of the APS mRNA are indicated by solid triangles. Within the 5' untranslated region, sequences such as ATATAA, CAAG---- are inderlined. The presumptive transcription termination signals are overlined in the 3' untranslated region.

moderately expressed genes. This is consistent with the observed amount of AspRS present in the yeast cell which is about 1/5000 of total proteins as can be calculated from purification data (3). Table 1 further shows that the preferred codons correspond to the anticodons of the major isoacceptor tRNA species of yeast (16) except for Alal and Gly3. These tRNAs may therefore be rate limiting factors for AspRS synthesis.

Table 1. Amino acid composition and codon usage of AspRS

Codon usage			
Phe TTT 12	Ser TCT 7	Tyr TAT 6	Cys TGT 3
Phe TTC 16	Ser TCC 10	Tyr TAC 12	Cys TGC 0
Leu TTA 6	Ser TCA 2	End TAA	End TGA 1
Leu TTG 35	Ser TCG 1	End TAG	Trp TGG 1
Leu CTT 3	Pro CCT 7	His CAT 4	Arg CGT 10
Leu CTC 2	Pro CCC 5	His CAC 8	Arg CGC 1
Leu CTA 7	Pro CCA 14	Gln CAA 18	Arg CGA 0
Leu CTG 2	Pro CCG 0	Gln CAG 5	Arg CGG 0
Ile ATT 16	Thr ACT 5	Asn AAT 4	Ser AGT 5
Ile ATC 12	Thr ACC 13	Asn AAC 12	Ser AGC 3
Ile ATA 0	Thr ACA 6	Lys AAA 16	Arg AGA 19
Met ATG 11	Thr ACG 1	Lys AAG 32	Arg AGG 6
Val GTT 12	Ala GCT 22	Asp GAT 10	Gly GGT 27
Val GTC 11	Ala GCC 13	Asp GAC 23	Gly GGC 1
Val GTA 4	Ala GCA 8	Glu GAA 43	Gly GGA 0
Val GTG 7	Ala GCG 0	Glu GAG 14	Gly GGG 4
Total codons used = 507			
Amino acid composition			
Ala 43	Glun 23	Leu 55	Ser 28
Arg 36	Glu 57	Lys 48	Thr 25
Asn 16	Gly 32	Met 11	Trp 1
Asp 33	His 12	Phe 28	Tyr 18
Cys 3	Ile 28	Pro 26	Val 34

5' and 3' end mapping of APS mRNA (Fig. 2)

The 5' and 3' ends of APS mRNA were mapped by S_1 nuclease digestion of hybrids between mRNA and 5' or 3' end labeled fragments : a 208 bp HinfI-RsaI and a 295 bp Sau3a-RsaI fragment respectively from the 5' and the 3' end. The length of the resulting protected DNA fragments were determined by fractionation on a 7 M urea 6% polyacrylamide DNA sequencing gel in parallel with the same fragments previously treated by the chemical Maxam and Gilbert method (11). The predominant protected bands corresponding to the 5' end of AspRS mRNA are localized at positions 86 and 87 from the translation initiation codon. The nucleotide coding for the major transcription termination site was mapped at 157 bp from the translation termination codon TGA. It must be noticed that several clusters of minor bands are also observed near the major 5' and 3' ends of mRNA. Multiple 5' ends have often been described in yeast. They may be due to degradation, or reflect numerous starting points for transcription or processing steps for the leader or primary transcript (17). Furthermore we observe (Fig. 2) that each major band is accompagned by several other minor bands spaced at one nucleotide

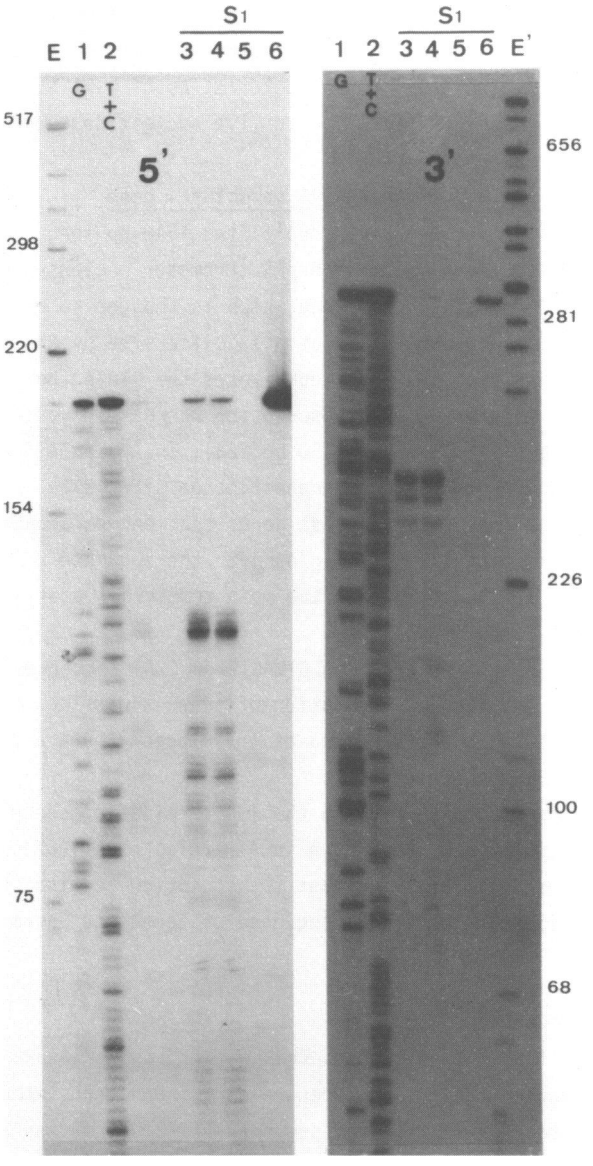

Fig. 2. S_1 nuclease analysis of the 3' and 5' termini of APS transcripts. DNA probes used in 3,4, 5 and 6 are a Hinf-Rsa 1 fragment for 5' and a Sau3a-Rsa 1 fragment for 3' end mapping. These fragments were hybridized to : RNA poly A^+ from F1100 (3), F1100-3.8 (4) ; tRNA (5) ; no RNA (6) and submitted to digestion by S_1 nuclease. Lanes 1 and 2 are Maxam and Gilbert sequencing reactions of the DNA probes. E and E' lanes contain respectively end labelled Hinf I and Taq_1-AluI fragments of PBR322.

intervalls. This is generally observed when S_1 nuclease-generated [^{32}P] DNA-RNA hybrids are analyzed on high resolution sequencing gels. This probably is due to two unfortunate characteristics of most S_1 nuclease preparations : a tendency to leave one to five nucleotide overhanging ends (18) and a tendency to end nibble.

3' and 5' flanking regions of the AspRS structural gene

We determined the primary structure of the 1548 bp long sequence upstream from the structural gene of AspRS. Promoter regions generally contain a ...TATA$^{A}_{A}$A consensus sequence which is thought to be required for proper positioning of the mRNA start at a specific site by RNA polymerase II (19-21). It was shown that in higher eukaryotes the "TATA" box is about 26-36 bp distant from initiation of transcription whereas in yeast this distance may vary between 39 bp and 150 bp (22). In APS an ATATAA sequence was found at 17-18 bp from the main transcription start site.

It was observed that several yeast genes that encode abundant mRNAs have their transcription start site in/or near the sequence CAAG preceeded by a CT rich block (23). In our case the main transcript starts in a CAAG sequence but there is no CT cluster.

Finally, we didn't find either ATGTGACTC or CAAT sequences which, as suggested (24) are involved in regulation of gene expression. No TAAGGACCT sequence complementary to the 3'OH end of 18S ribosomal RNA and found in the yeast LEU2 gene (25) was present.

The 544 bp long sequence between the TGA termination codon and the main transcription termination site is rich in AT nucleotides (66%). Concerning the model sequences implicated in yeast transcription termination, we identified the tripartite consensus sequence of Zaret and Sherman TAG---TAAGT---TTT (26).

DISCUSSION AND CONCLUSION

A 3.8 kb DNA fragment containing the gene coding for AspRS was fully sequenced. By computer analysis an open reading frame translated a protein sequence which matches well with the sequences of a set of AspRS peptides established by classical methods was detected. Furthermore, the length of this open reading frame (1671 bp) fits with the size of AspRS indicating that there are no intervening sequences in APS. The N-terminal third of AspRS contains more than half of the total amount of lysines, conferring a positive charge to this part of the molecule.

As observed (27), translation in eukaryotes is generally initiated at the first AUG after the 5' end of mRNA. However, in those cases where the

message contains several AUG near the transcription start it was proposed (28) that the flanking sequences of the triplets will determine the level at which each AUG is recognized by 40S ribosomal subunits. As a consequence, the efficiency of translation may be reduced by initiation at an internal AUG. It has been suggested that the optimal environment of AUG for efficient translation initiation requires a purine residue, usually an A, at position -3 and a purine at position +6 relative to the A of the initiation AUG triplet. In the case of APS we found that the ATG which opens the reading frame is the closest to the 5' terminus of APS mRNA. Its adjacent sequence pUxxATGxxT fits with the optimal consensus sequence proposed by Dobson et al. (23).

By S_1 analysis of APS mRNA we mapped one major 5' end 87 nucleotides from the translation initiation site. It is located in a CAAG sequence which was proposed to play a role in mRNA capping (23). An ATATAA promoter consensus sequence was found 17 bp upstream the main transcription start site. We have shown (2) that the APS gene inserted in the plasmid pFL1 expresses AspRS, although weakly, from its own promoter. This led us to check whether the upstream region of the AspRS gene contained a Shine and Dalgarno sequence which in the E.coli gene expression system was shown to bind near the 3' end of 16S ribosomal subunit (29). We detected a similar sequence (TTAAGG) 40 bp upstream from the AUG, which is however further away from the AUG than is found in E.coli.

S_1 mapping of the 3' non-coding region of AspRS mRNA revealed three clusters of 3' ends, the most frequent of whcih maps 39 bp from a TATAA sequence which ressembles the AAUAAA sequence implicated by Bennetzen et al. (27) in polyadenylation of higher eukaryotic messengers. However, Henikoff et al. (30) concluded that this sequence is implicated neither in poly A addition nor in transcription termination. Furthermore, a Sherman tripartite like sequence TAG---TAAGT---TTT is present 51 bp from the main transcription termination site. Henikoff et al. (30) have suggested that the sequence TTTTTATA is required for transcription termination in yeast. Although this sequence is not present, we observe several AT rich stretches in the 3' non-coding region of APS mRNA. However, to date one cannot define a strict sequence requirement for initiation or termination of transcription that is for capping or polyadenylation of mRNA ; probably more complex structural features are implicated in these mechanisms. Analysis of the 5' and 3' flanking regions by deletion and point mutagenesis may give a more precise answer to this question.

ACKNOWLEDGEMENTS

We thank M.L. Gangloff for skilfull technical assistance. This work was supported by grants from the Centre National de la Recherche Scientifique (ATP n°5684) and from INSERM (CR Externe 85/2009).

*To whom correspondence and reprint requests should be addressed

REFERENCES

1. Schimmel, P.R. and Söll, D. (1979) Ann. Rev. Biochem. 48, 601-648.
2. Gangloff, J. and Dirheimer, G. (1973) Biochim. Biophys. Acta 294, 263-272.
3. Lorber, B., Kern, D., Dietrich, A., Gangloff, J., Ebel, J.P. and Giegé, R. (1983) Biochem. Biophys. Res. Commun. 117, 259-267.
4. Dietrich, A., Giegé, R., Comarmond, M.B., Thierry, J.C. and Moras, D. (1980) J. Mol. Biol. 138, 129-135.
5. Hounwanou, N., Boulanger, Y. and Reinbold, J. (1983) Biochimie 65, 379-388.
6. Sellami, M., Prévost, G., Bonnet, J., Dirheimer, G. and Gangloff, J. (1985) Gene 40, 349-352.
7. Amiri, I., Mejdoub, H., Hounwanou, N., Boulanger, Y. and Reinbolt, J. (1985) Biochimie 67, 607-613.
8. Sanger, F., Nicklen, S. and Coulsen, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463-5467.
9. Russel, P.R. and Hall, B.D. (1982) Mol. Cell. Biol. 2, 106-116.
10. Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) Molecular Cloning, A laboratory Manual. Cold Spring Harbor Laboratory.
11. Maxam, A. and Gilbert, N. (1980) Methods Enzymol. 65, 449-560.
12. Smith, D.R. and Calvo, J.M. (1980) Nucl. Acids Res. 8, 2255-2274.
13. Berk, A.J. and Sharp, P.A. (1978) Proc. Natl. Acad. Sci. USA 75, 1274-1278.
14. Weaver, R.F. and Weissmann, C. (1979) Nucl. Acides Res. 7, 1175-1193.
15. Sanger, F. and Coulson, A.R. (1978) FEBS Lett. 87, 107-110.
16. Benetzen, J.L. and Hall, B.D. (1982) J. Biol. Chem. 257, 3026-3031.
17. Losson, R. and Lacroute, F. (1981) Mol. Gen. Genet. 184, 394-399.
18. Grosschedl, R. and Birnstiel, R.L. (1980) Proc. Natl. Acad. Sci. USA 77, 1432-1436.
19. Mathis, D.J. and Chambon, P. (1981) Nature 290, 310-315.
20. Benoist, C., O'Hare, K., Breathnach, R. and Chambon, P. (1958) Nucl. Acids Res. 8, 127-142.
21. Faye, G., Leung, D.W., Tatchell, K., Hall, B.D. and Smith, M. (1981) Proc. Natl. Acad. Sci. USA 78, 2258-2262.
22. Reynolds, P., Higgins, D., Prakash, L. and Prakash, S. (1985) Nucl. Acids Res. 13,
23. Dobson, M.J., Tuite, M.F., Robert, N.A., Kingsman, A.J., Kingsman, S.M., Perhins, R.E., Conroy, S.C., Dunbar, B. and Fothergill, L.A. (1982) Nucl. Acids Res. 10, 2625-2637.
24. Hinnenbusch, A.G. and Fink, G.R. (1983) J. Biol. Chem. 258, 5238-5247.
25. Andreadi, A., Hsu, Y.P., Kohlhaw, G.B. and Schimmel, P.R. (1982) Cell. 31, 319-325.
26. Zaret, K.S. and Sherman, F. (1982) Cell 28, 563-573.
27. Bennetzen, J.L. and Hall, B.D. (1982) J. Biol. Chem. 257, 3018-3025.
28. Kozak, M. (1984) Nucl. Acids Res. 9, 5233-5252.
29. Shine, J. and Dalgarno, L. (1975) Nature 254, 34-38.
30. Henikoff, S., Kelly, J.D. and Cohen, E.H. (1983) Cell. 33, 607-614.