Structure and sequence of a UDP glucose pyrophosphorylase gene of *Dictyostelium discoideum*

Jack A. Ragheb and Robert P. Dottin*

Department of Biology, The Johns Hopkins University, Baltimore, MD 21218, USA

Received December 3, 1986; Revised and Accepted April 10, 1987

Accession no. Y00145

ABSTRACT

Cell-cell contact and exogenous cAMP regulate the expression of uridine diphosphoglucose pyrophosphorylase (UDPGP) of *Dictyostelium discoideum* (B. Haribabu, A. Rajkovic and R. P. Dottin, 1986, *Dev. Biol.*, Vol. 113, 436-442). cAMP appears to regulate gene expression in *Dictyostelium* by transmembrane signal transduction (B. Haribabu and R. Dottin, 1986, *Mol. Cell. Biol.* 6, 2402-2408). To further characterize the mechanism of action of cAMP on the expression of this gene and the nature of the defects in UDPGP mutants that abort development, we sequenced the cDNA and the genomic DNA, including intervening and flanking sequences. The deduced amino acid sequence predicts a polypeptide of 57,893 d. molecular weight. Three short (100-200 nucleotides) A+T rich introns occur within the coding sequences but only one of them contains a sequence TAACTAAC, similar to the yeast lariat acceptor site. The 5' flanking sequences are also A+T rich and contain an oligo A tract (-14 to -24), a TATA box (-25 to -32), and a short G+C rich region (-63 to -101) which may be a control region. From -196 to -209 is a sequence AAAGTAGTATTCAA which matches in 11 of its 14 nucleotides, a sequence found upstream from the hormonally regulated P-enolpyruvate carboxykinase gene of rat.

INTRODUCTION

Dictyostelium discoideum represents one of the simplest organisms to investigate the regulation of eukaryotic gene expression during differentiation. Its developmental cycle is initiated when the unicellular amoeboid organisms aggregate under the chemotactic influence of cAMP to form multicellular slugs of approximately 10^5 cells. The formation of the multicellular aggregate involves specific cell-cell contacts which have been shown to induce expression of genes whose products accumulate at later stages of development. In addition to its role as a chemotactic agent, extracellular cAMP acts to regulate the expression of several genes at the level of gene induction and mRNA stability (1-6).

To study the mechanisms of temporal, cAMP, and cell contact mediated regulation of gene expression, we have focused on analyzing the expression of uridine diphosphoglucose pyrophosphorylase (UDPGP)(UTP: γ -D-glucose-1-

phosphate uridylyltransferase, EC 2.7.7.9), a well studied, enzyme in Dictyostelium. We chose this enzyme because UDPGP mutants in Dictyostelium discoideum abort the developmental cycle, suggesting that the enzyme is essential for the completion of development (7,8). It has long been known that cell contact plays a role in the regulation of specific activity of this enzyme (5). Although many other genes are regulated by cell contact and cAMP in Dictyostelium and several have been cloned, few have been identified. Discoidin is the only one for which mutants may have been isolated. We have previously shown that there are two mRNAs (UDPGP1,2) that encode several differentially regulated isoforms of the enzyme (9,10). By using the cDNA as a probe, we have also shown by Northern blot analysis that the ten fold rise in enzyme specific activity observed during development can be accounted for by the increase in UDPGP1 mRNA levels (10). We have recently demonstrated that both cell-cell contact and cAMP coordinately regulate the accumulation of the two UDPGP mRNAs (11). Disaggregation of slugs results in a rapid loss of UDPGP mRNAs. Addition of cAMP to disaggregated cells elevates the level of UDPGP mRNAs, although exogenous cAMP does not restore UDPGP mRNAs to their original level. Other factors may be required for maximal expression of the UDPGP1 gene. Unaggregated single cells, starved and shaken rapidly in suspension, do not accumulate UDPGP mRNAs. However, addition of cAMP to these cells caused UDPGP mRNAs to accumulate, suggesting that the requirement for cell-cell contact could be bypassed in part by cAMP addition. Furthermore, we have recently shown that exogenous cAMP analogues induce UDPGP1 expression and that of other genes in the same relative order of potency that the analogues bind to the cell surface cAMP receptor (12). This result suggests that the induction of UDPGP1 and other Dictyostelium genes by exogenous cAMP is mediated through the cell surface receptor, which may exert its effect on gene expression by transmembrane signal transduction using second messenger(s) (12,13). The nature of the second messenger(s) is unknown. The central role of UDPGP in Dictyostelium development, its importance as a marker for development and now for signal transduction, and the availability of mutations affecting UDPGP expression, make it an important gene to characterize at the nucleotide sequence level.

As a first step towards utilizing the available mutants in elucidating the mechanisms involved in the temporal and cAMP regulated expression of this gene, we isolated and characterized the UDPGP1 gene and determined the nucleotide sequence of both the UDPGP1 genomic and cDNA clones. The results show the presence of A+T rich flanking sequences and three A+T rich introns.

In the 5' flanking sequence there is a short G+C rich region which may have regulatory significance. A short 14 nucleotide sequence from -196 to -209 matches 11 nucleotides of a putative consensus cAMP response sequence in the hormone regulated P-enolpyruvate carboxykinase gene of rat (14). Since extracellular cAMP appears to behave as a hormone in Dictyostelium, similarities between peptide hormone action in higher eukaryotes and exogenous cAMP action in Dictyostelium are intriguing.

MATERIALS AND METHODS

DNA and RNA Isolation

Dictyostelium discoideum strain AX3 was used for all experiments. Conditions for growth and differentiation as well as the isolation of RNA are as described previously (8-10). To prepare nuclear DNA, cells were washed in 0.2% NaCl at 4°C, resuspended in ethidium bromide (300 µg/ml, 2.5×10^5 cells/ml) and lysed with 1.45% Triton X100 and Chemusol NP12 (Rhône-Poulenc). Nuclei were washed in one volume of the same solution. The nuclear pellet was resuspended at 2×10^6 nuclei/ml in NB [Tris·Cl (25 mM), Mg acetate (5 mM), EDTA (0.5 mM), sucrose (5%) pH 7.6]. EDTA was added to 250 mM at room temperature. Nuclei were gently lysed at 55° with 5% sarkosyl, and 1 gm/ml CsCl was added immediately. The refractive index was adjusted to 1.3965 and the nuclear DNA was banded by centrifugation at 40 K rpm for 40 hr in a Beckman Ty65 rotor.

Dictyostelium DNA Cloning Procedures

Restriction enzymes were purchased from BRL and New England Biolabs. DNA ligase, T₄ polynucleotide kinase, DNA polymerase, and its Klenow fragment were from BRL. DNase I, calf intestinal alkaline phosphatase and AMV reverse transcriptase were from Boehringer Mannheim. Radioactive nucleotides were from either Amersham Corp. or New England Nuclear. The M13 cloning and sequencing reagents were from either P-L Biochemicals or Bethesda Research Laboratories. The cloning methods used were basically as described by Maniatis et al. (15). One genomic library of Dictyostelium nuclear DNA in lambda was constructed and generously made available by Dr. Daphne Blumberg. The library was constructed using Dictyostelium DNA which was partially digested with Eco RI and size fractionated on a sucrose gradient. The DNA fragments were ligated into λgt wes-λB in place of the stuffer fragment of that vector. A second genomic library was constructed in the plasmid vector pUC12 by ligating a Bgl II, Hind III total double digest of genomic DNA to pUC12 that had been digested with Bam HI and Hind III. To enrich for clones

containing the UDPGP1 5' genomic fragment, the ligation products were then digested with Xho I and Pst I before being transfected into a recA host.

Plaque and Colony Screening, and DNA Blot Hybridization

Hybridization probes (10⁷ cpm/μg) were usually prepared by nick translation of isolated fragments from cDNA subclones (16). Occasionally probes were made by synthesizing the complementary strand of a recombinant M13 template. In these cases, the synthesis conditions used were the same as for dideoxy sequencing except that the ddNTPs were omitted, the cold dNTPs were used at 0.1 mM, and the reaction contained 100 μCi of [α-³²P]dATP. The labelled double-stranded region of interest was excised with the appropriate restriction enzymes and subsequently purified on an agarose gel.

Lambda plaques were transferred to either nitrocellulose (Schleicher & Schuell or Millipore Corp.) or nylon membranes (Biodyne) and screened by the method of Benton and Davis (17). Colonies were transferred to nylon membranes while still small (<1 mm) and then amplified on plates containing chloramphenicol (250 μg/ml) overnight. The screening was performed basically as described by Maniatis *et al.* (15). Restriction digests were fractionated by electrophoresis on agarose gels, transferred to nitrocellulose or nylon membranes, and hybridized to ³²P-labeled DNA probes as described by Southern (18).

DNA Sequencing and Primer Extension

DNA sequencing was usually carried out by the chain termination method of Sanger *et al.* (19,20) using the M13 vectors mp8,9,18, and 19. We used the chemical cleavage method of Maxam and Gilbert in cases where dideoxy sequencing reactions did not work because the highly A+T rich DNA caused nonspecific termination at oligo A or oligo T tracts. Ten nanograms of a [³²P] 5' end labeled (10⁶ cpm) synthetic oligonucleotide (5'-GATCCAGTTGATTGT-3'), complementary to a sequence in the UDPGP1 mRNA 49 nucleotides downstream from the initiating ATG, was annealed to 50 μg of total RNA from cells developed for 18 hr. The annealing reaction was performed in 10 μl of 10 mM Tris, pH 8.3, at 37°C, 2 mM MgCl₂ by heating at 75°C for 15 min and then cooling slowly to room temperature in a 13 x 100 mm test tube filled with water initially at 75°C. This reaction was incubated with 12.5 units of AMV reverse transcriptase (1000 U/ml) with or without Actinomycin D (25 μg/ml) in 4 mM DTT, 10 mM MgCl₂, 50 mM Tris, pH 8.3 at 37°C for 60 min at 42°C in the presence of 0.5 mM dNTPs.

RESULTS

Structure and Sequence of the UDPGP1 cDNA

We have previously reported the cloning of the UDPGP1 cDNA by dG·dC tailing into the Pst I site of pBR322 (10). A restriction map of the cDNA was constructed and fragments of it were subcloned into M13 vectors for sequencing by the dideoxy method of Sanger *et al.* (19,20). The restriction map and the sequencing strategy employed are shown in Fig.1. The direction of transcription had been determined by hybrid arrested translation using single strands of UDPGP1 cDNA (10) and was confirmed by hybridization of single-stranded probes to northern blots (data not shown).

The cDNA, excluding the tails, is 1,583 nucleotides long (Fig. 2). It contains a single open reading frame which is 1,525 nucleotides long and

Fig. 1. Structure of the UDPGP1 cDNA and genomic DNA. A restriction map of the 1.65 Kb cDNA insert in the Pst I site of pBR322 is shown. The cDNA map is aligned with the genomic restriction map, and some of the corresponding restriction sites in the two DNAs are indicated with dotted lines. The genomic map spans approximately 13 Kb. Restriction sites are designated as follows: B, Bgl II; D, Dra I; H, Hind III; N, Nsi I; P, Pst I; R, Eco RI; Rs, Rsa I; S, Sau3A I; Sc, Sca I; V, Eco RV. The sequencing strategy utilized for the two DNAs is indicated. Arrows indicate the direction of the sequencing reaction and the amount of information obtained from a given clone or restriction fragment. The 5' and 3' notations indicate the direction of transcription. The genomic segments contained within the various subclones are indicated in the bottom of the figure. The displayed map distances are in Kbp. Only those Rsa I, Sau3A I, and Dra I sites pertinent to the sequencing strategy are shown.

Sau 3A I

ATG ACA GAT ACA GCA ACA TCA AAA GCA ACA GTT GAA AGA CCA AAA TTA CAA TCA ACT CGA TCA TTA CAT AGT TTA
 MET Thr Asp Thr Ala Thr Ser Lys Ala Thr Val Glu Arg Pro Lys Leu Gln Ser Thr Gly Ser Leu His Ser Leu
 Dra I

TTT AAA GAT GTT GAT TTA TTT TCA GAG AAT CAT GAA GAA TTA TAT CCA CCA CTT CAA CAT GGT GCA AGA TTT GCA
 Phe Lys Asp Val Asp Leu Phe Ser Glu Asn Asp Glu Glu Leu Tyr Pro Pro Leu Gln His Gly Ala Arg Phe Ala

GCA CCA ATT GAA GAT AGT ACA TTA TTA GCA TTG GGT ATG AAA CCA GAT GAA CTT AAA GCA TTC CAA AAA CAA AGA
 Ala Pro Ile Glu Asp Thr Leu Leu Ala Leu Gly MET Lys Pro Asp Glu Leu Lys Ala Phe Gln Lys Gln Arg
 Nsi I

CAT GCA TAC ATT AAC AAG GAT CAA ATT TAC ACT GAT GAA ATT AAA ATT CCA AAT AAA ACT GAA ATG GTA GAT TAT
 His Ala Tyr Ile Asn Lys Asp Gln Ile Tyr Thr Asp Glu Ile Lys Ile Pro Asn Lys Thr Glu MET Val Asp Tyr

CAT CAA CTT CAT TTA GTC TCA CCA ATT GAC CAA TCA AAT GCT TGC AGA TTA TTA AAT AAA TTA GTT GTA ATT AAA
 His Gln Leu His Leu Val Ala His Ile Glu Gln Ile Asn Gln Ser Asp Tyr Asn Val Asp Val Pro Leu Val Ile Lys
 Rsa I

TTA AAT GGT GGT CTT GGT AAT AGT ATG GGT TGT AAA ACT GCT AAA ACT ACA ATG GAA ATA GCT CCA GGT GTT ACT
 Leu Asn Gly Gly Leu Gly Asn Ser MET Gly Cys Lys Thr Ala Lys Ser Thr MET Glu Ile Ala Pro Gly Val Thr

TTT TTA GAT ATG GCA GTT GCT CAT ATT GAA CAA ATT AAT CAA GAT TAT AAT GTT GAT GTC CCA TTG GTT ATT ATG
 Phe Leu Asp MET Ala Val Ala His Ile Glu Gln Ile Asn Gln Asp Tyr Asn Val Asp Val Pro Leu Val Ile MET
 Eco RI

AAT TCT TAT AAA ACT CAT AAT GAA ACT AAT AAG GTT ATT GAA AAG TAT AAA ACT CAT AAA GTT AGT ATT AAA ACT
 Asn Ser Tyr Lys Thr His Asn Glu Thr Asn Lys Val Ile Glu Lys Tyr Lys Thr His Lys Val Ser Ile Lys Thr

TTC CAA CAA TCA ATG TTC CCA AAG ATG TAT AAA CAT ACA TTA AAT TTA GTA CCA AAA CCA AAT ACA CCA ATG AAT
 Phe Gln Gln Ser MET Val Ser Pro Lys MET Tyr Lys Asp Thr Leu Asn Leu Val Pro Lys Pro Asn Thr Pro MET Asn
 Eco RV

CCA AAG GAA TGG TAT CCA CCA GGT TCA GGT CAT ATC TTT AGA TCA CTC CAA AGA TCT GGT TTG ATT CAT GAA TTT
 Pro Lys Glu Trp Tyr Pro Pro Gly Ser Gly Asp Ile Phe Arg Ser Leu Gln Arg Ser Gly Leu Ile Asp Glu Phe
 Bgl II

TTA GGT GCT GGT AAA GAA TAT ATT TGC ATT TCA AAT GTT CAA AAT TTA GGT TCA ATA ATT GAT CTT CAG GTA TTA
 Leu Ala Ala Gly Lys Glu Tyr Ile Phe Ile Ser Asn Val Glu Asn Leu Gly Ser Ile Ile Asp Leu Gln Val Leu

AAT CAT ATT CAT TTG CAA AAG ATT GAA TTT GGT TTA GAA GTC ACA AAT CGT ATT AAT ACT GAT TCA ACT GGT GGT
 Asn His Ile His Leu Gln Lys Ile Glu Phe Val Thr Asn Thr Asp Ser Asn Leu Ile Asn Thr Asp Thr Gly Gly
 Bgl II

ATT TTA ATG TCA TAT AAA GAT AAA CTT CAT CTT TTG GAA TTA TCT CAA GTT AAA CCA GAG AAA TTA AAG ATC TTT
 Ile Leu MET Ser Tyr Lys Asp Lys Leu His Leu Leu Glu Leu Ser Gln Val Lys Pro Glu Lys Leu Lys Ile Phe

AAA GAT TTT AAA CTT TGG AAT ACA AAT AAT ATT TGG GTT AAT TTG AAA TCA GTT TCA AAT TTA ATT AAA GAA GAT
 Lys Asp Phe Lys Leu Thr Asn Thr Asn Asn Ile Trp Val Asn Leu Lys Ser Asn Leu Ile Lys Thr Thr Gly Asp
 Eco RI

AAA TTA GAT TTA GAT TGG ATT GTT AAT TAT CCA CTT GAA AAT CAT AAA GCA ATG GTA CAA TTA GAA ACA CCA GCA
 Lys Leu Asp Leu Asp Trp Ile Val Asn Tyr Pro Leu Glu Asn His Lys Ala MET Val Gln Leu Glu Thr Pro Ala
 Eco RV

GGT ATG GGT ATT CAA AAT TTT AAG AAT TCA GTT GCA ATT TTT GTA CCA CGT GAT AGA TAT CGT CCA ATT AAA TCA
 Gly MET Gly Ile Gln Asn Phe Lys Asn Ser Val Ala Ile Phe Val Pro Arg Asp Arg Tyr Arg Pro Ile Lys Ser

ACA AGT CAA TTA TTG GTT GCA CAA TCA AAT ATT TTG CAA TTT CAT CAT GGT CAA GTT AAA TTA AAT TCA AAG AGA
 Thr Ser Gln Leu Leu Val Ala Gln Ser Asn Ile Phe Gln Phe Asp His Gly Gln Val Lys Leu Asn Ser Lys Arg

GAA GGT CAA GAT GTA CCA CTT GTT AAA TTG GGT GAA GAA TTT TCA ACA GTT TCA GAT TAT CAA AAG AGA TTT AAA
 Glu Gly Gln Asp Val Pro Leu Val Lys Leu Gly Glu Glu Phe Ser Thr Val Ser Asp Tyr Glu Lys Arg Phe Lys

TCA ATT CCA GAT TTA TTG GAA TTG GAT CAT CTT ACT GTT TCT GGT GAT GTT TAC TTT GGT TCA AGA ATT ACT CTT
 Ser Ile Pro Asp Leu Leu Glu Leu Asp His Leu Thr Val Ser Gly Asp Val Tyr Phe Gly Ser Arg Ile Thr Leu

AAA GGT ACA GTC ATT ATT GTA GCT AAT CAT GGT GAA CGT GTT GAT ATT CCA GAT GGT GTG GTT TTA GAA AAT AAA
 Lys Gly Thr Val Ile Ile Val Ala Asn His Gly Glu Arg Val Asp Ile Pro Asp Gly Val Val Leu Glu Asn Lys
 Sca I

GTA CTT TCT GGC ACT CTT AGA ATT TTG GAT CAT TAA att cta ctc aaa aaa tta att ggt caa aaa aaa aaa
 Val Leu Ser Gly Thr Ser Arg Ile Leu Asp His *
 aaa aaa aaa aaa aaa aaa a

Fig. 2. Nucleic acid sequence of the UDPGP1 cDNA and deduced amino acid sequence. The 1,583 bp cDNA sequence, excluding the dG•dC tails is shown. It begins with the first nucleotide to the right of the open triangle and ends with the terminus of the poly A tract. Only the sequence of the RNA sense strand is shown. Those restriction sites determined by restriction analysis are shown above the sequence and the recognition sequence underlined. The deduced amino acid sequence is shown immediately below the nucleic acid sequence. The identity of the four N terminal-most amino acids (the initiating Met, Thr, Asp, Thr) was deduced from the genomic DNA sequence. The termination codon is indicated by an asterisk below it. Solid triangles indicate the position of the intervening sequences in the genomic DNA.

terminates with a TAA codon. However, it lacks an initiating ATG and a 5' untranslated region. The 3' untranslated region is 25 nucleotides long and is followed by a 31 nucleotide long polyA tract. A canonical polyadenylation signal, AATAAA, is not found upstream from the polyA tract. A sequence, AAAAAA, which differs from it by a single nucleotide, is located 11 nucleotides 5' to the beginning of the polyA tract.

The portion of the open reading frame that is present in the cDNA is sufficient to encode a 57,373 d. polypeptide, which is close to the published molecular weight of UDPGP1, suggesting that only a small portion of the NH₂ terminal coding sequences are absent from the cDNA. Several interesting structural features of the polypeptide are revealed by analysis of the deduced amino acid sequence. Among these is the pattern of Lys usage: Twenty nine of forty five Lys residues occur in pairs or triplets with the structure Lys-(X)₁₋₄, -Lys-(X)₁₋₃-Lys. In one region of the polypeptide Lys residues constitute 5/14 amino acids, and in a second region 5/12 amino acids. There are 24 Pro residues, 6 of which occur within a 15 amino acid stretch in approximately the middle of the protein. There is only one Cys residue present in the predicted amino acid sequence, thus excluding the possibility of any intramolecular disulfide bridges in the UDPGP1 polypeptide.

Isolation and Structure of UDPGP1 Genomic Sequences

Using the UDPGP1 cDNA and subclones of cDNA fragments as hybridization probes, a restriction map of the gene was constructed by hybridization to Southern blots of genomic DNA (Fig. 1). Our data showed that a single copy of the gene hybridized to this cDNA and spanned three Eco RI fragments; 2.5, 0.9, and 9.5 Kb in size. Probes from the 5' end of the cDNA hybridized only to the 2.5 Kb genomic DNA fragment while probes from the 3' end of the cDNA hybridized only to the 9.5 Kb genomic fragment. The presence of a Hind III site in the genomic DNA that was absent in the cDNA suggested that the UDPGP1 gene contained an intervening sequence.

Using the UDPGP1 cDNA as a probe, a λ gt wes- λ B genomic library of nuclear DNA was screened by the method of Benton and Davis (17). After several rounds of screening, a clone, designated λ 3, was isolated and found to contain the 0.9 and 9.5 Kb Eco RI genomic fragments that constitute the central portion and the 3' end of the gene respectively (see Fig. 1). The 2.5 Kb Eco RI fragment containing the 5' end of the gene was subsequently isolated as part of a 2.9 Kb Hind III-Bgl II fragment from a genomic library constructed in the plasmid vector pUC12. This clone was designated pUC125.

The 0.9 Kb Eco RI fragment from λ 3 was subcloned into the M13 vector mp8.

This clone was designated mpR900. A 2.5 Kb portion of the 9.5 Kb Eco RI fragment, which contained all the 3' end sequences complementary to the cDNA, was subcloned into pUC12 as part of a 2.7 Kb Hind III-Bgl II fragment that overlaps at its 5' end sequences contained in the 0.9 Kb central Eco RI fragment. This clone was designated pUC123 (Fig. 1).

Restriction analysis of the genomic subclones and the cDNA revealed the presence of three small 100-200 bp introns within the translated regions of the UDPGP1 gene. These introns are located between the Nsi I and 5' Eco RI sites, the two central Bgl II sites, and the Bgl II and 3' Eco RI sites. DNA sequencing confirmed the presence of these introns.

Genomic sequences were determined by a combination of the dideoxy chain termination method and the chemical cleavage method of Maxam and Gilbert (21). The sequencing strategy is shown in Fig 1. The sequence composition of all three introns is >90% A+T, and the introns obey the GT/AG rule. The region of conservation of the donor sequence in the introns is actually somewhat longer, the hexanucleotide GTAAGT being present at the beginning of all three introns. The second intron contains a sequence, TAACTAAC, which is strikingly similar to the highly conserved TACTAAC box found in yeast introns (22) and is present 36 nucleotides before the splice site in the same relative position within that UDPGP1 intron as the conserved sequence found in yeast. This sequence is known to be involved in lariat formation during RNA splicing in yeast. Interestingly, the terminal AG dinucleotide of the third intron is the internal AG of the genomic Hind III site (AAGCTT), and thus the splicing event per se accounts for the absence of the Hind III site within the cDNA (Fig. 3). The first intron is located between codons 160 and 161, while the second intron splits codon 273 between the first and second nucleotide, and the third intron lies between codons 361 and 362 (Fig. 2). Partial sequencing of the pUC123 subclone demonstrated that there are no introns present in the gene downstream from the 3' Sca I site.

Partial sequencing of pUC125 provided the identity of the three N terminal-most amino acids and the initiating ATG. This ATG is preceded by thirteen A nucleotides, which is a pattern that has been observed with several other Dictyostelium translational start sites (23-26). A sequence, TATAAAAA, conforming to the consensus TATA box in Dictyostelium (TATAAAA^A)(25) was found about 80 nucleotides upstream from the initiating ATG and, as we shall show, between -25 and -32 from the start of transcription. For more than 300 bp upstream from the site of transcriptional initiation, the sequence is only 10% G+C, except for a region between -63 and -101 bp which is 45% G+C.

Fig. 3. Genomic DNA Sequence. Translated sequences are shown in upper case letters. The G+C rich region in the 5' flanking sequence is underlined. The so-called cAMP responsive sequence AAGTAGTATTCAA is boxed. The location of the TATAAAA sequence is underscored by a dashed line. The sites of transcriptional initiation are indicated by an arrow beneath the sequence. The first nucleotide of the shortest transcript is arbitrarily designated as the +1 nucleotide. The initiating ATG and the three N terminal-most amino acids are shown. The C terminal amino acid and termination codons are shown. The first poly A tract downstream from the termination codon corresponds to the poly A tract in the cDNA. The sequence of all three introns and their flanking codons are shown. The conserved sequences at the termini of the introns are underlined. The TAACTAAC box within the second intron is underscored with a dotted line.

Most of the G+C rich region consists a trimer ATC, which is repeated seven times in the sequence. Another noteworthy feature of this region is the sequence asymmetry of the strands; with the exception of one nucleotide, all of the G nucleotides are located on one strand and all the C nucleotides are on the other strand (Fig. 3). The considerably higher G+C content of this region and its unusual structure may signify that it has a role in gene regulation. A second sequence AAAGTAGTATTCAA (boxed) matches in 11 of its 14 nucleotides the sequence AAAGTTTAGTCAA in the hormone regulated rat PEP carboxykinase gene (14) which is postulated to be a cAMP responsive sequence

Fig. 4. Primer extension reaction products. A [32 P] 5' end labeled synthetic oligonucleotide (5'-GATCCAGTTGATTGT-3') complementary to a sequence in the UDPGP1 mRNA 49 nucleotides downstream from the initiating ATG was annealed to *Dictyostelium* 18 hr total RNA and extended with AMV reverse transcriptase in the presence (lane a) or absence (lane b) of Actinomycin D. The reaction products were ethanol precipitated, resuspended in 95% formamide, 10 mM EDTA, 0.1% bromophenol blue and xylene cyanol, and run adjacent to a DNA sequencing ladder of the 2.2 Kb Sau3A I-Hind III coding strand of genomic DNA which is labeled at the same terminal 5' G nucleotide as the primer used in the extension reaction. The sequence of this strand is shown alongside the gel. The 5' to 3' orientation of the sequence is indicated. Adjacent to this sequence is the complementary sequence of the DNA sense strand. Arrows indicate the nucleotide at which the extension products terminate.

only on the basis of its homology to sequences upstream from other cAMP regulated genes in prokaryotes and eukaryotes.

A 38 nucleotide poly A tract encoded in the genome occurs 25 nucleotides downstream from the translation termination site. The 3' end of the cDNA and genomic DNA are co-linear (Fig. 3).

Mapping the Site of Transcriptional Initiation

The DNA primer used for primer extension was a 5' end labeled synthetic pentadecanucleotide complementary to sequences near the 5' end of the mRNA beginning at the first Sau3A I site in the cDNA (Fig. 4). The template was total RNA from cells developed for 18 hrs. The reaction products were denatured and electrophoresed on a 7 M urea gel adjacent to a Maxam-Gilbert sequencing ladder of the 2.2 Kb Sau3A I-Hind III coding strand of genomic DNA which is 5' end labeled at the Sau3A I site, i.e. at the same position as the primer used in the extension reaction. This strategy allows direct determination of the complementary nucleotide sequence at which the mRNA is initiated. A doublet was observed corresponding to two T residues 49 and 50 nucleotides from the ATG (Fig. 4). This observation maps the major initiation product of transcription to a position about 49 nucleotides upstream from the A in the initiating ATG (Fig. 3). A fainter doublet occurs 53 and 54 nucleotides from the A in ATG. S1 nuclease analysis was also performed to map the site of transcriptional initiation. In general, the S1 nuclease resistant fragments were somewhat smaller than the products of the primer extension experiment (result not shown). The size of the resistant fragments was found to be inversely proportional to the temperature at which the S1 reaction was performed. This is presumably a consequence of breathing of the RNA-DNA hybrid at the A+T rich end of the mRNA.

DISCUSSION

UDPGP is a well studied enzyme that has long been used by developmental biologists as a marker for differentiation in Dictyostelium. The enzyme is essential for development and the UDPGP mutants that have been isolated all abort the developmental cycle at the same stage. We have recently shown that the expression of the UDPGP gene is regulated by cAMP through signal transduction (12,29). Here, we have described the cloning and sequencing of a cDNA complementary to the UDPGP1 mRNA, and the genomic DNA that encodes this transcript and its flanking sequences. These sequences will be useful in characterizing the UDPGP mutants and elucidating the mechanisms involved in

the regulation of UDGP1 gene expression. The cDNA contains a single open reading frame that is sufficient to encode nearly the entire UDGP1 polypeptide. Interesting structural features of the polypeptide were predicted from the deduced amino acid sequence. The polypeptide contains 49 Lys residues which are distributed in clusters made up of simple repeating units. The presence of a cluster of Pro residues in the middle of the polypeptide suggests that the protein may be divided into two large domains joined by a Pro rich, random coil joining segment.

Comparison of the two nucleic acids revealed the presence of three introns within the coding sequences of the genomic DNA. Like other introns sequenced from Dictyostelium they are 90% A+T, and are relatively short when compared to other eukaryotic introns. They all obey the GT/AG rule for intervening sequences, and actually contain a slightly longer conserved sequence, GTAAGT, at the 5' splice site. This level of conservation is unprecedented in Dictyostelium, but its functional or regulatory significance is unknown. The site at which lariat formation occurs is a highly conserved sequence (TACTAAC) in yeast (22). A variant of that consensus sequence is present in only one of the three introns (intron II) of this Dictyostelium gene. The presence of the initial GT and terminal AG dinucleotides suggest that intervening sequences may be processed from the UDGP1 mRNA via a lariat intermediate, as has been shown in other eukaryotes.

By utilizing the method of primer extension, we mapped the site of transcriptional initiation to approximately 50 nucleotides upstream from the initiating ATG. Although S1 nuclease mapping resulted in slightly smaller fragments the results are consistent with the conclusions drawn from the primer extension experiments.

Primer extension of the UDGP1 mRNA produced four products. A possible, though unsubstantiated, explanation for such microheterogeneity in the case of the UDGP1 mRNA can be seen upon examination of the sequences immediately upstream from the consensus TATA box. Since only the ATA at the second, third, and fourth positions of the TATA consensus sequence are invariant in Dictyostelium (25), it is possible to actually derive three additional TATA boxes from the sequence immediately upstream from the consensus TATA box. If this were the case, the four TATA boxes would overlap one another, each being staggered by two nucleotides from the other (TATATATATAAAAA). Since the position of the TATA box presumably determines the site of transcriptional initiation in eukaryotes, this duplication of TATA boxes could account for the microheterogeneity seen at the 5' end of the UDGP1 mRNA. Interestingly, the

most prominent of the primer extension products is the smallest one, which implies that the shortest UDPGP1 mRNA is the most abundant mRNA species. The transcriptional initiation of this message is presumably directed from the 3' most TATA box, i.e. the one that most closely matches the consensus sequence.

Dictyostelium genes in which a canonical polyadenylation signal cannot be readily identified have been reported previously (27). In the case of the Dictyostelium ras gene, as with the UDPGP1 gene, the absence of a canonical polyadenylation signal is accompanied by the presence of a poly A tract within the genome immediately downstream from the 3' untranslated region. The sequence, AAAAAA, is located where one would expect a polyadenylation signal to be found, but differs from the canonical sequence by a single nucleotide. Whether the poly A tail on the mRNA transcribed from these genes is derived from the encoded sequence in the genome, added post-transcriptionally, or both will require further investigation.

Unlike all other Dictyostelium genes that have been sequenced (23-26), the UDPGP1 gene does not have an oligo T stretch in the RNA sense strand of the gene just upstream from the site of transcriptional initiation. Rather, it has an oligo A stretch on that strand and of course an oligo T stretch on the complementary strand, suggesting that if the oligo T is important its orientation is not. The implication is that if the oligo T sequences do play a role in promoter function, they are capable of functioning bidirectionally. This hypothesis awaits testing in a Dictyostelium DNA transfection system. In addition to having a TATA box and an "oligo T element" that are not separated from one another, the UDPGP1 gene also differs from other known Dictyostelium genes in not having G residues flanking its TATA box. In fact, there are only six G residues in the 300 nucleotides that have been sequenced upstream from the TATA box, and the first of these occurs about 65 nucleotides upstream from it. A consensus TATA box (TATAAAAA) is found immediately after the oligo A tract between -25 and -32 (25,26). In addition, there is no CAAAT sequence in the 5' untranslated region of the mRNA. Flanking sequences upstream from the gene, like the introns, are 90% A+T. The presence of a G+C rich domain at -63 to -100 within this 90% A+T region is intriguing. Since UDPGP1 is developmentally regulated and its expression is modulated by cAMP, this G+C rich sequence may represent one of several possible regulatory regions that are presumably upstream from the gene. A G+C rich sequence upstream from the cysteine proteinase II gene of Dictyostelium has been identified by Datta *et al.* (28). There is evidence that this sequence coincides with a nuclease hypersensitive site (Pavlovic and Parish, personal communication).

Our previous experiments suggested that extracellular cAMP regulates gene expression by binding to a cell surface receptor and inducing the synthesis of an intracellular second messenger. Based on the pharmacological specificity of the target receptor molecule which is involved in gene expression, we and other workers concluded that the cAMP dependent protein kinase is not the direct target for extracellular cAMP (12,13). The nature of the second messenger(s) remain unknown. We have recently proposed a model to account for our current observations (29).

Peptide hormones are thought to regulate gene expression in higher eukaryotes by inducing the synthesis of intracellular second messengers such as cAMP. Hanson and Reich have independently suggested that intracellular cAMP may act directly via a "CAP-like" DNA binding protein to activate transcription (30,31). They have identified sequences upstream from several hormonally regulated genes that are partially homologous to the *E. coli* DNA CAP binding site. The functions of these sequences are being elucidated in other laboratories. The sequence identified upstream of the rat PEP carboxykinase gene by Hanson is conserved in the UDPGP1 gene and is located at a similar position (30). Though UDPGP1 is regulated by extracellular cAMP and it appears unlikely that intracellular cAMP is also involved in regulating its expression (13). Whether that sequence is required for cAMP regulation expression in rat PEP CK gene is also obscure. Therefore, the observation is interesting, but its significance remains to be determined by mutational analysis. The availability of mutants altered in the cAMP stimulus/response system, and a well-characterized receptor system should make *Dictyostelium* a useful model for studying how extracellular molecules regulate gene expression via transmembrane signal transduction.

ACKNOWLEDGEMENTS

We thank Dr. Daphne Blumberg for providing assistance in screening the lambda genomic library. We appreciate the help of Dorothy Regula and Jerry Keightley in preparing the manuscript. This work was submitted by J.A.R. in partial fulfillment of the requirements for a Ph.D. This work was supported by NIH GM2730. This publication is number XXXX from the Department of Biology.

*To whom reprint requests should be sent

REFERENCES

1. Blumberg, D. D., Margolskee, J. P., Barklis, S. N. Chung, S. N. Cohen, N. S. and Lodish, H. F. (1982) Specific cell-cell contacts are

- essential for induction of gene expression during differentiation of Dictyostelium discoideum. Proc. Natl. Acad. Sci. U.S.A. 79:127-131.
2. Chisholm, T., Barklis, E. and Lodish, H. F. (1984) Mechanism of sequential induction of cell-type specific mRNAs in Dictyostelium differentiation. Nature (London) 310:67-69.
 3. Chung, S., Landfear, S. N., Blumberg, D. D. Cohen, N. S. and Lodish, H. F. (1981) Synthesis and stability of developmentally regulated Dictyostelium mRNAs are affected by cell-cell contact and cAMP. Cell 24:785-797.
 4. Mehdy, M. C., Ratner, D. and Firtel, R. A. (1983) Induction and modulation of cell type specific gene expression in Dictyostelium. Cell 32: 763-771.
 5. Newell, P. C., Longlandt, M. and Sussman, M. (1971) Control of enzyme synthesis by cellular interaction during development of the cellular slime mold Dictyostelium discoideum. J. Mol. Biol. 58:541-554.
 6. Williams, J. G., Lloyd, M. M. and Devine, J. M. (1979) Characterization and transcription analysis of a cloned sequence derived from a major developmentally regulated mRNA of D. discoideum. Cell 17:903-913.
 7. Diamond, R. L., Farnsworth, P. A. and Loomis, W. F. (1976) Isolation and characterization of mutants affecting UDPG pyrophosphorylase activity in Dictyostelium discoideum. Dev. Biol. 50:169-181.
 8. Sussman, M., and Osborn, M. H. (1964) UDP-Galactose polysaccharide transferase in the cellular slime mold, Dictyostelium discoideum: Appearance and disappearance of activity during cell differentiation. Proc. Natl. Acad. Sci. USA 52:81-87.
 9. Fishel, B. R., Manrow, R. E. and Dottin, R. P. (1982) Developmental regulation of multiple forms of UDP glucose pyrophosphorylase of Dictyostelium. Dev. Biol. 92:175-187.
 10. Fishel, B. R., Ragheb, J. A., Rajkovic, A., Haribabu, B. Schweinfest, C. W. and Dottin, R. P. (1985) Molecular Cloning of a cDNA Complementary to a UDP-Glucose Pyrophosphorylase mRNA of Dictyostelium discoideum. Dev. Biol. 110:369-381.
 11. Haribabu, B., Rajkovic, A. and Dottin, R. P. (1985) Cell-Cell contact and cAMP regulate the expression of a UDP glucose pyrophosphorylase gene of Dictyostelium discoideum. Dev. Biol. 113:436-442.
 12. Haribabu, B. and Dottin, R. (1986) Pharmacological characterization of cyclic AMP receptors mediating gene regulation in Dictyostelium discoideum. Mol. Cell. Biol. 6:2402-2408.
 13. Oyama, M. and Blumberg, D. (1986) Interaction of cAMP with the cell surface receptor induces cell-type-specific mRNA accumulation in Dictyostelium discoideum. Proc. Natl. Acad. Sci. U.S.A. 83:4819-4823.
 14. Wynshaw-Boris, A., Lugo, T., Short, J., Fournier, R. and Hanson, R. (1984) Identification of a cAMP regulatory region in the gene for rat cytosolic phosphoenolpyruvate carboxykinase (GTP). J. Biol. Chem. 259:12161-12169.
 15. Maniatis, T., Fritsch, E. and Sambrook, J. (1982) Molecular cloning. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
 16. Rigby, P., Dieckmann, M., Rhodes, C. and Berg, P. (1977) Labelling deoxyribonucleic acid to high specific activity *in vitro* by nick translation with DNA polymerase I. J. Mol. Biol. 113:237-251.
 17. Benton, N., and Davis, R. W. (1977) Screening λ gt recombinant clones by hybridization to single plaques *in situ*. Science 196:180-183.
 18. Southern, E. (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. 98:503.
 19. Sanger, F., Nichen, S. and Colsen, A. (1977) DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA 74:5463-5467.

20. Kwiatkowski, R. W., Schweinfest, C. W. and Dottin, R. P. (1984) Molecular cloning and the complete nucleotide sequence of the creatine kinase-M cDNA from chicken. *Nucleic Acids Res.* **12**:6925-6934.
21. Maxam, A., and Gilbert, W. (1980) Sequencing end-labeled DNA with base-specific chemical cleavages. *Methods Enzymol.* **65**:499-560.
22. Rymond, B. C. and Rosbash, M. (1985) Cleavage of 5' splice site and lariat formation are independent of 3' splice site in yeast mRNA splicing. *Nature (London)* **317**:735.
23. Barklis, E., Pontius, B., Barfield, K. and Lodish, H. F. (1985) Structure of the promoter of the Dictyostelium discoideum prespore EB4 gene. *Mol. Cell. Biol.* **5**:1465-1472.
24. Barklis, E., Pontius, B. and Lodish, H. F. (1985) Structure of the Dictyostelium discoideum prestalk D11 gene and protein. *Mol. Cell. Biol.* **5**:1473-1479.
25. Kimmel, A. and Firtel, R. (1982) The organization and expression of the Dictyostelium genome, p.233-324. In W.F. Loomis (ed.), *The development of Dictyostelium discoideum*. Academic Press, Inc. New York.
26. Kimmel, A. and Firtel, R. (1983) Sequence organization in Dictyostelium: Unique structure at the 5' ends of protein coding genes. *Nucleic Acids Res.* **11**:541-552.
27. Reymond, C., Gomer, R., Mehdy, M. and Firtel, R. (1985) Developmental regulation of a Dictyostelium gene encoding a protein homologous to mammalian ras protein. *Cell* **39**:141.
28. Datta, S., Gomer, R. and Firtel, R. (1986) Spatial and temporal regulation of a foreign gene by a prestalk-specific promoter in transformed Dictyostelium discoideum. *Mol. Cell. Biol.* **6**:811-820.
29. Haribabu, B., Ragheb, J. and Dottin, R. (1986) ICN-UCLA Symposia Mol. Cell. Biol. New Series, Vol. 51, Editors Firtel, R. and Davidson, E. Alan Liss, Inc., New York, N. Y.
30. Wynshaw-Boris, A., Lugo, T., Short, T., Fournier, R. and Hanson, R. (1984) Identification of a cAMP regulatory region in the gene for rat cytosolic phosphoenolpyruvate carboxykinase (GTP). *J. Biol. Chem.* **259**: 12161-12169.
31. Nagamine, Y. and Reich, E. (1985) Gene expression and cAMP. *Proc. Natl. Acad. Sci., U.S.A.* **82**:4606-4610.