J. Genet., Vol. 76, Number 2, August 1997, pp. 147–159. © Indian Academy of Sciences

REVIEW ARTICLE

Cryptic genes: evolutionary puzzles

MITALI MUKERJI and S. MAHADEVAN*
Developmental Biology and Genetics Laboratory, Indian Institute of Science, Bangalore 560 012, India

MS received 8 May; revised 30 May 1997

Abstract. Many microorganisms carry genes that have the potential to code for specific functions but remain inactive during the normal lifetime of the organism. Such genes have been termed cryptic genes and their activation usually requires a mutational event. They are different from pseudogenes which arise as a result of duplication of a functional gene but remain inactivated because of the accumulation of multiple mutations. This review is an attempt to examine some of the well-characterized cryptic genetic systems in *Escherichia coli* in an effort to understand their functional and evolutionary significance.

Keywords. Silent genes; β-glucoside utilization; microbial evolution.

1. Introduction

Microorganisms have the capacity to survive in a multitude of environments such as extremes of pH and temperature, oxidative stress and presence of toxic substrates. Most bacteria were probably exposed to environments that fluctuated so often that they have evolved different strategies to meet these challenges (reviewed by Koch 1993). These include developing metabolically dormant states such as spores; activating previously evolved, but silent, genes; increasing the rate of mutations under stress conditions; and favouring recombination by promoting movement of exogeneous and endogenous genetic elements such as plasmids, phages and transposable elements. The strategy most relevant to this discussion is the activation of previously evolved but silent (cryptic) genes that are a part of the genetic repertoire of the organism.

Cryptic genes have been defined as phenotypically silent DNA sequences that are not expressed during the normal life cycle of an organism (Hall *et al.* 1983). They may be activated in a few individuals of a population by mutation, recombination, transposition or other genetic mechanisms. The presence of these silent genes becomes evident only upon mutation which results in a detectable phenotype. Therefore they differ from other silent DNA sequences, collectively referred to as selfish DNA, which spread in the genome by making additional copies of themselves and survive at the expense of the host (Dawkins 1976; Doolittle and Sapienza 1980; Orgel and Crick 1980), in most cases without conferring any phenotypic advantage to the organism. Cryptic genes also differ from pseudogenes which are homologous copies of an active gene but have a large number of accumulated mutations so that their reversion into an active form is not usually possible (Lauer *et al.* 1980; Nishioka *et al.* 1980; Li *et al.* 1981).

There are several examples of silent genes in prokaryotes (Hall *et al.* 1983). Genes for β-glucoside utilization are cryptic in *E. coli* and *Salmonella* (Schaefler and Malamy 1969). *E. coli*, which has been classified on the basis of its inability to utilize citrate, also

*email: mahi@serc.iisc.ernet.in.

possesses a silent gene for citrate utilization, detectable after mutational activation (Hall 1982). There are cryptic genes for various biosynthetic pathways in *Lactocillus* (Morishita *et al.* 1974). A cryptic gene for alcohol dehydrogenase (*Adh* IV) has been identified in yeast (Paquin and Williamson 1986).

If environmental conditions do not demand periodic selection of the active form of the gene, it will not contribute to the fitness of the organism. Therefore, in the absence of selection, mutations accumulating in the gene will render it nonfunctional and the gene will eventually be lost from the population (Dykhuizen 1978). However, most cryptic genes studied so far show that this is not the case. Unlike pseudogenes, they can be activated by a single mutational event in most of the cases studied, indicating that they have not accumulated multiple inactivating mutations. A study of the population dynamics of these genes and their molecular organization can give us important insights into the basis of their retention and maintenance in the face of evolutionary pressure.

It has been proposed that cryptic genes are maintained in a population through repeated cycles of activation and cryptification (Hall *et al.* 1983). The presence of an active allele may be disadvantageous in one environment, resulting in a strong selection for the cryptic state when the gene product is not required. However, in another environment, due to a distinct growth advantage, there might be a selection for the active state as it provides the ability to utilize additional substrates. Hence, depending on the environment, either the active or the cryptic state would be present under natural conditions. For example, an organism in which the genes for the metabolism of a specific compound are active will be handicapped in an environment in which toxic analogues of the compound are present. But conditions may exist where a nontoxic metabolite may be present as the sole nutritional source and it will be advantageous to activate the genes and be able to utilize that metabolite. Therefore activation of a cryptic gene will be more advantageous compared to the induction of a normal but repressed gene in such cases.

The differential selective advantage conferred by the cryptic and the active forms of a gene under different environmental conditions is exemplified in the case of the cryptic *ilvG* gene of *E. coli* K12. This gene codes for α-acetohydroxyacid synthase II, an isoenzyme that differs from isoenzymes I and III encoded by unlinked genes that are active in natural isolates. Activation of the *ilvG* gene is brought about by frameshift mutations and such mutants can be readily isolated (Lawther *et al.* 1981, 1982). Strains carrying the activated *ilvG* gene overproduce the *ilvE, ilvD* and *ilvA* products and wastefully excrete valine into the medium (Leavitt and Umbarger 1962). Therefore strains that have the active copy of the *ilvG* gene are at a disadvantage compared to wild-type strains under conditions where the genes encoding the isoenzymes I and III are active. Interestingly, strains carrying the activated *ilvG* gene are insensitive to inhibition by valine. When a moderate concentration of valine is present in the environment, growth of the wild-type strain, and not that of the mutant, is inhibited (Umbarger 1978; De Felice *et al.* 1979). Similarly, when auxotrophs are grown in competition with isogenic wild-type strains in glucose-limited chemostats containing excess of the required nutrient, they outgrow and displace the wild-type strain from the population. This phenomenon has been observed in *E. coli* and *Bacillus subtilis* strains (Zamenhoff and Eichorn 1967; Dykhuizen 1978). It has been speculated that such growth inhibition of the wild-type strain is due to accumulation of toxic intermediates of the anabolic pathway which is operational in the wild-type strain even when the end product is

already available in the medium (Hall *et al.* 1983). This may not apply to pathways that are regulated by feedback inhibition by the end product.

The basis for retention of a cryptic gene in the face of mutational pressures has been analysed using a mathematical model (Hall *et al.* 1983). The basic assumption in the model, as stated earlier, is that under one set of conditions, members of the population with a cryptic gene have higher fitness compared to those with the functional gene. Under an alternative set of conditions, those members which express the gene are at a stronger selective advantage. Thus in a population with three distinct classes of genes, cryptic, functional and nonfunctional, the equilibrium frequency of each of the classes depends on the rates of mutation from one class to the other and the relative fitness of each of the alleles in a particular environment. By considering different fitness values according to the conditions, the following conclusions were arrived at:

(i) In the absence of any selection between the cryptic and nonfunctional gene, the nonfunctional genes would get fixed in the population.

(ii) In the absence of any selective difference between the functional and cryptic genes and a low rate of mutation of these genes to a nonfunctional state the frequency of the cryptic gene depends mainly on the forward and backward mutation rates between the functional and cryptic states.

(iii) Under conditions where the cryptic and functional genes confer higher fitness compared to the nonfunctional gene, the frequency of the cryptic gene would be close to unity if the mutation rate from the functional state to the cryptic state is higher than vice versa.

These theoretical predictions were confirmed using numerical examples. It was shown that, even when its frequency is much lower and it is less fit than the functional allele, the cryptic allele persists in the population. Moreover, the cryptic gene could get fixed in a population even if it had only a slightly higher fitness advantage compared to the functional gene. Using computer simulations that included reasonable values for mutation rates, Li (1984) arrived at similar conclusions. A 10% advantage in expressing the gene under conditions where it is useful and a 0·10% advantage in being cryptic were considered. Under conditions where the gene is not useful Li found that a cryptic gene needs to be expressed is less than 8 out of 1000 generations for the gene to be retained in the population. The key feature in these two models is that conditions do exist in which the mutant allele is more favoured than the wild type.

Inactivation of a genetic system can be accomplished once selection is removed by a variety of genetic changes such as mutations in the regulatory elements and inactivation of the structural gene (s). However, for the evolutionary paradigm of cyclic activation and cryptification to work, the constraint that the inactivation process has to be reversible greatly resticts the choice of genetic mechanisms. Silencing/activation of a gene by mobile genetic units is distinctively advantageous compared to other kinds of mutations since insertion and excision of a transposable element are very precise or previse enough to regenerate the gene function anew. Moreover, the frequencies of insertion and excision of transposable elements are also quite reasonable for them to function in the activation or cryptification process. The use of a transpositional mechanism by the bacterium may offer important advantages in its competition with other bacteria. By examining cryptic genetic systems that have been well characterized at the structural and organizational levels, one can see whether any of the rationales discussed so far are encountered in nature.

2. Cryptic systems involved in β-glucoside utilization

The most common β-glucosides found in nature are the aryl β-glucosides arbutin and salicin and the disaccharide cellobiose, the repeating unit of cellulose. Metabolism of these sugars involves a complex set of permeases and hydrolases with different substrate specificities. Most β-glucosidases can catalyse the hydrolysis of only the phosphorylated derivative of the sugar (Fox and Wilson 1968). The sugars are transported by the phosphoenolpyruvate (PEP)-dependent transport system that phosphorylates the sugars at the c_6 position of glucose during transport and are hydrolysed by phospho-β-glucosidases which cleave it to produce glucose 6-phosphate and aglycone. The phospho-β-glucosidases are grouped into classes A and B based on their substrate specificity. The phospho-β-glucosidase A can cleave *p*-nitrophenyl glucoside and other aromatic glucosides except salicin. The phospho-β-glucosidase B is more thermolabile and can cleave salicin in addition to the other substrates.

Members of the family Enterobacteriaceae show a marked difference in their ability to utilize β-glucosides (Schaefler and Malamy 1969). *Klebsiella* can utilize all the three β-glucosides whereas *Salmonella* and *E. coli* cannot utilize any. However, they can spontaneously mutate at a frequency of 10^{-5} to utilize these sugars, *Salmonella* to Cel^+, and *E. coli* to Cel^+ and Bgl^+, indicating the presence of the corresponding genes in a cryptic state (Schaefler and Mintzer 1959; Schaefler and Schenkein 1968; Prasad and Schaefler 1974). *Citrobacter* can utilize cellobiose but not aryl β-glucosides, whereas *Proteus vulgaris* can utilize arbutin and salicin but not cellobiose.

Four cryptic genetic systems involved in the utilization of β-glucosides have been characterized in *Escherichia coli* (Prasad and Schaefler 1974; Kricker and Hall 1987; Parker and Hall 1988, 1990a, b; Hall and Xu 1992). These operons have been identified by their ability to mutate spontaneously to a phenotype detectable as papillae on indicator plates when supplemented with the specific β-glucoside. These papillae arise as a result of activating mutations in a few members of the population which have a growth advantage enabling them to utilize the specific carbon source present and hence outgrow the neighbouring cells within the colony. The cryptic systems involved in β-glucoside utilization are the *bgl* operon involved in the utilization of salicin and arbutin, the *cel* and *asc* operons involved in the utlilization of salicin, arbutin and cellobiose, and the *arbT* locus involved in the utilization of arbutin. Though there is a constitutive β-glucosidase involved in the utilization of arbutin, encoded by *bglA*, wild-wild *E. coli* strains show a Arb^- phenotype (Prasad *et al.* 1973). This is because the PEP-dependent arbutin-specific permease is encoded by the other cryptic operons. Hence activation of at least one of the operons is essential for an Arb^+ phenotype.

The different operons involved in β-glucoside utilization map to different locations on the *E. coli* chromosome. The *bgl* operon is located at 83·8 minutes (Bachmann 1990), *cel* at 37·8 (Bachmann 1990), *bglA* at 62·3 (Rudd *et al.* 1990), and the *asc* operon at 58·3 minutes (Hall *et al.* 1991) of the *E. coli* map. Though all the operons have similar functions and can catabolize at least two of the β-glucosides, there is no sequence homology between similar genes of the operon either at the DNA or at the protein level. They share considerable identity with genes which have related but not homologous functions. This indicates that the genes for β-glucoside utilization have not evolved from a common ancestor.

Despite the presence of four sets of genes for β-glucoside utilization in *E. coli*, it is unusual that all of them have been maintained in a cryptic state. It is quite possible that

the ecological niche of the organism might determine which of the cryptic genes may be activated. For example, among members of the Enterobacteriaceae such as the plant pathogens *Citrobacter* and *Erwinia* and soil microbes such as *Proteus* and *Klebsiella*, activation of the genes involved in cellobiose metabolism is probable since cellobiose is likely to be present in their environment. An understanding of the structure, function and molecular mechanism of activation of the operon along with the population studies of natural isolates may help us understand the basis of selection that maintains genes in a silent state in the face of mutational pressure. It might also shed some light on the existence of a common theme in the maintenance of cryptic genes.

3. The *bgl* operon of *E. coli*

The *bgl* operon, which is involved in the utilization of the aromatic β-glucosides salicin and arbutin, has been extensively studied at the molecular level. Wild-type strains of *E. coli* cannot grow in a medium containing salicin as sole carbon source and therefore exhibit a Bgl^- phenotype. However, mutants arise at a frequency of 10^{-5} to as high as 10^{-4} in some strains that show a Bgl^+ phenotype (Schaefler and Mintzer 1959; Schaefler and Schenkein 1968). The activating mutations have been mapped to a locus termed *bglR* which is present upstream of the structural genes (Prasad and Schaefler 1974). The predominant class of activating mutations consists of the insertion of IS*1* and IS*5* sequences within the *bglR* region (Reynolds *et al.* 1981). The observation that activation of the operon can be achieved by a single mutational event in the regulatory locus indicates that the structural genes are intact in the wild-type strain. The identification of three structural genes by genetic studies (Prasad and Schaefler 1974) has been confirmed by functional analysis (Mahadevan *et al.* 1987) and determination of the nucleotide sequence of the operon (Schnetz *et al.* 1987). The *bglG* gene encodes a positive regulatory protein and the *bglF* gene which encodes the PEP-dependent enzymeII^{bgl} also acts as a negative regulator of operon expression (Mahadevan *et al.* 1987). The *bglB* gene encodes a phospho-β-glucosidase that hydrolyses the phosphroylated substrates salicin and arbutin.

The *bgl* operon has evolved a very sophisticated mechanism to regulate expression in the presence of β-glucosides which involves antitermination of transcription (Mahadevan and Wright 1987; Schnetz and Rak 1988). The *bglG* gene is flanked by two terminators and the BglG protein acts at these sites to alleviate transcription termination. The two terminators share a highly conserved sequence motif proximal to and extending into their stems (Schnetz *et al.* 1987). It has been shown that BglG binds to this sequence motif at the mRNA level and prevents formation of the terminator structure, thereby allowing transcription to proceed (Houman *et al.* 1990). Negative regulation by BglF involves phosphorylation of BglG in the absence of the inducer (Amster–Choder *et al.* 1989; Schnetz and Rak 1990). Phosphorylation of BglG results in its monomerization, leading to loss of RNA binding (Amster–Choder and Wright 1992). The BglG––BglF combine is thus a very effective response regulator. The regulatory genes and the target sites near the terminators are highly conserved across species. The *Lactobacillus lactis* gene *bglR*, the *Bacillus subtilis* genes *sacT*, *sacY* and *licT*, and the *E. chrysanthemi arbG* gene are homologues of the *bglG* of *E. coli*. Products of all these genes show a strong conservation of amino acid residues and have been classified as the BglG family of antiterminator proteins (for reviews see Rutberg 1997;

Mahadevan 1997). The antiterminators are negatively regulated by the corresponding permease belonging to the phosphotransferase system (PTS) in each case.

The cryptic nature of the operon, in spite of possessing a highly efficient regulatory mechanism, is an enigma. As stated above, the operon is activated by a number of mechanisms. The most common activation event is the insertion of IS*1* and IS*5* elements at a well-defined locus in the *bglR* region (Reynolds *et al.* 1981, 1986; Schnetz and Rak 1992). Point mutations at the cAMP–CRP binding site which bring it closer to a core consensus sequence result in activation of the operon (Reynolds *et al.* 1986; Lopilato and Wright 1990). Besides, mutations in the unlinked *gyrA*, *gyrB*, *bglJ* and *hns* genes also lead to activation of the operon (Defez and DeFelice 1981; DiNardo *et al.* 1982; Hiffins *et al.* 1988; Giel *et al.* 1996). Transcriptional mapping studies have shown that, irrespective of the site of IS insertion, transcription always initiates from a unique site, indicating that the IS elements enhance transcription from a preexisting promoter (Reynolds *et al.* 1986).

Several studies have shown that negative elements present upstream of the promoter render the operon cryptic (Lopilato and Wright 1990; Schnetz 1995; Singh *et al.* 1995). Recent studies in our laboratory have shown that there are at least two negative elements down-regulating the *bgl* promoter. One of them is a structural element that can extrude under high superhelical densities and another is a sequence that interacts with the nucleoid protein H-NS leading to inhibition of transcription (Mukerji and Mahadevan 1997). These studies show that the former is involved in regulation mediated by supercoiling and the latter in regulation by CRP–cAMP. Most activating mutations result in disruption of the elements or their isolation from the promoter.

Interestingly, physiological conditions that alter DNA supercoiling or H-NS concentration in the cell may cause transient activation of the operon. This raises the question whether the operon is cryptic at all. However, the significance of this mode of regulation is not clear at present. One possibility is that specific substrates present in the environment can affect supercoiling or H-NS concentration. So far no such substrates have been observed in nature. If salicin and arbutin are considered as the natural substrates of the operon, insertion elements seem to be the predominant agents that lead to activation. Further information is needed before the significance of transient activation can be understood.

If the paradigm of cyclic activation and cryptification is considered, the activated operon is expected to revert to the cryptic state upon prolonged growth under nonselective conditions. Experiments carried out under laboratory conditions show that most Bgl^- revertants are unable to the reactivated to the Bgl^+ state as they carry either structural gene mutations or large deletions of the *bgl* genes (A. Ashtaputre and S. Mahadevan, unpublished). In the light of these observations, cycles of activation and silencing of the genes under laboratory conditions is unlikely and the persistence of the cryptic state is an enigma.

4. The *cel* operon

The *cel* operon, involved primarily in the catabolism of cellobiose, is another well-characterized cryptic genetic system. It is present at kilobase 1835 on the physical map of the *E. coli* chromosome and spans 4326 base pairs comprising 5 ORFs named *celA*,

celB, celC, celD and *celF*, specifying all the functions required for catabolism of β-glucosides, primarily cellobiose (Kricker and Hall 1984, 1987; Parker and Hall 1990a,b; Reizer *et al.* 1990). A 200-bp regulatory region is present upstream of the structural genes. The *celB* and *celC* genes are necessary for the transport of β-glucoside sugars. The protein encoded by *celB* is the cellobiose-specific enzII component of PTS involved in the concomitant phosphorylation and transport of cellobiose. The CelC protein resembles the PTS enzymeIII^{lac} of *Staphylococcus aureus* on the basis of sequence homology both at the DNA and at the protein level. CelD is a negative regulatory protein containing a helix-loop-helix (HLH) motif at its C-terminal end, a feature typical of DNA-binding proteins. It prevents expression of the wild-type *cel* operon and is trans-dominant. A plasmid containing the *cel* operon, deleted for *celD*, shows activation. CelF is the phospho-β-glucosidase which can cleave arbutin, salicin and cellobiose. The regulatory and the hydrolase proteins of the *cel* operon resemble those of the *mel* operon. However, there is no similarity between the transport proteins. The codon usage profile indicates that in the activated state the *cel* operon is expressed poorly. An activated *cel* operon has been found to be selectively disadvantageous when carried on a multicopy plasmid.

Two classes of mutations lead to activation of the *cel* operon: insertion of IS*1*, IS2, and IS5 is a defined 108-bp region upstream of the transcription start site, and point mutations in the *celD* locus (Parker and Hall 1990b). The nature of the activating mutation has been found to be dependent on the strain background from which it is isolated. Two natural isolates of *E. coli* which were Cel$^+$ were found to have been activated by an IS element and a point mutation. This showed that activation is not a laboratory artifact and both types of activation occur in nature. All the IS*1* insertions were observed to occur at an invariant locus between bp -71 and -72 though the position of insertion IS2 was variant. An IS2 insertion was detected at a downstream site. Activation mediated by point mutations however did not map in the *celR* region but were localized in the *celD* gene. Substitution of a lysine by threonine or glutamic acid resulted in activation. There was no such change in the IS-mediated activating mutants.

The wild-type *cel* operon in a multicopy plasmid showed a basal level of β-glucosidase activity and was induced three-fold in the presence of cellobiose. However, in the activated strain carrying the point mutation, the levels of induction ranged from 20-fold to 80-fold. Depending on the type of point mutation, the levels of induction varied. Wild-type *celD* on a multicopy plasmid could repress expression from a strain carrying an activating point mutation essentially to basal levels. This indicates that the wild-type repressor is dominant over the mutant, suggesting that it cannot interact with the inducer and therefore cannot be induced. The activating mutations enable recognization of the inducer by the repressor.

Mapping of the site of initiation indicated that integration of the IS element does not provide promoter elements. Irrespective of the site of insertion, transcription always initiated as the same site. Since the *cel* operon is regulated by the CelD repressor, the IS element could act by disrupting an operator site. However, in an IS-activated mutant, the *celD* repressor gene on a multicopy plasmid causes repression, showing that the IS insertion does not completely disrupt the binding site for the repressor. Since the *celD* repressor is itself a part of the operon, the system would autoregulate to maintain the cryptic state.

In the natural environment, the *cel* operon may not be cryptic but may be expressed in response to β-glucosides under certain conditions. This argument is substantiated by the following observation (Droffner and Yamamoto 1992). It has been demonstrated that when mesophilic *E. coli* strains were grown at 37°C on 1%-cellobiose-supplemented MacConkey agar, the colonies were colourless, implying that cellobiose was not catabolized. These colonies turned bright pink when shifted to 48°C for 3 to 4 hours. However, if rifampicin (200 μg/ml) was sprayed on these colonies before shifting to 48°C, the colonies remained colourless. Thus it appeared that mRNA synthesis was a prerequisite for expression of the *cel* operon at elevated temperatures. Expression of the *cel* operon in *E. coli* at the growth-refractory temperature was demonstrated by phospho-β-glucosidase activity. This enzyme activity was also detected at the growth-refractory temperature in *Salmonella typhimurium* and *Pseudomonas aeruginosa*. Thermotolerant and mesothermophilic mutants of *E. coli, S. typhimurium* and *P. aeruginosa* that are able to grow with generation times of 30 to 40 min at 48°C and 54°C exhibited phospho-β-glucosidase activity at growth temperatures of 48°C and 54°C. Thus the *cel* operon which has been categorized as a cryptic operon in *E. coli* and *S. typhimurium* is expressed at growth-refractory temperatures of the mesophilic parent and growth-permissive temperatures (48°C and 54°C) of the thermotolerant and mesothermophilic mutants. This shows that the cryptic state of an operon may actually depend on the environmental condition. More evidence comes from the observation that in a mixed-resource environment, when both cellobiose and glycerol were present, the cryptic *cel* operon was strongly favoured over active *cel* alleles, but in a single-resource environment where only cellobiose was present the active alleles were strongly favoured (Hall *et al.* 1986).

5. The *asc* operon

In a survey of natural isolates of *E. coli.* obtained from farm animals and African yellow baboons, it was found that though a majority of the isolates could not utilize any of the β-glucosides, a few members showed a positive phenotype (Hall and Faunce 1987). Molecular analysis revealed that most of the positive strains did not express RNA transcripts homologous to either the *bgl* or the *cel* operon. ECOR collection strains which could mutate spontaneously to Cel^+ or Bgl^+ phenotype did not express RNA homologous to the known *cel* and *bgl* genes (Hall and Betts 1987). This indicated the presence of another cryptic system for the utilization of these-β-glucosides. This locus was designated as the *sac* locus for salicin, arbutin and cellobiose utilization (Parker and Hall 1988). However, to avoid confusion between the *sac* gene involved in sucrose utilization in *B. subtilis* this system has been renamed as the *asc* operon (Hall and Xu 1992). The *asc* operon specifies a transport system and a hydrolase that acts on salicin and arbutin and to a lesser extent on cellobiose. It is inducible by arbutin and a secondary mutation permits some induction by salicin. Induction appears to be inhibited by cellobiose. The *asc* operon on a high-copy-number plasmid permits utilization of cellobiose. Wild-type strains exhibit a low level of syntesis of a β-glucosidase even when grown on glucose. This level of expression however does not support growth on cellobiose but suggests that, unlike the *bgl* and the *cel* operons, the *asc* operon is not completely silent in the wild-type state.

A major distinguishing feature of this operon from the other β-glucoside-utilizing operons is that while only a single mutation is required to activate the latter, three sequential mutations are required before the *asc* locus permits effective use of arbutin, salicin and cellobiose. The wild-type strain can mutate first to Arb^+, followed by Sal^+ and then Cel^+. The Cel^+ phenotype is apparent only on plates. Cells fail to grow in a liquid culture where cellobiose is supplemented as the sole carbon source.

The *asc* operon maps at 58·3 min on the *E. coli* genetic map and 2845 kb in the Kohara physical map. It encodes three genes *ascG, ascF* and *ascB*. The AscG protein has been identified as a repressor based on its similarity to the GalR repressor and the presence of a DNA-binding helix-turn-helix motif. Activation of the operon is mediated by disruption of *ascG* by IS*186* which results in semiconstitutive expression of the *asc* operon. AscF resembles the PTS enzymeII for β-glucosides, such as BglF. Strains carrying *ascF* on a plasmid are Arb^+ but remain Cel^- and Sal^-. When *ascF* is present in a high-copy-number plasmid, growth of cells is inhibited by salicin or cellobiose in minimal medium. This is correlated with accumulation of the phosphorylated form of sugars. AscB is the phospho-β-glucosidase which can cleave salicin, cellobiose and arbutin. AscB shows 70·5% sequence similarity to BglB.

The simplest model to explain the normally cryptic state of the *asc* operon is that the *ascG*-encoded repressor is insensitive to β-glucoside inducer and so the operon is expressed only after inactivation of the repressor by IS insertion (Hall and Xu 1992). It is surprising that activating mutations do not include missense, nonsense or frameshift mutations. The high specificity of the activating IS*186* insertion is also puzzling. The time taken for papillation to occur due to mutations in the *asc* locus is much longer (4–5 weeks) compared to other operons. Another puzzling observation was that though all the activating mutations were mediated by IS insertion, the stabilities of all the mutants were not similar.

The mode of activation of the *asc* operon by IS*186* within the 3′ region of the repressor gene *ascG* is similar to the mode of activation of the *cel* operon. However, in the *cel* operon the repressor gene is inactivated by point mutations in addition to insertions.

A comparative study of the mode of activation in all the β-glucoside utilization systems shows the following features:

(i) Activation involves the participation of IS elements in all three cases.

(ii) Inactivation of a repressor is also a common feature in the three systems. The only difference is that in the case of the *cel* and the *asc* operons, the repressor is specific to the operon, whereas in the case of the *bgl* operon, the repressor involved is the global repressor H-NS.

It is interesting to note that disruption of repressor interaction only serves to activate the operon. Following this, optimal expression of the operon requires the presence of an inducer and involves additional regulatory steps.

6. β-glucoside utilization in natural isolates

A screening of the ECOR collection of natural. *E. coli* isolates was undertaken to determine the proportion of strains that carried functional, cryptic and nonfunctional genes for utilization of the three β-glucoside sugars, arbutin, salicin and cellobiose (Hall and Betts 1987). None of the 71 natural isolates utilized any of the β-glucosides. When

the strains were subjected to selection for utilization of the sugars only five of the isolates were incapable of yielding spontaneous β-glucoside-utilizing mutants, indicating that they had completely lost the genes for β-glucoside utilization. Fortyfive strains yielded cellobiose$^+$ mutants, 61 yielded arbutin$^+$ mutants, and 58 strains yielded salicin$^+$ mutants. To determine whether the Cel$^+$ and the Bgl$^+$ phenotypes were specific to either *cel* or the *bgl* β-glucoside utilization operons in *E. coli* K-12, a subset of the mutants were screened by mRNA hybridization. Two cellobiose$^+$ and two arbutin$^+$-salicin$^+$ strains failed to express either of these known operons. This was later shown to be the *asc* operon of *E. coli.*

In addition to the observation that the frequency of the Cel$^+$ phenotype is very rare in natural isolates, a more surprising finding was that there seems to be a selection for deletions in the *bgl* operon in the Cel$^+$ isolates of *E. coli* since 95% of the cellobiose-utilizing strains had deletions of the *bgl* operon (Hall 1988). There were three different types of deletions in the *E. coli* population. In 70% of the cases the deleted *bgl* operon carried a replacement of DNA in the region normally occupied by *bgl.* The deletion spanned a region of 8 kb around the *bgl* region and was replaced by a 6·5 kb fragment. In another type, there was a deletion of similar length but it was not replaced by any DNA fragment. A third type was deletion of a smaller segment of 4 kb without replacement by any fragment. A complementary study to see whether the presence of active *bgl* operon results in deletions in the *cel* operon has not been carried out. The physiological significance of deletion of the *bgl* operon in the presence of an activated *cel* operon is not clear. There is no obvious basis for this selection since the *bgl* operon is silent in most wild-type *E. coli* and may be presumed to be silent in most Cel$^+$ *E. coli* also. If the active state of the *bgl* operon were to interfere with the functioning of the genes for cellobiose catabolism, then the presence of the *bgl* operon would constitute a genetic burden on Cel$^+$ strains. However, there is no evidence at present that the *bgl* genes interfere with the functioning of the *cel* operon genes.

7. Conclusions

The results reviewed above underscore the fact that cryptic genes are distinct from pseudogenes and are evolutionary puzzles in the true sense. Their activation in most cases requires a single genetic event. In a selective environment, the active form prevails in the population. This mode of their regulation is distinct from the physiological regulation seen in genes that are present in the active form.

Based on the observations on different cryptic genetic systems cited above, their presence in microorganisms can be rationalized in two different ways:

(i) The cryptic state is a transient state and oscillations between the cryptic and active states are directed by changing environments which impose different fitness values to the two states.

(ii) Cyptic genes are silent only under the conditions in which they are observed. They may be functional even in the absence of an activating mutation under specific environmental conditions that are not understood at present.

The features of the different cryptic systems examined above suggest that both of these possibilities may be true under different conditions.

Laboratory studies using the *bgl* operon of *E. coli* show that the cyclic process of activation and silencing is unlikely since revertants of active strains are unable to be

activated again as they carry large deletions. Combined with the possibility that the promoter can be activated under specific environmental conditions, these studies suggest that the operon may be expressed periodically even in the absence of mutational activation. This is true of the *cel* operon also. The observation that the sequence diversity seen in 12 naturally occurring alleles of the cryptic *celC* gene encoding the PTS enzIII^{cellobiose} is the same as that for alleles of of the functional *gut G* gene encoding a similar protein involved in glucitol transport (Hall and Xu 1992) suggests that cryptic genes are also subject to selection to the same degree as functional genes. Therefore, irrespective of whether they are cyclically activated or exressed under specific environmental conditions, it appears that cryptic genes confer selective advantage to the organisms that carry them. This is likely to be the major factor determining their continued presence in the genome even in the face of evolution.

Acknowledgements

We thank one of the referees who made several helpful suggestions for improving the manuscript. Work in the author's laboratory on the *bgl* operon was supported by a grant to S.M. from the Department of Biotechnology.

References

Amster-Choder O. and Wright A, 1992 Modulation of the dimerization of transcriptional antiterminator protein by phosphorylation. *Science* 257: 1395–1398

Amster-Choder O., Houman F. and Wright A. 1989 Protein phosphorylation regulates transcription of the β-glucoside utilisation operon in *E. coli*. *Cell* 58: 847–855

Bachmann B. J. 1990 Linkage map of *Escherichia coli* K-12, edition 8. *Microbiol. Rev.* 54: 130–197

Dawkins R. 1976 *The selfish gene* (Oxford: Oxford University Press)

De Felice M., Levinthal M., Iccarino M. and Guardiola A. 1979 Growth inhibition as a consequence of antagonism between related amino acids: effect of valine in *Escherichia coli*. *Microbiol. Rev.* 43: 42–58

Defez R. and DeFelice M. 1981 Cryptic gene for β-glucoside metabolism in *Escherichia coli* K-12: genetic evidence for a regulatory protein. *Genetics* 97: 11–25

DiNardo S., Voelkel K. A., Sternglanz R., Reynolds A. E. and Wright A. 1982 *Escherichia coli* DNA, topoisomerase I mutants have compensatory mutations in DNA gyrase genes. *Cell* 31: 43–51

Doolittle W. F. and Sapienza C. 1980 Selfish genes, the phenotype paradigm and genome evolution. *Nature* 284: 601–603

Droffner M. L. and Yamamoto N. 1992 Demonstration of *cel* operon expression of *Escherichia coli*, *Salmonella typhimurium* and *Pseudomonas aeruginosa* at elevated temperatures refractory to their growth. *Appl. Environ. Microbiol.* 58: 1784–1785

Dykhuizen D. 1978 Selection for tryptophan auxotrophy of *Escherichia coli* in glucose limited chemostat as a test for the energy conservation hypothesis. *Evolution* 32: 125–150

Fox C. F. and Wilson G. 1968 The role of phosphoenolpyruvate-dependent kinase system in β-glucoside catabolism in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* 59: 988–994

Giel M., Desnoyer M. and Lopilato J. 1996 A mutation in a new gene *bglJ* activates the *bgl* operon in *Escherichia coli* K-12. *Genetics* 43: 627–635

Hall B. G. 1982 A chromosomal mutation for citrate utilisation by *Escherichia coli* K-12. *J. Bacteriol.* 152: 269–273

Hall B. G. 1988 Widespread distribution of deletions of the *bgl* operon in natural isolates of *Escherichia coli*. *Mol. Biol. Evol.* 5: 456–467

Hall B. G. and Betts P. W. 1987 Cryptic genes for cellobiose utilisation in natural isolates of *Escherichia coli*. *Genetics* 115: 431–439

Hall B. G. and Faunce W. III 1987 Functional genes for cellobiose utilisation in natural isolates of *Escherichia coli*. *J. Bacteriol.* 169: 2713–2717

Hall B. G. and Xu L. 1992 Nucleotide sequence, function, activation and evolution of the cryptic *asc* operon of *Escherichia coli* K-12, *Mol. Biol. Evol.* 9: 688–706

Hall B. G., Yokoyama S. and Calhoun D. H. 1983 Role of cryptic genes in microbial evolution. *Mol. Biol. Evol.* 1: 109–124

Hall B. G., Betts P. W. and Kricker M. 1986 Maintenance of the cellobiose utilisation genes of *Escherichia coli* in a cryptic state. *Mol. Biol. Evol.* 3: 389–402

Hall B. G., Xu. L. and Ochman H. 1991 Physical map location of the *asc* (previously *sac*) operon of *Escherichia coli* K-12. *J. Bacteriol.* 173: 5250

Higgins C. F., Dorman C. J., Stirling D. A., Waddel L., Broth, I. R., May G. and Bremer E. 1988 Physiological role of DNA supercoiling in the osmotic regulation of gene expression in *S. typhimurium* and *E. coli. Cell* 52: 569–584

Houman F., Diaz-Torres M. R. and Wright A. 1990 Transcriptional antitermination in the *bgl* operon of *E. coli* is modulated by a specific RNA binding protein. *Cell* 62: 1153–1163

Koch A. L. 1993 Genetic response of microbes to extreme challenges. *J. Theor. Biol.* 160: 1–21

Kricker M. and Hall B. G. 1984 Directed evolution of cellobiose utilization in *Escherichia coli* K-12. *Mol. Biol. Evol.* 1: 171–182

Kricker M. and Hall B. G. 1987 Biochemical genetics of the cryptic gene system for cellobiose utilization in *Escherichia coli* K-12, *Genetics* 115: 419–429

Lauer J., Shen C. K. J. and Maniatis T. 1980 The chromosomal arrangement of human α-like globin genes: structure homology and α-globin gene deletions. *Cell* 20: 119–130

Lawther R. P., Calhoun D. H., Adams C. W., Hauser C. A., Gray J. and Hatfield W. 1981 Molecular basis of valine resistance in *Escherichia coli* K-12. *Proc. Natl. Acad. Sci. USA* 78: 922–925

Lawther R. P., Calhoun C. W., Hauser D. H., Adams C. A., Gray J. and Hatfield W. 1982 DNA fine structure analyses of *ilvG*(*ilvG*$^+$) mutation of *Escherichia coli* K-12. *J. Bacteriol.* 149: 294–298

Leavitt R. I. and Umbarger H. E. 1962 Isoleucine and valine metabolism in *Escherichia coli* XI: Valine inhibition of the growth of *Escherichia coli* strain K-12. *J. Bacteriol.* 83: 624–630

Li W.-H. 1984 Retention of cryptic genes in microbial population. *Mol. Biol. Evol.* 1: 213–219

Li W.-H., Gojobori T. and Nei M. 1981 Pseudogenes as a paradigm of neutral evolution. *Nature* 292: 237–239

Lopilato J. and Wright A. 1990 Mechanisms of activation of the cryptic *bgl* operon of *E. coli* K-12. In *The bacterial chromosome* (eds.) K. Drlica and M. Riley (Washington, DC: American Society for Microbiology) pp. 435–444

Mahadevan S. 1997 The BglG class of antiterminators: a growing family of bacterial regulators. *J. Biosciences.* (in press)

Mahadevan S. and Wright A. 1987 A bacterial gene involved in transcription antitermination: regulation at a rho-independent terminator in the *bgl* operon of *E. Coli. Cell* 50, 485–494

Mahadevan S., Reynolds A. E. and Wright A. 1987 Positive and negative regulation of the *bgl* operon in *Escherichia coli. J. Bacteriol.* 169: 2570–2578

Morishita T., Fukada T., Shirota M. and Yura T. 1974 Genetic basis of nutritional requirements in *Lactobacillus casei. J. Bacteriol.* 120: 1078–1084

Mukerji M. and Mahadevan S. 1997 Characterisation of the negative elements involved in silencing the *bgl* operon of *E. coli*: possible roles for DNA supercoiling, H-NS, and CRP-cAMP in regulation. *Mol. Microbiol.* 24: 617–627

Nishioka Y., Leder A. and Leder P. 1980 Unusual α-globin like gene that has clearly lost both globin intervening sequences. *Proc. Natl. Acad. Sci. USA* 77: 2806–2809

Orgel L. E. and Crick F. H. C. 1980 Selfish DNA—the ultimate parasite, *Nature* 284: 604–607

Paquin C. E. and Williamson V. M. 1986 Ty insertions at two loci account for most of the spontaneous antimycin A resistance mutations during growth at 15°C of *Saccharomyces cerevisiae* strains lacking ADH1. *Mol. Cell. Biol.* 6: 70–79

Parker L. L. and Hall B. G. 1988 A fourth *Escherichia coli* gene system with the potential to evolve β-glucoside utilisation. *Genetics* 119: 485–490

Parker L. L. and Hall B. G. 1990a Characterisation and nucleotide sequence of the cryptic *cel* operon of *Escherichia coli* K-12. *Genetics* 124: 455–471

Parker L. L. and Hall B. G. 1990b Mechanism of activatrion of the cryptic *cel* operon of *Escherichia coli* K-12. *Genetics* 124: 473–482

Prasad I. and Schaefler S. 1974 Regulation of the β-glucoside system in *Escherichia coli* K-12. *J. Bacteriol.* 120: 638–650

Prasad I., Young B. and Schaefler S. 1973 Genetic determination of the constitutive biosynthesis of phospho-β-glucosidase A in *Escherichia coli* K-12. *J. Bacteriol.* 114: 909–915

Reizer J., Reizer A. and Saier M. H. Jr 1990 The cellobiose permease of *Escherichia coli* consists of three proteins and is homologous to the lactose permease of *Staphylococcus aureus. Res. Microbiol.* 141: 1061–1067

Reynolds A. E., Felton J. and Wright A. 1981 Insertion of DNA activates the cryptic *bgl* operon of *E. coli* *Nature* 203: 625–629

Reynolds A. E., Mahadevan S., LeGrice S. F. J. and Wright A. 1986 Enhancement of bacterial gene expression by insertion elements or by mutation in a CAP-cAMP binding site. *J. Mol. Biol.* 191: 85–95

Rudd K. E., Miller W., Ostell J. and Benson D. A. 1990 Alignment of *Escherichia coli* K-12 DNA sequences to genomic restriction map *Nucl. Acids. Res.* 18: 313–321

Rutberg R. 1997 Antitermination of transcription of catabolic operons. *Mol. Microbiol.* 23: 413–421

Schaefler S. 1967 Inducible system for the utilisation of β-glucosides in *Escherichia coli* Active transport and utilisation of β-glucosides, *J. Bacteriol.* 93: 254–263

Schaefler S. and Malamy A. 1969 Taxonomic investigation on expressed and cryptic phospho-β-glucosidases in *Enterobacteriaceae. J. Bacteriol.* 99: 422–433

Schaefler S. and Mintzer L. 1959 Acquisition of lactose fermenting properties by *Salmonella.* I. Interrelationship between the fermentation of cellobiose and lactose. *J. Bacteriol.* 78: 159–163

Schaefler S. and Schenkein I. 1968 β-glucoside permeases and phospho-β-glucosidases in *Aerobacter aerogenes*: relationship with cryptic phospho-β-glucosidases in Enterobacteriaceae. *Proc. Natl. Acad. Sci. USA* 59: 285–292

Schnetz K. 1995 Silencing of *Escherichia coli bgl* promoter by flanking sequence elements. *EMBO J.* 14: 2545–2550

Schnetz K. and Rak B. 1988 Regulation of the *bgl* operon of *Escherichia coli* by transcription antitermination *EMBO J.* 7: 3271–3277

Schnetz K. and Rak B. 1990 β-glucoside permease represses the *bgl* operon of *Escherichia coli* by phosphorylation of the antiterminator protein and also interacts with glocuse-specific enzyme III, the key element in catabolite control. *Proc. Natl. Acad. Sci. USA* 87: 5074–5078

Schnetz K. and Rak B. 1992 A mobile enhancer of transcription in *Escherichia coli. Proc. Natl. Acad. Sci. USA* 89: 1244–1248

Schnetz K., Toloczyki L. and Rak B. 1987 β-glucoside (*bgl*) operon of *Escherichia coli* K-12: Nucleotide sequence, genetic organisation and possible evolutionary relationship to regulatory components of two *Bacillus subtilis* genes. *J. Bacteriol.* 169: 2579–2590

Singh J., Mukerji M. and Mahadevan S. 1995 Transcriptional activation of the *Escherichia bgl* operon: negative regulation by DNA structural elements near the promoter, *Mol. Microbiol.* 17: 1085–1092

Umbarger H. E. 1978 Amino acid biosynthesis and its regulation. *Annu. Rev. Biochem.* 47: 533–606

Zamenhoff S. and Eichorn H. H. 1967 Studies of microbial evolution through loss of biosynthetic functions: establishment of defective mutants. *Nature* 216: 455–458