379
N81d
No. 377C

REGULATION, EVOLUTION, AND PROPERTIES OF THE *ATO* OPERON AND ITS GENE PRODUCTS IN *ESCHERICHIA COLI*

DISSERTATION

Presented to the Graduate Council of the
University of North Texas in Partial
Fulfillment of the Requirements

For the Degree of·

DOCOTR OF PHILOSOPHY

By

Chaw-Yuan Chen, B.S., M.S.
Denton, Texas

August 1993

379
N81d
No. 377C

REGULATION, EVOLUTION, AND PROPERTIES OF THE *ATO* OPERON AND ITS GENE PRODUCTS IN *ESCHERICHIA COLI*

DISSERTATION

Presented to the Graduate Council of the
University of North Texas in Partial
Fulfillment of the Requirements

For the Degree of·

DOCOTR OF PHILOSOPHY

By

Chaw-Yuan Chen, B.S., M.S.
Denton, Texas

August 1993

Chen, Chaw-Yuan, Regulation, Evolution, and Properties of the *ato* Operon and its Gene Products in *Escherichia coli*. Doctor of Philosophy (Molecular Biology), August, 1993, 158 pp., 27 Illustrations, 7 tables, bibliography.

The regulation of short chain fatty acid metabolism has been examined. Metabolism of acetoacetate, and short chain fatty acids such as butyrate and valerate, is predicated upon the expression of genes of the *ato* operon. Acetoacetate induces expression of a CoA transferase (encoded by the *atoDA* genes) and expression of a thiolase (encoded by the *atoB* gene). Metabolism of saturated short chain fatty acids requires the activities of the transferase and thiolase and enzymes of ß-oxidation as well. Spontaneous mutant strains were isolated that were either constitutive or that were inducible by valerate or butyrate instead of acetoacetate.

The *ato* operon is positively regulated by the *atoC* gene product. The *atoC* gene maps next to the *ato* operon. By deduction, AtoC is a 447 amino acid protein with a molecular weight of 51,000 Daltons. The AtoC protein shares extensive sequence homology with the NtrC/NifA family of proteins, including XylR from *Pseudomonas putida*. In pair-wise comparison with the NtrC protein from *Klebsiella pneumoniae, the* AtoC regulatory protein has identical residues at over 150 positions. Even greater homology is obtained if

allowance is made for conservative substitutions. Preceding the *ato* operon is a consensus RpoN (σ^{54}) sequence indicating that the mechanism of activation of the promoter for the *ato* operon may involve recognition by σ^{54} RNA polymerase coupled to interaction with AtoC.

The nucleotide sequence of the entire *ato* operon has been determined. The operon contains four genes in the order *atoDAXB*, which are transcribed coordinately. A new open reading frame was identified that encodes a hydrophobic protein that may participate in transport of short chain fatty acids and/or acetoacetate.

Table of Contents

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER I

INTRODUCTION

Fatty Acid Metabolism

The common laboratory bacterial strain *Escherichia coli* K-12 utilizes long-chain (C12 to C18) fatty acids as sole carbon and energy sources (Klein et al., 1971). *E. coli* cells can be easily cultivated on a simple salts medium, supplemented with common fatty acids such as palmitic acid (palmitate: C16) or stearic acid (stearate: C18)(Overath et al., 1967; 1969). No other carbon source is required. Thus *E. coli* can use inorganic nitrogen and sulfur (in the form of ammonium sulfate), inorganic phosphorus (as phosphate) and inorganic metal halides (calcium, ferric and magnesium chlorides) in combination with long chain fatty acids to produce all the constituents of the cell. Growing aerobically and using the enzymes of fatty acid degradation, *E. coli* can synthesize carbohydrates, nucleic acid bases, amino acids and vitamins from fatty acids. In addition, fatty acids can be incorporated directly into complex lipids, especially membrane phospholipids (Esfahani et al., 1971).

Thus there exist pathways for the degradation and synthesis of lipids, especially fatty acids (Nunn 1968).

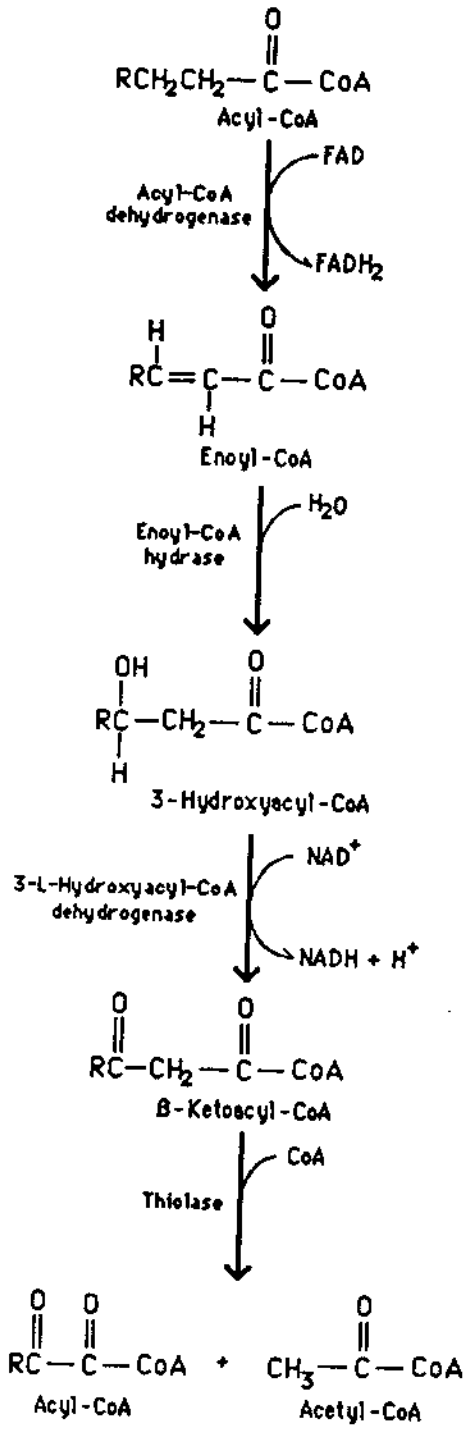
Fatty acids are degraded through the enzymes of the ß-oxidation pathway. In this metabolic pathway, the long chain fatty acids are broken down into two carbon units of acetyl-CoA. The acetyl-CoA so produced can enter the TCA cycle under aerobic growth conditions to provide energy as well as carbon skeletons for compounds derived from the Krebs cycle. Carbohydrates are also produced and the cycle is replenished by the action of the enzymes of the glyoxylate cycle (Kornberg, 1966). Growth of *E. coli* on long chain fatty acids induces all of the enzymes required for fatty acid degradation (the enzymes of ß-oxidation), as well as those required for gluconeogenic growth on oxidative growth substrates (the enzymes of the glyoxylate cycle)(Maloy & Nunn, 1981, 1982). Included among the activities induced are enzymes for transport (Frerman & Bennett, 1973; Klein, et al. 1971), acylation, and regulation as well as ß-oxidation itself (57). The enzymes of fatty acid degradation are encoded by the *fad* genes, while the enzymes of the glyoxylate cycle are encoded by the *ace* operon. Enzyme synthesis from the *fad* genes and from the *ace* genes is induced by long chain fatty acids directly or from intermediates in their metabolism (Nunn, 1986; Weeks, et al., 1969).

The ß-oxidation of exogenously supplied fatty acids in *E. coli* begins with the transport of the free fatty acid into

the cell (Nunn, 1986). Long chain fatty acids are translocated across the outer membrane by a protein encoded by the *fadL* gene. Its exact role is uncharacterized, but the product of the *fadL* gene also transports medium chain fatty acids across the membrane. In addition, medium chain fatty acids can diffuse across the outer membrane of *E. coli* in the absence of functional *fadL* protein (Frerman & Bennett, 1973). Short chain fatty acid transport appears to be carrier mediated (Frerman, 1973), but does not require a functional *fadL* gene product. To enter the cell and be transported across the cytoplasmic (inner membrane) medium and long chain fatty acids must be activated by acyl-CoA synthetase and raised to the level of CoA. This reaction occurs on the inner membrane and requires free CoA (reduced CoA actually, CoASH) and ATP.

The ß-oxidation pathway results in the degradation of fatty acids into acetyl-CoA and a fatty acyl chain shortened by two carbon atoms as a result of the reactions. In the first step (Figure 1.1), the fatty acyl-CoA is oxidized, using FAD as the oxidant to give the enoyl-CoA derivative which possesses the *trans* configuration about the double bond. Next, the double bond between the second and third carbon atoms (α and ß carbons) is hydrated by enoyl CoA hydratase. Next the enoyl-CoA is subjected to a second oxidation by a hydroxyacyl-CoA dehydrogenase. In this reaction $NAD^+ + H^+$ is reduced to $NADH_2$ as the fatty acyl-CoA

Figure 1.1 ß-oxidation pathway for the degradation of fatty acids. Fatty acids are first activated by conversion to the CoA derivative. Then the fatty acyl-CoA is subjected to successive rounds of oxidation, hydration, oxidation and thiolysis. The end products of the pathway are acetyl-CoA and a fatty acyl-CoA that has a chain length decreased by two carbon atoms. The degradation involves sequential oxidation of the ß carbon to a keto group.

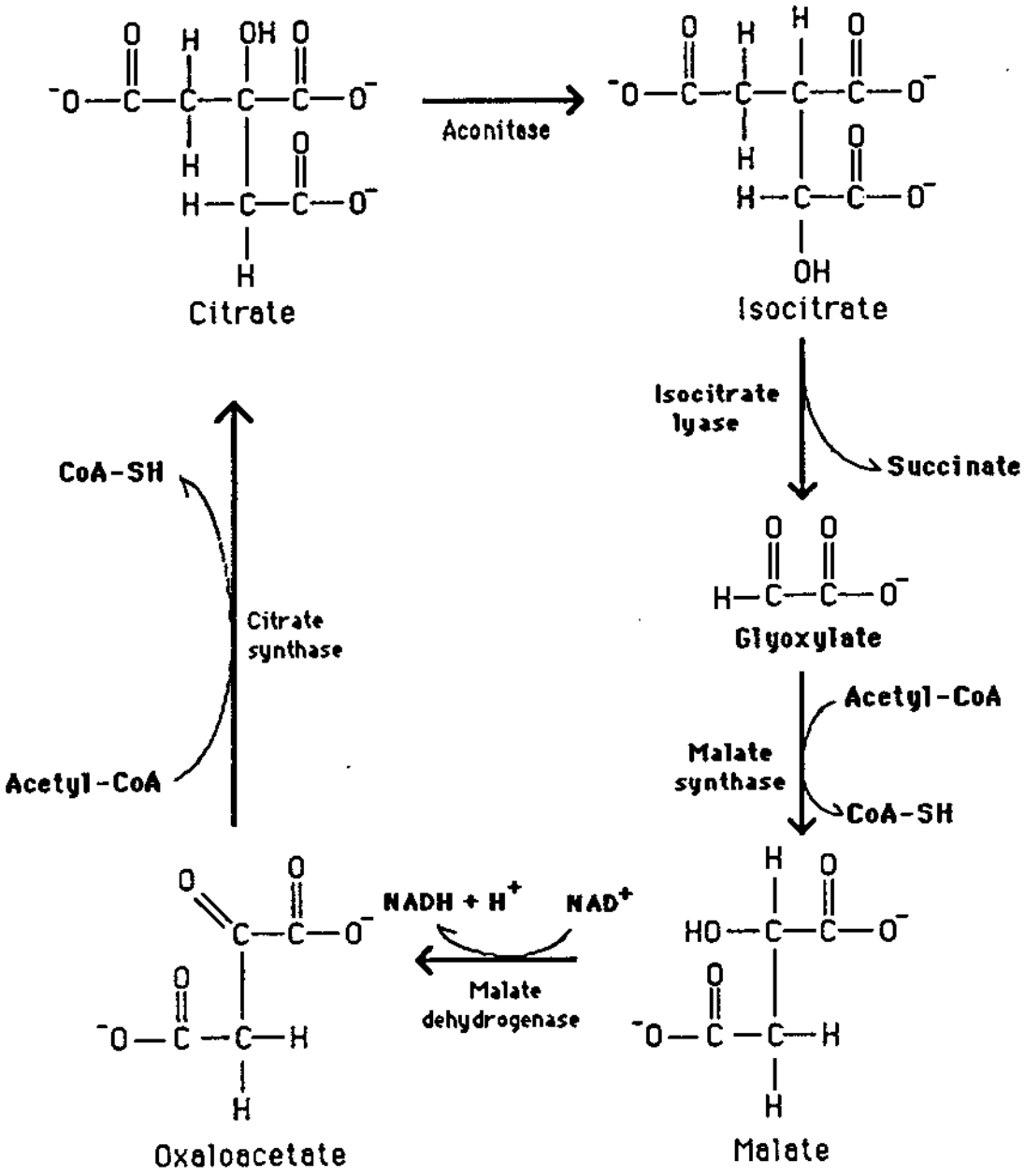

RCH2CH2—C—CoA
Acyl-CoA
Acyl-CoA dehydrogenase
FAD
FADH2
RC=C—C—CoA
Enoyl-CoA
Enoyl-CoA hydrase
H2O
RC—CH2—C—CoA
3-Hydroxyacyl-CoA
3-L-Hydroxyacyl-CoA dehydrogenase
NAD+
NADH + H+
RC—CH2—C—CoA
β-Ketoacyl-CoA
Thiolase
CoA
RC—C—CoA
Acyl-CoA
+
CH3—C—CoA
Acetyl-CoA

is further oxidized to the 3-keto (ß-keto) acyl-CoA derivative. In the final reaction of the ß-oxidation cycle, a thiolase thiolytically cleaves the acyl-CoA using free CoA (CoASH) to give acetyl-CoA and an acyl-CoA reduced in chain length by two carbon atoms (Kornberg, 1966). The acetyl-CoA can then enter the TCA cycle and intermediary metabolism.

Though fatty acids provide a great deal of energy for the cell, only in plants and bacteria can fatty acids, acetate, and compounds that yield acetate be used as sole sources of carbon. The acetyl-CoA produced by fatty acid degradation is metabolized through the TCA cycle. Every two carbon acetyl-CoA condensed with oxaloacetate to form citrate is metabolized through the reactions of the cycle where two carbon atoms are given off as CO_2. Thus there is no net gain in carbon when fatty acids are metabolized. The glyoxylate cycle in *E. coli* functions as an anaplerotic reaction sequence to replenish carbon and evade the TCA reactions that evolve CO_2 (43). There are two enzymatic steps in the glyoxylate cycle (Figure 1.2).

First, isocitrate is formed and cleaved by isocitrate lyase to give succinate and the compound glyoxylate. Thus the CO_2 evolving steps in the conversion of isocitrate to succinate in the TCA cycle are bypassed. Isocitrate lyase in *E. coli* is encoded by the *aceA* gene. Glyoxylate is next condensed with acetyl-CoA to form malate and free CoA. The reaction is catalyzed by the enzyme malate synthase which is

Figure 1.2. The glyoxylate cycle. Acetate and compounds that are metabolized to acetate can be used as sole carbon and energy sources in cells with a functional glyoxylate cycle. The two enzymes of the glyoxylate cycle (on the right side in bold print) bypass the decarboxylation reactions of the TCA cycle allowing compounds to flow through the cycle with no loss of carbon. Enzyme activities of the TCA cycle are listed in regular print. Of course the glyoxylate cycle and the TCA cycle must operate simultaneously.


Citrate
Aconitase
Isocitrate
Isocitrate lyase
Succinate
Glyoxylate
Acetyl-CoA
Malate synthase
CoA-SH
Malate
NADH + H+
NAD+
Malate dehydrogenase
Oxaloacetate
CoA-SH
Citrate synthase
Acetyl-CoA

encoded by the *aceB* gene. The genes are organized into an operon and are subject to catabolite repression and induction by growth on fatty acids (Maloy & Nunn, 1981). In addition, the enzyme isocitrate dehydrogenase which competes for isocitrate with the isocitrate lyase is regulated by a reversible phosphorylation and dephosphorylation catalyzed by the product of the *aceK* gene which maps with the other genes in the *ace* operon (Nunn, 1986).

Growth of *E. coli* on medium chain fatty acids (C6 to C11) requires the same enzymatic activities as does growth on fatty acids with longer chain lengths. But wild type cells do not grow on the medium chain fatty acids such as decanoate (C10). Even though the C10 and other medium chain homologs are handled by the ß-oxidation enzymes encoded by the *fad* genes, medium chain fatty acids do not induce the synthesis of the enzymes for fatty acid degradation (Overath et al., 1967; 1969; Weeks et al., 1969). Only fatty acids with a chain length greater that C12 will induce the *fad* regulon. Normally in wild type ($fadR^+$) strains, the *fadR* gene product exerts negative control over the *fad* genes for fatty acid degradation (Overath et al., 1969). Genes for enzymatic activities of transport, activation and ß-oxidation are repressed by binding of FadR protein to the corresponding genes in the absence of inducer. In a *fadR* mutant strain, the repressor is inactive or not synthesized, and all of the activities are expressed constitutively. Only *E. coli fadR*

strains can grow on decanoate and other medium chain fatty acids as sole sources of carbon and energy.

Growth of *E. coli* on short chain fatty acids also does not occur in wild type strains. As with long and medium chain fatty acids, short chain fatty acids such as valerate and butyrate require the enzymes of ß-oxidation. But the short chain fatty acids are not activated to the level of CoA by the ATP dependent acyl-CoA synthetase like the medium and long chain fatty acids are. Instead, the short chain fatty acids like butyrate and valerate are raised to the level of CoA by a transferase. So, besides the enzymes of ß-oxidation, in order to grow on short-chain fatty acids (C4 to C5), *E. coli* needs to turn on the *ato* operon, which is controlled by *atoC* regulatory protein and contains three structural genes, *atoDA* and *atoB*, encoding the enzymes acetyl-CoA:acetoacetate coenzyme A transferase (EC2.8.3.5., CoA transferase) and acetyl-CoA:acetyl-CoA C-acetyl transferase (EC2.3.1.9., thiolase)(Frerman, 1973; Frerman and Bennett 1973, Pauli and Overath 1972). Wild type *E. coli* grows on acetoacetate as the sole carbon source because acetoacetate is the inducer that binds to the *atoC* product to stimulate the expression of the *atoDAB* genes.

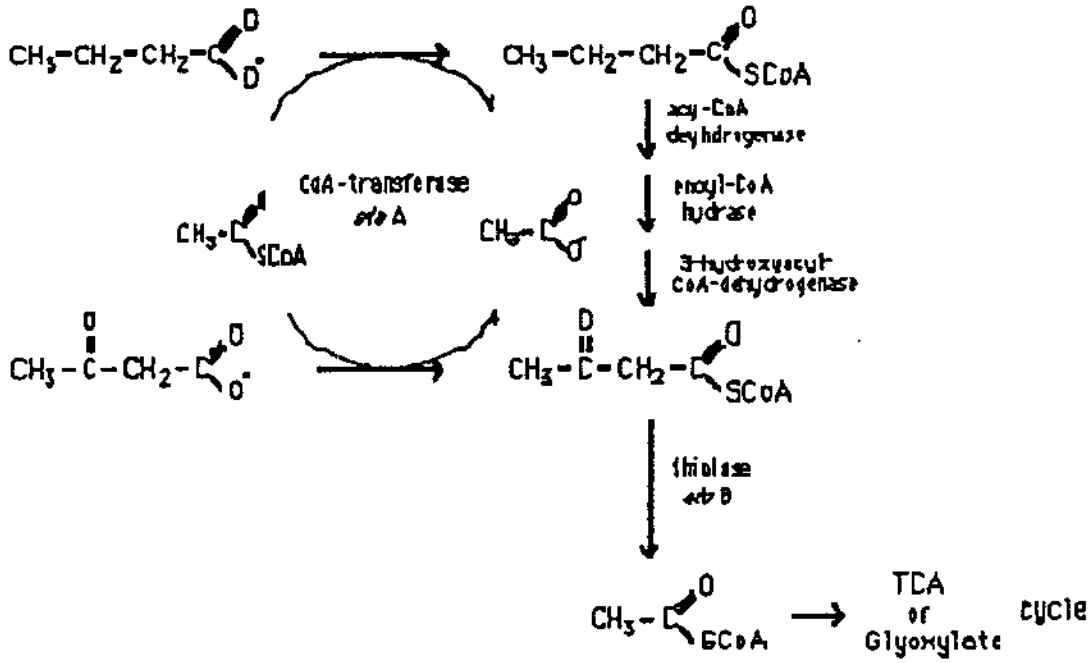
The purpose of this study is to investigate and characterize the genes, proteins, enzymes, and regulatory elements of the *ato* operon for the degradation of short chain fatty acids in the bacterium *Escherichia coli*. In addition,

some of the studied CoA transferases from the different organisms will be compared to each other from an evolutionary view point.

Short Chain Fatty Acid Metabolism

Frerman (1973) demonstrated that *E. coli* can transport short-chain fatty acids solely if the enzyme acetoacetyl-CoA transferase is present in cells. This conclusion was obtained by comparing the uptake of C4 into membrane vesicles of *E. coli* strains taken from the presence and the absence of acetoacetate. Although these results suggest a role for acetoacetyl-CoA transferase in short-chain fatty acid uptake, they can not strongly support that this enzyme is solely responsible for short-chain fatty acid transport. In many bacteria, acetoacetate is metabolized *via* two enzymatic steps to two molecules of acetyl-CoA and thence to the TCA cycle as a source of energy and carbon (using the attendant glyoxylate cycle). In *Escherichia coli* acetoacetate is raised to the level of a CoA thioester by the enzyme acetyl-CoA:acetoacetate coenzyme A transferase. Acetoacetyl-CoA is subsequently cleaved thiolytically by acetyl-CoA:acetyl-CoA C-acetyl transferase to yield two molecules of acetyl-CoA (Figure 3). Growth of *E. coli* on other short chain fatty acids like butyrate and valerate is also accomplished through the action of these two enzymes in combination with activities the enzymes of ß-oxidation (Figure 1.3).

Figure 1.3. The metabolism of short chain fatty acids. In *E. coli*, acetoacetate is converted to acetoacetyl-CoA by a CoA transferase using acetyl-CoA as the CoA donor. The transferase is a two subunit enzyme encoded by the *atoDA* genes. Butyrate and valerate are converted to their corresponding CoA derivatives by the same CoA transferase. However, butyrate and valerate must be converted to the ß-keto derivatives by the enzymes of ß-oxidation. Acetoacetyl-CoA (and any other short chain ß-keto acyl-CoA) is thiolytically cleaved by the *atoB* thiolase.

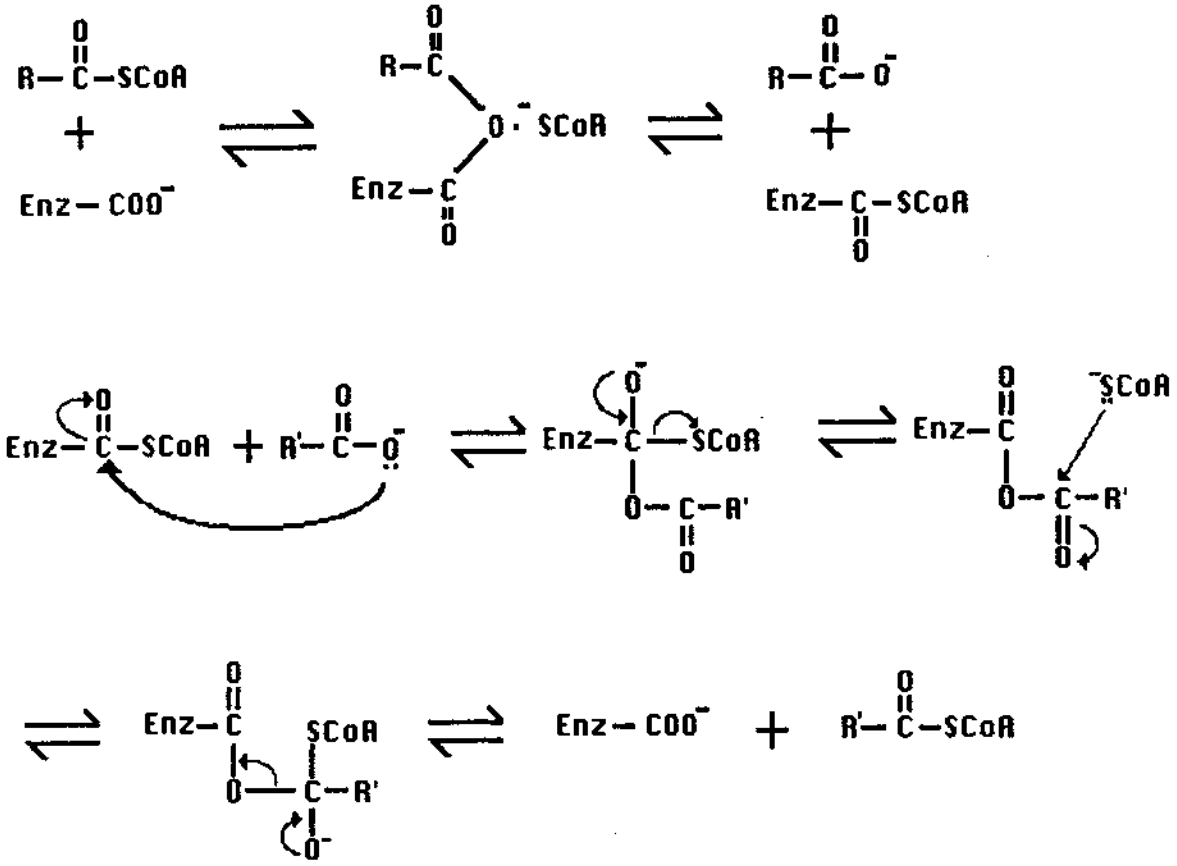

CoA-transferase
atoA
acyl-CoA dehydrogenase
enoyl-CoA hydrase
3-Hydroxyacyl-CoA-dehydrogenase
thiolase
atoB
TCA
or
Glyoxylate
cycle

CoA Transferases

CoA transferases in general function directly to preserve the energy available in thioester linkages and transfer it between different thioesters. The problem faced by these enzymes is to retain the high group transfer potential of the acyl group while transferring the CoA moiety between the substrates. The reaction catalyzed by CoA transferases is unusual in the sense that the enzymes catalyze the transfer of the leaving group rather than the displacement of the leaving group (Jenks, 1973). In the reaction, the activating compound, or the donor-CoA thioester, acts as a dehydrating agent that extracts an oxygen atom from the carboxylate ion of the CoA acceptor at the same time it donates the CoA group that replaces the oxygen atom to form the activated CoA product (Figure 1.4). The extraction of oxygen from the acceptor carboxylate ion and its transfer to the CoA donor carboxyl group has been demonstrated experimentally by the use of isotopically labeled compounds for the related enzyme succinyl-CoA:3-ketoacid Coenzyme A-transferase (EC 2.8.3.5) from pig heart (Flacone & Boyer, 1959).

This eukaryotic CoA transferase is a member of a family of CoA transferases and studies from several laboratories suggest that the enzymes possess similar structures and catalytic mechanisms. Best characterized of all these

Figure 1.4. The reaction catalyzed by CoA transferase. The CoA moiety of the CoA donor is transferred to a γ-glutamyl residue on the enzyme (line 1). Then the CoA is transferred to the ketoacid acceptor (lines 2 and 3).



enzymes is the aforementioned succinyl-CoA: 3-ketoacid CoA transferase of pig heart (Falcone & Boyer, 1959; Hersh & Jenks, 1967a;1967b). The native enzyme is a dimer with an apparent molecular weight of 92,000 Daltons (Stern et al. 1956). The two subunits are identical, with a monomeric molecular weight of 45,600. An enzyme-CoA intermediate is formed through the covalent attachment of the CoA moiety of succinyl-CoA to the γ-carboxyl group of a glutamyl residue on the enzyme (Hersh & Jenks, 1967a;1967b). There is also a sensitive thiol group whose modification results in inactivation of the enzyme. The covalent enzyme-CoA intermediate, and not the free enzyme, is most susceptible to this inactivation indicating a conformational change upon formation of the enzymatically active intermediate (Hersh & Jenks, 1967a;1967b). A cationic lysyl residue has been identified by acylation, and is presumably responsible for neutralizing the charge of the substrate, succinyl-CoA, during binding (Hersh & Jenks, 1967a;Jenks, 1973). A related eukaryotic succinyl-CoA: 3-ketoacid CoA transferase has been partially characterized from sheep kidney (Sharp & Edwards, 1978;1983) and possesses a dimeric structure with a subunit molecular weight of 55-56,000 Daltons. Like the pig heart enzyme, this transferase displays a ping-pong kinetic mechanism. While these enzymes are valuable models for determination of kinetic mechanisms and active site properties, structural studies of the enzymes have not been

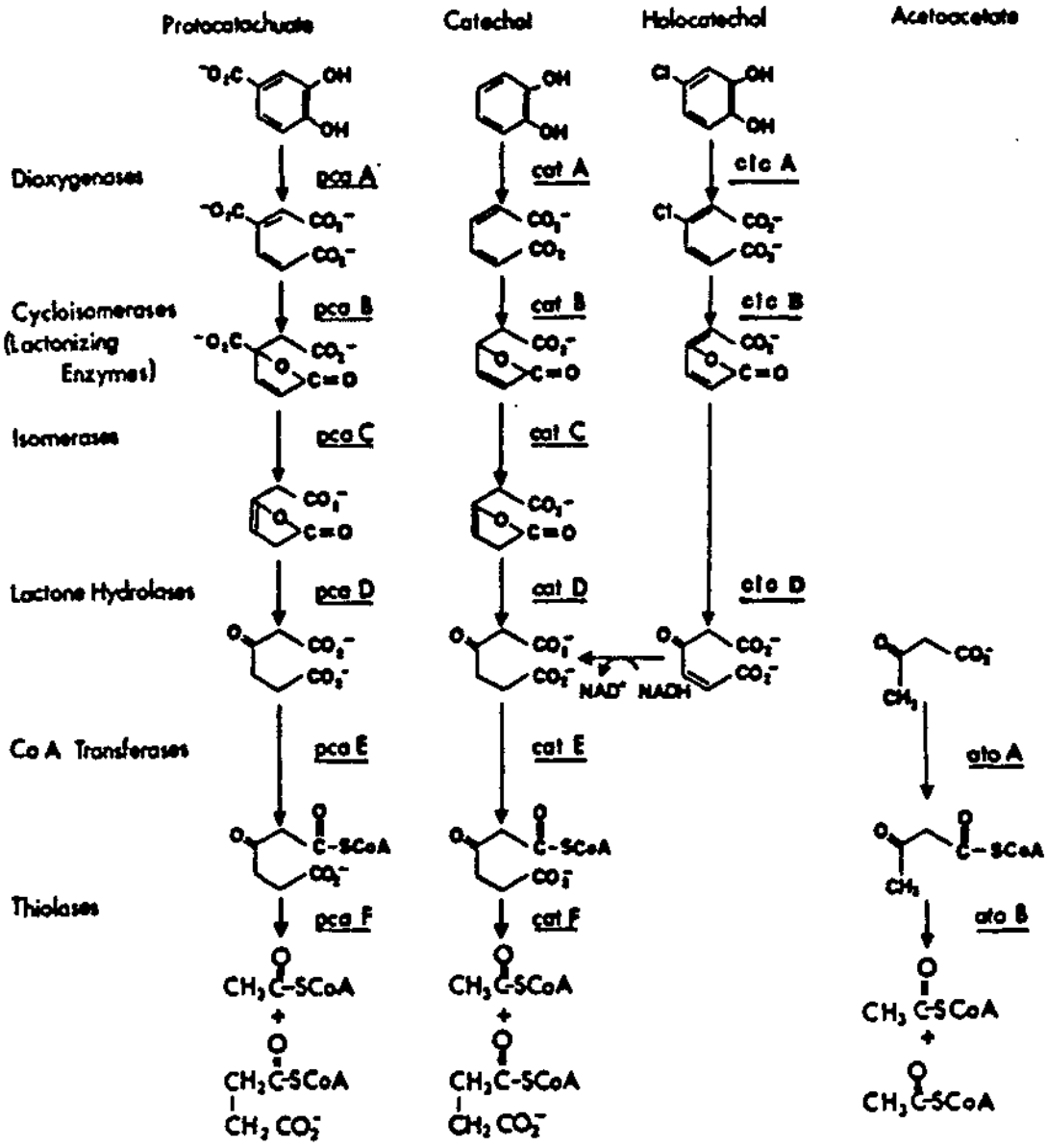
performed. The purification procedures are complex and genetic manipulation of the genes is not practical. Thus while the gross subunit architecture and the amino acid composition of the pig heart enzyme have been determined (White & Jenks, 1976), the primary sequence and higher order structure have not been described.

An additional question of the proposed research concerns the related acetyl-CoA:acetoacetate coenzyme A-transferase of *E. coli* (Sramek & Frerman, 1975a, 1975b; Sramek et al., 1977). This prokaryotic CoA transferase catalyzes the transfer of the coenzyme A-moiety from the activated thioester substrate, acetyl-CoA, to the a-carboxyl group of the 3-keto acid, acetoacetate. The enzyme is a tetramer composed of two different subunit types, α and ß, with subunit molecular weights of 26,000 and 23,000 Daltons respectively (Sramek & Frerman, 1975a). Presumably, an αß protomer corresponds to a single subunit in the eukaryotic enzyme. This might represent a case of domain fusion typically observed for eukaryotic enzymes and enzyme complexes. Substantiation awaits detailed knowledge of the primary structure of the enzymes. In a manner similar to the eukaryotic enzymes, the *E. coli* transferase forms a covalent enzyme-CoA intermediate through a γ-carboxyl group of a glutamyl residue on the ß subunit as identified by formation of the covalent intermediate and subsequent reduction by tritiated borohydride (Sramek & Frerman, 1975b). The enzyme

has a reactive lysine residue subject to acetylation by a metal catalyzed transfer of the acyl group of the substrate acetyl-CoA (Sramek & Frerman, 1975a). Covalent modification of a particularly reactive class of cysteinyl residues (one per αß protomer) results in an 80 per cent loss in activity. The role of the 26,000 Dalton α subunit is unknown (Sramek & Frerman, 1975a).

The homologous succinyl-CoA: ß-ketoadipate CoA transferases (EC 2.8.3.6) from *Acinetobacter calcoaceticus* (Yeh & Ornston, 1981) and *Pseudomonas putida* (Stanier & Ornston, 1973) catalyze the penultimate reaction of the ß-ketoadipate pathway for the dissimilation of aromatic compounds such as benzoate and *para*-hydroxybenzoate. The ß-ketoadipate pathway has two branches. The catechol branch has six enzymatic steps for the conversion of catechol to TCA cycle intermediates via the compound ß-ketoadipate (Figure 5) (Stanier & Ornston, 1973). The protocatechuate branch consists of six enzymes that convert protocatechuate to citric acid cycle intermediates by reactions analogous or identical to those catalyzed in the catechol branch. In *P. putida*, a single transferase is elaborated (Hemp & Hegeman 1968), whereas in *A. calcoaceticus* two separate forms of the transferase are made: transferase I is expressed specifically for the catabolism of protocatechuate, and transferase II is expressed specifically in the catechol branch (Canovas & Stanier, 1967).

Figure 1.5. The ß-ketoadipate pathway. The ß-ketoadipate pathway has three branches, one for catechol, one for protocatechuate and one for halocatechol. The three branches catalyze analogous reactions and share the common intermediate ß-ketoadipate. The ß-ketoadipate is raised to the level of CoA by a CoA transferase and subsequently cleaved to succinyl-CoA and acetyl-CoA by a thiolase. The acetoacetate degradative pathway from *E. coli* is shown for comparison.


Protocatechuate
Catechol
Halocatechol
Acetoacetate
Dioxygenases
Cycloisomerases (Lactonizing Enzymes)
Isomerases
Lactone Hydrolases
Co A Transferases
Thiolases
pca A
pca B
pca C
pca D
pca E
pca F
cat A
cat B
cat C
cat D
cat E
cat F
clc A
clc B
clc D
ato A
ato B
NAD+
NADH

Like the *E. coli* acetoacetate CoA transferase, the ß-ketoadipate CoA transferases from *P. putida* and *A. calcoaceticus* have been purified to homogeneity (Yeh & Ornston, 1981; 1984). The enzymes each have an approximate molecular weight of 108,000 Daltons and they cross react with one another immunologically. They have homologous $\alpha_2ß_2$ oligomeric structures formed by the association of nonidentical subunits of about 25,000 Daltons. They are inactivated on a molar basis with *para*-chloromercuribenzoate (1:1 per αß protomer) indicating a selectively sensitive cysteinyl residue near the active site (Yeh & Ornston, 1984). Like all of the related enzymes they show a high degree of specificity towards the CoA donor while the structural requirements for the CoA acceptor substrate are less specific. The *E. coli atoDA* CoA transferase can not use succinyl-CoA as the CoA-donor, perhaps a reflection of the low level of the succinyl-CoA pool in *E. coli* (Jackowski & Rock, 1986). The *catE* gene encoding the succinyl-CoA: ß-ketoadipate CoA transferase II from *A. calcoaceticus* has been cloned (Shanley et al., 1986) and sequenced (unpublished) and the primary structure of both subunits of the enzyme have been deduced. The genes encoding the isofunctional succinyl-CoA: ß-ketoadipate CoA transferase I from *A. calcoaceticus*, specific for the catabolism of protocatechuate, have also been cloned (Doten, et al., 1987). The related ketopadipate

CoA transferase from *P. putida* has been cloned and sequenced (Erlich and Harwood, 1992).

Thiolases

Thiolase catalyzes the last reaction in the pathway for the degradation of short chain fatty acids, namely the thiolytic cleavage of ß-ketoacyl-CoA derivatives in the presence of Coenzyme A as a cofactor. As in most organisms where there is a multiplicity of thiolases, in *E. coli* there are at least two thiolases (Feigenbaum & Schulz, 1975). In general thiolases work to cleave ß-ketoacyl derivatives of CoA by displacement with a thiol group of another molecule. The classical example is the chain cleavage step in the degradation of fatty acids in the ß-oxidation sequence (Figure 1.1). It has been suggested and experimentally demonstrated that a thiol group on the enzyme reacts initially with the ß-carbonyl group to give an enzyme bound S-acyl intermediate. The acyl group (in the case of thiolase II, the acetyl group) is then transferred to CoA in a second step (Duncombe & Frerman, 1976).

Thiolase II from *E. coli* is active on acetoacetyl-CoA and is specific for that compound. In contrast, thiolase I, the *fadA* thiolase, isolated from *E. coli* is active on CoA-derivatives of chain length of 4 to 16 carbons. It exhibits optimal activity with medium-chain substrates (Feigenbaum & Schulz, 1975). The thiolase II, the *atoB* thiolase, from *E.*

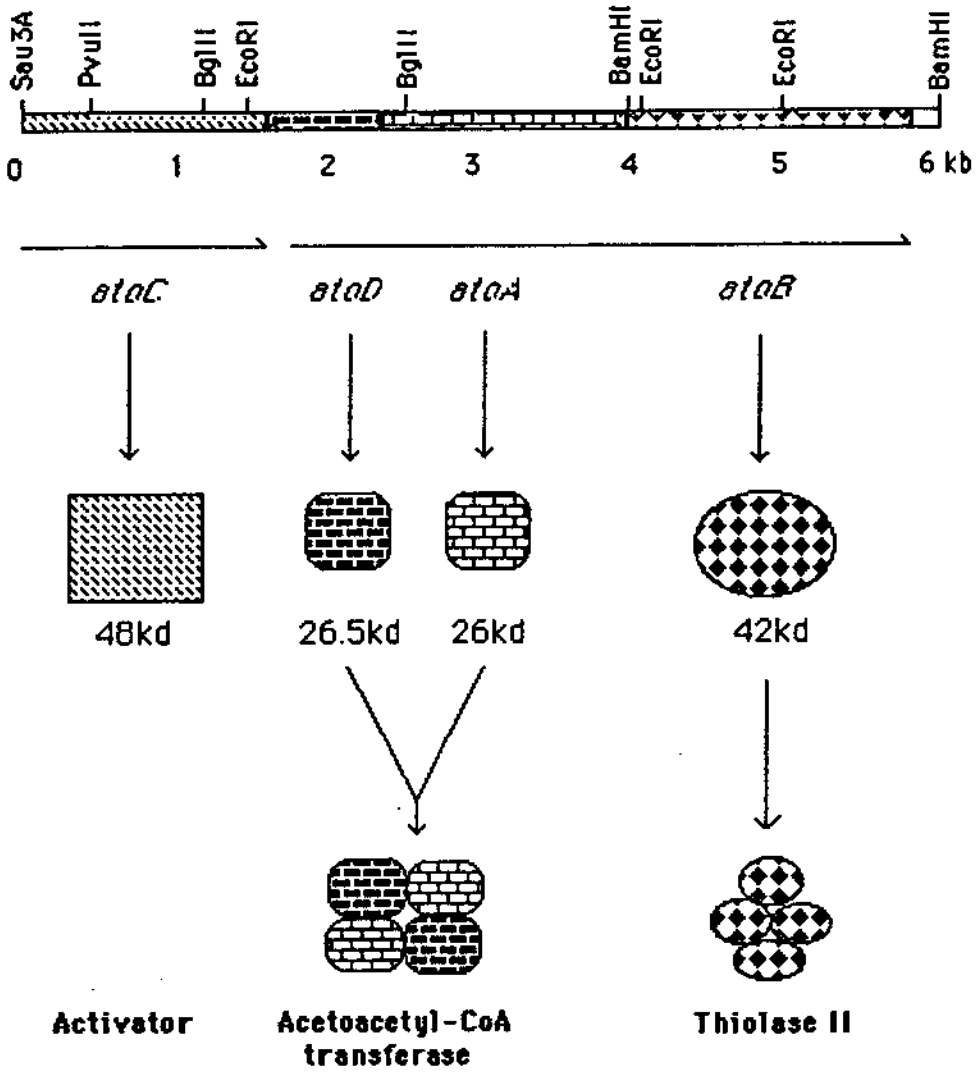
coli has been purified to homogeneity and extensively characterized (Duncombe & Frerman, 1976). The enzyme has a molecular weight of 166,000 and is composed of four identical protomers with a subunit molecular weight of 41,500. The enzyme has two free sulfhydryl residues per subunit. The acetyl-enzyme covalent intermediate has been trapped and demonstrated by reduction with borohydride after incubation with [^{14}C]acetyl-CoA. The acetyl groups are bound to the enzyme by a thioester bond (Duncombe & Frerman, 1976).

Thiolytic cleavage of CoA thioesters is a common theme is biological systems. The related CoA thiolase from *A. calcoaceticus* has been cloned (Shanley et al., 1987). Acetoacetyl-CoA is in equilibrium with acetyl-CoA in the reaction mediated by thiolase. Acetoacetate in higher organisms can be "synthesized" from two molecules of acetyl-CoA in the reverse reaction of the thiolase to give acetoacetyl-CoA. Not only can acetoacetyl-CoA be cleaved to two molecules of acetyl-CoA and enter the citric acid cycle, but it is also a precursor for the synthesis of polyisoprenoid compounds including cholesterol. More significantly, free acetoacetate is a ß-keto acid that can be easily and spontaneously decarboxylated to acetone and reduced by an NADH-dependent dehydrogenase to hydroxybutyrate. The three compounds acetoacetate, acetone and hydroxybutyrate are known collectively as the ketone bodies.

The regulation of the *ato* operon in *E. coli*

As is known, the saturated short-chain fatty acids butyrate (BUT) and valerate (VAL) cannot directly serve as substrates for the acetoacetate (ATO) metabolizing enzymes of the *ato* operon. Two mutations causing constitutive expression of the *fad* and *ato* structural genes are necessary for the metabolism of these saturated short-chain fatty acids. Mutants able to utilize butyrate as a sole carbon source are readily selected by plating *fadR* strains on *E. coli* minimum medium containing butyrate. Pauli and Overath (1972) first showed that most BUT^+ mutants obtained in this way were constitutive for the ATO enzymes. They also identified a regulatory locus, *atoC*, which is also located at 47 min.on the chromosome and is located adjacent to the *ato* structural genes (Figure 1.6). The *atoC* gene appeared to encode a positive regulatory element. Jenkins and Nunn (1987a;1987b) demonstrated genetically that *atoC* mutants resulted in the pleiotropic loss of acetoacetyl-CoA transferase and thiolase II activities. They also proved that the absence of the 48,000 Dalton *atoC* encoded protein resulted in the complete loss of the *ato* encoded enzyme activities. The conclusion indicated that the *atoC* encodes a 48,000 Dalton molecular weight protein that is required for the expression or activity of the *ato* encoded enzymes.

Figure 1.6. The *ato* operon of *E. coli*. The genes for the *atoDA* transferase and the *atoB* thiolase are arranged in an operon and are coordinately controlled from a single promoter in the order *atoDAB*. Lying immediately adjacent to the *ato* operon is the gene for the *atoC* regulatory protein.


Sau3A
PvuII
BglII
EcoRI
BglII
BamHI
EcoRI
EcoRI
BamHI
0
1
2
3
4
5
6 kb
atoC
atoD
atoA
atoB
48kd
26.5kd
26kd
42kd
Activator
Acetoacetyl-CoA transferase
Thiolase II

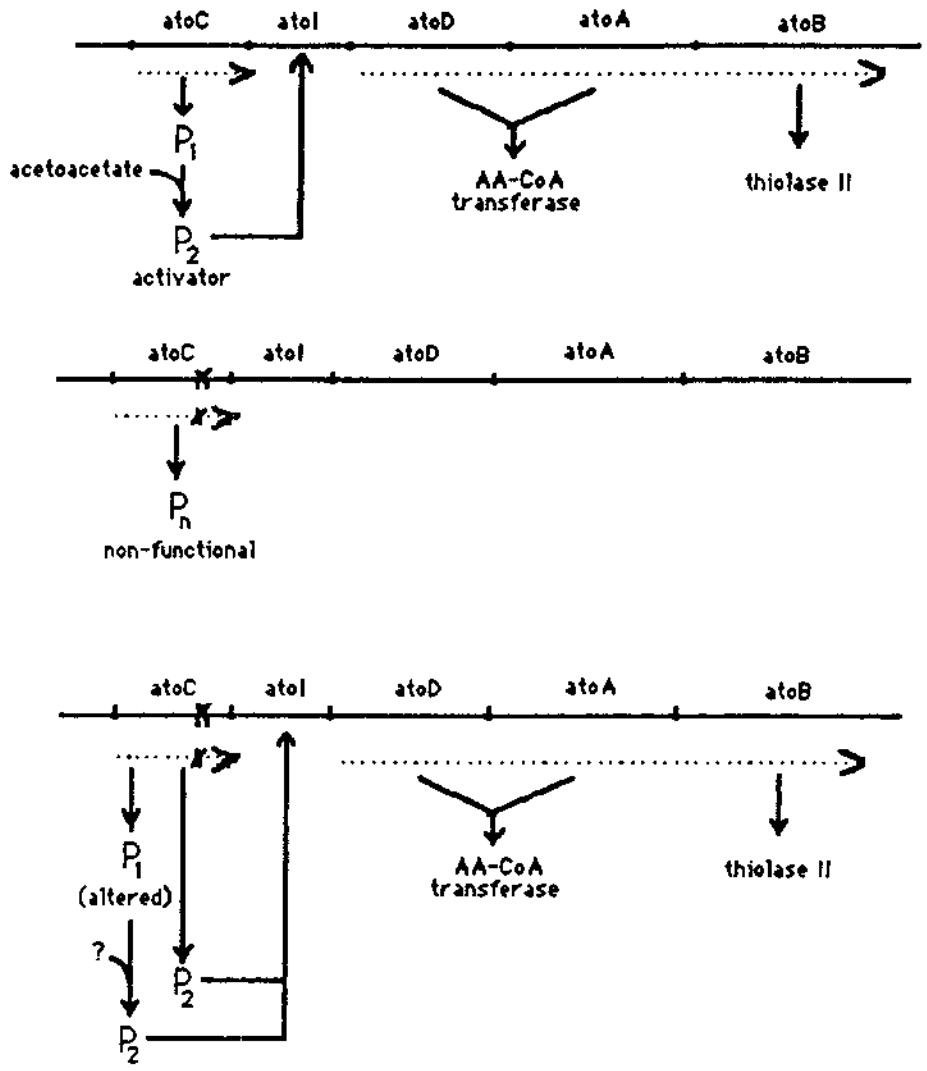
Jenkins and Nunn (1987a;1987b) believed that like the regulatory properties of *malT*, *atoC* was not autoregulated but was catabolite repressible. Among several positive regulators (e.g.. *araC* ,*dsdC* and *malT*) the *atoC* gene product might be the only positive regulatory gene system which has the possibility to mutate spontaneously at high frequency to a constitutive phenotype (i.e., high level expression in the absence of the inducer actoacetate or $atoC^{C}$) .

Pauli and Overath (1972) used one *atoC* mutant to examine the regulatory mechanism controlling *ato* gene expression. Normal wild type $atoC^{+}$ strains do not grow on butyrate medium. They found that all merodiploid cells which were $atoC^{+}A^{+}B^{+}$ for the structural gene products expressed the *ato* genes constitutively in an *atoC* mutant ($atoC^{C}$-49). They also found that all merodiploids which had $atoC^{+}A^{+}B^{+}$ on the episome and $atoC^{C}A^{+}B^{-}$ or $atoC^{C}A^{-}B^{+}$ on the chromosome had the But^{+} phenotype (ability to grown on butyrate minimal medium in the absence of inducer acetoacetate) and constitutive levels of acetoacetyl-CoA transferase and thiolase II activities. These data supported the proposal the $atoC^{C}$ mutation is trans-dominant to $atoC^{+}$. Salanitro and Wegener (1971) isolated two types of spontaneous But^{+} mutants. One type was the $atoC^{C}$ mutants which synthesized the ATO enzymes constitutively. The other mutants were inducible for butyrate uptake (whereas butyrate uptake was constitutive in $atoC^{C}$ mutants). They proposed that it might have a

regulatory mutation that alters the inducer binding site of the regulatory proteins. Since the inducible But$^+$ phenotype was never mapped by them, it is not clear whether the mutation is on the *atoC* gene directly or somewhere else than the *ato* regulatory locus. By gene fusion techniques, the *atoAB* promoter has been linked into the promoterless *galK* gene and the expression of the fused *galK* gene has been studied in the different *atoC* mutants. Expression of *galK* was inducible by the acetoacetate substrate in the *atoC*$^+$ host, noninducible in the *atoC* strain, and constitutive in the *atoC*^{C} mutant. Also, the *atoC*^{C} has two fold higher expression levels than the inducible wild type *atoC*$^+$ mutant. These results provided the evidence for a model proposing the mechanism of the *ato* system by Jenkins and Nunn (Figure 1.7). In the *atoC*$^+$ strain, the *atoC* protein has to be activated by acetoacetate to turn on the operon and the *atoC*^{C} protein can switch on the transcription unit without acetoacetate in the *atoC*^{C} mutant.

The work described in this dissertation addresses a number of factors concerning short chain fatty acid metabolism in *E. coli*. Different classes of regulatory mutants were isolated and characterized. In addition, the entire *ato* operon encoding the *atoDA* transferase and *atoB* thiolase has been sequenced. The cloned *ato* transferase has been expressed in constitutive mutant strains at high levels and purified. And the neighboring region encoding the *atoC*

Figure 1.7. Proposed regulatory scheme for the *ato* operon of *E. coli*. The genes for the *atoDA* transferase and the *atoB* thiolase are arranged in an operon and are coordinately controlled from a single promoter in the order *atoDAB*. Lying immediately adjacent to the *ato* operon is the gene for the *atoC* regulatory protein. The *atoC* gene acts as a *trans*-acting positive activator of gene expression. In the presence of the inducer acetoacetate, the *atoC* gene product binds to the *atoDAB* promoter to stimulate transcription. The direction of transcription of the *atoDAB* operon and the *atoC* gene are indicated by broken arrows. The initiator for the *atoDAB* operon is noted as *atoI*. No mutation affecting *atoI* has been described. Also unknown are the locations of the promoter with respect to the initiator (Jenkins and Nunn, 1987b)


atoC
atoI
atoD
atoA
atoB
P1
acetoacetate
P2
activator
AA-CoA
transferase
thiolase II
atoC
atoI
atoD
atoA
atoB
Pn
non-functional
atoC
atoI
atoD
atoA
atoB
P1
(altered)
?
P2
P2
AA-CoA
transferase
thiolase II

regulatory element has been sequenced and its primary structure determined. The promoter and upstream regulatory region where the *atoC* protein binds has also been sequenced. It is hoped that the research described in this work will provide additional information about and insight into short chain fatty acid metabolism and extend our knowledge of this process in *E. coli*.

CHAPTER II

CHARACTERIZATION OF REGULATORY MUTANTS AFFECTING CONSTITUTIVE EXPRESSION OF THE *ATO* OPERON

In *Escherichia coli*, acetoacetate is metabolized *via* two enzymatic steps to two molecules of acetyl-CoA (Figure 3) and thence to the TCA cycle as a source of energy and carbon (using the attendant glyoxylate cycle, Figure 1.2). A CoA transferase, encoded by the *atoDA* genes, and a thiolase, encoded by the *atoB* gene, working sequentially and in combination raise acetoacetate to the level of CoA and subsequently thiolytically cleave the acetoacetyl-CoA to two molecules of acetyl-CoA (Pauli & Overath, 1972). The structural genes for these two enzymes are arranged in an operon and are coordinately expressed (Jenkins & Nunn, 1987a). The genes are induced by the binding of acetoacetate to a *trans*-acting positive regulator encoded by the closely linked *atoC* gene (Jenkins & Nunn, 1987b). Two classes of constitutive mutants have been previously described and we have isolated both types in this study (Salamitro & Wegener, 1971). In related studies the *atoDAB* genes have been sequenced (Chapter IV), as has been the *atoC* gene (Chapter

III). The characterization of the different mutant classes affecting *ato* gene expression is a subject of this study.

Growth of *E. coli* on short chain fatty acids such as butyrate and valerate is also accomplished through the action of the *ato* CoA transferase and thiolase (Figure 1.3). But, additional enzymes of ß-oxidation for the degradation of fatty acids are required to further metabolize butyryl-CoA (and valeryl-CoA). Since butyrate will not induce the fatty acid degradative genes (*fad* genes), growth on butyrate requires the presence of a long chain fatty acid (greater than C10, or decanoate). Alternately, a regulatory mutation giving constitutive expression of the *fad* genes (*fadR*) can be used to obtain growth of *E. coli* on the short chain fatty acid (SCFA) butyrate. Regulatory *fadR* mutants with constitutive expression of the enzymes of ß-oxidation can be isolated simply by selecting for *E. coli* strains capable of growth on minimal decanoate medium (Pauli & Overath, 1972).

As with the *fad* genes, butyrate and other SCFAs cannot induce the *atoDAB* transferase and thiolase. Only acetoacetate acts and inducer of the ato operon. Constitutive *ato* mutants can be isolated, in a *fadR* background, that permit the growth of *E. coli* on butyrate. Such strains have been designated *atoC*^{C}. We used a rapid selection of regulatory mutant strains to facilitate the study of the mechanism of *ato* gene regulation. Such strains allow the identification of specific *cis* and *trans* acting

mutations that allow for *ato* gene expression in the absence of the inducer acetoacetate.

Methods and Materials

Growth media and conditions. All strains grown for determination of enzyme activity were cultured in *E. coli* minimal medium, prepared with the following ingredients per liter: 10.5 g K_2HPO_4, 4.5 g KH_2PO_4, 1.0 g $(NH_4)_2SO_4$, and 0.5 g trisodium citrate. All media and solutions were formulated with distilled deionized water. Then solutions was sterilized by autoclaving. After the solution was cool, 1 ml of 1M $MgSO_4$ and 1 ml of 0.1% thiamin (vitamin B1)were added. Succinate or short chain fatty acid were added to a final concentration of 10 mM.

Cells were grown in rich medium for storage and preservation. The rich medium used is these studies was LB medium. LB medium contains per liter of distilled deionized water: 10 g yeast extract, 5 g tryptone, and 10 g sodium chloride. To prepared solid medium, 1.8 per cent agar (w/v) was added. All cultures were incubated at 37°C, and liquid cultures were provided with shaking at 200 rpm.

Mutant construction. Regulatory mutants with altered expression properties for the *atoDAB* genes were isolated for these experiments. The parent strains *E. coli* K12 (wild type) and JM101 (D(*lac-pro*), *sup*E44, F′::*traD36*, *proAB*,

lacI^{q}*lacZ*ΔM15) grow on acetoacetate, but not butyrate or valerate. Cells were grown on LB rich media overnight and plated first onto decanoate plates for 8-10 days incubation at 37°C to isolate spontaneous *fadR* strains. The spontaneous mutation rate was approximately 10^{-7}. Colonies which appear are termed *fadR* and subsequently show regular rapid growth on decanoate minimal medium. Strains that are *fadR* grow up on decanoate minimal plates to give visible growth after two days. The resulting *fadR* strains were then simply plated onto minimal medium plates containing a short chain fatty acid other than acetoacetate, such as butyrate and valerate. Big colonies were visible on the second day and small colonies on the fourth day of incubation at 37°C.

Enzyme assays. CoA transferase and thiolase were measured in cell extracts by measuring the decrease or increase in absorbance of the acetoacetyl-CoA-magnesium ion complex at 310 nm. Cultures for enzyme assays were grown in 250 ml volumes in 1 liter baffled growth flasks. Overnight cultures were collected by centrifugation, washed once and resuspended in one fiftieth the original volume (in most cases 5 ml) of 100 mM potassium phosphate buffer (pH 7.2) with 1 mM 2-mercaptoethanol. Cell extracts were prepared by explosive decompression. Cell suspensions were passed twice through a French Pressure Cell mounted in a Carver Laboratory Hydraulic Press (SLM Aminco Corp., Illinois) with the

pressure maintained at 10-15,000 psi. Unbroken calls and cell debris were removed by centrifugation at 15,000 xg for 15 minutes at 4°C. Cells were broken and used immediately for assay, or the cells were centrifuged and the resulting cell pellets stored at -70°C for latter use. The enzymes were found to be stable in frozen cell pellets for periods of up to one year (Shanley and Chen, unpublished observation).

Acetoacetate acetyl-CoA transferase assay. The CoA transferase was assayed with minor modifications as described by Sramek and Frerman (1975). The assay mixture contains 100 µM acetoacetyl-CoA, 10 mM sodium acetate and 5 mM magnesium ion in 67 mM Tris-chloride buffer (pH 8.1). Reaction was initiated by addition of enzyme.

Acetoacetyl-CoA is an unstable compound and was routinely synthesized immediately before use. The compound diketene is a liquid and an anhydride of acetoacetate. In aqueous solution it rapidly is hydrolyzed to acetoacetate. In the presence of CoA acetoacetyl-CoA can be produced from diketene. A quantity of 10 mg of coenzyme A trilithium salt was dissolved in 1.0 ml of 100 mM Tris-HCl (pH 7.5) and incubated with 2 µl diketene at room temperature for 5 minutes. An additional 1 µl diketene was added and the solution stored on ice or frozen until use. The solution of acetoacetyl-CoA was stable for several months when stored frozen.

The enzyme is freely reversible and was measured in the direction of formation of acetyl-CoA from acetoacetyl-CoA and acetate. Thus the decrease in absorbance at 310nm due to disappearance of the acetoacetyl-CoA-magnesium ion complex per unit time is a measure of CoA transferase activity. A unit of enzyme activity is defined as the amount of enzyme required to convert 1 nmole acetoacetyl-CoA to acetoacetate per minute under the above conditions at 25°C.

The procedure included the following steps. A reaction mix was made by adding 0.67 ml of Tris-HCl (pH 8.1), 0.05 ml of magnesium chloride and 9.08 ml of distilled water in a 50 ml Erlenmeyer flask. Into a 1.5 ml self-masking semi-micro quartz cuvette (10mm path length) was added 0.98 ml of the above assay mixture. Next 10 µl of acetoacetyl-CoA and 10 µl of acetate were added and mixed. The cuvette was immediately placed into the spectrophotometer. The decrease in absorbance at 310nm for was recorded for one minute in order to determine the background rate (ΔA_{310} per min.). Next cell free extract was added (1 to 20 µl) and mixed. The cuvette was returned to the spectrophotometer and the decrease in absorbance at 310nm was determined for 2 min. The rate of absorbance change (ΔA_{310} per min.) was calculated.

Acetoacetyl-CoA thiolase Assay. The acetyl-CoA thiolase thiolytically cleaves acetoacetate into two molecules of acetyl-CoA. As for the assay of CoA transferases, the

disappearance of the substrate is measured to determine enzyme activity. The decrease in absorbance at 310 nm due to acetoacetyl-CoA-magnesium ion complex per unit time is a measure of acetoacetyl-CoA thiolase activity. The assay conditions are the same as previously described for the CoA transferase except that CoA is added to the reaction mix instead of acetate. In the assay, 10 µl of a 10 mM CoA solution are added to give a final concentration of 0.1 mM CoA in the reaction mix.

Results

In order to select for mutants altered in their ability to utilize short chain fatty acids as sole carbon sources, *E. coli* strains first had to be made constitutive for expression of the enzymes of ß-oxidation. Strains that are *fadR* were isolated by growth on minimal decanoate (C10) plates. No mutagen was used and all mutants isolated arose spontaneously. These cells arise frequently, yielding up to one hundred colonies per plate when as little as 10^8 cells were plated. Such strains arise only after a week or more of incubation. Colonies arising that are *fadR* show no further lag when grown in minimal decanoate medium and have normal, although quite slow growth. Growth on minimal decanoate plates often requires 48 hour incubation before cell growth becomes visible. Such strains grow readily on long and medium chain fatty acids (and acetoacetate), but still are

medium chain fatty acids (and acetoacetate), but still are unable to grow on the short chain fatty acids butyrate or valerate.

The resulting *fadR* strains were streaked for isolation and used for construction of cells capable of growth on valerate or butyrate. Val^+ or But^+ cells are isolated simply by plating a *fadR* strain onto minimal butyrate or valerate plates (Figure 2.1). When valerate (or butyrate) positive cells are selected, two possible Val^+ phenotypes are observed in a *fadR* background. First some strains exhibit constitutive high levels of the transferase and thiolase which we term $atoC^c$. Second, some strains show induction not by acetoacetate, but by valerate (or butyrate), and are termed $atoC^i$. Screening was performed simply by observation of colony morphology and by enzyme assays for the gene products of the *ato* operon.

When cells are plated onto butyrate plates colonies arise spontaneously that are capable of growth on butyrate as the sole source of carbon. Colonies of But^+ cells are visible after two days. After four days more colonies arise that are all different sizes (Figure 2.1). The colony size ranges from pin point colonies to normal (2-3 mm) diameter colonies. When all colonies arising (regardless of size) are counted, the spontaneous mutation rate for conversion to the But^+ phenotype ranges between 10^{-5} and 10^{-6}. The natural predilection in the laboratory is to pick the first colonies

Figure 2.1. Growth of *E. coli* on butyrate. When *fadR* strains are plated onto a plate where the sole source of carbon is the short chain fatty acid valerate, mutant colonies arise. In this case the plate has been incubated 4 days at 37°C. Note the different colony sizes apparent on the plate.

arising or the larger colonies. But when the different colonies are examined individually, they can be classified based on their regulatory response to the growth substrate. These differences in regulatory properties correlate to differences in colony morphology in the original isolation.

Initial studies were performed using large, rapidly growing clones that first arose on selective media. *E. coli* But$^+$ colonies were picked and streaked for isolation. Large colonies were initially screened for regulatory differences in the expression of the *atoDA* transferase and *atoB* thiolase. One strain EATO1 had a large colony morphology and a 55 minute generation time on liquid butyrate minimal medium. When assayed for the levels of the enzymes under various conditions of growth, high expression of both the transferase and the closely linked thiolase was observed (Table 2.1). The levels of these two enzymes cells was similar to the fully induced levels seen in cells grown on the natural inducer acetoacetate. Incubation of the parent *fadR* strain on butyrate (the cells did not grow) gave undetectable levels of the *ato* gene products. When grown on butyrate, the But$^+$ cells produce up to 3 per cent of their total soluble protein as the CoA transferase based upon the level of expression and the properties of the purified protein.

That the enzymes were expressed constitutively was demonstrated by the high level expression in the absence of the natural inducer acetoacetate. However, butyrate is

naturally converted to acetoacetyl-CoA and it is possible that some interconversions occurred that yielded significant intracellular levels of the natural inducer. In order to test for constitutivity, EATO1 cells were grown in succinate minimal medium, with and with out the short chain fatty acid butyrate. High level expression was observed for the levels of enzymes expressed in either case, though the level of transferase and thiolase was invariably higher in the presence of butyrate. High level expression of transferase and thiolase was observed on succinate alone (Table 2.1, line 1), demonstrating the constitutive expression of the *ato* operon.

Table 2.1. Gene expression in But$^+$ regulatory mutants of *E. coli*.

	Transferase[a] (*atoDA*)	Thiolase[a] (*atoB*)
EATO1 (grown on succinate)	2510	3400
EATO2 (grown on succinate + butyrate)	2330	3060
EATO2 (grown on succinate)	225	410

[a]Enzyme activity is reported as nmoles per minute per mg protein.

In contrast, isolate EATO2 was originally picked as a small colony. Upon subculturing, it retained its small colony phenotype. In addition, EATO2 had a much longer generation time in minimal butyrate liquid medium. The EATO2 strain had a 130 minute generation time, as compared to the 55 minute generation time for EATO1. The strain EATO2 also showed high level expression in the presence of the selective metabolite butyrate (Table 2.1, line 2).

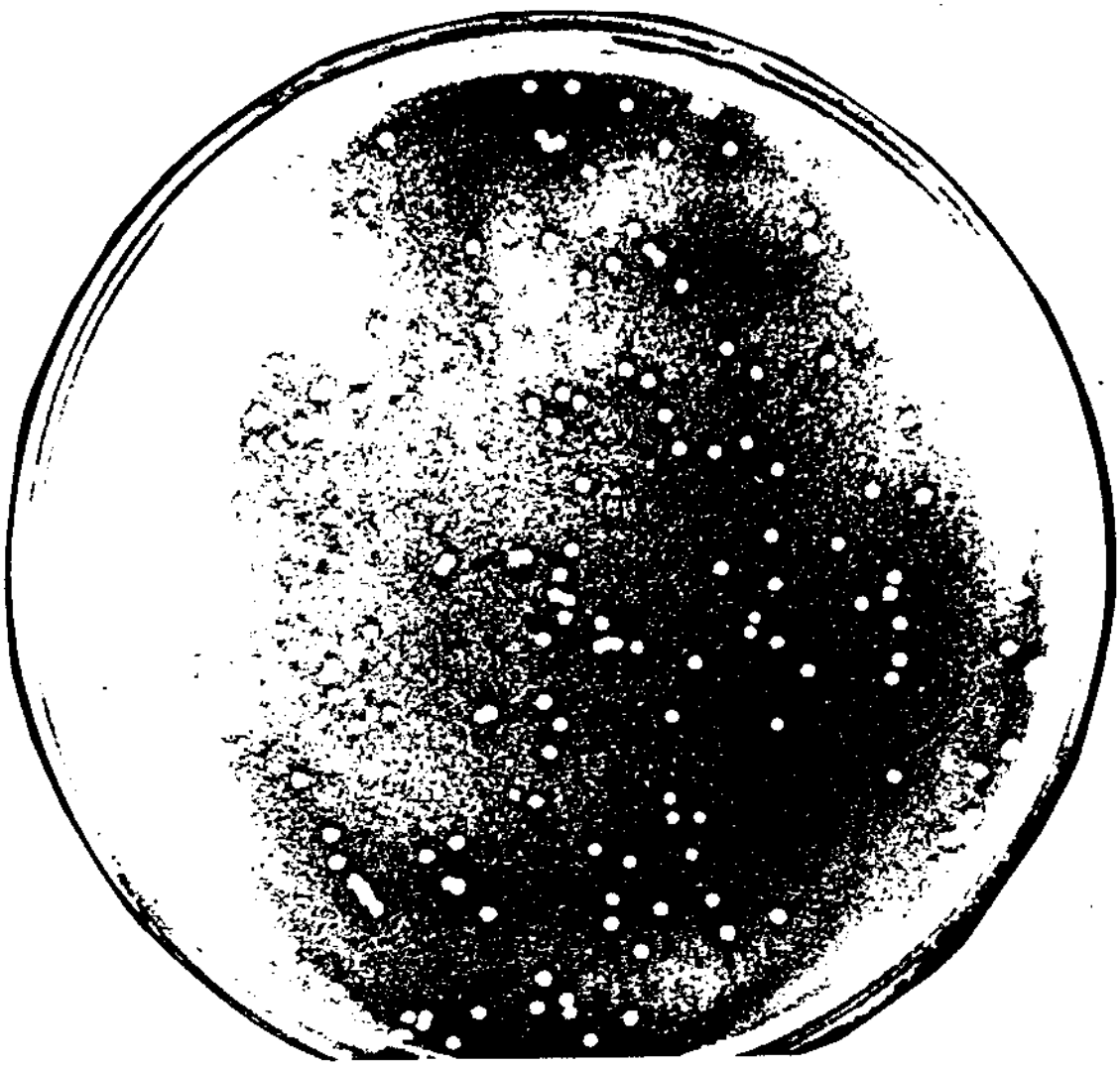
In a manner different to the previously described constitutive strain, EATO2 exhibits only low level expression in the absence of inducer and presence of succinate alone (Table 2.1, line 3). Only about 10 per cent of the isolated But^+ strains exhibit this phenotype. But the phenotypic property of butyrate inducibility (as opposed to true constitutivity) showed complete correlation with colony morphology. Large colonies from the original butyrate selection showed high level expression on succinate minimal medium, small colonies showed high level expression only when butyrate was present. The addition of succinate to EATO2 cultures did not affect expression. That is to say, EATO2 exhibited the level of enzyme specific activity when grown on butyrate medium as on butyrate plus succinate medium.

The small colonies that are inducible by butyrate were termed *atoC*^{i} and the large colonies that appear constitutive under all conditions were designated *atoC*^{c}. The butyrate inducible colonies (*atoC*^{i}) that arise spontaneously on

butyrate selective media are always smaller than the truly constitutive strains (*atoC*^{c}), and they have a much slower growth rate (55 versus 130 minute generation time). Thus those colonies selected on the minimal plate in Figure 2.1 consist of two types. The small colonies are those that now show induction by an alternate inducer instead of acetoacetate, butyrate in this case. The type strain for this phenotype is EATO2. The big colonies show true high level constitutive expression in the absence of inducer like strain EATO1.

Constitutive *atoC*^{c} strains are very stable and retain their big colony phenotype even after growth under non-selective conditions (Figure 2.2). However plating out the slower growing *atoC*^{i} strains onto minimal valerate or butyrate plates selects for the faster growing fully induced phenotype, and small and big colonies are observed. When the difference between large and small colonies was observed, plates were reexamined of differences in the colony morphology on the original selective plates. After lengthy incubation, 4 or more days, there are actually three distinct colony morphs. The large colonies appear after two days. By four days small and pin point size colonies appear. Further incubation of the pin point colonies does not give any increase in their size. The same effect was observed with the small sized colonies. When struck out for isolation and incubated until individual colonies were apparent, the pin

Figure 2.2. Growth of *atoC*^{c} strains. When cells purified from the large colonies are plated onto minimal medium, the cells retain their characteristic rapid growth rate and large colony morphology. Note the consistent size of the colonies.



point and small types retain their respective size colony property. Though lengthy incubation under selective conditions of the pin point colony strains gave rise to variants on the plates.

In order to investigate the stability of the individual phenotypes, cells were grown up and were spread onto selective medium. Consistent results on the size of the colonies and the retention of the phenotypic colony traits was observed. Big colonies retained their large size during plate transfers, whether streaked out or plated out onto selective medium. Small colonies also kept their size and did not show any variation, even after incubation for extended periods under selective conditions. In contrast to the stability of the large and small sized colony types, plating of cells from pin point colonies showed all kinds of size colonies on each transferred plate.

In order to investigate if the cells had some mutation that directly affected colony morphology, the three different morphology types were grown on different media. Large, small and pin point colony strains were plated on minimum medium plus glucose and rich LB medium to investigate if there are other factors causing size morphology differences. All three kinds of colonies had the identical (large) size on both media. The result supports the suggestion that *ato* structural gene expression is the only factor affecting cell

growth. These data indicated that the size difference among the regulatory mutants only existed when short chain fatty acids were used as the sole carbon source in the media. This makes more likely the possibility that the regulatory participants of the *ato* system, which control the expression of enzymes for utilization of short chain fatty acids in *E. coli*, play the main role in this phenomenon.

Though initial studies were performed with butyrate as the selective SCFA, valerate can also be used to isolate *ato* regulatory mutants. When *fadR* cells were plated onto minimal valerate, the same results were obtained. Large colonies arise that show true high level constitutive expression of the CoA transferase and thiolase under all conditions of growth. These Val^+ strains with large colony morphology were termed $atoC^c$ as before. Small colonies were termed $atoC^i$ like the small $atoC^i$ strains isolated on butyrate.

Differences in enzyme expression level were investigated in the different size classes of Val^+ mutants. Clones which had been cultivated through several passages and which had visibly distinct sizes on plate culture were chosen. Three different size strains from the three types of colonies were cultured on *E. coli* minimal medium with different carbon sources. Cells were harvested during the late log phase and cell-free crude extracts were made by passage of cells through the French pressure cell. The enzyme assays were done by the method described previously.

As before, the large colonies were constitutive for expression of the *atoDA* transferase and the *atoB* thiolase. Growth on succinate yield high level expression of the enzymes, even in the absence of inducer (Table 2.2). Growth on valerate alone or in combination with succinate gave some what higher levels of expression. In addition, the catabolic *ato* operon was examined for effects of catabolite repression by glucose. Since the *ato* genes are synthesized for the utilization of acetoacetate, it is possible that the genes were subject to control by carbon source availability. When cultured on glucose minimal medium, there was a 75 per cent drop in the activity compared to the constitutive level in the succinate grown cells where no catabolite repression was expected. The parent *fadR* and wild type strains have undetectable levels of the transferase and thiolase when cells are grown on glucose.

As expected, small colonies of Val^+ cells had distinctive patterns of gene expression for the enzymes of the *ato* operon. When the small Val^+ cells are grown on succinate alone, low levels of the enzymes are present. On valerate or succinate plus valerate, the enzymes are expressed at fully induced levels. These cells now responded to the presence of valerate as an inducer of the enzymes of the *ato* operon. The presence of glucose caused catabolite repression and a decrease in the specific activity of the transferase and thiolase. Though low level expression is

observed on succinate alone, there was a four fold or greater induction of enzyme synthesis in the presence of valerate and so these cells were classified as *atoC*^{i}.

TABLE 2.2. Gene expression of the *atoDA* transferase in regulatory mutants of *E. coli*.

Colony Size	Carbon Source	Enzyme Activity[a]
Large	Succinate	1300
	Glucose	380
Small	Valerate	2400
	Succinate	680
	Glucose	260
Pin Point	Succinate	0
	Succinate + Valerate	1550
	Glucose + Valerate	0

[a]Enzyme activity is expressed as nmoles per minute per mg protein

More dramatic and complete induction was observed in the pin point colonies. These cells had undetectable levels of enzyme when grown on succinate. However, when the inducer valerate was added to growing cultures of the cells, there was a large increase in the level of transferase expression (Table 2.2). Constitutive expression of the CoA transferase and thiolase was not observed. These cells had no background level of expression of the enzymes when cultivated on

succinate and so were also classified as *atoC*^{i}. Growth of these cells on glucose and the inducer butyrate also gave little or no *ato* gene expression.

Discussion

Wild type *E. coli* can grow on the compound acetoacetate using the enzymatic activities encoded by the *ato* operon. A CoA transferase converts acetoacetate to acetoacetyl-CoA using acetyl-CoA as the CoA donor. Subsequently a thiolase cleaves the acetoacetyl-CoA into two molecules of acetyl-CoA which can then enter intermediary metabolism. However, native strains of *E. coli* can not grow on the short chain fatty acid molecules butyrate and valerate, even though acetoacetate is a common intermediate in their degradation. Additional enzymes of ß-oxidation are required and these are not induced by SCFA molecules. Nor are the activities of the *ato* operon induced by any other compound that acetoacetate. Plating a *fadR* parent strain of *E. coli* (constitutive for the expression of the ß-oxidation pathway) onto selective medium containing either valerate or butyrate as the sole source of carbon and energy allows for the selection of mutant strains that express the enzymes of the *ato* operon in the absence of the normal inducer acetoacetate.

Several types of strains of *E. coli* can be observed based on colony morphology after selection of a *fadR* parent for growth on the short chain fatty acids butyrate and

valerate. Similar results are observed when no matter which of the two SCFA compounds are used for selection. Some cells are truly constitutive and have a large colony phenotype. These cells have a generation time of 55 minutes on the selective carbon source. They express the *atoDA* transferase and the *atoB* thiolase in the absence of the inducer acetoacetate and even under growth conditions where the pathway is not even utilized, for example during growth on succinate. The enzymes are expressed at high constitutive levels and are each present in the cell at approximately three per cent of the total cellular protein. The enzymes are still expressed in the presence of glucose, though the level of expression is repressed significantly to about a third of the level in the absence of glucose. These cells are termed *atoC*^{c} and except when catabolite repressed, exhibit high level expression of the CoA transferase and thiolase, in the absence of any inducer. These mutants are presumed to be *cis*-acting promoter mutants that no longer bind the repressor protein gene product of the *atoC* gene. It is also possible that this class of mutants represents mutants in the *atoC* gene product's DNA binding region, making the protein unable to bind to the promoter to affect regulation of the *atoDAB* operon. Though the *ato* gene region (*atoDAB* operon and *atoC* gene) from these mutants has been cloned, the mutants have not been sequenced to precisely

locate the mutation and correlate a change in the nucleotide sequence with a change in the regulatory properties.

Another class of regulatory mutants have a smaller colony size than the *atoC*^{c} mutants. They may be termed partly inducible or partly constitutive, though neither term is satisfactory. These small colonies have low but significant levels of the enzymes of the *ato* operon when grown on succinate in the absence of any inducer. But when these cells are grown in the presence of valerate, they exhibit high level expression of the inducer. Thus under non-induced conditions, they exhibit low level (constitutive) gene expression. When inducer is added (now valerate) the cells respond with an increase in *ato* gene expression to fully induced levels similar to those found in the wild type strain (Table 2.2). These mutants have been termed *atoC*^{i}.

Most significant about this class of mutants is the regulatory response to inducer. Though there is always present a background level of enzyme activity, even in succinate minimal medium with no SCFA present, addition of valerate gives full induction of the *ato* operon. In the wild type cells, valerate has no effect. In Val^{+} cells, valerate acts as an inducer. Thus these cells have an altered regulatory response. Instead of acetoacetate acting as the inducer of the pathway, valerate induces high level gene expression. It is tempting to speculate that the *atoC* gene product has been altered so that it now binds valerate

instead of acetoacetate to affect the allosteric transition that results in gene expression. The binding site of the *atoC* product has not yet been identified. Sequence analysis of the mutants may provide insight into the location of the inducer binding site in the primary sequence of the AtoC regulatory protein. Since there is some low level expression of the *ato* operon in these mutants even in the absence of inducer, it is possible that the protein is also affected in its DNA binding ability. Whether different mutants within this class harbor distinct mutations is unknown.

Another class of mutants always produced pin point colonies on medium with SCFA as the carbon source. These colonies regularly gave rise to large colonies when struck onto and maintained on selective medium. This class of mutants also termed *atoC*^{i} showed induction of the *ato* operon's gene products in the presence of inducing growth substrate. Those strains selected for growth on butyrate responded to butyrate as the inducer and those strains selected on valerate showed induction by growth on valerate. When these *atoC*^{i} strains were grown up in liquid medium with succinate (or glucose) as the carbon source no basal level of *ato* gene expression could be measured. Transferase and thiolase activity were undetectable. Growth in minimal valerate medium, of valerate inducible strains, gave high level expression of the *ato* operon (Table 2.2). Since there was complete repression of the operon in the absence of

inducer, it is possible that this class of mutants is altered solely in the specificity of inducer binding by the *atoC* gene product. These *atoC*^{i} strains may be *trans*-acting mutants of the *atoC* protein that now respond to valerate (or butyrate as the case may be), but still normally and tightly bind the operator under uninduced conditions.

All of the mutant classes have been loaded with transposons in the *ato* operon. Selection was made against ability to utilize SCFA or acetoacetate for growth. Clones were then constructed for future studies. Sequence analysis may reveals differences in these different *atoC* mutant strains, differences that correlate to changes in the regulatory properties of the strains. There are three distinct classes of mutants. The *atoC*^{c} strains exhibit constitutive high level expression in the absence of inducer. Inducible *atoC*^{i} strains exhibit altered patterns of induction. Some *atoC*^{i} strains have a basal level of *ato* enzyme synthesis, but are inducible by a new small molecule inducer. Specificity of the AtoC protein has been altered from recognition of acetoacetate to either valerate or butyrate. Some *atoC*^{i} strains show normal lack of expression in the absence of inducing metabolites, but likewise have altered specificity of induction. It is tempting to speculate that the difference in succinate background expression in these mutants may represent a spectrum of altered regulatory mutants. Once the precise genotype and

primary sequence identity of a number of mutants is known, then specific changes may be quantized and correlated with specific regulatory regions; specifically the *atoC* gene product DNA binding site, the *atoC* gene product inducer binding site, or the operator of the *atoDAB* operon, *i.e.* the DNA sequence binding the *atoC* gene product itself.

CHAPTER III

REGULATORY SEQUENCES CONTROLLING SHORT CHAIN FATTY ACID METABOLISM IN *E. COLI*

It was first proposed by Garen and Echol that a regulator gene could specify "the formation of an endogenous inducer for alkaline phosphatase synthesis" in 1962. A year later, Englesberg suggested a similar hypothesis for the regulation of enzymes involved in L-arabinose metabolism. Both of these groups got little immediate acceptance on their concepts by the scientific community. It took the extensive study of Englesberg's team on the L-arabinose system (Wilcox et al., 1971), Raibaud's group on the maltose system (Gutierrez & Raibaud, 1984; Debarbouille et al., 1978; Raibaud et al., 1983; 1984), and independent confirmation by other researchers on the function of catabolic activator protein (Anderson et al., 1971; DeCrombrugghe et al., 1984; Riggs et al., 1971) to convince people of the concept of positive control of gene expression. Historically it has always been easier to destroy than to create, to disrupt than to enhance, and to turn off than to turn on. Now positive

control and gene activation is widely accepted and commonly observed, thirty years after it was first proposed.

A gene that can be positively controlled is often identified by the finding that mutations located outside the structural gene prevent or greatly decrease its expression. For positive controlled systems, such mutations should be recessive to the wild type allele; be reasonably frequent; comprise deletions, insertion, or nonsense mutations; and affect this gene in the initiation of transcription but not other steps in gene expression (such as processing, translation or post-translational modification). Positive control is defined as a requirement of transcription initiation at a given promoter for the participation of a protein factor that is either not always present, or not always active, in exponentially growing cells. The protein factor is called an activator, as originally proposed by Englesberg *et al*. An activator protein is able to initiate the transcription and to effect expression of the genes under its control. Raibaud (Raibaud & Schwartz, 1984) pointed out that the functions of activator proteins could be a) to act as an accessory factor that allows RNA polymerase to initiate transcription at specific promoter; b) to participate as a factor that replaces one of the subunits of RNA polymerase, thereby altering its promoter-recognition specificity; or c) to be an entirely new RNA polymerase.

by different chain length fatty acids (Clark, 1981; Dirusso, 1988; Dirusson & Nunn, 1984; Nunn, 1986).

The *ato* operon in *E. coli* is induced by acetoacetate (Pauli & Overath, 1972). Cells respond to the presence of the inducer acetoacetate with a dramatic increase in the level of expression of the acetoacetyl:acetylCoA transferase and AcetoacetylCoA thiolase. This increase occurs only under conditions of glucose limitation. In the presence of glucose, no expression of the *ato* operon (*atoDAB* structural genes) is observed. Presumably, when glucose is limiting, cAMP levels in the cell rise and bind to catabolite activator protein (CAP; also know as catabolite repressor protein or CRP and cAMP binding protein). The cAMP-CAP complex binds to specific sequences in the DNA and allow for further regulation of the genes under control catabolite repression. These genes are normally inducible (either positively or negatively inducible) and are not normally expressed in the absence of a specific small molecule inducer. Catabolite repressible operons are observed to have poor -35 regions that lack consensus DNA sequences and this lack of homology prevents efficient recognition by RNA polymerase (Adhya & Garges, 1990; Gottesman, 1984). Binding of the cAMP-CAP complex under conditions of glucose limitation is a necessary but not sufficient requirement for gene expression. Binding of cAMP-CAP to the CAP binding site primes these catabolite repressible operons to become active, if inducer subsequently

is present. The CAP site is a specific sequence in the DNA with a consensus sequence of AANTGTGANNT, though considerable divergence is apparently tolerated compared to other DNA recognition sequences. Binding of cAMP-CAP allows promoters with poor -35 recognition sequences to bind RNA polymerase (Von Hippel et al., 1984). Then gene expression may subsequently controlled by positive or negative control (Adhya & Garges, 1990; DeCrombrugghe, et al., 1984).

The *ato* operon is positively inducible by acetoacetate. When grown on acetoacetate (in the absence of glucose), *E. coli* will turn on expression of the *ato* operon. This induction is achieved by a *trans*-acting protein the AtoC protein, product of the *atoC* gene. AtoC works in *trans* to stimulate expression of the *ato* operon. Deletion of *atoC* results in a loss of induction of the ato operon and a phenotypic loss of the ability to utilize acetoacetate. Constitutive mutants have been isolated (Chapter II) that have not yet been precisely mapped, so that they have not yet been identified as *cis*- or *trans-acting* (promoter or *atoC* respectively) mutants. This work sought to characterize the nature of the regulatory sequences affecting regulation of the *ato* operon. Specifically, the sequence of the *atoC* gene was obtained in order to compare the predicted amino acid sequence to other regulatory proteins. The sequence of the promoter operator region of the *ato* operon was determined so that important regulatory sequences and protein binding sites

might be identified. It was hoped that sequence identity of the regulatory protein and regulatory regions upstream of the *ato* operon would provide insight into the regulation of gene expression in this system.

Materials and Methods

Bacterial strains and plasmids. The plasmid pLJ10 was obtained from Laren Sallus Jenkins (Jenkins & Nunn, 1987a). This plasmid carries a 6.0 kbp fragment and contains the *ato* operon and the neighboring *atoC* regulatory gene. The fragment has *Sau*3A and *Bam*HI ends (Figure 1.6), but since this clone was isolated from a Sau3A partial digest, it has internal BamH1 and Sau3A sites. The DNA containing the *ato* genes was subcloned into the plasmid pUC19 (Messing, 1983). All clones were manipulated in the host strain DH5α. DH5α was obtained from Gibco-BRL laboratories. It has the genotype F^-, λ^-, *recA1*, Δ(*lacZYA-argF*) U169, *hsdR17*, thi-1, *gyrA96*, *supE44*, *endA1*, *relA1*, ϕ80d*lacZ*ΔM15.

Growth and maintenance of strains. Cells were grown on YT medium. YT medium contains 5 g per 1 yeast extract, 8 g per 1 tryptone and 5 g per 1 sodium chloride. The medium was solidified for agar plates by the addition of 15 g per 1 agar. Cells were grown at 37°C with shaking. In order to maintain selection for the plasmid, 75 to 100 μg per ml ampicillin was added to the medium.

Plasmid purification. Plasmids were purified as described in Maniatis et al. (1982). The plasmid preparation was one of two types. For rapid plasmid analysis of transformants, a rapid plasmid preparation was performed. For preparation of large amounts of plasmid DNA a bulk plasmid preparations from amplified cells was performed. The methods for rapid plasmid preparation was as follows. Cells were grown in 5 ml YT medium at 37°C incubator on a New Brunswick TB1 tube roller at 50 rpm. Cells in a total volume of 3 ml were lysed by alkaline buffer solution to obtain enough sample for at least, three sequencing reactions or ten samples restriction enzyme digestion (Kraft et al., 1988; McClelland et al., 1988).

DNA sequencing. DNA sequencing was performed as single or double stranded DNA sequencing by Sanger's dideoxynucleotide method (1983) using the ^{35}S-labeled dATP (Biggin et al., 1983) and Sequenase from USB Corporation. Buffer gradient polyacrylamide gels (TBE concentration from 2.5X to 0.5X) was employed in order to have the better sequencing resolution (Biggin et al., 1983). Oligonucleotides were either synthesized by a Pharmacia Gene Assembler DNA synthesizer or purchased from Bio-Synthesis, Inc. After desalting by running through a Pharmacia Sephadex G-25 column, crude oligonucleotide solutions were used directly without further purification.

DNA sequence analysis. After the DNA sequencing gels were completely dry, the gels were exposed to a sheet of Kodak XAR5 X-ray film. After 1 to 5 days incubation, the films were developed. Since ^{35}S will not penetrate the film support to expose the emulsion on the distal side of the film, this emulsion was removed by stripping with a 1 to 10 dilution of commercial bleach. Care was taken not to wet the exposed side of the film with the bleach. The film was rinsed, dried and illuminated. Sequences were read and entered into file on the computer. Typically, the sequences were entered into a sequence file using the program DNA Inspector. The sequences were independently confirmed by reading the film twice, reading back the sequence using the computer and following along on the gel and having a different investigator read the DNA sequence. Since the sequence was determined on both strands, it was logged into the final sequence file only after opposite strand confirmation.

Results

A plasmid containing the *atoC* gene sequence was constructed from the parent plasmid pLJ10. Plasmid pMCYC1 was digested with *Sau*3A and *Bgl*II to produce a 2.5 kbp subclone that carried *atoC* and part of the *atoDAB* operon.

The presence of *atoC* on the subcloned fragment was assumed based on previous map results from the work of Jenkins and Nunn (1987a). The resulting subclone was used for subsequent sequencing.

The sequence of the first 1720 bp of the *atoC* gene region beginning from the Sau3A site on the 5' side of *atoC* is shown in the next illustration (Figure 3.1). Since complementational and mutational studies were not performed, the exact location of the *atoC* coding sequence was not known. In addition, none of the amino acid sequence of the protein was known, so no examination of the deduced amino acid sequence could provide exact clues for the actual location of the *atoC* coding region. The *atoC* coding sequence was identified by open reading frame analysis of the completed sequence. Only one reading frame of sufficient size to accommodate a polypeptide of the reported 48,000 Dalton molecular weight could be identified by inspection.

An open reading frame with an approximate correct size was situated almost immediately interior to the leftmost *Sau3A* site of the clone. It was difficult to identify the precise start of the clone since the open reading frame extended from four different methionine residues (Figure 3.2) that all exisit in the same open reading frame. The first reading frame is the only frame that when extended gave a coding sequence of approximately the correct size for the *atoC* gene product. It has been reported from gel

Figure 3.1. The nucleotide sequence from Sau3A to position 1720. The plus strand (coding) sequence for the region is shown.

```
        10         20         30         40         50         60
GATCAACCCG CAGGAAATCA GACTGTATGA CTGCTATTAA TCGCATCCTT ATTGTGGATG
                                                              120
ATGAAGATAA TGTTGCGCCG TATGCTGAGC ACCGCTTTTG CACTACAAGG ATTCGAAACA
                                                              180
CATTGTGCGA ACAACGGACG CACAGCATTA CACCTGTTTG CCGATATTCA CCCTGATGTG
                                                              240
GTGTTGATGG ATATCCGCAT GCCAGAGATG GACGGCATCA AGGCACTAAA GGAGATGCGC
                                                              300
AGCCATGAGA CCCGGACACC CGTTATTCTG ATGACGGCCT ATGCGGAAGT GGAAACCGCC
                                                              360
GTCGAAGCGC TACGCTGCGG AGCCTTCGAC TATGTTATTA AACCGTTTGA TCTCGATGAG
                                                              420
TTGAATTTAA TCGTTCAGCG CGCTTTACAA CTCCAGTCAA TGAAAAAAGA GATCCGTCAT
                                                              480
CTGCACCAGG CACTGAGCAC CAGCTGGCAA TGGGGGCACA TTCTCACCAA CAGCCCGGCG
                                                              540
ATGATGGACA TCTGCAAAGA CACCGCCAAA ATTGCCCTTT CTCAGGCCAG CGTCTTGATT
                                                              600
AGCGGTGAAA GCGGCACCGG GAAAGAGTTG ATTGCCAGAG CGATTCACTA CAATTCGCGG
                                                              660
CGGGCAAAGG GGCCGTTCAT TAAAGTCAAC TGCGCGGCGC TGCCGGAATC GTTGCTCGAA
                                                              720
AGTGAACTGT TTGGTCATGA AAAAGGTGCA TTTACTGGTG CACAAACCTT GCGTCAGGGA
                                                              780
TTATTTGAAC GAGCCAACGA AGGTACTCTG CTCCTCGACG AAATTGGCGA AATGCCGCTG
                                                              840
GTACTACAAG CCAAATTACT ACGCATTCTA CAGGAACGGG AATTTGAACG GATTGGCGGC
                                                              900
CATCAGACCA TAAAAGTTGA TATCCGCATC ATTGCTGCCA CCAACCGCGA CTTGCAGGCA
                                                              960
ATGGTAAAAG AAGGCACCTT CCGTGAAGAT CTCTTTTATC GCCTTAACGT TATTCATTTA
                                                             1020
ATACTGCCGC CTCTGCGCGA TCGCCGGGAA GATATTTCCC TGTTAGCTAA TCACTTTTTG
                                                             1080
CAAAAATTCA GTAGTGAGAA TCAGCGCGAT ATTATCGACA TCGATCCGAT GGCAATGTCA
                                                             1140
CTGCTTACCG CCTGGTCATG GCCGGGAAAT ATTCGAGAGC TTTCCAACGT TATTGAACGC
                                                             1200
GCCGTCGTGA TGAATTCAGG CCCGATCATT TTTTCTGAGG ATCTTCCGCC ACAGATTCGT
                                                             1260
CAGCCAGTCT GTAATGCTGG CGAGGTAAAA ACAGCCCCTG TCGGTGAGCG TAATTTAAAA
                                                             1320
GAGGAAATTA AACGCGTCGA AAAACGCATC ATTATGGAAG TGCTGGAACA ACAAGAAGGA
                                                             1380
AACCGAACCC GCACTGCTTT AATGCTGGGC ATCAGTCGCC GTGCATTGAT GTATAAACTC
                                                             1440
CAGGAATACG GTATCGATCC GGCGGATGTA TAACACCAAA ACTTGCTATG CAGAAATTTG
                                                             1600
CACAGTGCGC AATTTTCTGC ATAGCCGCTC ATTCTCCTTA TAAATCCCCA TCCAATTTAT
                                                             1660
CCCTTCATAT TCAATTAGTT AAATAACTAA ATCCAATAAT CTCATTCTGG CACTCCCCTT
                                                             1720
GCTATTGCCT GACTGTACCC ACAACGGTGT ATGCAAGAGG GATAAAAAAT GAAAACAAAA
```

Figure 3.2. The nucleotide sequence and predicted translation from the Sau3A cut site to nucleotide position 120. This corresponds to 40 amino acids worth of coding. All three frames are translated. Stop codons are denoted by the asterisks. The *atoC* open reading frame is continuous from position 31. The first (top) frame contains this open reading frame and extends to nucleotide position 1413 to a TAA stop. The actual start of the *atoC* gene is ambiguous since four AUG methionine starts are evident (underlined).

```
                                                          30
GAT CAA CCC GCA GGA AAT CAG ACT GTA TGA
Asp Gln Pro Ala Gly Asn Gln Thr Val ***
 Ile Asn Pro Gln Glu Ile Arg Leu Tyr Asp
  Ser Thr Arg Arg Lys Ser Asp Cys Met Thr

                                                          60
CTG CTA TTA ATC GCA TCC TTA TTG TGG ATG
Leu Leu Leu Ile Ala Ser Leu Leu Trp Met
 Cys Tyr *** Ser his Pro Tyr Cys Gly ***
  Ala Ile Asn Arg Ile Leu Ile Val Asp Asp

                                                          90
ATG AAG ATA ATG TTG CGC CGT ATG CTG AGC
Met Lys Ile Met Leu Arg Arg Met Leu Ser
 *** Arg *** Cys Cys Ala Val Cys *** Ala
  Glu Asp Asn Val Ala Pro Tyr Ala Glu His

                                                         120
ACC GCT TTT GCA CTA CAA GGA TTC GAA ACA
Thr Ala Phe Ala Leu Gln Gly Phe Glu Thr
 Pro Leu Leu His Tyr Lys Asp Ser Lys
  Arg Phe Cys Thr Thr Arg Ile Arg Asn
```

electrophoresis of labels proteins from maxicells (Sancar et al., 1979) that *atoC* is approximately 48,000 Daltons in size (Jenkins & Nunn, 1987a). Translation of this reading frame from the different AUG starts gave a protein molecular weight for the predicted products ranging between 50,048 and 51,109 Daltons. The shortest protein had a sequence eight, seven, or four amino acids short than the longer products. Considering the accuracy of the size estimate from maxicell analysis, none of the possible start sites could be confirmed or eliminated.

However, it is possible to predict the actual AUG translational start of the *atoC* open reading frame from the location of a consensus Shine-Dalgarno sequence (Weyens et al., 1988). The only methionine codon with a consensus ribosome binding site, at the proper position (four to eight nucleotides distant), is the third methionine shown in Figure 3.3. Much weaker homology with the consensus Shine Dalgarno sequence can also be demonstrated for the fourth methionine. This may be a possible translational start site and would yield a protein with four fewer amino acids on the amino terminus.

When the predicted translation is performed *atoC* encodes a protein that is448 amino acids long. The complete DNA and amino acid sequences are shown in Figure 3.4. This translation yields a protein with a predicted molecular

Figure 3.3. Shine Dalgarno sequence and predicted AUG start for *atoC*. The DNA sequence from the previous figures is shown with predicted translation showing all possible translational start sites. Only one AUG start is correctly positioned after a consensus (5/7) ribosome binding site and it is denoted with underlying asterisks.

```
          10          20          30          40          50
GATCAACCCG CAGGAAATCA GACTGTATGA CTGCTATTAA TCGCATCCTT

        AgG AgGAnnnnnA TG        80          90          100
ATTGTGGATG ATGAAGATAA TGTTGCGCCG TATGCTGAGC ACCGCTTTTG
        Met MetLysIleM etLeuArgAr gMetLeuSer ThrAlaPheA
                    * **
         110         120
CACTACAAGG ATTCGAAACA
laLeuGlnGl yPheGluThr
```

Figure 3.4. Nucleotide sequence and translation of *atoC*. The plus strand DNA sequence of the *atoC* gene and the predicted amino acid sequence are shown. Numbers refer to the codon or amino acid position. The DNA is 1344 nucleotides long including the ATG start and TAA stop and it encodes a protein of 1341 amino acids. The sequence is continued on the next page.

```
              9          18          27          36          45          54
5' ATG TTG CGC CGT ATG CTG AGC ACC GCT TTT GCA CTA CAA GGA TTC GAA ACA CAT
   Met Leu Arg Arg Met Leu Ser Thr Ala Phe Ala Leu Gln Gly Phe Glu Thr His

             63          72          81          90          99         108
   TGT GCG AAC AAC GGA CGC ACA GCA TTA CAC CTG TTT GCC GAT ATT CAC CCT GAT
   Cys Ala Asn Asn Gly Arg Thr Ala Leu His Leu Phe Ala Asp Ile His Pro Asp

            117         126         135         144         153         162
   GTG GTG TTG ATG GAT ATC CGC ATG CCA GAG ATG GAC GGC ATC AAG GCA CTA AAG
   Val Val Leu Met Asp Ile Arg Met Pro Glu Met Asp Gly Ile Lys Ala Leu Lys

            171         180         189         198         207         216
   GAG ATG CGC AGC CAT GAG ACC CGG ACA CCC GTT ATT CTG ATG ACG GCC TAT GCG
   Glu Met Arg Ser His Glu Thr Arg Thr Pro Val Ile Leu Met Thr Ala Tyr Ala

            225         234         243         252         261         270
   GAA GTG GAA ACC GCC GTC GAA GCG CTA CGC TGC GGA GCC TTC GAC TAT GTT ATT
   Glu Val Glu Thr Ala Val Glu Ala Leu Arg Cys Gly Ala Phe Asp Tyr Val Ile

            279         288         297         306         315         324
   AAA CCG TTT GAT CTC GAT GAG TTG AAT TTA ATC GTT CAG CGC GCT TTA CAA CTC
   Lys Pro Phe Asp Leu Asp Glu Leu Asn Leu Ile Val Gln Arg Ala Leu Gln Leu

            333         342         351         360         369         378
   CAG TCA ATG AAA AAA GAG ATC CGT CAT CTG CAC CAG GCA CTG AGC ACC AGC TGG
   Gln Ser Met Lys Lys Glu Ile Arg His Leu His Gln Ala Leu Ser Thr Ser Trp

            387         396         405         414         423         432
   CAA TGG GGG CAC ATT CTC ACC AAC AGC CCG GCG ATG ATG GAC ATC TGC AAA GAC
   Gln Trp Gly His Ile Leu Thr Asn Ser Pro Ala Met Met Asp Ile Cys Lys Asp

            441         450         459         468         477         486
   ACC GCC AAA ATT GCC CTT TCT CAG GCC AGC GTC TTG ATT AGC GGT GAA AGC GGC
   Thr Ala Lys Ile Ala Leu Ser Gln Ala Ser Val Leu Ile Ser Gly Glu Ser Gly

            495         504         513         522         531         540
   ACC GGG AAA GAG TTG ATT GCC AGA GCG ATT CAC TAC AAT TCG CGG CGG GCA AAG
   Thr Gly Lys Glu Leu Ile Ala Arg Ala Ile His Tyr Asn Ser Arg Arg Ala Lys

            549         558         567         576         585         594
   GGG CCG TTC ATT AAA GTC AAC TGC GCG GCG CTG CCG GAA TCG TTG CTC GAA AGT
   Gly Pro Phe Ile Lys Val Asn Cys Ala Ala Leu Pro Glu Ser Leu Leu Glu Ser

            603         612         621         630         639         648
   GAA CTG TTT GGT CAT GAA AAA GGT GCA TTT ACT GGT GCA CAA ACC TTG CGT CAG
   Glu Leu Phe Gly His Glu Lys Gly Ala Phe Thr Gly Ala Gln Thr Leu Arg Gln

            657         666         675         684         693         702
   GGA TTA TTT GAA CGA GCC AAC GAA GGT ACT CTG CTC CTC GAC GAA ATT GGC GAA
   Gly Leu Phe Glu Arg Ala Asn Glu Gly Thr Leu Leu Leu Asp Glu Ile Gly Glu

            711         720         729         738         747         756
   ATG CCG CTG GTA CTA CAA GCC AAA TTA CTA CGC ATT CTA CAG GAA CGG GAA TTT
   Met Pro Leu Val Leu Gln Ala Lys Leu Leu Arg Ile Leu Gln Glu Arg Glu Phe

            765         774         783         792         801         810
   GAA CGG ATT GGC GGC CAT CAG ACC ATA AAA GTT GAT ATC CGC ATC ATT GCT GCC
   Glu Arg Ile Gly Gly His Gln Thr Ile Lys Val Asp Ile Arg Ile Ile Ala Ala

            819         828         837         846         855         864
   ACC AAC CGC GAC TTG CAG GCA ATG GTA AAA GAA GGC ACC TTC CGT GAA GAT CTC
   Thr Asn Arg Asp Leu Gln Ala Met Val Lys Glu Gly Thr Phe Arg Glu Asp Leu
```

```
        873         882         891         900         909         918
TTT TAT CGC CTT AAC GTT ATT CAT TTA ATA CTG CCG CCT CTG CGC GAT CGC CGG
Phe Tyr Arg Leu Asn Val Ile His Leu Ile Leu Pro Pro Leu Arg Asp Arg Arg

        927         936         945         954         963         972
GAA GAT ATT TCC CTG TTA GCT AAT CAC TTT TTG CAA AAA TTC AGT AGT GAG AAT
Glu Asp Ile Ser Leu Leu Ala Asn His Phe Leu Gln Lys Phe Ser Ser Glu Asn

        981         990         999        1008        1017        1026
CAG CGC GAT ATT ATC GAC ATC GAT CCG ATG GCA ATG TCA CTG CTT ACC GCC TGG
Gln Arg Asp Ile Ile Asp Ile Asp Pro Met Ala Met Ser Leu Leu Thr Ala Trp

       1035        1044        1053        1062        1071        1080
TCA TGG CCG GGA AAT ATT CGA GAG CTT TCC AAC GTT ATT GAA CGC GCC GTC GTG
Ser Trp Pro Gly Asn Ile Arg Glu Leu Ser Asn Val Ile Glu Arg Ala Val Val

       1089        1098        1107        1116        1125        1134
ATG AAT TCA GGC CCG ATC ATT TTT TCT GAG GAT CTT CCG CCA CAG ATT CGT CAG
Met Asn Ser Gly Pro Ile Ile Phe Ser Glu Asp Leu Pro Pro Gln Ile Arg Gln

       1143        1152        1161        1170        1179        1188
CCA GTC TGT AAT GCT GGC GAG GTA AAA ACA GCC CCT GTC GGT GAG CGT AAT TTA
Pro Val Cys Asn Ala Gly Glu Val Lys Thr Ala Pro Val Gly Glu Arg Asn Leu

       1197        1206        1215        1224        1233        1242
AAA GAG GAA ATT AAA CGC GTC GAA AAA CGC ATC ATT ATG GAA GTG CTG GAA CAA
Lys Glu Glu Ile Lys Arg Val Glu Lys Arg Ile Ile Met Glu Val Leu Glu Gln

       1251        1260        1269        1278        1287        1296
CAA GAA GGA AAC CGA ACC CGC ACT GCT TTA ATG CTG GGC ATC AGT CGC CGT GCA
Gln Glu Gly Asn Arg Thr Arg Thr Ala Leu Met Leu Gly Ile Ser Arg Arg Ala

       1305        1314        1323        1332        1341
TTG ATG TAT AAA CTC CAG GAA TAC GGT ATC GAT CCG GCG GAT GTA TAA  3'
Leu Met Tyr Lys Leu Gln Glu Tyr Gly Ile Asp Pro Ala Asp Val ***
```

weight of 50,605 Daltons. The amino acid composition is shown in Table 3.1.

Table 3.1. Amino acid content of the *atoC* gene product.

Amino Acid			Number	Mole %
Alanine	Ala	A	38	8.48
Arginine	Arg	R	35	7.81
Asparagine	Asn	N	17	3.79
Aspartic Acid	Asp	D	21	4.69
Cysteine	Cys	C	5	1.12
Glutamine	Gln	Q	20	4.46
Glutamic Acid	Glu	E	38	8.48
Glycine	Gly	G	25	5.58
Histidine	His	H	12	2.68
Isoleucine	Ile	I	37	8.26
Leucine	Leu	L	53	11.83
Lysine	Lys	K	20	4.46
Methionine	Met	M	18	4.02
Phenylalanine	Phe	F	15	3.35
Proline	Pro	P	18	4.02
Serine	Ser	S	23	5.13
Threonine	Thr	T	21	4.69
Tryptophan	Trp	W	4	0.89
Tyrosine	Tyr	Y	6	1.34
Valine	Val	V	22	4.91

The first three columns list each amino acid and its three letter and single letter abbreviations, in that order. The number of amino acids given is out of a total amino acid chain length of 448. This value was also used to determine the mole per cent of each amino acid.

The *atoC* open reading frame peptide is predicted to have a molecular weight of 50,605 Daltons with the amino

terminal methionine. This assumes that the methionine (or any other residue) is not processed of posttranslationally. Approximately fifty percent of all proteins in *E. coli* have the amino terminal methionine removed, but the actual case for the *atoC* gene product in the cell is not known. The predicted amino acid sequence of the *atoC* protein yields a protein with an isoelectric point of 6.52. Information about the physical properties of the protein may aid in future attempts at purification.

When the region upstream of the *ato* promoter is examined, no consensus promoter sequences are obvious. The open reading frame that we have identified for *atoC* is only 70 base pairs from the end of our clone. It is possible that the promoter lies partly or completely outside of this region. Or it is possible that the atoC protein has a weak promoter sequence or one that is not typical of other consensus promoters. If the stringency of the homology is lowered for this region, several possible promoters become plausible. Figure 3.5 shows two possible promoters in the region upstream for the *atoC* gene. Figure 3.5 top shows one possible promoter with homology to the typical sigma 70 (*rpoD*) promoter (Collado-Vides et al., 1991; McClure, 1985). In each case the homology is low. For example, the sigma 70 promoter sequence has a four out of six base match (75 per cent) in the -10 region, and a three out of six base match (50 per cent) in the -35 region. The obligatory 16 to 18

Figure 3.5. Upstream region of *atoC*. The first 100 nucleotides of the clone are shown with the predicted translation of *atoC* beginning at position 70. Possible promoter sequences lying upstream of the *atoC* open reading frame are shown with the sequence of the 5′ region of *atoC*. Consensus promoter sequences are shown above the DNA sequence. Lower case letters indicate no match, upper case letters indicate a match. The P stands for any purine A or G. The top figure illustrates a possible sigma 70 promoter and the bottom figure shows a possible *rpoN* promoter.

```
            ttGAcAnnn n 14-16 nn nnnnntaTAA T
GATCAACCCG CAGGAAATCA GACTGTATGA CTGCTATTAA TCGCATCCTT

P        60         70         80         90         100
ATTGTGGATG ATGAAGATAA TGTTGCGCCG TATGCTGAGC ACCGCTTTTG
                    M etLeuArgAr gMetLeuSer ThrAlaPheA
```

```
           CtGGNAnnnn nntTGcA         A  40         50
GATCAACCCG CAGGAAATCA GACTGTATGA CTGCTATTAA TCGCATCCTT

         60         70         80         90         100
ATTGTGGATG ATGAAGATAA TGTTGCGCCG TATGCTGAGC ACCGCTTTTG
                    M etLeuArgAr gMetLeuSer ThrAlaPheA
```

base pair distance between the sigma 70 promoter recognition sites is preserved, as is the position of the transcriptional start at an A or G residue, in this case an A. However if other possible promoter sites are considered, a consensus sigma 54 site is also present. This promoter for the *rpoN* gene has nine out of twelve positions identical (75 per cent) with the consensus promoter sequence (Beck & Warren, 1988).

Discussion

The pathways of short chain fatty acid metabolism are regulated at the level of transcription by the regulatory protein AtoC in response to binding the pathway substrate and inducer acetoacetate. The *atoC* protein is presumed to be a trans-acting regulatory protein since deletion of the *atoC* gene, and mutations in it, yield the cells unable to express the genes of the *ato* operon. Thus an *atoC* mutant is unable to utilize acetoacetate. And since the *atoDA* transferase and *atoB* thiolase are required for metabolism of other short chain fatty acids such as valerate and butyrate, *atoC* mutants unable to induce the expression of these enzymes are also unable to metabolize other short chain fatty acids. In addition, we have shown (Chapter II) that it is possible to construct regulatory mutants that respond to the presence of other inducing metabolites, specifically other short chain fatty acids. The *atoC* gene for the AtoC protein has been cloned and sequenced.

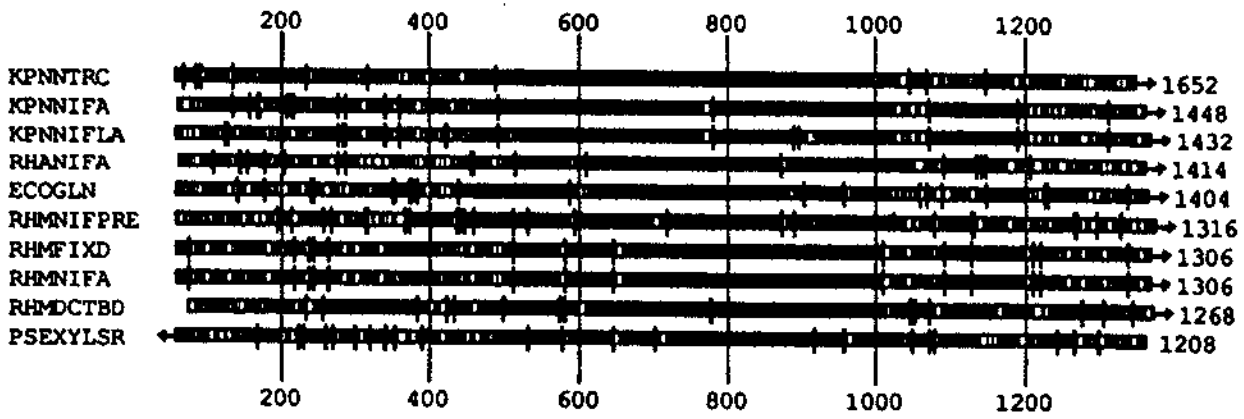
The gene for the *trans*-acting positive regulatory element encoded by *atoC* was sequenced as described in the Materials and Methods and as shown in the Results. It is a protein of molecular weight 50,600 Daltons. This compares with a previous estimate of 48,000 Daltons as determined by polyacrylamide gel electrophoresis of maxicell labeled proteins (Jenkins and Nunn, 1987a). This work also confirms the direction of transcription of the *atoC* protein. Deletions and polar mutations suggested that the *atoC* gene was transcribed rightward from the *Sau*3a site and that this gene was not part of the *ato* operon that it controlled (Jenkins and Nunn, 1987b). We have confirmed this hypothesis too.

The amino acid sequence of AtoC is not homologous with other *trans*-acting positive regulators that belong to the LysR or AraC families of proteins (Adhya & Garges, 1990; Aldrich & Chakrabarty, 1988; Cai et al., 1989, Chang & Crawford, 1990, Chang et al., 1989; Heincz & McFall, 1978; Liljestrom et al., 1982;Little & Mount, 1982; Ogden et al., 1980; Ramos et al., 1990; Rothmel et.al., 1990; Wilcox et al, 1971). However a homology search for amino acid homology between AtoC and a protein data base gave results that show that AtoC belongs to a family of proteins involved in the regulation of genes that respond to a form of RNA polymerase that has a different sigma recognition factor (Christman et al., 1989; Drummond et al., 1986; Inouye et al., 1988).

As shown in the next figure (Figure 3.6), the AtoC protein is homologous to a class of proteins referred to as the NifA family of protein regulators. This family includes the *ntrC* gene product from *Klebsiella pneumoniae*, as well as the *nifA* gene product from the same organism. Other NifA proteins include the *ntrC*(*glnF*) regulatory protein from *E. coli*. This protein controls many genes involved in nitrogen metabolism. Upon starvation or limitation for nitrogen, the level of gene expression for a number of genes increases 10 to 100 fold. The genes induced encode amino acid transport proteins, amino acid degradative enzymes and glutamine synthetase. A protein of similar structure and the identical designation (NtrC) functions in *K. pneumoniae* to regulate nitrogen metabolism. The *ntrC* gene product in *K. pneumoniae* works in concert with the gene product of the *nifA* gene, which specifically activates nitrogen fixation genes. These two proteins, NtrC and NifA, from *K. pneumoniae* share extensive homology. And NifA defines the family of these proteins (Drummond et al., 1986).

All NifA proteins are trasncriptional activators. All share most homology with their central domains. In addition, homology is observed in the carboxyl terminal domain. The carboxyl end of these proteins contains the DNA binding site, though each binds to specific but different regulatory sequences. Each protein thus binds in a similar manner (through interactions between the DNA and the carboxyl

Figure 3.6. Homology search of AtoC. Amino acid sequence homology was measured between AtoC and all other proteins of bacterial origin. The top figure represents an alignment where high levels of homology between the protein listed and AtoC are designated by a darkened bar. Thus open areas of the bar represent regions of low homology between the proteins. The organism, gene and gene product for each comparison in the figure is listed with its designation in the list below.



	Sequence	Opt.	Init.	Description
1.	KPNNTRC	1652	748	Klebsiella pneumoniae ntrC gene.
2.	KPNNIFA	1448	908	Klebsiella pneumoniae nifA gene.
3.	KPNNIFLA	1432	812	K.oxytoca nifLA operon encoding nitrogen fixation proteins
4.	RHANIFA	1414	812	Azorhizobium caulinodans nifA gene.
5.	ECOGLN	1404	962	E.coli glnALG operon with glnA gene for glutamine synthetas
6.	RHMNIFPRE	1316	240	Rhizobium leguminosarum fixX (partial), nifA, and nifB (par
7.	RHMFIXD	1306	566	Rhizobium meliloti regulatory nitrogen fixation gene fixD
8.	RHMNIFA	1306	566	Rhizobium meliloti nifA gene.
9.	RHMDCTBD	1268	224	Rhizobium leguminosarum dctB and dctD genes involved in C4-
10.	PSEXYLSR	1208	294	TOL plasmid (from P.putida) xylS and xylR genes, complete c

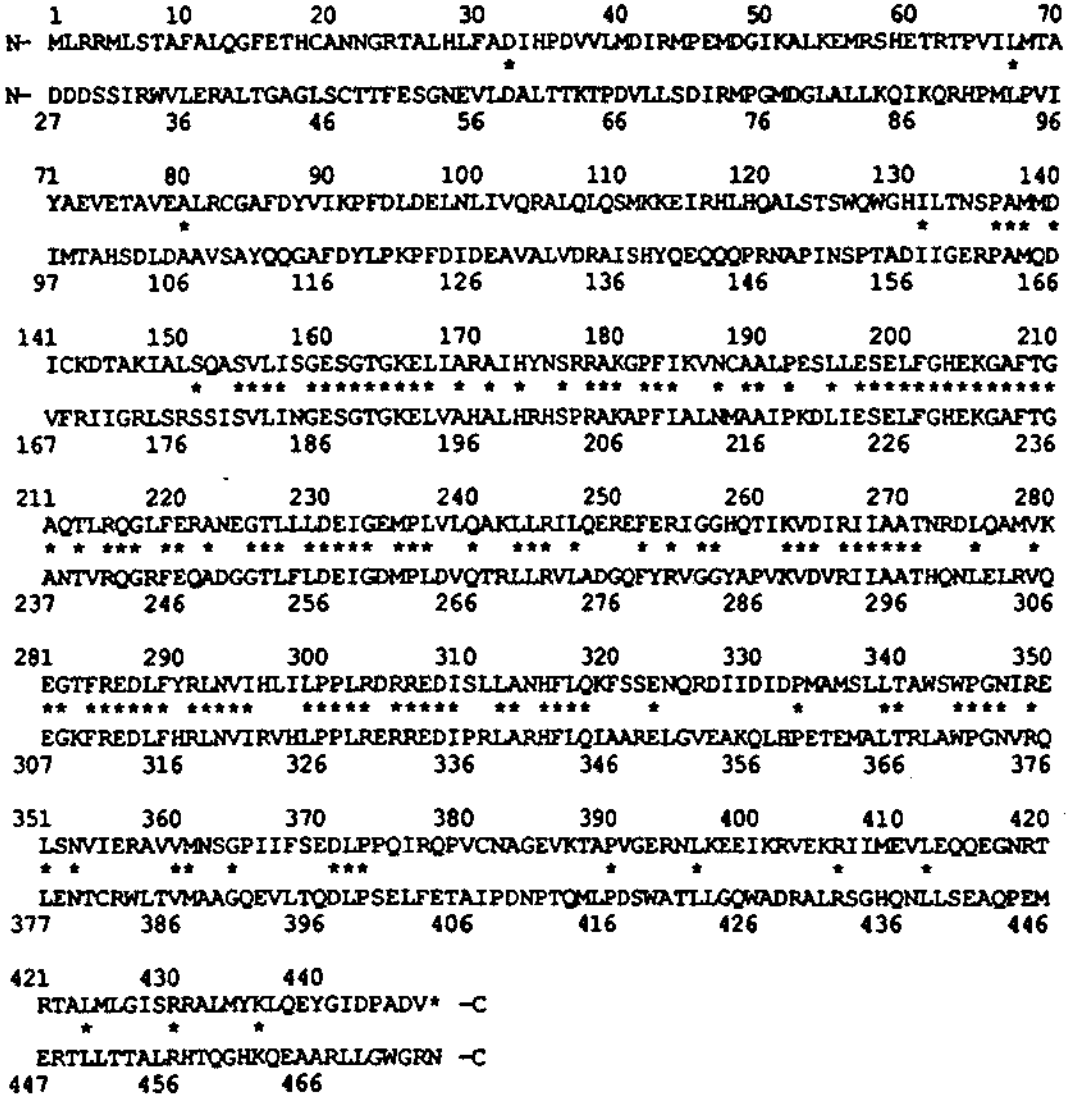
terminus) but to binding sites distinct for each promoter or class of promoters. Presumably, the regulatory small molecule is bound at an effector binding site in the amino terminal end of these proteins. NifA protein activators all act in concert with a form of RNA polymerase that possesses a special sigma subunit (Merrick et al., 1987). The sigma subunit is designated sigma 54 in reference to its molecular weight in Daltons and is encoded by the *rpoD* (*ntrA*) gene. All NifA proteins act at sigma 54 promoters. Their binding sites when identified lie at least at position -85, and for the most part are centered at position -110 (Morett & Buck, 1988; Siebenlist et al., 1980). All sigma 54 promoters are subject to activation and none is subject to repression. These promoters do not normally exhibit typical "up mutations" in the sense that the requirement for the activator is never completely overcome. Sigma 54 promoters are expected to operate as follows.

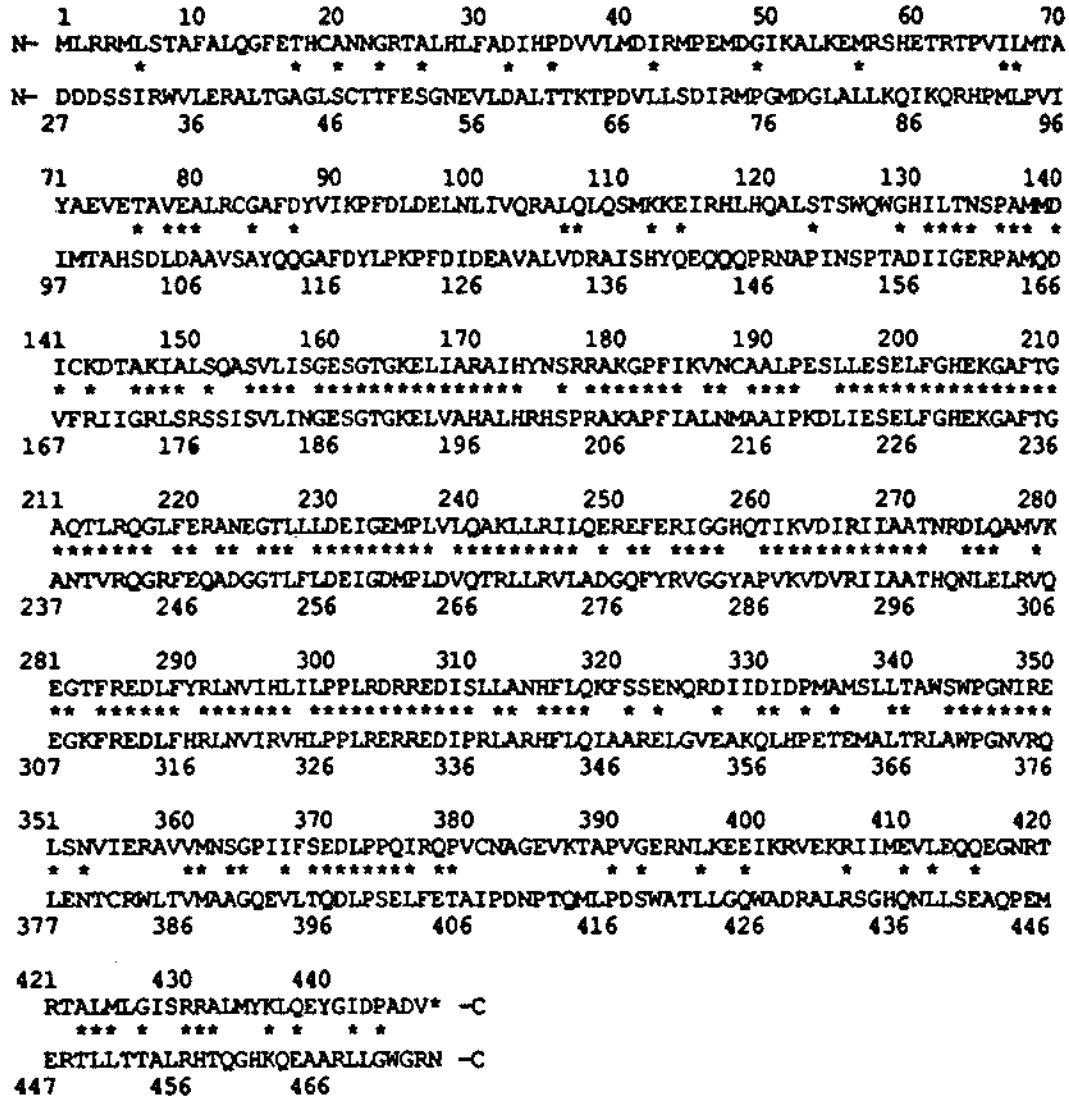
Promoters that utilize polymerase with sigma 54 are not recognized by the RNA polymerase holoenzyme with the typical sigma 70 sigma factor. There is no consensus -35 and -10 region. Instead an RNA polymerase with sigma 54 bound recognizes the sequence CTGGCA followed 6 bases later by TTGCA followed 8 bases later by a trasncriptional start at an A. After binding however, there is no further progress made toward DNA melting, formation of the open complex and initiation of transcription. Instead the polymerase remains

bound in a stable, but inactive and closed complex. Only when the trasncriptional activator is bound does the polymerase make the transition to the open complex and initiate mRNA synthesis,. This explains the dependence that sigma 54 promoters have for the presence of an activator protein. Evidently from the distance involved, the footprint of the polymerase does not overlap that for the transcriptional activators (in cases where such information is known). The region between the bound NifA and RNA polymerase remains unbound as determined by DNA footprinting. The absolute requirement for the presence and binding of the activator is theorized to explain why no sigma 54 promoters are regulated by repressors. The distance between the RNA polymerase binding site and the activator binding site explains why the presence of integration host factor (IHF) plays an important role in the regulation of these operons (Abril et al., 1991; Hotel et al., 1990; 1992; Hoover et al., 1990; Lorenzo et al., 1991; Su et al., 1990). Integration host factor is a DNA binding protein that holds the DNA in a certain configuration so that other proteins may bind and act. Binding of IHF may stabilize bends or loops in the DNA of the promoter and favor activation.

The AtoC protein has extensive homology with NifA. Depending on the conditions imposed, the amino acid sequence homology in a pairwise comparison exceeds fifty per cent (Figure 3.7). If allowance is made for limited substitutions

Figure 3.7. NtrC and AtoC comparison. A pairwise comparison of AtoC is shown . This alignment is very simple and does not allow for limited substitutions and deletions. In this comparison the homology is 31.92 percent. The second comparison illustrates a modification of this comparison. If allowance is made for conservative substitutions, then the homology increased to 51.12 per cent.


1 10 20 30 40 50 60 70
N- MLRRMLSTAFALQGFETHCANNGRTALHLFADIHPDVVLMDIRMPEMDGIKALKEMRSHETRTPVILMTA
N- DDDSSIRWVLERALTGAGLSCTTFESGNEVLDALTTKTPDVLLSDIRMPGMDGLALLKQIKQRHPMLPVI
27 36 46 56 66 76 86 96
71 80 90 100 110 120 130 140
YAEVETAVEALRCGAFDYVIKPFDLDELNLIVQRALQLQSMKKEIRHLHQALSTSWQWGHILTNSPAMMD
IMTAHSDLDAAVSAYQQGAFDYLPKPFDIDEAVALVDRAISHYQEQQQPRNAPINSPTADIIGERPAMQD
97 106 116 126 136 146 156 166
141 150 160 170 180 190 200 210
ICKDTAKIALSQASVLISGESGTGKELIARAIHYNSRRAKGPFIKVNCAALPESLLESELFGHEKGAFTG
VFRIIGRLSRSSISVLINGESGTGKELVAHALHRHSPRAKAPFIALNMAAIPKDLIESELFGHEKGAFTG
167 176 186 196 206 216 226 236
211 220 230 240 250 260 270 280
AQTLRQGLFERANEGTLLLDEIGEMPLVLQAKLLRILQEREFERIGGHQTIKVDIRIIAATNRDLQAMVK
ANTVRQGRFEQADGGTLFLDEIGDMPLDVQTRLLRVLADGQFYRVGGYAPVKVDVRIIAATHQNLELRVQ
237 246 256 266 276 286 296 306
281 290 300 310 320 330 340 350
EGTFREDLFYRLNVIHLILPPLRDRREDISLLANHFLQKFSSENQRDIIDIDPMAMSLLTAWSWPGNIRE
EGKFREDLFHRLNVIRVHLPPLRERREDIPRLARHFLQIAARELGVEAKQLHPETEMALTRLAWPGNVRQ
307 316 326 336 346 356 366 376
351 360 370 380 390 400 410 420
LSNVIERAVVMNSGPIIFSEDLPPQIRQPVCNAGEVKTAPVGERNLKEEIKRVEKRIIMEVLEQQEGNRT
LENTCRWLTVMAAGQEVLTQDLPSELFETAIPDNPTQMLPDSWATLLGQWADRALRSGHQNLLSEAQPEM
377 386 396 406 416 426 436 446
421 430 440
RTAIMLGISRRALMYKLQEYGIDPADV* -C
ERTLLTTALRHTQGHKQEAARLLGWGRN -C
447 456 466

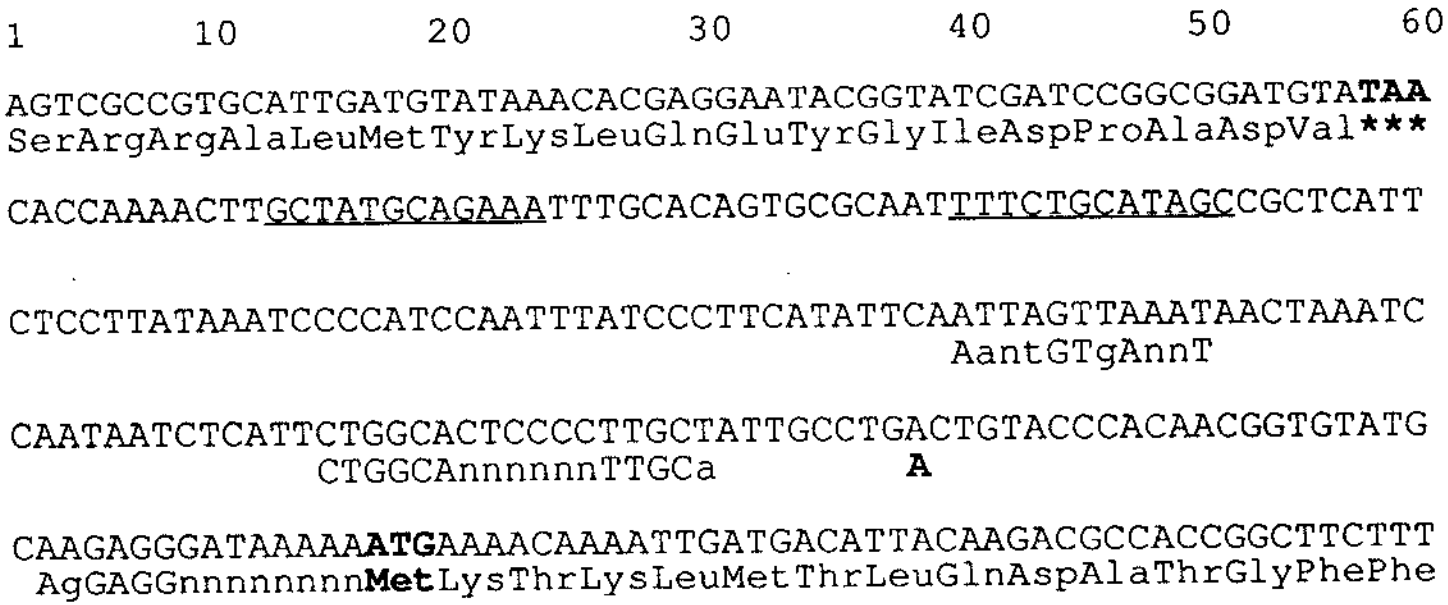

1 10 20 30 40 50 60 70
N- MLRRMLSTAFALQGFETHCANNGRTALHLFADIHPDVVLMDIRMPEMDGIKALKEMRSHETRTPVILMTA
N- DDDSSIRWVLERALTGAGLSCTTFESGNEVLDALTTKTPDVLLSDIRMPGMDGLALLKQIKQRHPMLPVI
27 36 46 56 66 76 86 96
71 80 90 100 110 120 130 140
YAEVETAVEALRCGAFDYVIKPFDLDELNLIVQRALQLQSMKKEIRHLHQALSTSWQWGHILTNSPAMMD
IMTAHSDLDAAVSAYQQGAFDYLPKPFDIDEAVALVDRAISHYQEQQQPRNAPINSPTADIIGERPAMQD
97 106 116 126 136 146 156 166
141 150 160 170 180 190 200 210
ICKDTAKIALSQASVLISGESGTGKELIARAIHYNSRRAKGPFIKVNCAALPESLLESELFGHEKGAFTG
VFRIIGRLSRSSISVLINGESGTGKELVAHALHRHSPRAKAPFIALNMAAIPKDLIESELFGHEKGAFTG
167 176 186 196 206 216 226 236
211 220 230 240 250 260 270 280
AQTLRQGLFERANEGTLLLDEIGEMPLVLQAKLLRILQEREFERIGGHQTIKVDIRIIAATNRDLQAMVK
ANTVRQGRFEQADGGTLFLDEIGDMPLDVQTRLLRVLADGQFYRVGGYAPVKVDVRIIAATHQNLELRVQ
237 246 256 266 276 286 296 306
281 290 300 310 320 330 340 350
EGTFREDLFYRLNVIHLILPPLRDRREDISLLANHFLQKFSSENQRDIIDIDPMAMSLLTAWSWPGNIRE
EGKFREDLFHRLNVIRVHLPPLRERREDIPRLARHFLQIAARELGVEAKQLHPETEMALTRLAWPGNVRQ
307 316 326 336 346 356 366 376
351 360 370 380 390 400 410 420
LSNVIERAVVMNSGPIIFSEDLPPQIRQPVCNAGEVKTAPVGERNLKEEIKRVEKRIIMEVLEQQEGNRT
LENTCRNLTVMAAGQEVLTQDLPSELFETAIPDNPTQMLPDSWATLLGQWADRALRSGHQNLLSEAQPEM
377 386 396 406 416 426 436 446
421 430 440
RTAIMLGISRRALMYKLQEYGIDPADV* -C
ERTLLTTALRHTQGHKQEAARLLGWGRN -C
447 456 466

and deletions, the homology increases dramatically. In addition, only sequence identity is marked in this figure. Figure 3.7 shows two pairwise comparisons of the AtoC protein with that of NtrC from *E. coli*. If allowance is made for conservative substitutions, then the two proteins are clearly closely related. Most of the homology is evident in the middle portion of the two proteins and at the carboxyl terminus. Much lower homology is found when the amino terminal end of the proteins are compared.

Since the AtoC protein belongs to the NifA family of proteins, the possibility that the *atoDAB* operon contains a sigma 54 promoter was investigated. Since AtoC regulates atoDAB expression and since it is theorized to work in concert with a sigma 54 bound RNA polymerase, a search for consensus structural motifs was performed. The promoter for the *atoDAB* operon is illustrated in the next figure (Figure 3.8).

Figure 3.8 shows a 300 base pair length of DNA that encompasses the region between the terminus of the *atoC* gene and the beginning of the *atoD* gene. The *atoD* gene encodes one subunit of the CoA transferase. The last 19 amino acids of the *atoC* gene's open reading frame are shown with the TAA stop. Then approximately 190 base pairs later the ATG/methionine start of the transferase gene is located. Immediately upstream of the transferase gene is a typical ribosome binding site that is a five out of six base pair

Figure 3.8. The *atoDAB* promoter. This 300 base pair fragment represents the DNA sequence between the 3′ end of atoC and the 5′ end of *atoD*. The sequence contains the promoter for *atoDAB* expression. The DNA sequence is written above the predicted amino acid sequence for the *atoD* transferase subunit and the *atoC* activator protein. The TAA stop and the ATG start for the *atoC* and atoD genes respectively is shown in bold. The sigma 54 binding sequence CTGGCANNNNNNTTGCA is written underneath its location in the promoter. Capital letters depict identity, whereas lower case letters indicate divergence from the consensus sequence. The trasncriptional start site for the sigma 54 promoter is shown in bold. Upstream is a potential CAP site and a region of dyad symmetry is underlined in the sequence.


1 10 20 30 40 50 60
AGTCGCCGTGCATTGATGTATAAACACGAGGAATACGGTATCGATCCGGCGGATGTATAA
SerArgArgAlaLeuMetTyrLysLeuGlnGluTyrGlyIleAspProAlaAspVal***
CACCAAAACTTGCTATGCAGAAATTTGCACAGTGCGCAATTTTCTGCATAGCCGCTCATT
CTCCTTATAAATCCCCATCCAATTTATCCCTTCATATTCAATTAGTTAAATAACTAAATC
AantGTgAnnT
CAATAATCTCATTCTGGCACTCCCCTTGCTATTGCCTGACTGTACCCACAACGGTGTATG
CTGGCAnnnnnnTTGCa
A
CAAGAGGGATAAAAAATGAAAACAAAATTGATGACATTACAAGACGCCACCGGCTTCTTT
AgGAGGnnnnnnnnnMetLysThrLysLeuMetThrLeuGlnAspAlaThrGlyPhePhe

match for the consensus sequence AGGAGG, lying the requisite four to eight base pairs upstream of the ATG start codon. Lying upstream of the atoD start and Shine Dalgarno sequence is a consensus sigma 54 promoter. Indeed the match is the best in the literature with only a single base deviation (out of eleven bases). A number of other possible regulatory features are evident. Figure 3.8 also shows a possible CAP binding site with a five out of eight base pair match. If no level of homology greater than this is sought, two other CAP sites are possible. The *ato* operon is n subject to catabolite repression and as such may contain one or more cap sites. Though in the case of sigma 54 regulation, no role for CAP has been previously proposed. It is possible that a CAP-cAMP complex may play a role in proximal coactivation, playing a role similar to IHF in controlling ato gene expression. Upstream at approximate position -110 to -140 is a perfect 12 base pair inverted repeat. The two arms of the palindromic DNA are separated by a 17 base pair spacer, that may itself experience further internal binding, though not perfect. This stem loop structure lies immediately 3′ to the end of the *atoC* open reading frame. It may play a role in termination, though it is not typical of any known terminators except for its general hairpin appearance. This region of dyad symmetry is positioned in the location as the binding site for other proteins in the NifA family. Further studies are required to exactly define the role of the 12

base pair dyad. In addition there are possible IHF binding sites in this short DNA segment (Figure 3.8).

In conclusion, the nucleotide sequence of the *atoC* gene has been determined. The *atoC* gene maps next to the *ato* operon and has been precisely located. Its 3' end (translational stop) is 300 base pairs preceding the translational start of the *atoD* gene of the *atoDAB* operon. The genes are arranged in an operon and are inducible by the substrate acetoacetate. By deduction, AtoC is a 447 amino acid protein with a molecular weight of 51,000 Daltons. The AtoC protein shares extensive sequence homology with the NtrC/NifA family of proteins. In pair-wise comparison with the NtrC protein from *Klebsiella pneumoniae, the* AtoC regulatory protein has identical residues at over 150 positions. Even greater homology is obtained if allowance is made for conservative substitutions. Preceding the *ato* operon is a consensus RpoN (s^{54}) sequence indicating that the mechanism of activation of the *ato* promoter may involve recognition by s^{54} RNA polymerase coupled to interaction with AtoC. The sigma 54 promoter mechanism is more flexible than that for sigma 70 promoters. While a sigma 54 promoter is much less compact that a sigma 70 promoter, it may offer advantages that outweigh the added expense of more nucleotide sequence to maintain. Sigma 54 type promoters would be expected to operate and not be selected against in eukaryotic organisms where larger amounts of DNA can be tolerated. The

sigma 54 of E. coli does share some sequence homology with the yeast TFIID. The transcription factor TFIID is responsible for binding to elements of RNA polymerase II promoters in eukaryotes. It seems that sigma 54 promoters may be constructed like eukaryotic promoters, with different elements that function together but in response to different environmental signals (Dynan & Tjian, 1985; Kadonaga et al., 1987; Singh et al., 1988; 1989; Tang & Taylor, 1990). The presence of a sigma binding site, IHF binding sites, CAP sites and AtoC binding sites certainly makes the *atoDAB* promoter complex. But the complexity may arise out of the accumulation of regulatory modules or cassettes that introduce different levels of control into an evolving promoter so that it functions more finely to a variety of regulatory signals and metabolic demands in the organism.

CHAPTER IV

SEQUENCING AND ANALYSIS OF THE ATO STRUCTURAL GENES AND CORRESPONDING PROTEINS

DNA sequence analysis of predicted translation products of open reading frames can provide a wealth of information. The predicted amino acid sequence yields information about the molecular weight, subunit architecture, hydropathy profile, and even the physical structure of the enzyme. Sequence analysis of the primary amino acid structure can be used to predict the secondary and tertiary structure of the enzyme. Features that identify alpha helices, zinc fingers, ß-bends and ß-pleated sheets are relatively straight forward. Prediction of tertiary structure is more difficult, but still possible. For example, taken together with hydropathy profiles, membrane proteins can be identified and the regions inserted or crossing the membrane predicted.

Comparison of the predicted amino acid sequence with other known sequences can prove immensely valuable in predicting structure function relationships within an enzyme. Conserved amino acids may provide clues for the location of the active site, as well as possible amino acids involved in

catalysis. Identification of the residues participating in the enzyme catalyzed reaction allow predictions of enzymatic mechanisms. Predictions of an amino acid residue's participation in catalysis, and therefore the mechanism, can be confirmed by site-directed mutagenesis. Predictions of secondary and tertiary structure can be confirmed by X-ray crystallography. X-ray crystallography of proteins will be facilitated by a well characterized purification procedure. And the purification will be best if performed using cloned genes from overproducing strains yielding high levels of enzyme production. With long term goals of performing comparative studies of amino acid and nucleotide sequences, identifying important residues in catalysis, and predicting physical properties of the gene products, the sequences of the structural genes of the *ato* operon were sought.

The ato operon has been reported to contain three genes for two enzymes (Pauli & Overath, 1972; Jenkins & Nunn, 1987a;1987b). The acetoacetylCoA transferase is a tetrameric enzyme with two different protein subunits (Sramek & Frerman, 1975a). The *alpha* subunit is encoded by the *atoD* gene. The *beta* subunit is encoded by *atoA*. The thiolase has been reported to be coordinately expressed with the *atoDA* transferase. The thiolase is encoded by the *atoB* gene. The *ato* operon has the gene arrangement *DAB* as determined by genetic mapping of deletion and transposon insertion mutants. To date several transferases have been cloned and sequenced

(Doten et al., 1987; Parales & Harwood, 1992; Shanley et al., 1986), as have several thiolases (Igual et al., 1992; Yang et al., 1990). This work reports the first sequence analysis of a transferase and thiolase that work in concert and are encoded by coordinately controlled genes. Several unique findings were discovered that require new analysis of *ato* gene configuration and arrangement. In addition to the characterization of the physical, chemical and structural properties of a protein, analysis of the primary sequence can provide important clues as to the path and mechanism by which these genes evolved.

Materials and Methods

Bacterial strains and plasmids. The plasmid pLJ10 was obtained from Laren Sallus Jenkins (Jenkins & Nunn, 1987a). This plasmid carries a 6.0 kbp fragment and contains the *ato* operon and the neighboring *atoC* regulatory gene. The fragment has *Sau*3A and *Bam*HI ends (Figure 1.6), but since this clone was isolated from a Sau3A partial digest, it has internal BamH1 and Sau3A sites. The DNA containing the *ato* genes was subcloned into the plasmid pUC19 (Messing, 1983). All clones were manipulated in the host strain DH5α. DH5α was obtained from Gibco-BRL laboratories. It has the genotype F^-, λ^-, *recA1*, Δ(*lacZYA-argF*) U169, *hsdR17*, thi-1, *gyrA96*, *supE44*, *endA1*, *relA1*, ϕ80d*lacZ*ΔM15.

Growth and maintenance of strains. Cells were grown on YT medium. YT medium contains 5 g per l yeast extract, 8 g per l tryptone and 5 g per l sodium chloride. The medium was solidified for agar plates by the addition of 15 g per l agar. Cells were grown at 37°C with shaking. In order to maintain selection for the plasmid, 75 to 100 µg per ml ampicillin was added to the medium.

Plasmid purification. Plasmids were purified as described in Maniatis et al. (1982). The plasmid preparation was one of two types. For rapid plasmid analysis of transformants, a rapid plasmid preparation was performed. For preparation of large amounts of plasmid DNA a bulk plasmid preparations from amplified cells was performed. The methods for rapid plasmid preparation was as follows. Cells were grown in 5 ml YT medium at 37°C incubator on a New Brunswick TB1 tube roller at 50 rpm. Cells in a total volume of 3 ml were lysed by alkaline buffer solution to obtain enough sample for at least, three sequencing reactions or ten samples restriction enzyme digestion (Kraft et al., 1988; McClelland et al., 1988).

DNA sequencing. DNA sequencing was performed as single or double stranded DNA sequencing by Sanger's dideoxynucleotide method (1983) using the ^{35}S-labeled dATP

(Biggin et al., 1983) and Sequenase from USB Corporation. Buffer gradient polyacrylamide gels (TBE concentration from 2.5X to 0.5X) was employed in order to have the better sequencing resolution (Biggin et al., 1983). Oligonucleotides were either synthesized by a Pharmacia Gene Assembler DNA synthesizer or purchased from Bio-Synthesis, Inc. After desalting by running through a Pharmacia Sephadex G-25 column, crude oligonucleotide solutions were used directly without further purification.

DNA sequence analysis. After the DNA sequencing gels were completely dry, the gels were exposed to a sheet of Kodak XAR5 X-ray film. After 1 to 5 days incubation, the films were developed. Since ^{35}S will not penetrate the film support to expose the emulsion on the distal side of the film, this emulsion was removed by stripping with a 1 to 10 dilution of commercial bleach. Care was taken not to wet the exposed side of the film with the bleach. The film was rinsed, dried and illuminated. Sequences were read and entered into file on the computer. Typically, the sequences were entered into a sequence file using the program DNA Inspector. The sequences were independently confirmed by reading the film twice, reading back the sequence using the computer and following along on the gel and having a different investigator read the DNA sequence. Since the

sequence was determined on both strands, it was logged into the final sequence file only after opposite strand confirmation.

Results

The *ato* operon was obtained as a *Sau*3A to *Bam*H1 clone in plasmid pBR322. It had the reported restriction map and gene arrangement as depicted in Figure 1.6 (Chapter I). It had a size of 6.2 kbp as determined by restriction endonuclease digestion and agarose gel electrophoresis (Jenkins & Nunn, 1987a). The entire fragment has been completely sequenced. Both strands were sequenced in order to confirm the sequence of the genes. The positions of the genes were identified by open reading frame analysis and inspection for sequence homology to related proteins whose amino acid sequences, or deduced amino acid sequences are known. The entire sequence is shown in the figure beginning on the following page (Figure 4.1).

The size of the fragment is 6,082 base pairs confirming previous reports (Jenkins & Nunn, 1987a) and providing a more accurate value for the length. The positions of all the reported restriction sites have been confirmed and precisely positioned in the sequence (data not shown). The sequence and restriction pattern will aid in construction of subclones in future studies.

Figure 4.1. The *atoC-atoDAB* nucleotide sequence. All 6,082 bases of the Sau3A to HindIII *ato* gene region are shown. Only the top strand (plus) strand is depicted. All of the open reading frames for the *ato* genes are contained in this strand. The sequence continues on for five pages. The sequence has been submitted to GenBank and is awaiting assignment of an accession number.

```
        10          20          30          40          50          60
GATCAACCCG CAGGAAATCA GACTGTATGA CTGCTATTAA TCGCATCCTT ATTGTGGATG

        70          80          90         100         110         120
ATGAAGATAA TGTTGCGCCG TATGCTGAGC ACCGCTTTTG CACTACAAGG ATTCGAAACA

       130         140         150         160         170         180
CATTGTGCGA ACAACGGACG CACAGCATTA CACCTGTTTG CCGATATTCA CCCTGATGTG

       190         200         210         220         230         240
GTGTTGATGG ATATCCGCAT GCCAGAGATG GACGGCATCA AGGCACTAAA GGAGATGCGC

       250         260         270         280         290         300
AGCCATGAGA CCCGGACACC CGTTATTCTG ATGACGGCCT ATGCGGAAGT GGAAACCGCC

       310         320         330         340         350         360
GTCGAAGCGC TACGCTGCGG AGCCTTCGAC TATGTTATTA AACCGTTTGA TCTCGATGAG

       370         380         390         400         410         420
TTGAATTTAA TCGTTCAGCG CGCTTTACAA CTCCAGTCAA TGAAAAAAGA GATCCGTCAT

       430         440         450         460         470         480
CTGCACCAGG CACTGAGCAC CAGCTGGCAA TGGGGGCACA TTCTCACCAA CAGCCCGGCG

       490         500         510         520         530         540
ATGATGGACA TCTGCAAAGA CACCGCCAAA ATTGCCCTTT CTCAGGCCAG CGTCTTGATT

       550         560         570         580         590         600
AGCGGTGAAA GCGGCACCGG GAAAGAGTTG ATTGCCAGAG CGATTCACTA CAATTCGCGG

       610         620         630         640         650         660
CGGGCAAAGG GGCCGTTCAT TAAAGTCAAC TGCGCGGCGC TGCCGGAATC GTTGCTCGAA

       670         680         690         700         710         720
AGTGAACTGT TTGGTCATGA AAAAGGTGCA TTTACTGGTG CACAAACCTT GCGTCAGGGA

       730         740         750         760         770         780
TTATTTGAAC GAGCCAACGA AGGTACTCTG CTCCTCGACG AAATTGGCGA AATGCCGCTG

       790         800         810         820         830         840
GTACTACAAG CCAAATTACT ACGCATTCTA CAGGAACGGG AATTTGAACG GATTGGCGGC

       850         860         870         880         890         900
CATCAGACCA TAAAAGTTGA TATCCGCATC ATTGCTGCCA CCAACCGCGA CTTGCAGGCA

       910         920         930         940         950         960
ATGGTAAAAG AAGGCACCTT CCGTGAAGAT CTCTTTTATC GCCTTAACGT TATTCATTTA

       970         980         990        1000        1010        1020
ATACTGCCGC CTCTGCGCGA TCGCCGGGAA GATATTTCCC TGTTAGCTAA TCACTTTTTG

      1030        1040        1050        1060        1070        1080
CAAAAATTCA GTAGTGAGAA TCAGCGCGAT ATTATCGACA TCGATCCGAT GGCAATGTCA

      1090        1100        1110        1120        1130        1140
CTGCTTACCG CCTGGTCATG GCCGGGAAAT ATTCGAGAGC TTTCCAACGT TATTGAACGC

      1150        1160        1170        1180        1190        1200
GCCGTCGTGA TGAATTCAGG CCCGATCATT TTTTCTGAGG ATCTTCCGCC ACAGATTCGT

      1210        1220        1230        1240        1250        1260
CAGCCAGTCT GTAATGCTGG CGAGGTAAAA ACAGCCCCTG TCGGTGAGCG TAATTTAAAA
```

```
      1270       1280       1290       1300       1310       1320
GAGGAAATTA AACGCGTCGA AAAACGCATC ATTATGGAAG TGCTGGAACA ACAAGAAGGA

      1330       1340       1350       1360       1370       1380
AACCGAACCC GCACTGCTTT AATGCTGGGC ATCAGTCGCC GTGCATTGAT GTATAAACTC

      1390       1400       1410       1420       1430       1440
CAGGAATACG GTATCGATCC GGCGGATGTA TAACACCAAA ACTTGCTATG CAGAAATTTG

      1450       1460       1470       1480       1490       1500
CACAGTGCGC AATTTTCTGC ATAGCCGCTC ATTCTCCTTA TAAATCCCCA TCCAATTTAT

      1510       1520       1530       1540       1550       1560
CCCTTCATAT TCAATTAGTT AAATAACTAA ATCCAATAAT CTCATTCTGG CACTCCCCTT

      1570       1580       1590       1600       1610       1620
GCTATTGCCT GACTGTACCC ACAACGGTGT ATGCAAGAGG GATAAAAAAT GAAAACAAAA

      1630       1640       1650       1660       1670       1680
TTGATGACAT TACAAGACGC CACCGGCTTC TTTCGTGACG GCATGACCAT CATGGTGGGC

      1690       1700       1710       1720       1730       1740
GGATTTATGG GGATTGGCAC TCCATCCCGC CTGGTTGAAG CATTACTGGA ATCTGGTGTT

      1750       1760       1770       1780       1790       1800
CGCGACCTGA CATTGATAGC CAATGATACC GCGTTTGTTG ATACCGGCAT CGGTCCGCTC

      1810       1820       1830       1840       1850       1860
ATCGTCAATG GTCGAGTCCG CAAAGTGATT GCTTCACATA TCGGCACCAA CCCGGAAACA

      1870       1880       1890       1900       1910       1920
GGTCGGCGCA TGATATCTGG TGAGATGGAC GTCGTTCTGG TGCCGCAAGG TACGCTAATC

      1930       1940       1950       1960       1970       1980
GAGCAAATTC GCTGTGGTGG AGCTGGACTT GGTGGTTTTC TCACCCCAAC GGGTGTCGGC

      1990       2000       2010       2020       2030       2040
ACCGTCGTAG AGGAAGGCAA ACAGACACTG ACACTCGACG GTAAAACCTG GCTGCTCGAA

      2050       2060       2070       2080       2090       2100
CGCCCACTGC GCGCCGACCT GGCGCTAATT CGCGCTCATC GTTGCGACAC ACTTGGCAAC

      2110       2120       2130       2140       2150       2160
CTGACCTATC AACTTAGCGC CCGCAACTTT AACCCCCTGA TAGCCCTTGC GGCTGATATC

      2170       2180       2190       2200       2210       2220
ACGCTGGTAG AGCCAGATGA ACTGGTCGAA ACCGGCGAGC TGCAACCTGA CCATATTGTA

      2230       2240       2250       2260       2270       2280
CCCCCTGGTG CCGTTATCGA CCACATCATC GTTTCACAGG AGAGCAAATA ATGGATGCGA

      2290       2300       2310       2320       2330       2340
AACAACGTAT TGGCCGGTGT GGCGCAAGAG CTTCGTGATG GTGACATCGT TAACTTAGGG

      2350       2360       2370       2380       2390       2400
ATCGGTTTAC CCACAATGGT CGCCAATTAT TTACCGGAGG GTATTCATAT CACTCTGCAA

      2410       2420       2430       2440       2450       2460
TCGGAAAACG GCTTCCTCGG TTTAGGCCCG GTCACGACAG CGCATCCAGA TCTGGTGAAC

      2470       2480       2490       2500       2510       2520
GCTGGCGGGC AACCGTGCGG TGTTTTACCC GGTGCAGCCA TGTTTGATAG CGCCATGTCA
```

```
      2530        2540        2550        2560        2570        2580
TTTGCGCTAA TCCGTGGCGG TCATATTGAT GCCTGCGTGC TCGGCGGTTT GCAAGTAGAC

      2590        2600        2610        2620        2630        2640
GAAGAAGCAA ACCTCGCGAA CTGGGTAGTG CCTGGGAAAA TGGTGCCCGG TATGGGTGGC

      2650        2660        2670        2680        2690        2700
GCGATGGATC TGGTGACCGG GTCGCGCAAA GTGATCATCG CCATGGAACA TTGCGCCAAA

      2710        2720        2730        2740        2750        2760
GATGGTTCAG CAAAAATTTT GCGCCGCTGC ACCATGCCAC TCACTGCGCA ACATGCGGTG

      2770        2780        2790        2800        2810        2820
CATATGCTGG TTACTGAACT GGCTGTCTTT CGTTTTATTG ACGGCAAAAT GTGGCTCACC

      2830        2840        2850        2860        2870        2880
GAAATTGCCG ACGGGTGTGA TTTAGCCACC GTGCGTGCCA AAACAGAAGC TCGGTTTGAA

      2890        2900        2910        2920        2930        2940
GTCGCCGCCG ATCTGAATAC GCAACGGGGT GATTTATGAT TGGTCGCATA TCGCGTTTTA

      2950        2960        2970        2980        2990        3000
TGACGCGTTT TGTCAGCCGG TGGCTTCCCG ATCCACTGAT CTTTGCCATG TTGCTGACAT

      3010        3020        3030        3040        3050        3060
TGCTAACATT CGTGATCTCG CTTTGGTTAA CACCACAAAC GCCGATCAGC ATGGTGAAAA

      3070        3080        3090        3100        3110        3120
TGTGGGGTGA CGGTTTCTGG AACTTGCTGG CGTTTGGTAT GCAGATGGCG CTTATCATCG

      3130        3140        3150        3160        3170        3180
TTACCGGTCA TGCCCTTGCC AGCTCTGCTC CGGTGAAAAG TTTGCTGCGT ACTGCCGCCT

      3190        3200        3210        3220        3230        3240
CCGCCGCAAA GACGCCCGTA CAGGGCGTCA TGCTGGTCAC TTTCTTCGGT TCAGTCGCTT

      3250        3260        3270        3280        3290        3300
GTGTCATCAA CTGGGGATTT GGTTTGGTTG TCGGCGCAAT GTTTGCCCGT GAAGTCGCCC

      3310        3320        3330        3340        3350        3360
GGCGAGTCCC CGGTTCTGAT TATCCGTTGC TCATTGCCTG CGCCTACATT GGTTTTCTCA

      3370        3380        3390        3400        3410        3420
CCTGGGGTGG CGGCTTCTCT GGATCAATGC CTCTGTTGGC TGCAACACCG GGCAACCCGG

      3430        3440        3450        3460        3470        3480
TTGAGCATAT CGCCGGGCTG ATCCCGGTGG GCGATACTCT GTTCAGTGGT TTTAACATTT

      3490        3500        3510        3520        3530        3540
TCATCACTGT GGCGTTGATT GTGGTGATGC CATTTATCAC CCGCATGATG ATGCCAAAAC

      3550        3560        3570        3580        3590        3600
CGTCTGACGT GGTGAGTATC GATCCAAAAC TACTCATGGA AGAGGCTGAT TTTCAAAAGC

      3610        3620        3630        3640        3650        3660
AGCTACCGAA AGATGCCCCA CCATCCGAGC GACTGGAAGA AAGCCGCATT CTGACGTTGA

      3670        3680        3690        3700        3710        3720
TCATCGGCGC ACTCGGTATC GCTTACCTTG CGATGTACTT CAGCGAACAT GGCTTCAACA

      3730        3740        3750        3760        3770        3780
TCACCATCAA TACCGTCAAC CTGATGTTTA TGATTGCGGG TCTGCTGCTA CATAAAACGC
```

```
      3790       3800       3810       3820       3830       3840
CAATGGCTTA TATGCGTGCT ATCAGCGCGG CAGCACGCAG TACTGCCGGT ATTCTGGTGC

      3850       3860       3870       3880       3890       3900
AATTCCCCTT CTACGCTGGG ATCCAACTGA TGATGGAGCA TTCCGGTCTG GGCGGACTCA

      3910       3920       3930       3940       3950       3960
TTACCGAATT CTTCATCAAT GTTGCGAACA AAGACACCTT CCCGGTAATG ACCTTTTTTA

      3970       3980       3990       4000       4010       4020
GTTCTGCACT GATTAACTTC GCCGTTCCGT CTGGCGGCGG TCACTGGGTT ATTCAGGGAC

      4030       4040       4050       4060       4070       4080
CTTTCGTGAT ACCCGCAGCC CAGGCGCTGG GCGCTGATCT CGGTAAATCG GTAATGGCGA

      4090       4100       4110       4120       4130       4140
TCGCCTACGG CGAGCAATGG ATGAACATGG CACAACCATT CTGGGCGCTG CCAGCACTGG

      4150       4160       4170       4180       4190       4200
CAATCGCCGG ACTCGGTGTC CGCGACATCA TGGGCTACTG CATCACTGCC CTGCTCTTCT

      4210       4220       4230       4240       4250       4260
CCGGTGTCAT TTTCGTCATT GGTTTAACGC TGTTCTGACG GCACCCCTAC AAACAGAAGG

      4270       4280       4290       4300       4310       4320
AATATAAAAT GAAAAATTGT GTCATCGTCA GTGCGGTACG TACTGCTATC GGTAGTTTTA

      4330       4340       4350       4360       4370       4380
ACGGTTCACT CGCTTCCACC AGCGCCATCG ACCTGGGGGC GACAGTAATT AAAGCCGCCA

      4390       4400       4410       4420       4430       4440
TTGAACGTGC AAAAATCGAT TCACAACACG TTGATGAAGT GATTATGGGT AACGTGTTAC

      4450       4460       4470       4480       4490       4500
AAGCCGGGCT GGGGCAAAAT CCGGCGCGTC AGGCACTGTT AAAAAGCGGG CTGGCAGAAA

      4510       4520       4530       4540       4550       4560
CGGTGTGCGG ATTCACGGTC AATAAAGTAT GTGGTTCGGG TCTTAAAAGT GTGGCGCTTG

      4570       4580       4590       4600       4610       4620
CCGCCCAGGC CATTCAGGCA GGTCAGGCGC AGAGCATTGT GGCGGGGGGT ATGGAAAATA

      4630       4640       4650       4660       4670       4680
TGAGTTTAGC CCCCTACTTA CTCGATGCAA AAGCACGCTC TGGTTATCGT CTTGGAGACG

      4690       4700       4710       4720       4730       4740
GACAGGTTTA TGACGTAATC CTGCGCGATG GCCTGATGTG CGCCACCCAT GGTTATCATA

      4750       4760       4770       4780       4790       4800
TGGGGATTAC CGCCGAAAAC GTGGCTAAAG AGTACGGAAT TACCCGTGAA ATGCAGGATG

      4810       4820       4830       4840       4850       4860
AACTGGCGCT ACATTCACAG CGTAAAGCGG CAGCCGCAAT TGAGTCCGGT GCTTTTACAG

      4870       4880       4890       4900       4910       4920
CCGAAATCGT CCCGGTAAAT GTTGTCACTC GAAAGAAAAC CTTCGTCTTC AGTCAAGACG

      4930       4940       4950       4960       4970       4980
AATTCAACGG CTGAAGCGTT AGGTGCATTG CGCCCGGCCT TCGATAAAGC AGGAACAGTC

      4990       5000       5010       5020       5030       5040
ACCGCTGGGA ACGCGTCTGG TATTAACGAC GGTGCTGCCC GCTCTGGTGA TTATGGAAGA
```

```
      5050        5060        5070        5080        5090        5100
ATCTGCGGCG CTGGCAGCAG GCCTTACCCC CCTGGCTCGC ATTAAAAGTT ATGCCAGCGG

      5110        5120        5130        5140        5150        5160
TGGCGTGCCC CCCGCATTGA TGGGTATGGG GCCAGTACCT GCCACGCAAA AAGCGTTACA

      5170        5180        5190        5200        5210        5220
ACTGGCGGGG CTGCAACTGG CGGATATTGA TCTCATTGAG GCTAATGAAG CATTTGCTGC

      5230        5240        5250        5260        5270        5280
ACAGTTCCTT GCCGTTGGGA AAAACCTGGG CTTTGATTCT GAGAAAGTGA ATGTCAACGG

      5290        5300        5310        5320        5330        5340
CGGGGCCATC GCGCTCGGGC ATCCTATCGG TGCCAGTGGT GCTCGTATTC TGGTCACACT

      5350        5360        5370        5380        5390        5400
ATTACATGCC ATGCAGGCAC GCGATAAAAC GCTGGGGCTG GCAACACTGT GCATTGGCGG

      5410        5420        5430        5440        5450        5460
CGGTCAGGGA ATTGCGATGG TGATTGAACG GTTGAATTAA TCAATAAAAA CACCCGATAG

      5470        5480        5490        5500        5510        5520
CGAAAGTTAT CGGGTGTTTT CTTGAACATC GACGGCGAAG GTAACCCCAT TAATCACCAG

      5530        5540        5550        5560        5570        5580
TCAAAACTTT TCACCAGCGT CAGCTCGCCA GCATTACGCA TCGGTACAAT AAATGTTTCC

      5590        5600        5610        5620        5630        5640
TGTTTCTCAT TGACCGATCC TTCATCGGTG ATCAGCGTCA GTTGGGCGGT GGTTAATTCC

      5650        5660        5670        5680        5690        5700
GTTCTGGCGC CCACCATAGT AGTTGATATA CACCTGATAG CGCCCGTGAA TTGGCGCGGG

      5710        5720        5730        5740        5750        5760
CATGGCGAAA ATCTCGGGTC CGTACCCCGT CGTGACATCC ATATCCAGTG CACCACTGTT

      5770        5780        5790        5800        5810        5820
TTTCAGCACG TGTTACGTAC CAGCGTGTTG CATCGGCGTA ACAACGTGAA GGTCGAGGTC

      5830        5840        5850        5860        5870        5880
GGTATTGTCC GTATCCCACG AGAGAACCAG CCGTAAACGT GCACGAATCG TTCCTGTACC

      5890        5900        5910        5920        5930        5940
CGGCGTTGAG TAAAACTGCA TTTTTTGTCG GCTTTCGATC CGGGCTGATG ACCTGCACGC

      5950        5960        5970        5980        5990        6000
TGTTGCTGCC TTCAGTGAAT ATAAGGGCGT GCAAATGAAC CATCGGATTC AATTCTCTGC

      6010        6020        6030        6040        6050        6060
GGCATACTGG CACCGTTGAC CACCAGCCGG CCCTGCTGAA CTTTACCCGC CGCAGCATAA

      6070        6080        6090        6100        6110        6120
TTCTTTATTT TGCCGCGGAT CC..... ..... ..... .....
```

Open reading frame analysis of the 6,082 base pair sequence reveals five open reading frames. The first open reading frame is separated from the second by approximately 200 base pairs. The first open reading frame corresponds to the *atoC* gene encoding the activator protein for the *ato* operon and was reported and discussed in the previous chapter (Chapter III). Extending from the region past *atoC* are four open reading frames that appear to be transcribed contiguously from a single promoter.

The first two open reading frames encode the *alpha* and *beta* subunits of the *atoDA* transferase. The *atoD* gene encodes the *alpha* subunit and the *atoA* gene encodes the *beta* subunit. The two genes are transcribed together and are separated by only three bases. The *atoD* gene extends from nucleotide 1605 to nucleotide position 2271 in the complete sequence. The *atoD* open reading frame is 663 nucleotides long including the ATG start and the TAA stop (Figure 4.2). The predicted amino acid sequence yields a protein that is 220 amino acids in length. Using the deduced amino acid composition (Table 4.1), the protein has a molecular weight of 23,502 Daltons.

The *atoA* gene for the alpha subunit of the transferase is translated in the same reading frame as the *atoD* gene. The *atoA* gene specifies an open reading frame that extends from position 2274 to position 2919 in the original 6,082 base pair sequence. The open reading frame is 645

Figure 4.2. The *atoD* nucleotide sequence. The nucleotide sequence of *atoD* is depicted with the deduced amino acid sequence for the alpha subunit of the CoA transferase.

```
                9          18          27          36          45          54
5' ATG AAA ACA AAA TTG ATG ACA TTA CAA GAC GCC ACC GGC TTC TTT CGT GAC GGC
   Met Lys Thr Lys Leu Met Thr Leu Gln Asp Ala Thr Gly Phe Phe Arg Asp Gly

               63          72          81          90          99         108
   ATG ACC ATC ATG GTG GGC GGA TTT ATG GGG ATT GGC ACT CCA TCC CGC CTG GTT
   Met Thr Ile Met Val Gly Gly Phe Met Gly Ile Gly Thr Pro Ser Arg Leu Val

           117         126         135         144         153         162
   GAA GCA TTA CTG GAA TCT GGT GTT CGC GAC CTG ACA TTG ATA GCC AAT GAT ACC
   Glu Ala Leu Leu Glu Ser Gly Val Arg Asp Leu Thr Leu Ile Ala Asn Asp Thr

           171         180         189         198         207         216
   GCG TTT GTT GAT ACC GGC ATC GGT CCG CTC ATC GTC AAT GGT CGA GTC CGC AAA
   Ala Phe Val Asp Thr Gly Ile Gly Pro Leu Ile Val Asn Gly Arg Val Arg Lys

           225         234         243         252         261         270
   GTG ATT GCT TCA CAT ATC GGC ACC AAC CCG GAA ACA GGT CGG CGC ATG ATA TCT
   Val Ile Ala Ser His Ile Gly Thr Asn Pro Glu Thr Gly Arg Arg Met Ile Ser

           279         288         297         306         315         324
   GGT GAG ATG GAC GTC GTT CTG GTG CCG CAA GGT ACG CTA ATC GAG CAA ATT CGC
   Gly Glu Met Asp Val Val Leu Val Pro Gln Gly Thr Leu Ile Glu Gln Ile Arg

           333         342         351         360         369         378
   TGT GGT GGA GCT GGA CTT GGT GGT TTT CTC ACC CCA ACG GGT GTC GGC ACC GTC
   Cys Gly Gly Ala Gly Leu Gly Gly Phe Leu Thr Pro Thr Gly Val Gly Thr Val

           387         396         405         414         423         432
   GTA GAG GAA GGC AAA CAG ACA CTG ACA CTC GAC GGT AAA ACC TGG CTG CTC GAA
   Val Glu Glu Gly Lys Gln Thr Leu Thr Leu Asp Gly Lys Thr Trp Leu Leu Glu

           441         450         459         468         477         486
   CGC CCA CTG CGC GCC GAC CTG GCG CTA ATT CGC GCT CAT CGT TGC GAC ACA CTT
   Arg Pro Leu Arg Ala Asp Leu Ala Leu Ile Arg Ala His Arg Cys Asp Thr Leu

           495         504         513         522         531         540
   GGC AAC CTG ACC TAT CAA CTT AGC GCC CGC AAC TTT AAC CCC CTG ATA GCC CTT
   Gly Asn Leu Thr Tyr Gln Leu Ser Ala Arg Asn Phe Asn Pro Leu Ile Ala Leu

           549         558         567         576         585         594
   GCG GCT GAT ATC ACG CTG GTA GAG CCA GAT GAA CTG GTC GAA ACC GGC GAG CTG
   Ala Ala Asp Ile Thr Leu Val Glu Pro Asp Glu Leu Val Glu Thr Gly Glu Leu

           603         612         621         630         639         648
   CAA CCT GAC CAT ATT GTA CCC CCT GGT GCC GTT ATC GAC CAC ATC ATC GTT TCA
   Gln Pro Asp His Ile Val Pro Pro Gly Ala Val Ile Asp His Ile Ile Val Ser

           657
   CAG GAG AGC AAA TAA  3'
   Gln Glu Ser Lys ***
```

nucleotides long (Figure 4.3) and it is terminated by a TGA stop codon. The predicted amino acid sequence gives a protein that is 214 amino acids long. The molecular weight of the *atoA* transferase subunit is 22,569 Daltons. The amino acid content of the protein is shown in Table 4.2.

Table 4.1. Amino acid content of the *atoD* gene product.

Amino Acid			Number	Mole %
Alanine	Ala	A	14	6.33
Arginine	Arg	R	13	5.88
Asparagine	Asn	N	6	2.71
Aspartic Acid	Asp	D	13	5.88
Cysteine	Cys	C	2	0.90
Glutamine	Gln	Q	7	3.17
Glutamic Acid	Glu	E	13	5.88
Glycine	Gly	G	26	11.76
Histidine	His	H	4	1.81
Isoleucine	Ile	I	17	7.69
Leucine	Leu	L	27	12.22
Lysine	Lys	K	6	2.71
Methionine	Met	M	7	3.17
Phenylalanine	Phe	F	6	2.71
Proline	Pro	P	11	4.98
Serine	Ser	S	7	3.17
Threonine	Thr	T	21	9.50
Tryptophan	Trp	W	1	0.45
Tyrosine	Tyr	Y	1	0.45
Valine	Val	V	18	8.14

The first three columns list each amino acid and its three letter and single letter abbreviations, in that order. The number of amino acids given is out of a total amino acid chain length of 220. This value was also used to determine the mole per cent of each amino acid.

Figure 4.3. The *atoA* nucleotide sequence. The nucleotide sequence of *atoA* is depicted with the deduced amino acid sequence for the beta subunit of the CoA transferase.

```
            9          18          27          36          45          54
5' ATG CGA AAC AAC GTA TTG GCC GGT GTG GCG CAA GAG CTT CGT GAT GGT GAC ATC
   Met Arg Asn Asn Val Leu Ala Gly Val Ala Gln Glu Leu Arg Asp Gly Asp Ile

           63          72          81          90          99         108
   GTT AAC TTA GGG ATC GGT TTA CCC ACA ATG GTC GCC AAT TAT TTA CCG GAG GGT
   Val Asn Leu Gly Ile Gly Leu Pro Thr Met Val Ala Asn Tyr Leu Pro Glu Gly

          117         126         135         144         153         162
   ATT CAT ATC ACT CTG CAA TCG GAA AAC GGC TTC CTC GGT TTA GGC CCG GTC ACG
   Ile His Ile Thr Leu Gln Ser Glu Asn Gly Phe Leu Gly Leu Gly Pro Val Thr

          171         180         189         198         207         216
   ACA GCG CAT CCA GAT CTG GTG AAC GCT GGC GGG CAA CCG TGC GGT GTT TTA CCC
   Thr Ala His Pro Asp Leu Val Asn Ala Gly Gly Gln Pro Cys Gly Val Leu Pro

          225         234         243         252         261         270
   GGT GCA GCC ATG TTT GAT AGC GCC ATG TCA TTT GCG CTA ATC CGT GGC GGT CAT
   Gly Ala Ala Met Phe Asp Ser Ala Met Ser Phe Ala Leu Ile Arg Gly Gly His

          279         288         297         306         315         324
   ATT GAT GCC TGC GTG CTC GGC GGT TTG CAA GTA GAC GAA GAA GCA AAC CTC GCG
   Ile Asp Ala Cys Val Leu Gly Gly Leu Gln Val Asp Glu Glu Ala Asn Leu Ala

          333         342         351         360         369         378
   AAC TGG GTA GTG CCT GGG AAA ATG GTG CCC GGT ATG GGT GGC GCG ATG GAT CTG
   Asn Trp Val Val Pro Gly Lys Met Val Pro Gly Met Gly Gly Ala Met Asp Leu

          387         396         405         414         423         432
   GTG ACC GGG TCG CGC AAA GTG ATC ATC GCC ATG GAA CAT TGC GCC AAA GAT GGT
   Val Thr Gly Ser Arg Lys Val Ile Ile Ala Met Glu His Cys Ala Lys Asp Gly

          441         450         459         468         477         486
   TCA GCA AAA ATT TTG CGC CGC TGC ACC ATG CCA CTC ACT GCG CAA CAT GCG GTG
   Ser Ala Lys Ile Leu Arg Arg Cys Thr Met Pro Leu Thr Ala Gln His Ala Val

          495         504         513         522         531         540
   CAT ATG CTG GTT ACT GAA CTG GCT GTC TTT CGT TTT ATT GAC GGC AAA ATG TGG
   His Met Leu Val Thr Glu Leu Ala Val Phe Arg Phe Ile Asp Gly Lys Met Trp

          549         558         567         576         585         594
   CTC ACC GAA ATT GCC GAC GGG TGT GAT TTA GCC ACC GTG CGT GCC AAA ACA GAA
   Leu Thr Glu Ile Ala Asp Gly Cys Asp Leu Ala Thr Val Arg Ala Lys Thr Glu

          603         612         621         630         639
   GCT CGG TTT GAA GTC GCC GCC GAT CTG AAT ACG CAA CGG GGT GAT TTA TGA  3'
   Ala Arg Phe Glu Val Ala Ala Asp Leu Asn Thr Gln Arg Gly Asp Leu ***
```

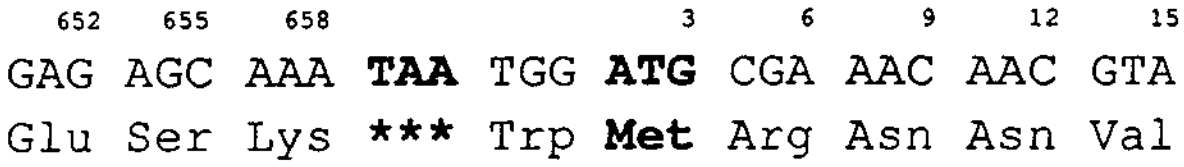
Table 4.2. Amino acid content of the *atoA* gene product.

Amino Acid			Number	Mole %
Alanine	Ala	A	25	11.63
Arginine	Arg	R	10	4.65
Asparagine	Asn	N	9	4.19
Aspartic Acid	Asp	D	13	6.05
Cysteine	Cys	C	5	2.33
Glutamine	Gln	Q	6	2.79
Glutamic Acid	Glu	E	10	4.65
Glycine	Gly	G	25	11.63
Histidine	His	H	6	2.79
Isoleucine	Ile	I	11	5.12
Leucine	Leu	L	23	10.70
Lysine	Lys	K	6	2.79
Methionine	Met	M	11	5.12
Phenylalanine	Phe	F	6	2.79
Proline	Pro	P	9	4.19
Serine	Ser	S	5	2.33
Threonine	Thr	T	12	5.58
Tryptophan	Trp	W	2	0.93
Tyrosine	Tyr	Y	1	0.47
Valine	Val	V	19	8.84

The first three columns list each amino acid and its three letter and single letter abbreviations, in that order. The number of amino acids given is out of a total amino acid chain length of 214. This value was also used to determine the mole per cent of each amino acid.

The *atoD* and *atoA* genes are translated from the same reading frame. Lying between the two genes is a single TGG triplet (the codon for tryptophan, Figure 4.4). Overlapping the stop signals in the *atoA* gene is a start sequence for another open reading frame. This frame lies in the position

Figure 4.4. The *atoAD* intergenic region. The nucleotide sequence for the 3′ end of *atoA* and the 5′ end of *atoX* is shown with the predicted amino acid sequence underlying. The ATG start of *atoX* and TGA stop of *atoA* are underlined. The *** represents a nonsense codon. Numbers represent the nucleotide position for each gene.


652 655 658 3 6 9 12 15
GAG AGC AAA TAA TGG ATG CGA AAC AAC GTA
Glu Ser Lys *** Trp Met Arg Asn Asn Val

reported for the thiolase gene *atoB*. However no homology with other known thiolases has been identified. And the *atoB* gene was positioned in the sequence further down stream. Until a formal gene designation can be assigned (presumably atoE) this gene will be designated *atoX*. The overlap is significant and is much more than a TGA stop sharing a common A with an ATG start. There is an actual overlap of one nucleotide between *atoA* and *atoX*, four nucleotides if the TGA stop is included. Of course this means that the two proteins are translated in different reading frames (Figure 4.5).

The *atoX* gene has no known homology with any other protein to any significant degree when searched against a data base of all known protein sequences. The *atoX* gene extends from nucleotide position number 2915 to position 4238 to give a total length (including the TGA stop) of 1323 nucleotides (Figure 4.6). The open reading frame then contains 441 amino acids and gives a predicted molecular weight of 47,506 Daltons. The amino acid content is given in Table 4.3.

The thiolase gene, *atoB*, was identified by inspection and extends from position number 4269 to nucleotide position number 5422 in the original 6,082 base sequence. The thiolase was first identified as an open reading frame in the sequence. After close inspection, the translated protein was found to have homology with other reported thiolase amino acid sequences. The thiolase gene is 1173 nucleotides long

Figure 4.5. The *atoAX* intergenic region. The nucleotide sequence for the 3′ end of *atoA* and the 5′ end of *atoX* is shown with the predicted amino acid sequence underlying. The ATG start of *atoX* and TGA stop of *atoA* are underlined. The *** represents a nonsense codon. Numbers represent the nucleotide position for each gene. The syncopated depiction of the amino acid sequence for the *atoX* gene reflects its translation in a different reading frame.

```
  636  639  642   3    6    9   12   15   18   21
GGT GAT TTA TGA TTG GTC GCA TAT CGC GTT
Gly Asp Leu ***
          M etI leG lyA rgI leS erA rgP
```

Figure 4.6. The *atoX* nucleotide sequence. The nucleotide sequence of *atoX* is depicted with the deduced amino acid sequence below. The sequence extends for two pages.

```
              9          18          27          36          45          54
5' ATG ATT GGT CGC ATA TCG CGT TTT ATG ACG CGT TTT GTC AGC CGG TGG CTT CCC
   Met Ile Gly Arg Ile Ser Arg Phe Met Thr Arg Phe Val Ser Arg Trp Leu Pro

             63          72          81          90          99         108
   GAT CCA CTG ATC TTT GCC ATG TTG CTG ACA TTG CTA ACA TTC GTG ATC GCG CTT
   Asp Pro Leu Ile Phe Ala Met Leu Leu Thr Leu Leu Thr Phe Val Ile Ala Leu

            117         126         135         144         153         162
   TGG TTA ACA CCA CAA ACG CCG ATC AGC ATG GTG AAA ATG TGG GGT GAC GGT TTC
   Trp Leu Thr Pro Gln Thr Pro Ile Ser Met Val Lys Met Trp Gly Asp Gly Phe

            171         180         189         198         207         216
   TGG AAC TTG CTG GCG TTT GGT ATG CAG ATG GCG CTT ATC ATC GTT ACC GGT CAT
   Trp Asn Leu Leu Ala Phe Gly Met Gln Met Ala Leu Ile Ile Val Thr Gly His

            225         234         243         252         261         270
   GCC CTT GCC AGC TCT GCT CCG GTG AAA AGT TTG CTG CGT ACT GCC GCC TCC GCC
   Ala Leu Ala Ser Ser Ala Pro Val Lys Ser Leu Leu Arg Thr Ala Ala Ser Ala

            279         288         297         306         315         324
   GCA AAG ACG CCC GTA CAG GGC GTC ATG CTG GTC ACT TTC TTC GGT TCA GTC GCT
   Ala Lys Thr Pro Val Gln Gly Val Met Leu Val Thr Phe Phe Gly Ser Val Ala

            333         342         351         360         369         378
   TGT GTC ATC AAC TGG GGA TTT GGT TTG GTT GTC GGC GCA ATG TTT GCC CGT GAA
   Cys Val Ile Asn Trp Gly Phe Gly Leu Val Val Gly Ala Met Phe Ala Arg Glu

            387         396         405         414         423         432
   GTC GCC CGG CGA GTC CCC GGT TCT GAT TAT CCG TTG CTC ATT GCC TGC GCC TAC
   Val Ala Arg Arg Val Pro Gly Ser Asp Tyr Pro Leu Leu Ile Ala Cys Ala Tyr

            441         450         459         468         477         486
   ATT GGT TTT CTC ACC TGG GGT GGC GGC TTC TCT GGA TCA ATG CCT CTG TTG GCT
   Ile Gly Phe Leu Thr Trp Gly Gly Gly Phe Ser Gly Ser Met Pro Leu Leu Ala

            495         504         513         522         531         540
   GCA ACA CCG GGC AAC CCG GTT GAG CAT ATC GCC GGG CTG ATC CCG GTG GGC GAT
   Ala Thr Pro Gly Asn Pro Val Glu His Ile Ala Gly Leu Ile Pro Val Gly Asp

            549         558         567         576         585         594
   ACT CTG TTC AGT GGT TTT AAC ATT TTC ATC ACT GTG GCG TTG ATT GTG GTG ATG
   Thr Leu Phe Ser Gly Phe Asn Ile Phe Ile Thr Val Ala Leu Ile Val Val Met

            603         612         621         630         639         648
   CCA TTT ATC ACC CGC ATG ATG ATG CCA AAA CCG TCT GAC GTG GTG AGT ATC GAT
   Pro Phe Ile Thr Arg Met Met Met Pro Lys Pro Ser Asp Val Val Ser Ile Asp

            657         666         675         684         693         702
   CCA AAA CTA CTC ATG GAA GAG GCT GAT TTT CAA AAG CAG CTA CCG AAA GAT GCC
   Pro Lys Leu Leu Met Glu Glu Ala Asp Phe Gln Lys Gln Leu Pro Lys Asp Ala

            711         720         729         738         747         756
   CCA CCA TCC GAG CGA CTG GAA GAA AGC CGC ATT CTG ACG TTG ATC ATC GGC GCA
   Pro Pro Ser Glu Arg Leu Glu Glu Ser Arg Ile Leu Thr Leu Ile Ile Gly Ala

            765         774         783         792         801         810
   CTC GGT ATC GCT TAC CTT GCG ATG TAC TTC AGC GAA CAT GGC TTC AAC ATC ACC
   Leu Gly Ile Ala Tyr Leu Ala Met Tyr Phe Ser Glu His Gly Phe Asn Ile Thr
```

```
        819         828         837         846         855         864
ATC AAT ACC GTC AAC CTG ATG TTT ATG ATT GCG GGT CTG CTG CTA CAT AAA ACG
Ile Asn Thr Val Asn Leu Met Phe Met Ile Ala Gly Leu Leu Leu His Lys Thr

        873         882         891         900         909         918
CCA ATG GCT TAT ATG CGT GCT ATC AGC GCG GCA GCA CGC AGT ACT GCC GGT ATT
Pro Met Ala Tyr Met Arg Ala Ile Ser Ala Ala Ala Arg Ser Thr Ala Gly Ile

        927         936         945         954         963         972
CTG GTG CAA TTC CCC TTC TAC GCT GGG ATC CAA CTG ATG ATG GAG CAT TCC GGT
Leu Val Gln Phe Pro Phe Tyr Ala Gly Ile Gln Leu Met Met Glu His Ser Gly

        981         990         999        1008        1017        1026
CTG GGC GGA CTC ATT ACC GAA TTC TTC ATC AAT GTT GCG AAC AAA GAC ACC TTC
Leu Gly Gly Leu Ile Thr Glu Phe Phe Ile Asn Val Ala Asn Lys Asp Thr Phe

       1035        1044        1053        1062        1071        1080
CCG GTA ATG ACC TTT TTT AGT TCT GCA CTG ATT AAC TTC GCC GTT CCG TCT GGC
Pro Val Met Thr Phe Phe Ser Ser Ala Leu Ile Asn Phe Ala Val Pro Ser Gly

       1089        1098        1107        1116        1125        1134
GGC GGT CAC TGG GTT ATT CAG GGA CCT TTC GTG ATA CCC GCA GCC CAG GCG CTG
Gly Gly His Trp Val Ile Gln Gly Pro Phe Val Ile Pro Ala Ala Gln Ala Leu

       1143        1152        1161        1170        1179        1188
GGC GCT GAT CTC GGT AAA TCG GTA ATG GCG ATC GCC TAC GGC GAG CAA TGG ATG
Gly Ala Asp Leu Gly Lys Ser Val Met Ala Ile Ala Tyr Gly Glu Gln Trp Met

       1197        1206        1215        1224        1233        1242
AAC ATG GCA CAA CCA TTC TGG GCG CTG CCA GCA CTG GCA ATC GCC GGA CTC GGT
Asn Met Ala Gln Pro Phe Trp Ala Leu Pro Ala Leu Ala Ile Ala Gly Leu Gly

       1251        1260        1269        1278        1287        1296
GTC CGC GAC ATC ATG GGC TAC TGC ATC ACT GCC CTG CTC TTC TCC GGT GTC ATT
Val Arg Asp Ile Met Gly Tyr Cys Ile Thr Ala Leu Leu Phe Ser Gly Val Ile

       1305        1314        1323
TTC GTC ATT GGT TTA ACG CTG TTC TGA  3'
Phe Val Ile Gly Leu Thr Leu Phe ***
```

including a TAA stop. This gives a protein with a predicted molecular weight of 40,472 Daltons for its 390 amino acid length. The *atoB* nucleotide sequence and deduced thiolase amino acid sequence are shown in Figure 4.7. The amino acid content is given in Table 4.4.

Table 4.3. Amino acid content of the *atoX* gene product.

Amino Acid			Number	Mole %
Alanine	Ala	A	49	11.11
Arginine	Arg	R	14	3.17
Asparagine	Asn	N	11	2.49
Aspartic Acid	Asp	D	11	2.49
Cysteine	Cys	C	3	0.68
Glutamine	Gln	Q	11	2.49
Glutamic Acid	Glu	E	11	2.49
Glycine	Gly	G	41	9.30
Histidine	His	H	6	1.36
Isoleucine	Ile	I	38	8.62
Leucine	Leu	L	50	11.34
Lysine	Lys	K	10	2.27
Methionine	Met	M	27	6.12
Phenylalanine	Phe	F	33	7.48
Proline	Pro	P	27	6.12
Serine	Ser	S	25	5.67
Threonine	Thr	T	24	5.44
Tryptophan	Trp	W	9	2.04
Tyrosine	Tyr	Y	8	1.81
Valine	Val	V	32	7.26

The first three columns list each amino acid and its three letter and single letter abbreviations, in that order. The number of amino acids given is out of a total amino acid chain length of 440. This value was also used to determine the mole per cent of each amino acid.

Figure 4.7. The *atoB* nucleotide sequence. The nucleotide sequence of *atoB* which encodes the acetoacetyl CoA thiolase is depicted with the deduced amino acid sequence below. The sequence extends for two pages.

```
            9          18          27          36          45          54
5' ATG AAA AAT TGT GTC ATC GTC AGT GCG GTA CGT ACT GCT ATC GGT AGT TTT AAC
   Met Lys Asn Cys Val Ile Val Ser Ala Val Arg Thr Ala Ile Gly Ser Phe Asn

           63          72          81          90          99         108
   GGT TCA CTC GCT TCC ACC AGC GCC ATC GAC CTG GGG GCG ACA GTA ATT AAA GCC
   Gly Ser Leu Ala Ser Thr Ser Ala Ile Asp Leu Gly Ala Thr Val Ile Lys Ala

          117         126         135         144         153         162
   GCC ATT GAA CGT GCA AAA ATC GAT TCA CAA CAC GTT GAT GAA GTG ATT ATG GGT
   Ala Ile Glu Arg Ala Lys Ile Asp Ser Gln His Val Asp Glu Val Ile Met Gly

          171         180         189         198         207         216
   AAC GTG TTA CAA GCC GGG CTG GGG CAA AAT CCG GCG CGT CAG GCA CTG TTA AAA
   Asn Val Leu Gln Ala Gly Leu Gly Gln Asn Pro Ala Arg Gln Ala Leu Leu Lys

          225         234         243         252         261         270
   AGC GGG CTG GCA GAA ACG GTG TGC GGA TTC ACG GTC AAT AAA GTA TGT GGT TCG
   Ser Gly Leu Ala Glu Thr Val Cys Gly Phe Thr Val Asn Lys Val Cys Gly Ser

          279         288         297         306         315         324
   GGT CTT AAA AGT GTG GCG CTT GCC GCC CAG GCC ATT CAG GCA GGT CAG GCG CAG
   Gly Leu Lys Ser Val Ala Leu Ala Ala Gln Ala Ile Gln Ala Gly Gln Ala Gln

          333         342         351         360         369         378
   AGC ATT GTG GCG GGG GGT ATG GAA AAT ATG AGT TTA GCC CCC TAC TTA CTC GAT
   Ser Ile Val Ala Gly Gly Met Glu Asn Met Ser Leu Ala Pro Tyr Leu Leu Asp

          387         396         405         414         423         432
   GCA AAA GCA CGC TCT GGT TAT CGT CTT GGA GAC GGA CAG GTT TAT GAC GTA ATC
   Ala Lys Ala Arg Ser Gly Tyr Arg Leu Gly Asp Gly Gln Val Tyr Asp Val Ile

          441         450         459         468         477         486
   CTG CGC GAT GGC CTG ATG TGC GCC ACC CAT GGT TAT CAT ATG GGG ATT ACC GCC
   Leu Arg Asp Gly Leu Met Cys Ala Thr His Gly Tyr His Met Gly Ile Thr Ala

          495         504         513         522         531         540
   GAA AAC GTG GCT AAA GAG TAC GGA ATT ACC CGT GAA ATG CAG GAT GAA CTG GCG
   Glu Asn Val Ala Lys Glu Tyr Gly Ile Thr Arg Glu Met Gln Asp Glu Leu Ala

          549         558         567         576         585         594
   CTA CAT TCA CAG CGT AAA GCG GCA GCC GCA ATT GAG TCC GGT GCT TTT ACA GCC
   Leu His Ser Gln Arg Lys Ala Ala Ala Ala Ile Glu Ser Gly Ala Phe Thr Ala

          603         612         621         630         639         648
   GAA ATC GTC CCG GTA AAT GTT GTC ACT CGA AAG AAA ACC TTC GTC TTC AGT CAA
   Glu Ile Val Pro Val Asn Val Val Thr Arg Lys Lys Thr Phe Val Phe Ser Gln

          657         666         675         684         693         702
   GAC GAA TTC AAC GGC TTG AAG CGT TTG GTG CAT TGC GCC CGG CCT TCG ATA AAG
   Asp Glu Phe Asn Gly Leu Lys Arg Leu Val His Cys Ala Arg Pro Ser Ile Lys

          711         720         729         738         747         756
   CAG GAA CAG TCA CCG CTG GGA ACG CGT CTG GTA TTA ACG ACG GTG CTG CCC GCT
   Gln Glu Gln Ser Pro Leu Gly Thr Arg Leu Val Leu Thr Thr Val Leu Pro Ala

          765         774         783         792         801         810
   CTG GTG ATT ATG GAA GAA TCT GCG GCG CTG GCA GCA GGC CTT ACC CCC CTG GCT
   Leu Val Ile Met Glu Glu Ser Ala Ala Leu Ala Ala Gly Leu Thr Pro Leu Ala
```

```
        819         828         837         846         855         864
CGC ATT AAA AGT TAT GCC AGC GGT GGC GTG CCC CCC GCA TTG ATG GGT ATG GGG
Arg Ile Lys Ser Tyr Ala Ser Gly Gly Val Pro Pro Ala Leu Met Gly Met Gly

        873         882         891         900         909         918
CCA GTA CCT GCC ACG CAA AAA GCG TTA CAA CTG GCG GGG CTG CAA CTG GCG GAT
Pro Val Pro Ala Thr Gln Lys Ala Leu Gln Leu Ala Gly Leu Gln Leu Ala Asp

        927         936         945         954         963         972
ATT GAT CTC ATT GAG GCT AAT GAA GCA TTT GCT GCA CAG TTC CTT GCC GTT GGG
Ile Asp Leu Ile Glu Ala Asn Glu Ala Phe Ala Ala Gln Phe Leu Ala Val Gly

        981         990         999        1008        1017        1026
AAA AAC CTG GGC TTT GAT TCT GAG AAA GTG AAT GTC AAC GGC GGG GCC ATC GCG
Lys Asn Leu Gly Phe Asp Ser Glu Lys Val Asn Val Asn Gly Gly Ala Ile Ala

       1035        1044        1053        1062        1071        1080
CTC GGG CAT CCT ATC GGT GCC AGT GGT GCT CGT ATT CTG GTC ACA CTA TTA CAT
Leu Gly His Pro Ile Gly Ala Ser Gly Ala Arg Ile Leu Val Thr Leu Leu His

       1089        1098        1107        1116        1125        1134
GCC ATG CAG GCA CGC GAT AAA ACG CTG GGG CTG GCA ACA CTG TGC ATT GGC GGC
Ala Met Gln Ala Arg Asp Lys Thr Leu Gly Leu Ala Thr Leu Cys Ile Gly Gly

       1143        1152        1161        1170
GGT CAG GGA ATT GCG ATG GTG ATT GAA CGG TTG AAT TAA  3'
Gly Gln Gly Ile Ala Met Val Ile Glu Arg Leu Asn ***
```

Table 4.4. Amino acid content of the *atoB* gene product.

Amino Acid			Number	Mole %
Alanine	Ala	A	58	14.83
Arginine	Arg	R	16	4.09
Asparagine	Asn	N	14	3.58
Aspartic Acid	Asp	D	13	3.32
Cysteine	Cys	C	6	1.53
Glutamine	Gln	Q	20	5.12
Glutamic Acid	Glu	E	18	4.60
Glycine	Gly	G	41	10.49
Histidine	His	H	7	1.79
Isoleucine	Ile	I	25	6.39
Leucine	Leu	L	43	11.00
Lysine	Lys	K	18	4.60
Methionine	Met	M	12	3.07
Phenylalanine	Phe	F	9	2.30
Proline	Pro	P	12	3.07
Serine	Ser	S	22	5.63
Threonine	Thr	T	19	4.86
Tryptophan	Trp	W	0	0.00
Tyrosine	Tyr	Y	6	1.53
Valine	Val	V	31	7.93

The first three columns list each amino acid and its three letter and single letter abbreviations, in that order. The number of amino acids given is out of a total amino acid chain length of 390. This value was also used to determine the mole per cent of each amino acid.

Discussion

The entire *ato* operon and the accompanying regulatory gene has been sequenced. The *ato* operon lies on a Sau3A to BamH1 fragment that is 6,082 base pairs in length. Both

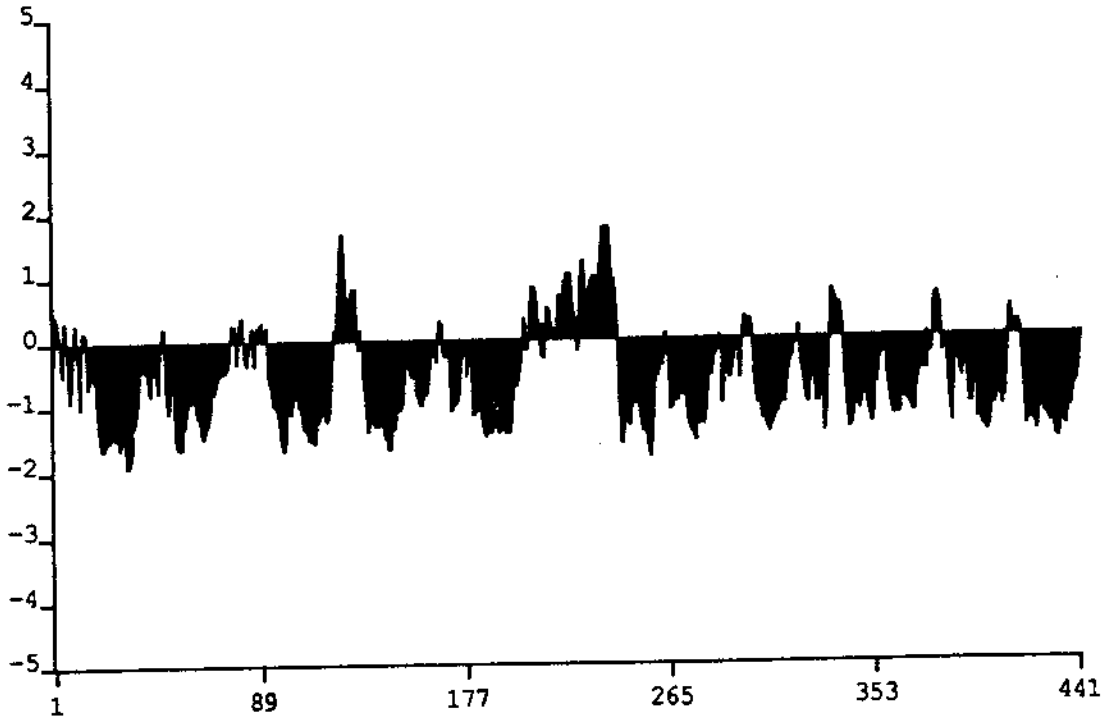
strands were sequenced to ensure accuracy and resolve any ambiguities in the sequence. The contiguous nature of the genes of the *ato* operon was confirmed. Accurate measures of their molecular weights and primary amino acid sequences were deduced.

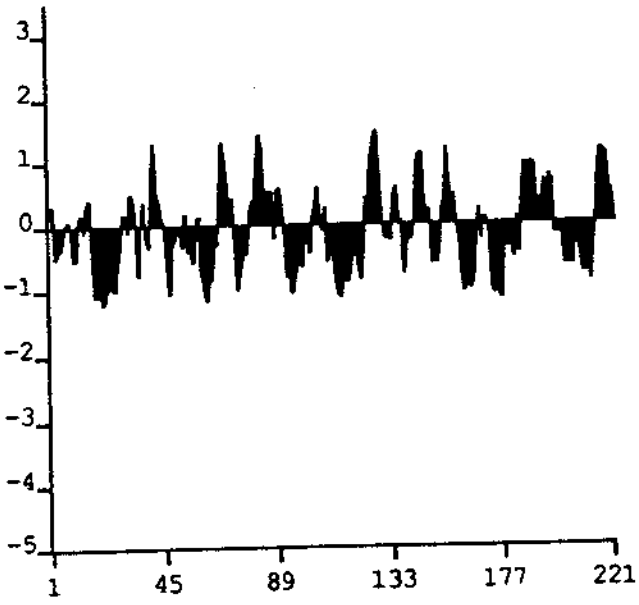
The *atoDAB* operon contains the genes for the two enzymes of acetoacetate metabolism. These enzymes are also required for the metabolism of short chain fatty acids in *E. coli*. The two genes are separated by a third gene termed *atoX*. Its function is unknown and the *atoX* translated product shows little homology to other known protein sequences. It is possible that *atoX* represents a transport protein. No know or reported protein that is responsible for the transport of short chain fatty acids nor acetoacetate has been reported (Nunn, 1986). Acetoacetate is thought to enter the cell by facilitated diffusion, conversion to the CoA derivative ensuring its one way transport into the cell (Frerman, 1973; Frerman & Bennett, 1973; Salnitro & Wegener, 1971). Similarly, short chain fatty acids are thought to be freely diffusible into the cell. This is contrast to medium and long chain fatty acids which enter via specific transport proteins, specific for chain length (Esfahani et al., 1971; Klein et al., 1971; Sallus et al., 1983). While short chain fatty acids such as butyrate and valerate may be freely diffusible, acetoacetate a polar carbonyl oxygen in addition to a negative charge. Displacement of this polar charged

molecule may be accomplished by a transport protein. The amino acid content (Table 4.3) shows that there are fewer that expected charged residues for a protein the size of the *atoX* gene product. A hydropathy plot (Figure 4.8) shows that the translation product is extremely hydrophobic and probably would not exist in solution, but must be an integral or trans-membrane protein. Previous studies with the cloned *ato* operon failed to detect labeled proteins corresponding to this open reading frame (Jenkins & Nunn, 1987a). Perhaps the ^{35}S labeled proteins from the maxicell labeling experiment did not include the membrane fraction from the cell. All of the label appearing in an *atoX* protein would be immediately inserted into the cellular membrane fraction and might be removed by centrifugation prior to gel electrophoresis of sample. It would seem unlikely that this protein is not translated given its overlap with the CoA transferase preceding the gene. Translation of these two proteins is probably tightly coupled, with little dissociation of the ribosome and polar decrease of expression.

Comparison of the *ato* transferase with other transferase proteins shows that these proteins share common structural properties. The *E. coli atoDA* transferase is homologous to the ß-ketoadipate CoA transferases from both *P. putida* (Parales & Harwood, 1992) and *A. calcoaceticus* (Shanley et al., 1986; unpublished). The *E. coli atoD* alpha subunit of the enzyme is 39.9 per cent homologous with *pcaI* transferase

Figure 4.8. Hydropathy plot of *atoX*. The hydropathy profile of the *atoX* protein is shown in the top figure. The horizontal axis refers to amino acid position in the protein. The Y axis gives a relative indication of that region of the protein's tendency to sequester itself from water, either by burying those residues in the interior of the protein or into lipid bilayers. The lower plot is a similar analysis for the *atoD* transferase ß subunit for comparison. The transferase is a soluble enzyme.


5
4
3
2
1
0
-1
-2
-3
-4
-5
1
89
177
265
353
441


3
2
1
0
-1
-2
-3
-4
-5
1
45
89
133
177
221

subunit from *P. putida*. And the homology is 39.6 per cent when compared to the *catI* alpha subunit for the transferase from *A. calcoaceticus*. These three organisms represent a broad swath through the Gram negative genera. *P. putida* is 60 per cent GC, *E. coli* is 50 per cent GC and *A. calcoaceticus* is 40 per cent GC. The *E. coli* beta subunit encode by *atoA* is 34.7 and 36.7 per cent homologous with the *A. calcoaceticus catJ* gene and *P. putida pcaJ* gene respectively. The two *pcaJ/catJ* and *pcaI/catI* transferases are 62.1 and 66.7 per cent homologous to each other, indicating their identical catalytic functions. The glutamate involved in the formation of the covalent enzyme substrate intermediate (Hersh & Jenks, 1967b) was identified by inspection and comparison to be residue glu144 of the beta subunit encoded by the *atoA* gene (Figure 4.9).

The identification of the *atoX* gene has added a new gene to the gene map for the *ato* operon (Jenkins & Nunn, 1987a). Figure 4.10 depicts the revised map of the *ato* operon. All of the genes have been precisely located and ordered with in the sequence of the chromosomal DNA fragment that carries them. With the sequence of these genes it will be possible to construct subclones that can be used to study the proteins as well as the regulation of their expression. With the sequence, a comparative basis is established for examination of mutations that affect catalysis or that alter the regulatory properties of the operon. This work has provided

Figure 4.9. Sequence alignment of beta subunits of CoA transferases. The beta subunits of the *E. coli*, *A. calcoaceticus* and *P. putida* transferases are shown. The *P.putida* subunit is encoded by the *pcaI* gene and the *A. calcoaceticus* subunit is encoded by the *catI* gene. Both of these enzymes transfer CoA to ß-ketoadipate. The common glutamate residue is emboldened.

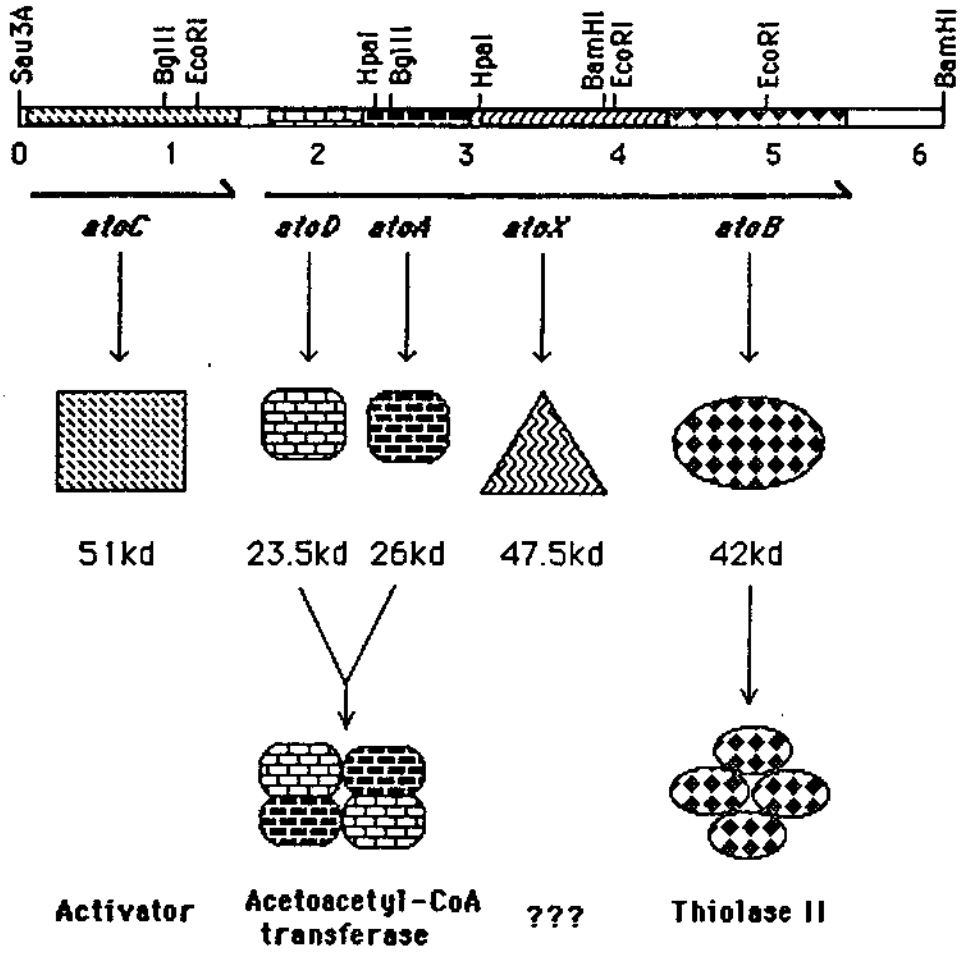
Figure 4.9. Sequence alignment of beta subunits of CoA transferases. The beta subunits of the *E. coli*, *A. calcoaceticus* and *P. putida* transferases are shown. The *P.putida* subunit is encoded by the *pcaI* gene and the *A. calcoaceticus* subunit is encoded by the *catI* gene. Both of these enzymes transfer CoA to ß-ketoadipate. The common glutamate residue is emboldened.

E.c. Leu Gln Ser **Glu** Asn Gly Phe Leu

A.c. Leu His Ser **Glu** Asn Gly Leu Leu

P.p. Leu His Ser **Glu** Asn Gly Leu Leu

Figure 4.10. The *ato* operon. A revised depiction of the genetic and physical map of the *atoDAXB* operon and the neighboring *atoC* gene. All genes are transcribed rightward. The map depicts the genes in their proper orientation with their corrected molecular weights.


Sau3A
BglII
EcoRI
HpaI
BglII
HpaI
BamHI
EcoRI
EcoRI
BamHI
0
1
2
3
4
5
6
atoC
atoD
atoA
atoX
atoB
51kd
23.5kd
26kd
47.5kd
42kd
Activator
Acetoacetyl-CoA transferase
???
Thiolase II

a basis for further study of the regulation and evolution of the *ato* operon and its gene products.

REFERENCES

1. **Abril, M.-A., M. Buck, and J.L. Ramos.** 1991. Activation of the *Pseudomonas* TOL plasmid upper pathway operon: Identification of binding sites for the positive regulator XylR and for integration host factor protein. J. Biol. Chem. **266**(24):15832-15838.

2. **Adhya, S., and S. Garges.** 1990. Positive control. J. Biol. Chem. **265**:10797-10800.

3. **Aldrich, T.L., and A.M. Chakrabarty.** 1988. Transcriptional regulation, nucleotide sequence, and localization of the promoter of the *catBC* operon in *Pseudomonas putida*. J. Bacteriol. **170**:1297-1304.

4. **Anderson, W.B., A.B. Schneider, M. Emmer, L.R. Perlman, and I. Pastan.** 1971. Purification of and properties of the cyclic adenosine 3′, 5′-monophosphate receptor protein which mediates cyclic adenosine 3′, 5′-monophosphate-dependent gene transcription in *Escherichia coli*. J. Biol. Chem. **246**:5929-5937.

5. **Beck, C.F. and R.A.J. Warren.** 1988. Divergent promoters, a common form of gene organization. Microbiol. Rev. **52**:318-326.

6. **Biggin, M.D., Gibson, T.J. and Hong, G.F.:** Buffer gradient gels and ^{35}S as an aid to rapid DNA sequence determination. Proc. Natl. Acad. Sci. USA 80 (1983) 3963-3965.

7. **Cai, X.-Y., M.E. Maxon, B. Redfield, R. Glass, N. Brot, and H. Weissbach.** 1989. Methionine synthesis in *Escherichia coli*: effect of the MetR protein on *metE* and *metH* expression. Proc. Natl. Acad. Sci. USA **86**:4407-4411.

8. **Canovas, J.L., and R.Y. Stanier.** 1967. Regulation of the enzymes of the ß-ketoadipate pathway in *Moraxella calcoacetica*. 1. General aspects. Eur. J. Biochem. **1**:289-300.

9. **Chang, M. and I.P. Crawford.** 1990. The roles of indoleglycerol phosphate and the TrpI protein in the expression of *trpBA* from *Psuedomonas aeruginosa*. Nucl. Acids Res. **18**:979-987.

10. **Chang, M., A. Hadero, and I.P. Crawford.** 1989. Sequence of the *Pseudomonas aeruginosa trpI* activator gene and relatedness of *trpI* to other procaryotic regulatory genes. J. Bacteriol. **171**:172-183.

11. **Christman, M.F., G. Storz, and B.N. Ames.** 1989. OxyR, a positive regulator of hydrogen peroxide-inducible genes in *Escherichia coli* and *Salmonella typhimurium*, is homologous to a family of bacterial

regulatory proteins. Proc. Natl. Acad. Sci. USA **86**:3484-3488.

12. **Clark, D.** 1981. Regulation of fatty acid degradation in *Escherichia coli*: Analysis by operon fusion. J. Bacteriol. **148**:521-526.

13. **Collado-Vides, J., B. Magasanik, and J.D. Gralla.** 1991. Control site location and transcriptional regulation in *Escherichia coli*. Microbiol. Rev. **55**(3):371-394.

14. **Comeau, D.E., and M. Inouye.** 1988. A novel method for the cloning of chromosomal mutations in a single step: isolation of two mutant alleles of *envZ*, an osmoregulatory gene from *Escherichia coli*. Mol. Gen. Genet. **213**:166-169.

15. **De Crombrugghe, B., S. Busby, and H. Buc.** 1984. Cyclic AMP receptor protein: Role in transcription activation. Science **224**:831-838.

16. **Debarbouille, M., H.A. Shuman, T.J. Silhavy, and M. Schwartz.** 1978. Dominant constitutive mutations in *malT*, the positive regulator gene of the maltose regulon in *Escherichia coli*. J. Mol. Biol. **124**:359-371.

17. **DiRusso, C.C.** 1988. Nucleotide sequence of the *fadR* gene, a multifunctional regulator of fatty acid metabolism in *Escherichia coli* . Nucl. Acids. Res. **16**:7995-8009.

18. **DiRusso, C.C., and W. D. Nunn.** 1984. Cloning and characterization of a gene (*fadR*) involved in regulation of fatty acid metabolism in *Escherichia coli*. J. Bacteriol. **161**:583-588.

19. **Doten, R.C., K.-L. Ngai, D.J. Mitchell, and L.N. Ornston.** 1987. Cloning and genetic organization of the *pca* gene cluster from *Acinetobacter calcoaceticus*. J. Bacteriol. **169**:3168-3174.

20. **Drummond, M., P. Whitty, and J, Wootton.** 1986. Sequence and domain relationships of *ntrC* and nifA from *Klebsiella pneumoniae*: homologies to other regulatory proteins. EMBO J. 5(2):441-447.

21. **Duncombe, G. R., and F. E. Frerman.** 1976. Molecular and catalytic properties of the acetoacetyl-coenzyme A thiolase of *Escherichia coli*. Arch. Biochem. Biophys. **176**:159-170.

22. **Dynan, W. S., and R. Tjian.** 1985. Control of eukaryotic messenger RNA synthesis by sequence-specific DNA-binding proteins. Nature **316**:774-778.

23. **Esfahani, M., T. Ioneda, and S. J. Wakil.** 1971. Studies on the control of fatty acid metabolism: Incorporation of fatty acids into phospholipids and regulation of fatty acid synthetase of *Escherichia coli*. J. Biol. Chem. **246**:50-56.

24. **Falcone, A.B., and P.D. Boyer.** 1959. Arch. Biochem. Biophys. **83**:337-344.

25. **Feigenbaum, J., and H. Schulz.** 1975. Thiolases of *Escherichia coli*: Purification and chain length specificities. J. Bacteriol. **122**:407-411.

26. **Frerman, F.E.** 1973. The role of acetyl-coenzyme A in the transferase uptake of butyrate by isolated membrane vesicles of *Escherichia coli*. Arch. Biochem. Biophys. **159**:444-452.

27. **Frerman, F.E., and W. Bennett.** 1973. Studies on the uptake of fatty acids by *Escherichia coli*. Arch. Biochem. Biophys. **159**:434-443.

28. **Frerman, F.E., P. Andreone, and D. Mielke.** 1977. Reaction of pyridoxal 5'-phosphate with *Escherichia coli* CoA transferase: Evidence for an essential lysine residue. Arch. Biochem. Biophys. **181**:508-515.

29. **Fried, M., and D. M. Crother.** 1981. Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. Nucl. Acids Res. **9**:6505-6523.

30. **Gottesman, S.** 1984. Bacterial regulation: global regulatory networks. Ann. Rev. Genet. **18**:415-441.

31. **Gutierrez, C. and O. Raibaud.** 1984. Point mutations that reduce the expression of *malPQ*, a positively controlled operon of *Escherichia coli*. J. Mol. Biol. **177**:69-86.

32. **Heincz, M and E. McFall.** 1978. Role of the *dsdC* activator in regulation of D-serine deaminase synthesis. J. Bacteriol. **136**:96-103.

33. **Hemsley, A., N. Arnheim, M. D. Toney, G. Cortopassi, and D. J. Galas.** 1989. A simple method for site-directed mutagenesis using the polymerase chain reaction. Nucl. Acids Res. **17**:6545-6551.

34. **Hersh, L. B. and W. P. Jencks.** 1967a. Coenzyme A transferase. Kinetics and exchange reactions. J. Biol. Chem. **242**:3468-3480.

35. **Hersh, L. B. and W. P. Jencks.** 1967b. Coenzyme A transferase. Isolation and properties of an enzyme-coenzyme A intermediate. J. Biol. Chem. **242**:3481-3486.

36. **Ho, Y.-S., M. Lewis, and M. Rosenberg.** 1982. Purification and properties of a transcriptional activator: The cII of phage lambda. J. Biol. Chem. **257**:9128-9134.

37. **Hochschild, A., N. Irwin and M. Ptashne.** 1983. Repressor structure and the mechanism of positive control. Cell **32**:319-325.

38. **Holtel, A., K.N. Timmis, and J.L. Ramos.** 1992. Upstream binding sequences of the XylR activator protein and integration host factor in the *xylS* gene

promoter region of the *Pseudomonas* TOL plasmid. Nucleic Acids Res. **20**:1755-1762.

39. **Holtel, A.,M.-A. Abril S. Marques, K. N. Timmis, and J. L. Ramos.** 1990. Promoter-upstream activator sequences are required for expression of the *xylS* gene and upper-pathway operon on the *Pseudomonas* TOL plasmid. Mol. Microb. **4**(9):1551-1556.

40. **Hoover, T. R., E. Santero, S. Porter, and S. Kustu.** 1990. The integration host factor stimulates interaction of RNA polymerase with NIFA, the transcriptional activator for nitrogen fixation operons. Cell **63**:11-22.

41. **Igual, J. C.,C. Gonzalezbosch, J. Dopazo, and J. E. Perezortin.** 1992. Phylogenetic analysis of the thiolase family-implications for the evolutionary origin of peroxisomes. J. Molec. Evolution. **35**:147-155.

42. **Inouye, S., A. Nakazawa, and T. Nakazawa.** 1988. Nucleotide sequence of the regulatory gene *xylR* of the TOL plasmid from *Pseudomonas putida*. Gene **66**:301-306.

43. **Jackowski, S., and C.O. Rock.** 1986. Consequences of reduced intracellular coenzyme A content in *Escherichia coli*. J. Bacteriol. **166**:866-871.

44. **Jenkins, L. S. and W. D. Nunn.** 1987a. Genetic and molecular characterization of the genes involved in

short chain fatty acid degradation in *Escherichia coli*. J. Bacteriol. **169**:42-52.

45. Jenkins, L.S. and W.D. Nunn. 1987b. Regulation of the *ato* operon by the *atoC* gene in *Escherichia coli*. J. Bacteriol. **169**:2096-2102.

46. Jenks, W.P. 1973. Coenzyme A trasferases, p.483-496. *In* P.D. Boyer (ed.), The Enzymes, Vol. IX Group Transfer, Part B Phosphoryl Transfer, One-Carbon Group Transfer, Glycosyl Transfer, Amino Group Trasnfer, Other Transferases. Academic Press, New York.

47. Kadonaga, J. T., K. R. Carner, F. R. Masiarz, and R. Tjian. 1987. Isolation of cDNA encoding transcription factor Sp1 and functional analysis of the DNA binding domain. Cell. **51**:1079-1090.

48 Kemp, M.B., and G.D. Hegeman. 1968. J. Bacteriol. **96**:1488-1499.

49. Klein, K., R. Steinberg, B. Fiethen, and P. Overath. 1971. Fatty acid degradation in *Escherichia coli*. An inducible system for the uptake of fatty acids and further characterization of old mutants. Eur. J. Biochem. **19**:442-450.

50. Komeda, Y and T. Iino. 1979. Regulation of expression of the flagellin gene (*hag*) in *Escherichia coli* K-12: Analysis of *hag-lac* gene fusions. J. Bacteriol. **139**:721-729.

51. **Kornberg, H. L.** 1966. The role and control of the glyoxylate cycle in *Escherichia coli*. Biochem. J. **99**:1-11.

52 **Köhler, T., S. Harayama, J.-L. Ramos, and K. N. Timmis**. 1989. Involvement of *Pseudomonas putida* RpoN sigma factor in regulation of various metabolic functions. J. Bacteriol. **171**:4326-4333.

53. **Kraft, R,. J. Tardiff, K.S. Krauter, and L.A. Leinwand**. 1988. Using mini-prep plasmid DNA for sequencing double stranded template with Sequenase. BioTechniques. **6**:544-546.

54. **Liljestrom, P., P.L. Maattanen, and E.T. Palva**. 1982. Cloning of the regulatory locus *ompB* of *Salmonella typhimurium* LT-2: Isolation of the *ompR* gene and identification of its gene product. Mol. Gen. Genet. **188**:184-189.

55. **Little, J.W., and D.W. Mount**. 1982. The SOS regulatory system of *Escherichia coli*. Cell **29**:11-22.

56. **Lorenzo, V., M. Herrero, M. Metzke, and K.N. Timmis**. 1991. An upstream XylR- and IHF-induced nucleoprotein complex regulates the sigma^{54} -dependent Pu promoter of TOL plasimd. EMBO. **10**(5):1159-1167.

57. **Maloy, S.R. and W.D. Nunn**. 1981. Role of gene *fadR* in *Escherichia coli* acetate metabolism. J. Bacteriol. **148**:83-90.

58. **Maloy, S.R. and W.D. Nunn.** 1982. Genetic regulation of the glyoxylate shunt in *Escherichia coli* K-12. J. Bacteriol. **149**:173-180.

59. **Maniatis, T., E.T. Fritsch, and J. Sambrook.** 1982. Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

60. **McClelland, M., J. Hanish, M. Nelson, and Y. Patel.** 1988. KGB: a single buffer for all restriction endonucleases. Nucl. acids Res. 16:364-364.

61. **McClure, W. R.** 1985. Mechanism and control of transcription initiation in prokaryotes. Ann. Rev. Biochem. **54**:171-204.

62. **Merrick, M., J. Gibbins, and A. Toukdarian.** 1987. The nucleotide sequence of the sigma factor gene *ntrA* (*rpoN*) of *Azotobacter vinelandii*: Analysis of conserved sequences in NtrA proteins. Mol. Gen. Genet. **210**:323-330.

63 **Messing, J.**: New M13 vectors for cloning, pp. 20-79. *In* Wu, R., Grossman, L. and Moldave, K. (eds.), Methods in Enzymology, Volume 101, Recombinant DNA, Part C. Academic Press, New York, 1983.

64. **Miyada, C.G., X. Soberon, K. Itakura, and G. Wilcox.** 1982. The use of synthetic oligodeoxyribonucleotides to produce specific deletions

in the *araBAD* promoter of *Escherichia coli* B/r. Gene. **17**:167-177.

65. **Morett, E., and M. Buck.** 1988. NifA-dependent *in vivo* protection demonstrates that the upstream activator sequence of *nif* promoters is a protein binding site. Proc. Natl. Acad. Sci. USA 85:9401-9405.

66. **Nunn, W.D.** 1986. A molecular view of fatty acid catabolism in *Escherichia coli*. Microbiol. Rev. **50**:179-192.

67. **Ogden, S., D. Haggerty, C.M. Stoner, D. Kolodrubetz, and R. Schleif.** 1980. The *Escherichia coli* L-arabinose operon: Binding sites of the regulatory proteins and a mechanism of positive and negative regulation. Proc. Natl. Acad. Sci. USA **77**:3346-3350.

68. **Overath, P., E. Raufuss, W. Stoffel, and W. Ecker.** 1967. The induction of the enzymes of fatty acid degradation in *Escherichia coli*. Biochem. Biophys. Res. Commun. **29**:28-33.

69. **Overath, P., G. Pauli, and H.U. Schairer.** 1969. Fatty acid degradation in *Escherichia coli*. An inducible acyl-CoA synthetase, the mapping of *old*-mutations, and the isolation of regulatory mutants. Eur. J. Biochem. **7**:559-574.

70. **Pabo, C.O., and R. T. Sauer.** 1984. Protein-DNA recognition. Ann. Rev. Biochem. **53**:293-321.

71. **Parales, R.E.and C.S. Harwood.** 1992. Characterization of the genes encoding beta-ketoadipate:succinyl-Coenzyme-A transferase in *Pseudomonas putida*. J. Bacteriol. **174**:4657-4666.

72. **Pauli, G. and P. Overath.** 1972. *ato* operon: A highly inducible system for acetoacetate and butyrate degradation in *Escherichia coli*. Eur. J. Biochem. **29**:553-562.

73. **Ptashne, M.** 1967. Isolation of the lambda phage repressor. Genetics. **57**:306-313.

74. **Raibaud, O and M. Schwartz.** 1984. Positive control of transcription initiation in bacteria. Ann. Rev. Genet. **18**:173-205.

75. **Raibaud, O., M. Debarbouille, and M. Schwartz.** 1983. Use of deletions created *in vitro* to map transcriptional regulatory signals in the *mal*A region of *Escherichia coli*. J. Mol. Biol. **163**:395-408.

76. **Raibaud, O., M. Mock, and M. Schwartz.** 1984. A technique for integrating any DNA fragment into the chromosome of *Escherichia coli*. Gene **29**:231-241.

77. **Ramos, J.L., C. Michan, F. Rojo, D. Dwyer, and K.N. Timmis.** 1990. Signal-regulator interactions. Genetic analysis of the effector binding site of *xylS*, the benzoate-activated positive regulator of

Pseudomonas TOL plasmid *meta*-cleavage pathway operon. J. Mol. Biol. **211**:373-382.

78. **Ramos, J.L., F. Rojo, L. Zhou and K.N. Timmis.** 1990. A family of positive regulators related to the *Pseudomonad putida* TOL plasmid XylS and the *Escherichia coli* AraC activators. Nucl. Acids Res. **18**:2149-2152.

79. **Reznikoff, W. S., D. A. Siegele, D. W. Cowing, and C. A. Gross.** 1985. The regulation of transcription initiation in bacteria. Ann. Rev. Genet. **19**:355-87.

80. **Riggs, A. D., G. Reiness, and G. Zubay.** 1971. Purification and DNA-binding properties of the catabolite gene activator protein. Proc. Nat. Acad. Sci. USA **68**:1222-1225.

81. **Rothmel, E.K., T. Aldeich, J.E. Houghton, W.M. Coco, L.N. Ornston, and A.M. Chakrabarty.** 1990. Nucleotide sequencing and characterization of Pseudomonas putida *catR*: a positive regulator of the *catBC* operon is a member of the LysR family. J. Bacteriol. **172**:922-931.

82. **Salanitro, J.P. and W.S. Wegener.** 1971. Growth of *Escherichia coli* on short-chain fatty acids: Growth characteristics of mutants. J. Bacteriol. **108**:885-892.

83. **Salanitro, J.P. and W.S. Wegener.** 1971. Growth of *Escherichia coli* on short-chain fatty acids:

Nature of the uptake system. J. Bacteriol. **108**:893-901.

84. **Sallus, L, R.J. Haselbeck, and W.D. Nunn.** 1983. Regulation of fatty acid transport in *Escherichia coli*: Analysis by operon fusion. J. Bacteriol. **155**:1450-1454.

85. **Sancar, A., A.M. Hack, and W.D. Rupp.** 1979. Simple method for identification of plasmid-coded proteins. J. Bacteriol. **137**:692-693.

86. **Schell, M.A., and E.F. Poser.** 1989. Demonstration, characterization, and mutational analysis of NahR protein binding to *nah* and *sal* promoters. J. Bacteriol. **174**(2):837-846.

87. **Schell, M. A., and M. Sukordhaman.** 1989. Evidence that the transcription activator encoded by the *Psuedomonas putida nahR* gene is evolutionarily related to the transcription activator encoded by the *Rhizobium nodD* genes. J. Bacteriol. **174**(4):1952-1959.

88. **Shanley, M.S., E.L. Neidle, R.E. Parales, and L.N. Ornston.** 1986. Cloning and expression of *Acinetobacter calcoaceticus catBCDE* genes in *Pseudomonas putida* and *Escherichia coli*. J. Bacteriol. **165**:557-567.

89 **Sharp, J.A. and M.R. Edwards.** 1978. Purification and properties of succinoyl-coenzyme A-3-oxo acid

coenzyme A-transferase from sheep kidney. Biochem. J. **173**:759-765.

90 **Sharp, J.A. and M.R. Edwards.** 1983. Initial velocity kinetics of succinoyl-coenzyme A-3-oxo acid coenzyme A-transferase from sheep kidney. Biochem. J. **213**:179-185.

91. **Shen, W. ,and M. Y. Waye.** 1988. A novel method for generating a nested set of unidirectional deletion mutants using mixed oligodeoxynucleotides. Gene **70**:205-211.

92. **Siebenlist, U., R.B. Simpson, and W. Gilbert.** 1980. *E. coli* RNA polymerase interacts homologously with two different promoters. Cell **20**:269-281.

93. **Silhavy, T.J., and J.R. Beckwith.** 1985. Uses of *lac* fusions for the study of biological problems. Microbiol. Rev. **49**:398-418.

94. **Simons, R. W., P. A. Egan, H. T. Chute, and W. D. Nunn.** 1980. Regulation of fatty acid degradation in *Escherichia coli*: isolation and characterization of strains bearing insertion and temperature-sensitive mutations in *fadR*. J. Bacteriol. **142**:621-632.

95. **Singh, H., J. H. Lebowitz, A. S. Baldwin, Jr., and P.A. Sharp.** 1988. Molecular cloning of an enhancer binding protein: isolation by screening of an expression library with a recognition site DNA. Cell **52**:415-423.

96. **Singh, H., R. G. Clerc, and J. H. LeBowitz.** 1989. Molecular cloning of seqence-specific DNA binding proteins using recognition site probes. BioTechniques. **7**:252-261.

97. **Spratt, S. K., C. L. Ginsburgh, and W. D. Nunn.** 1981. Isolation and genetic characterization of *Escherichia coli* mutants defective in propionate metabolism. J. Bacteriol. **146**:1166-1169.

98. **Sramek, S. J. and F. E. Frerman.** 1975a. Purification and properties of *Escherichia coli* coenzyme A-transferase. Arch. Biochem. Biophys. **171**:14-26.

99. **Sramek, S. J. and F. E. Frerman.** 1975b. *Escherichia coli* coenzyme A-transferase: Kinetics, catalytic pathway and structure. Arch. Biochem. Biophys. **171**:27-35.

100. **Sramek, S. J., F. E. Frerman, and M. B. Adams.** 1977a. Sulfhydryl group reactivity in the *Escherichia coli* CoA transferase. Arch. Biochem. Biophys. **181**:516-524.

101. **Sramek, S.J., F.E. Frerman, D.J. McCormick, and G.R. Duncombe.** 1977b. Substrate-induced conformational changes and half-the-sites reactivity in the *Escherichia coli* CoA transferase. Arch. Biochem. Biophys. **181**:525-533.

102. Stanier, R.Y., and L.N. Ornston. 1973. The ß-ketoadiapte pathway. Adv. Microb. Physiol. **9**:89-151.

103. Su, W., S. Porter, S. Kustu, and H. Echols. 1990. DNA-looping and enhancer activity: association between DNA-bound NtrC activator and RNA polymerase at the bacterial *glnA* promoter. Biochemistry **87**:5504-5508.

104. Tang D.-C., and M.W. Taylor. 1990. Transcriptional activation of the adenine phosphoribosyltransferase promoter by an upstream butyrate-induced Moloney murine sarcoma virus enhancer-promoter element. J. Virology. **64**:2907-2911.

105. Tommassen, J., P. De Geus, B. Lugtenberg. 1982. Regulation of the *pho* regulon of *Escherichia coli* K-12: Cloning of the regulatory genes *phoB* and *phoR* and identification of their gene products. J. Mol. Biol. **157**:265-274.

106. Von Hippel, P.H., D.G. Bear, W.D. Morgan, and J.A. McSwiggen. 1984. Protein-nucleic acid interactions in transcription: A molecular analysis. Ann. Rev. Biochem. **53**:389-446.

107. Weeks, G., M. Shapiro, R.O. Burns, and S.J. Wakil. 1969. Control of fatty acid metabolism. I. Induction of the enzymes of fatty acid oxidation in *Escherichia coli*. J. Bacteriol. **97**:827-836.

108. **Weiss, D. S., J. Batut, K. E. Klose, J. Keener, and S. Kustu.** 1991. The phosphorylated form of the enhancer-binding protein NTRC has an ATPase activity that is essential for activation of transcription. Cell **67**:155-167.

109. **Weiss, V., F. Claverie-Martin, and B. Magasanik.** 1992. Phosphorylation of nitrogen regulator I of *Escherichia coli* induces strong cooperative binding to DNA essential for activation of transcription. Proc. Natl. Acad. Sci. USA **89**:5088-5092.

110. **Weyens, G., D. Charlier, M. Roovers, A. Piérard and N. Glansdorff.** 1988. On the role of the Shine-Dalgarno sequence in determining the efficiency of translation initiation at a weak start codon in the *car* operon of *Escherichia coli* K12. J. Mol. Biol. **204**:1045-1048.

111. **White, H., and W.P. Jencks.** 1976. Properties of succinyl-CoA:3-ketoacid coenzyme A trasnferase. J. Biol. Chem. **251**:1708-1711.

112. **Wilcox, G., K. J. Clemetson, D. V. Santi, and E. Englesberg.** 1971. Purification of the *araC* protein. Proc. Nat. Acad. Sci. USA **68**:2145-2148.

113. **Yang, S.-Y., X.-Y.H. Yang, G. Healy-Louie, H. Schulz, and M. Elzinga.** 1990. Nucleotide sequence of the *fadA* gene. J. Biol. Chem. **265**:10424-10429.

114. Yeh, W.-K., and L.N. Ornston. 1981. Evolutionarily homologous $\alpha_2ß_2$ oligomeric structures in ß-ketoadipate succinyl-CoA transferases from *Acinetobacter calcoaceticus* and *Pseudomonas putida*. J. Biol. Chem. **256**:1565-1569.

115. Yeh, W.-K., and L.N. Ornston. 1984. *p*-Chloromercuribenzoate specifically modifies thiols associated with the active sites of ß-ketoadipate enol-lactone hydrolase and succinyl CoA:ß-ketoadipate CoA transferases. Archiv. Microbiol. **138**:102-105.

116. You, I.-S., D. Ghosal, and I. C. Gunsalus. 1988. Nucleotide sequence of plasmid NAH7 gene *nahR* and DNA binding of the *nahR* product. J. Bacteriol. **170**:5409-5415.