Functional genomics, analysis of adaptation in and applications of models to the metabolism of engineered *Escherichia coli*

by

William A Bryant

UCL

PhD in Biochemical Engineering

I, William A Bryant confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

In order to examine the metabolism of bacteria in the genus *Enterobacteriaceae* tools for gene complement comparison and stoichiometric model building have been developed to take advantage of both the number of complete bacterial genome sequences currently available and the relationship between genes and metabolism.

A functional genomic approach to improving knowledge of the metabolism of *Escherichia coli* CFT073 (a uropathogen) has been undertaken taking into account not only its genome sequence, but its close relationship to *E. coli* MG1655. A fresh comparison of *E. coli* CFT073 has been done with *E. coli* MG1655 to identify all those genes in CFT073 that are not present in MG1655 and may have metabolic characteristics. These genes have further been bioinformatically assessed to determine whether they might encode enzymes for the metabolism of chemicals commonly found in human urine, and one set of such genes has been experimentally confirmed to encode an L-sorbose utilisation pathway.

Little experimental work has been done as yet to elucidate how bacteria adaptively respond to the introduction of heterologous metabolic genes. To investigate how bacteria respond to such DNA, genes encoding the L-sorbose utilisation and uptake operon from CFT073 have been cloned and transformed into DH5α and a selective pressure (minimal medium with L-sorbose as sole carbon source) has been applied over 100 generations of growth of this strain in serial passage to investigate the change in its behaviour.

The availability of large numbers of completely sequenced genomes, along with the development of a stoichiometric metabolic model with very high coverage of *E. coli* metabolism (iAF1260 [1]) have made possible the analysis of the core metabolism of large numbers of bacteria to investigate gene essentiality in these bacteria. A novel way of assessing gene complement has been developed using BLAST and DiagHunter to improve reliability of gene synteny comparisons with contextual information about the genes and to extend work by others to cover all *E. coli* and *Shigella* genome sequences with available sequences on GanBank (as of 1st June 2009) in order to bioinformatically investigate essential genes in these bacteria and the heterogeneity of their metabolic networks. Further to this a metabolic model has been constructed for DH5α with an added L-sorbose pathway and for CFT073 and these models have been used to investigate behavioural changes during adaptation of bacteria to novel heterologous genes.

Contents

The Supplementary Tables and Files are available on the CD located inside the back cover of this Thesis.

Chapter 1

Introduction

1.1 Bacterial metabolism and its genomic basis

All bacteria must eat to survive. They must take in an array of relevant molecules then through the controlled breakdown and use of these molecules they must maintain their structural integrity, respond to changes in the environment, grow and ultimately reproduce. At the same time, they must keep out a whole tranche of deleterious molecules, and delicately control the uptake and rate of metabolism of those molecules that are to be used because too much of almost any molecule is dangerous. Further when competing with other bacteria, even sister cells, efficiency is the key to outcompeting and ultimately dominating in the evolutionary arms race. This competition in metabolic terms consists of optimising the enzymes at the bacterium's disposal to increase activity but also, and perhaps more importantly when many enzymes are well optimised already, of tuning the regulatory response of the bacterium such that in any particular environment or conditions only relevant enzymes are produced, in the right amounts and in the correct proportions, so that metabolic capacity is not wasted in making too much of one product or not enough of another.

Ultimately through a delicate balance of levels of all enzymes in the bacterium efficient use can be made of the molecules both outside and inside the cell. Transport across the cell boundary and the intricate web of interdependent reactions that carry on inside the cell combine to produce a well ordered and ultimately successful living organism. The gene centred model of bacterial metabolic function began in the Nineteenth Century with Mendel's experiments on heredity providing the first evidence of links between ancestry and what is now called phenotype [2]. Louis Pasteur's work on fermentation of sugar to ethanol by yeast cells was seminal in the association of unicellular organisms and chemical processes. A further step was required, however, to link enzymatic metabolism with inheritance, and this was taken by Beadle and Tatum [3] when their observations that a single mutation in a gene of *Neurospora crassa*'s seemed to produce dependence of that organism on vitamin B6. So it was established that in the majority of cases a single gene accounts for a single enzymatic function so in effect each gene (or set of genes in the case of multi-protein enzymes) corresponds to the ability of an organism to drive a particular chemical reaction or transport process within the cells of the organism. Thus the problem of understanding how bacteria adapt to their environment, carry out their day-to-day maintenance and reproduce has become one in a great part of genetics and, in the post-Genomic era, of genomics. Through the one-to-one association between genes and enzymatic proteins the intimate link between genomes and metabolic function is established.

There is more information about metabolism in the genome than just the gene-enzyme correspondence. Genes are grouped in co-expressed groups, operons, which often indicate an association of function. Operons are controlled by various mechanisms related to the specific DNA sequence directly upstream of the first gene in the operon, which regulate the rate of production of each enzyme and indeed whether an enzyme is expressed at all. For instance possibly the best studied and certainly the most famous regulatory mechanisms are those of the *lac* operon, studied by Monod [4], which repress the production of the lactose transport and metabolism genes in the presence of glucose or in the absence of lactose, and up-regulate the expression of these genes when glucose is lacking and lactose is present. Indeed regulatory mechanisms extend beyond control of single operons to global regulation of gene expression and they co-ordinate regulation of large sets of genes by specific transcription factors (TFs) which are themselves proteins encoded by genes and under the control of further TFs and of other stimuli, for instance presence or absence of specific molecules, as with the *lac* operon.

This tangled web of interactions between proteins and chemicals which extends to thousands of specified components and permeates all of life controls every aspect of cells' lives; response to the environment, growth and reproduction all require high precision timing and control of thousands of biological actions which are achieved through these complex networks of interacting regulators and enzymatic proteins. To fully understand the metabolic network of a bacterium and ultimately to adapt it to practical ends, for instance the production of useful chemicals (e.g. [5]), in pharmaceuticals (e.g. [6]), or the clean-up of toxic waste, the components of this network must be identified, characterised and eventually understood in the context of the whole cell. One great step towards this goal is the sequencing of whole bacterial genomes, which gives the composition of all proteins involved in the life of a bacterium and, when it can be deciphered, the details of many of the mechanisms of control of expression of these proteins.

1.2 Genomic Sequencing and Comparisons

1.2.1 Genome Sequencing Techniques

The bacterial genome, if correctly interpreted, is a complete list of tools for a bacterium. It is therefore a vital step in the understanding of bacteria to sequence their genomes. The double helical structure [7], linear nature and small alphabet coding system of DNA [8] make it convenient to represent in a linear way using four characters: C, A, T and G; efficient genome sequencing also relies on these characteristics. The first feasible systems for sequencing large amounts of DNA were developed by Sanger *et al* in Cambridge, UK, in the mid-1970s ([9, 10]), at the same time as Maxam and Gilbert developed their own method in at Harvard, Cambridge, Massachusetts [11].

Sanger's dideoxy method was widely adopted over Maxam and Gilbert's chemical sequencing method and was used, once full-scale automation of this genome sequencing technique had been developed (for instance by Leroy Hood's laboratory in Caltech [12]), for the sequencing of increasing amounts of DNA. In the mid 1990s The Institute for Genomic Research (TIGR) was set up at NIH by Venter with a large number of automated sequencing machines from ABI producing one of the first large-scale sequencing centres [13]. It was here that the first complete bacterial genome sequence, that of *Haemophylus influenza* [14], was completed. Further development of automation based on the dideoxy technique involved the replacement of a slab gel for running samples with more efficient capillary electrophoresis, pioneered by ABI in the form of the ABI Prism 300 automated sequencer in 1996. At this point in the development of DNA sequencing the idea of sequencing the human genome became feasible and this was pursued both publicly and privately culminating in two draft human genomes published in 2001 [15, 16].

Capillary array electrophoresis (CAE), where many capillaries in parallel are used for sequencing, is currently the standard technique for DNA sequencing. However, in the past few years several commercial systems have been developed that do not rely on the dideoxy method for sequencing, but instead use novel techniques which sacrifice length of individual reads of DNA for very high numbers of DNA fragments read. The machines of 454 Life Sciences [17] use pyrosequencing [18] to produce sequences of length approximately 250 bp, those of Solexa use a single molecule sequencing (SMS) technique to produce reads. CAE currently routinely produces read lengths of approximately 1000 bp [19] in comparison to these.

New sequencing techniques are being developed at a great pace (reviewed by Chan [20]) and the ones mentioned above are still in their infancy, so still being improved very rapidly. Developments of several different approaches to SMS are currently being researched and as this type of technique, using a single DNA molecule, solves several of the problems associated with traditional sequencing techniques, such as PCR bias, and has the potential to give long reads at high precision quickly (for a review of SMS technology and research see Gupta [21]).

1.2.2 BLAST

The Basic Local Alignment Search Tool (BLAST) [22, 23] was developed in response to the increasing amount of sequence data produced by sequencing projects. Gene sequence data were proliferating through the 1990s and although databases of sequences were set up the full potential of genome, gene and genetic product sequences was not yet realisable. What was required was a system for assessing relationships between sequences, specifically a way to determine common ancestry. Since all life is related, similarities in genes and genomes range across the whole spectrum from single celled bacteria to trees to humans, from hyperthermophilic organisms clustered around deep sea thermal vents to polar bears. Quantification of relationships between genes and determination of function through comparisons with already functionally characterised genes required a tool for comparisons and BLAST was that tool.

BLAST uses variations on a single algorithm for comparing different types of sequences: such as protein sequences against other protein sequences or DNA sequences against DNA sequences. Since genome sequence comparisons are of interest the DNA versus DNA algorithm will be described. For similarity measures there is a simple matrix of scores for base pair identities and differences between the query and reference sequences. Since identities score positively and differences score negatively, there will be two lengths of DNA (not necessarily the full length of either of the sequences being tested), one in each tested sequence, that give the highest score when compared with each other. This is called the maximal segment pair (MSP) and the score derived from this MSP is then used as an indicator of the extent of similarity between the two sequences. This similarity is quantitative and can be used to find closely related sequences, and lower scoring pairs of sequences that may be bear random similarities even though they are unrelated or be distantly related.

There are various techniques used by the BLAST algorithm which compromise between accuracy and time taken for a comparison to be done, but generally these have a very low impact on the reliability of the results produced [22]. The limitations in terms of gaps in alignments, where perhaps a pair of sequences had more than a single region of high similarity, separated by an unrelated or poorly conserved region of DNA, was eliminated by the introduction of gapped BLAST in 1997 [23] and this allowed the application of BLAST to ever more sequences in various circumstances. There are very many articles which rely on BLAST results since its creation and for genome sequence comparisons (discussed below) all rely on BLAST for their comparative power. Not only can BLAST be used directly between two sequences of interest, but it can be used to survey the entirety of any sequence database to find homologs and consequently infer gene function. The databases used most frequently are accessed through the National Center for Biotechnology Information (NCBI) Entrez system [24] which incorporates the GenBank [25] database of nucleotide sequences for more than 300,000 organisms, including 944 complete published genome sequences according to the Genomes Online Database (GOLD) [26] as well as many complete genome sequences without an associated publication.

BLAST is therefore useful in two situations in analysing complete genome sequences: firstly to identify putative (and perhaps even specific) functions for genes and secondly to survey differences between complete genome sequences. There are however limitations to BLAST, of course it cannot predict the function of a gene where there are no homologous genes in the interrogated database or where none of the homologs have an experimentally confirmed function. Also, even when genes have almost identical sequences this does not guarantee that the genes perform the same function. It is possible that a single point mutation might completely inactivate the product of a mutated gene and has been shown to occur in some cases (for instance in [27]). This does not invalidate the BLAST approach to identifying gene function through sequence comparison, since in the vast majority of cases point mutation does not significantly affect gene function. If a gene is inactivated in a viable organism and does not affect the organism's viability the gene no longer has any evolutionary pressure on it and will quickly accumulate other mutations (on an adaptive timescale), moving the gene sequences away from homology. Thus although care is needed when inferring conservation of function, BLAST is an excellent tool in most cases.

According to its NCBI protein table a recently sequenced *E. coli* genome (that of *E. coli* O127 H6 [28]) contains 843 labelled hypothetical proteins out of a total of 4554 protein coding genes in the genome. This means that the proteins are predicted to be expressible but there were no other sequences at the point of sequence analysis that could have been homologs and have the same function, that were functionally annotated in GenBank. This is a fairly typical number in whole genome sequences that roughly 20 % of the genes have unknown function. A notable exception to this state of affairs is *E. coli* K-12 strain MG1655 and very closely related strains: MG1655 has 26 proteins labelled as hypothetical, in large part due to its position as model commensal *E. coli* in laboratories around the world and the huge number of published articles on its properties. This does not mean that every other gene in MG1655 has a known function and others are labelled with such descriptions as 'inner membrane protein' and other general labels, but it indicates the current knowledge of *E. coli* as a whole.

When BLAST fails to find homologs for genes in current sequence databases there are several other programs to aid the search if not for a specific function then at least for a class description or other general indication of the coded protein's function. These are often also based on sequence similarities, but of translated nucleotide amino acid sequences, the inferred protein sequences of the uncharacterised genes. One example of this approach is the NCBI's Conserved Domains [29] which compares the inferred protein sequence of a gene with conserved protein domains in an NCBI domain database. These conserved domains are based on a pre-run multiple sequence alignment of the relevant part of various proteins that contain the same domain, coming from several different organisms, to give a more robust comparison where a consensus sequence derived from these previous comparisons shows which parts of the sequence are conserved most in these different contexts and therefore which parts of the sequence are most important for conserving function.

This sort of analysis can then potentially be used to assign putative functions to proteins of otherwise unknown function. Another technique for the elucidation of protein function is *de novo* protein chain folding models and 3-dimensional structure comparisons, often aided by genome sequences as in subgenome analysis of functionally and evolutionarily linked genes in a particular genome (a technique reviewed by Man *et al* [30]) although structure and function are related in a complex way and the difficulties of inferring function from structure are very similar to those of inferring function from homology, as discussed by Punta and Ofran [31]. In practice attempts to get better functional annotations of novel proteins combine several methods of analysis in to synthesise from these a more specific functional annotation. Where proteins are unlike any other thus far functionally characterised combinations of methods are also used to find clues about potential function [32, 33].

1.2.3 Genome Comparisons

The first complete DNA genome to be sequenced was that of bacteriophage Φ X 176 which contains 5386 bases [34]. The first organism to have its genome completely sequenced was *Haemophilus influenzae* [14], which is 1.8 Mb (megabases) in length. The first *E. coli* to be sequenced was strain K12 MG1655 [35], 4.6 Mb in length, and since then 45 more strains of bacteria in genus *Escherichia* have been fully sequenced according to the NCBI Taxonomy Browser `http://www.ncbi.nlm.nih.gov/Taxonomy/`. Table 1.1 (p. 42) shows most of these strains and their accession number or source if they have not yet been assigned an accession number. Where the genome sequences have associated publications these have been indicated.

It is also shown in Table 1.1 whether the genome sequence is in the form of a complete single contig, or still in multiple contigs from the shotgun sequencing, and not yet fully pieced together. This has consequences for the analysis of whole genomes so the feasibility of using these whole genome shotgun (WGS) sequences will be discussed here. The nature of the change of bacterial genomes is plastic, that is, they are quite capable of large-scale rearrangements (for instance the inversion of the segment of DNA between loci *rrnD* and *rrnE* in *E. coli* W3110 [58] relative to MG1655), so even large segments of sequenced DNA cannot necessarily be lined up and joined using another closely related strain as a template. However, this does not completely preclude these multi-contig shotgun sequences from use: that large scale rearrangements can be accommodated by bacteria is shown by W3110 and MG1655, with only very little phenotypic change, so in some applications it is irrelevant to bacterial function how large tracts of the genome sequence fit together, only which genes and operons are intact and functional.

The WGS sequences listed in Table 1.1 have varying numbers of contigs, but typically they have around 100. The coverage of these contigs seems good: lengths of shotgun sequences appear to be of a similar length to complete sequences, suggesting that there is not much left unsequenced. If it is assumed that almost all of the genome covered by the shotgun sequences, that all of it is made up of DNA related to one operon or another and that operons are randomly sized then approximately 100 will be incorrectly sequenced, since they will lie across the ends of the contigs. Indeed, due to the size of the smallest contigs (meaning that both their ends might well lie in the same operon) the actual number of incompletely sequenced operons may well be less than 100. This compares to 2670 operons overall in *Escherichia coli* K12 MG1655 according to RegulonDB [59], so under 5 % of operons are incompletely sequenced. So for the purposes of whole genome comparisons these WGS sequences can at least be considered for addition to any whole genome comparison, albeit with the proviso that sequences at the edges of the contigs should be treated carefully.

Comparisons of bacterial genomes have occurred more-or-less since the second bacterial genome was sequenced, that of *Mycoplasma genitalium* [60] in 1995. Many investigations now take subsets of genomes for comparison, but do it across all sequenced organisms, giving a short but wide set of data to analyse (for instance [61]). BLAST [23] is the workhorse of such comparisons, providing confidence levels that genes or operons are conserved between bacteria. This comparison of genome sequences is of vital importance when investigating how and why genome complements differ between bacteria, even very closely related ones, and eventually for elucidating some insight into how genomes change in time. The actual process of comparison can be as simple as single BLAST searches against whole genomes of interest to find homologs using an identity cutoff [62] or very intricate, using many computer programs to assess alignment and phylogenetic distance [61], though many systems, however complex, often still require visual inspection before final results are obtained. Human pattern recognition has yet to be rendered entirely redundant, but computers vastly increase the amount of data that can be processed and ultimately enable analyses of this sort.

Comparisons are not limited to inferring conserved functions in the newer bacterial genome, but also in studying evolution and evolutionary dynamics. They have been used in trying to find a minimal set of genes required for a self-sufficient organism [63], which is closely related to the search for essential genes, in inferring not only recent (100 million years) evolutionary dynamics [64] but also in gaining insights into the characteristics of the evolution of early life [65, 61] and even in supporting the claim that there was life supported by an RNA genome before DNA became the ubiquitous material for genomes as it is with all modern terrestrial life [66]. Evolution on the timescale of the existence of animals, particularly mammals, is of particular interest since it sheds light on how commensal and pathogenic bacteria that inhabit humanity have developed in interaction with their hosts who would in many respects have been quite similar to modern humans, and with other contemporaneous bacteria, and in this genome comparisons are invaluable. They shed light on those aspects of chromosome organisation and complement that are essential and those which are unimportant in coexistence with a host and with competing organisms, which particularly aids attempts to understand and combat bacterial pathogenic infection [67, 62].

Comparisons are often carried out over multiple genome sequences to build up patterns of similarities and differences. For instance the library of genome sequences of genus *Escherichia* has been used to investigate the acquisition of toxin genes across sequenced enterotoxigenic *E. coli* (ETEC) [68]. Various methods for making the most of bacterial genome sequence libraries have been developed, such as robust inference of transcriptional regulatory networks for newly sequenced bacteria [69] and discerning the dynamics of large-scale clustering of bacterial operons by sequence comparisons [70]. Whole genome sequence comparisons have also enabled the investigation of those aspects of organisms' lifecycles that are intrinsically related to the whole genome or large parts of it, such as the metabolic network, which typically in *E. coli* involves of the order of 1000 genes, and although mostly only a small subset of those will be expressed at a single time tight control of every single gene is essential for competitive advantage and even survival in certain conditions.

The challenge for Biologists when given all these new data for analysis is to extract as much information as is possible from these genome sequences with as little benchwork as possible (to save time, energy and money) and to allow experimental time to be used effectively and in a targeted manner. While undertaking this task there is a payoff to be had using these sequences between general results, rich in explanatory power but of limited immediate practical use (such as the investigation of bacterial operon clustering [70]) and specific results, of limited applicability but very useful in that small area of research (such as developing the knowledge of ETEC with the hopes of pointing the way to the development of more effective vaccines [68]). Research into bacterial metabolism, including genome-scale metabolic models (for instance iAF1260 [1]) and genome-scale transcriptional regulatory networks [69], tries to bridge the scale divides between general insight into how bacterial networks are built and specific questions about how bacteria respond to changes in environment, how they respond to natural genetic perturbations and increasingly how they respond to genetic engineering.

1.2.4 Successive inferred annotation

In essence the normal method of annotation is to find mutual best BLAST hits in the GenBank non-redundant database for each gene found in the new genome. The problem is that once this new gene has been annotated and submitted to GenBank as part of the newly sequenced genome it becomes a valid target for other BLAST runs for newer sequenced genes. So for example if gene A (the function of which has been experimentally verified) is a mutual best BLAST hit for gene B and gene B is a mutual best BLAST hit for gene C then gene C can be inferred to have the same function (or at least the same annotation) as gene A despite it not necessarily being close enough in sequence homology to gene A to merit this inference. One of the potential mechanisms of evolution is gene duplication and consequent genetic drift of one of the genes to produce novel function. Paralogs produced in this fashion will be particularly susceptible to the successive BLAST annotational inference error due to their identical genetic origin.

1.3 Bacterial metabolic models

The metabolic network of a bacterium can be modelled as a system of differential equations, each equation representing a reaction that is present in a particular system (though not necessarily with any flux through it), however a realistic model governing reaction rates is often intransigent to solution when there are a large number of simultaneous differential equations to solve, so often just the network of reactions and metabolites is used [1], or a small subset of metabolites and reactions (eg. limited to central metabolism) is used, but even then a simplified regulatory model is usually used [71, 72].

In the evolution of modern metabolic networks many factors will have impacted on how they developed, including their network nature. One model of this evolution is that it has occurred as a stepwise construction of the metabolism one reaction at a time - this model has led Handorf *et al* to conduct simulated network creation in this manner and to the conclusion that the expansion of metabolic networks is robust to small perturbations (eliminations of a few reactions), but some reactions when lost hamper the growth of these metabolic networks considerably [73]. A common attribute of all biological metabolic networks is their scale-free nature, that there is a power law distribution of connectivity for the metabolites [74], and although the reason for this is by no means settled it seems that it is a consequence of the way in which metabolic networks have been constructed and that even given very few assumptions about the mechanisms of evolution of these networks the log-normal dynamics of the growth of metabolic networks produce scale-free networks and metabolic networks that are not scale-free at any particular time tend towards this characteristic as they develop [75, 76].

The huge amount of data on metabolic pathways and genome sequences available have produced ever more complex and integrative tools for metabolic network reconstruction, such as the SEED tool curated by the Fellowship for the Interpretation of Genomes (FIG) [77], which breaks up the whole metabolic network into smaller component networks which are analysed against a database of components to get robust verification of these components independently, then reconnects the whole network of the given organism. Also, information about gene functions can often be extracted from metabolic gene annotations since these annotations include the name of the reaction catalysed by the enzyme coded for by the gene (for instance this is the basis of MetaCyc's PathoLogic algorithm for inferring metabolic networks [78]), but annotations do not always wholly define the gene function. It would be more convenient thereofore to infer a quantitative metabolic function from a manually created and curated model of a closely related organism, which could then be integrated with all other metabolic functions of all other coded proteins in a particular organism to infer a stoichiometrically complete metabolic model.

1.3.1 Whole genome metabolic models in metabolic engineering

Whole genome metabolic models are being increasingly used in metabolic engineering for assessing the impact of gene complement changes in bacteria and directing improvement by identifying gene deletions that could improve production of, say, recombinant protein by redirection of metabolic flux. For instance Oddone *et al* [79] have recently used such a model to qualitatively predict improved flux distribution in *L. lactis* engineered to overproduce a target protein. The construction of this model was done by initial inference from gene annotation data and manual assessment of each gene functional assignment [80].

Approaches like this have also been used to improve succinate production in *E. coli*, by modification of genes identified by flux analysis of a whole genome metabolic model [81]. These uses of metabolic models in engineering are becoming increasingly common, as they provide ways of directing genetic modification for improvement of production of recombinant proteins and small molecules.

1.3.2 Reconstructing metabolic networks

One resource for metabolic reactions that takes advantage of genome sequencing and annotation is MetaCyc's Pathway Tools software [82] which reconstructs metabolic networks and currently (as of June 2009) holds predicted gene complements of a large number of organisms, including 8 *E. coli* and 6 *Shigella* strains, including MG1655 and uropathogenic *E. coli* CFT073. The Pathologic algorithm for producing metabolic networks for MetaCyc relies on gene annotations from sequenced bacteria, requiring an initial annotation step before comparison. Further, steps towards fully automated reconstruction of metabolic networks have been taken using Genomatica's proprietary Simpheny$^{\text{TM}}$software, for instance to produce an inferred metabolic network for *Lactococcus lactis* IL1403 [83] using previously constructed networks from 3 other metabolic networks, including that of MG1655.

It has been noted by others that the annotation of newly sequenced genomes are often poor, despite the large number of bioinformatic tools available for the task [84]. Therefore in the production of new metabolic models it would be useful not to rely on such annotations. Further, there is often a lack of physicochemical characteristics of reactions (such as their reversibility in physiological conditions and the physical locations of the reactions in the cell) which must be found by analysis of individual reactions [85]. There have been metabolic model reconstructions that have collected such data for several reference organisms, which rely on biochemical characterisation (from literature) for their characteristics, thus represent repositories of knowledge specifically relevant to the metabolism of those reference organisms.

On such model is of *E. coli* MG1655, iAF1260 [1], which accounts for 1260 open reading frames (ORFs) of the 4244 genes in that organism. It is a constraint-based stoichiometric model which means that it comprises stoichiometrically balanced reactions with absolute flux constraints on each reaction, but is otherwise only constrained by reaction directions determined by thermodynamic considerations. It does not take into account any regulatory information about the bacterium, even though much is known about specific regulatory mechanisms (for instance as available from RegulonDB), since adding regulation to this size of model would make it computationally difficult to solve. Instead solutions produced by the linear optimisation of the model with an objective of maximising the so-called ‘Biomass’ reaction represent boundaries between physically possible and physically impossible flux distributions. The ‘Biomass’ reaction is balanced such that in the solution of flux values for each reaction, flux through the Biomass reaction is equal to the growth rate of the organism in those conditions. It is not assumed that bacteria in their natural habitat will achieve this rate of growth, indeed the idea that maximising growth rate or molar yield is a goal of bacteria has been shown not to be universal [86], though it has been shown that MG1655 approaches this rate when grown in conditions designed to select for high growth rates in a constant environment [87].

Importantly for reconstruction of other related metabolic networks, the model iAF1260 includes gene-protein-reaction relationships [1]. These relationships are very important in the inference of new metabolic models since they link genomic comparisons, in the form of BLAST hits between genes or some other bioinformatic method for gene inferences, with conserved metabolic reactions between MG1655 and the query bacterium. The stoichiometric nature of the model means that the inference of a new model is equivalent to deletions of the columns corresponding to unconserved reactions in a matrix of reactions available in the original model. It is immediately apparent that there might be limitations to the use of a single model for inferring other bacterial models - any reactions not present in the model strain must be added by some other method to a new model of a different bacterium.

There are currently many genome sequences of bacteria of genus *Escherichia* listed in Table 1.1, as well as those of *Shigella* bacteria, which are very closely related to genus *Escherichia*. With this large number of genome sequences it should be possible to reconstruct and compare many metabolic models, running them according to the conditions set out for iAF1260, that is, using the same model parameters and initial conditions to compare the models. The principles applied to comparing these closely related bacteria could also be used with more distantly related bacteria, and the limits of the applicability of one bacterial model to genomically disparate organisms could be tested.

1.4 Gene essentiality in *E. coli*

Gene essentiality knowledge is potentially a useful tool in assessing validity of gene function inferences and will be considered in Chapter 6. Several experimental attempts have been made to identify the essential genes in *E. coli*, that is, the genes which when disrupted in otherwise wild-type *E. coli* render the bacterium incapable of growth in rich media. The Keio collection [88] was constructed as a complete set of viable in-frame single gene-knockout mutants of *E. coli* K-12 BW25113. The process by which this was achieved produced in principle knockouts for every gene in that bacterium, but those which were not viable did not grow for their deletions to be confirmed. Therefore all of the genes for which a knockout could not be found are candidate essential genes.

The strain BW25113 is closely related to MG1655, having been derived from the same wild-type parent strain and its genotype with respect to the wild-type K-12 EMG2 is: *rrnB3* Δ*lacZ4787 hsdR514* Δ(*araBAD*)*567* Δ(*rhaBAD*)*568 rph-1*. The genotype of MG1655 is *rph-1*, so the differences only account for a few well-defined genes; in both strains the F plasmid and phage λ are absent. MG1655 gene equivalents in BW25113 for which no viable deletion mutants could be produced, 296 in total, are listed Supplementary Table 3. Four other genes inferred in this study have since had their entries discontinued in GenBank. Essentiality of genes inferred by Baba *et al* [88] was determined in LB medium. They also did a comparison of the number of essential genes in three other *E. coli* strains and found that 282 of these genes were also present (according to contemporary annotations) in these other three strains.

Gerdes *et al* [89] have made a survey of essential genes using a genetic fingerprinting technique (using transposon mutagenesis), revealing 620 genes found to be essential in the conditions set out in the survey - these genes are also indicated in Supplementary Table 3. The 'Profiling of *E. coli* Chromosome' database [90] contains a list of essential genes according to literature surveys and is included in Supplementary Table 3. Finding gene essentiality is a reductive enterprise - if a gene can be shown to be non-essential in a particular environment then it can be removed from the list of essential genes. The two mutagenesis studies mentioned above both use LB rich media, which was designed for growth of *E. coli* [91], and is therefore excellent for elucidating truly essential genes from media-specific essential genes. The set of essential genes used to investigate the genome comparisons presented in this thesis is the intersection of the sets of genes that are essential according to the three lists described above, 191 in total shown in Supplementary Table 3.

1.5 Uropathogenic *E. coli* and Metabolism

Uropathogenic *E. coli* (UPEC) are the causal agents of 80% of all community-acquired urinary tract infections (UTIs) [92]; other bacteria such as *Klebsiella* spp. and *Pseudomonas* spp. as well as many others can also infect the urinary tract. UPEC typically cause disease in one of several parts of the urinary tract: urine (bacteriuria), the kidneys (pyelonephritis), the bladder (cystitis), the ureter (ureteritis) and the prostate (prostatitis), for instance

UPEC CFT073 was isolated from a patient with acute pyelonephritis [93] (though it was isolated from the blood) and F11 from a patient with cystitis [45] - these are the most common forms of UTIs, thus bacteria causing these have been the most extensively investigated. Although these two strains are very closely related, inklings of the sensitivity of these bacteria to gene complement can be seen in the fact that the courses of pyelonephritis and cystitis caused by CFT073 and F11 (and several others of each type) have been shown to be quite different [94]. It should be noted, however, that there are commonalities between the infections: they are generally rising infections from the lower urinary tract and thus must survive in very similar conditions as they colonise their host's urinary tract.

One of the fascinating and most important discoveries resulting from widespread whole genome sequencing of pathogenic bacteria is the discovery that many of these virulence factors are physically close on the bacterial genome and that they generally fall into Pathogenicity-Associated Islands (PAIs) which are flanked by directly repeated sequences which are implicated in the evolution of the pathogenic bacteria via horizontal gene transfer. PAIs were first recognised by Blum *et al* [95] in *E. coli* 536, another UPEC, but it was on full genome sequencing of these strains that the paradigm of multiple unstable sites of PAI integration and the ability of bacteria to exchange these PAIs in single (and very common) excision/insertion events became clear, with transfer RNA genes being the target sites for excision (and insertion). These events are presumed to be mediated by integrases or IS elements which are usually present in the PAIs themselves, though often they are cryptic, indicating that many PAIs were once unstable, but have stabilised due to mutations in those genes (e.g. as reviewed by Hacker *et al* [96]).

E. coli CFT073 was selected, presumably because it was one of the most virulent UPEC isolated at that time, as the first UPEC to have its complete genome sequenced [42]. It already had several of its PAIs sequenced ([97, 98]) which contained many of its virulence factors - those genes contributing significantly to virulence - such as those coding for P fimbriae which are implicated in bacterial adhesion in the urinary tract and those coding for iron chelation proteins. Iron chelation is a common preoccupation of bacteria invading human hosts as generally the host's iron is closely bound to large complexes (such as haemoglobin) and there is little free for the use of bacterial cells [99] - this iron paucity is especially pronounced in the urinary tract. CFT073 also contains genes encoding a haemolysin (*hly* [98]) and a secreted autotransporter cytotoxin (*Sat* [100]) which both may help not only in iron acquisition but also acquisition of other scarce molecules through the disruption and lysis of host cells.

Several UPEC have had their complete genomes sequenced (CFT073, UTI89, 536, F11, see Table 1.1), along with twenty other or so *E. coli*, giving an excellent resource for discovering and investigating these PAIs. The nature of the PAIs and how bacteria come to acquire them is of course of much interest to those wishing to combat UTIs: much research has focused on these parts of the UPEC genomes. For instance, Welch *et al* [42] discussed PAIs found further to those found previously in CFT073 and compared the genome with (commensal) MG1655 and (EHEC) *E. coli* strain O157 H7 EDL933 and showed that only a small fraction of the genes that were absent from the commensal strain were contained in both pathogenic strains (3 % of the total CFT073 genes compared to 24 % of the CFT073 genes not contained in MG1655), though these genes included iron uptake systems, adhesins and other potentially virulence related genes.

Further to this it was found by Brzuszkiewicz *et al* [62] on comparing the complete sequences of CFT073 and *E. coli* 536, another UPEC, that although there was a significant overlap of 180 genes implicated as virulence factors (including about 20 implicated in metabolism: some encoding cobalamine biosynthesis and others of putative metabolic function) there are also more than 200 genes specific to 536 that are in regions that may contribute to virulence. This lack of consensus in the complement of virulence related genes carried by individual strains of UPEC can be seen over a large range of bacterial isolates from people with UTIs and the comparison of 125 commensal *E. coli* and UPEC by PCR-based detection used by Brzuszkiewicz *et al* [62] illustrates this heterogeneity in virulence factor gene complement. However, it should be borne in mind that although specific genes may not be conserved at the same position on the genomes of two bacteria, homologous and functionally equivalent genes may be present at different positions, or even non-homologous functional equivalents.

Genome sequence comparisons such as those mentioned above are useful since exact sequences for all genes in each strain are known. However, it is possible to experimentally determine approximate gene complement overlaps using several different approaches, including that used by Brzuszkiewicz *et al* [62]. Comparative genomic hybridisation (CGH) is also a common technique used for such analyses. This technique takes advantage of the hybridisation of query DNA from a sample (say, from a bacterium of interest) to a pre-prepared array of DNA from a reference species and consequent positional detection of fluorescence. CGH has been used by Lloyd *et al* [101] to investigate the presence of known virulence determinants in CFT073 in seven UPEC and several faecal isolates.

This type of investigation is incredibly valuable in helping to examine the contributions of various known virulence determinants by their presence or absence in isolates from people with various specific urinary tract infections, though it has the limitation that it will not by itself find further virulence determinants that are not present in the sequenced bacterium used as the hybridisation subject (in the case above, CFT073). An advantage of genome sequence comparisons and correspondences is that if an unexpected gene or set of genes is newly characterised (say, as a virulence determinant) then testing all sequenced bacteria for the gene or genes is a matter of conducting a BLAST search.

The traits commonly attributed with enabling *E. coli* to colonise the urinary tract are production of various specific fimbriae (sometimes called pili), the presence of a chemically modulated and smooth polysaccharide coat, iron sequestration mechanisms, serum resistance, production of haemolysins [102]. However, infections of the urinary tract by certain *E. coli* have been found which do not produce symptoms in the host, a state of so-called asymptomatic bacteriuria (ABU). One such bacterium is *E. coli* 83972, a strain that was isolated from a Swedish girl [103] who had been stably infected for 3 years with no symptoms associated with a urinary tract infection. This bacterium has non-functional fimbriae genes, and does not produce a functional haemolysin [104]; this indicates that there may be an alternative colonisation strategy for bacteria in the urinary tract to virulence related aggression. An unexplained property of this bacterium is that it outcompetes UPEC (including CFT073, 536 and *E. coli* NU14 [105]) in human urine and when in competition with NU14 (the strain which it outcompeted by the least amount in human urine) in a murine UTI model [106]. *E. coli* 83972 grows faster than these UPEC and the data from Roos *et al* [106] indicate that it somehow arrests the progress of such bacteria in the urinary tract (as has also been observed in human urinary tracts where 83972 has been used as a prophylactic [107]), perhaps by inhabiting the niche otherwise vulnerable to UPEC, or by some other mechanism yet to be elucidated.

It is unknown why UPEC do not have the capability to grow as quickly as 83972. Its ancestors were almost certainly pathogenic and the non-functional fimbrial genes present in its genome attest to that, but at some point the host-pathogen interaction may have promoted a mutually beneficial system where the bacteria somehow remained, but with the loss of function of several antigenic genes. If 83972 does not have several important virulence factors then it must have an alternative colonisation strategy, which may purely be by way of growth rate. The competitive advantage that 83972 has could be due to modified regulation of metabolic genes to enable faster growth; it has been shown that bacteria can improve growth rates in constant environments [108]. It could be that given 83972's continuous colonisation of a single host for 3 years (and possibly in previous human hosts), it has optimised its growth in the urinary tract environment, but chronic symptomatic urinary tract infections are common so might provide evidence against this possibility. A second possibility is that there are uncharacterised metabolic genes increasing 83972's ability to grow in urine, though given the frequency of horizontal gene transfer the lack of a "superstrain" containing both functional virulence genes and these metabolic genes seems indicate that this is not the case. A third potential reason is that just removal of virulence genes improves 83972's fitness, but then the existence of any uropathogenic *E. coli* is a puzzle.

The ability of UPEC to colonise the urinary tract has been studied in-depth in terms of virulence factors such as pili and haemolysins [109], however the ability to utilise the metabolites available in this environment has not been fully investigated to date. While studies of growth rates in urine have been conducted (for instance by Gordon and Riley [110]), it has never been established exactly what is used as the primary carbon source (or if there is a single one) for growth of UPEC, and which genes enable the utilisation of this carbon source. The growth of UPEC in this environment, however, indicate that it is certainly possible for bacteria to use purely urine for their metabolic needs rather than, for instance, the mucus produced by the epithelial cells of the urinary tract to which the bacteria adhere. There has been discussion of metabolic genes contributing to uropathogenicity, considering two UPEC: CFT073 and 536 [62]. D-serine has been posited as a potential carbon and nitrogen source, due to the presence of D-serine catabolic genes in *E. coli* CFT073 and *E. coli* 536 (both UPEC), the requirement for these genes or their involvement in the course of a UTI has yet to be investigated.

By separating the genome sequences of those bacteria that successfully colonise the urinary tract and those that do not it might be possible to discern a set of genes more common in the first set than the second, therefore which might be a contributing factor to successful colonisation. Indeed some genes have been looked for in this very way, for instance by Brzuszkiewicz *et al* [62], who compared 5 complete genome sequences in search of conserved and non-conserved genes between UPEC and non-UPEC which might contribute specifically to virulence. There is huge potential for widening this type of approach to a large number of completely sequenced bacteria, and since the pace of genomic sequencing is increasing exponentially [111] comparisons of this sort will be increasingly useful. The application to discovering whether metabolic genes are implicated in pathogen survival, as shown in the work presented here, is viable and informative.

Parts of the metabolism of UPEC have been studied for their importance in the course of urinary tract infections: guaA and argC are known to be of importance due to the requirement for guanine related products for survival and growth in urine [112] and recent work has focused on the role of D-serine uptake and metabolism during colonisation [113, 114, 115, 116]. UPEC have also developed several mechanisms for scavenging metals, for instance zinc [117], but especially iron which is incredibly scarce in the urinary tract (as reviewed by Wiles *et al* [99]).

1.5.1 *E. coli* CFT073 as a model of uropathogen metabolism

CFT073 contains 5379 predicted ORFs, only about half of which are bioinformatically characterised, let alone have an experimentally verified function. However, a closely related strain, *E. coli* K-12 MG1655 is very well characterised, being the standard laboratory strain of *E. coli* in use around the world. Given the high homology between corresponding genes and synteny conservation between CFT073 and MG1655 it seems likely that many of the bioinformatically inferred gene functions of CFT073, which rely on MG1655 annotation, are correct. As well as the genes with some annotation, even as unspecific as 'putative oxidoreductase' (c4981), there are 2516 hypothetical proteins, that is, open reading frames in the genome sequence with no similarity to any characterised or partially characterised gene so far discovered. With this number of genes of unknown function some sort of screening process is required to obtain a feasible number of genes to investigate for metabolic capabilities.

1.6 Aims

Broadly, the aims of this project have been to use genome sequences of *E. coli* and other bacteria to aid the discovery of uncharacterized metabolic genes by relative occurrences of these genes in various bacteria, to investigate the method of adaptation of bacteria to the addition of heterologous metabolic genes and to assess the applicability of a stoichiomteric model of MG1655 for inferring models of closely related bacteria.

1.6.1 Genome comparisons for the characterisation of novel metabolic genes

An initial gene synteny comparison of bacteria will be carried out in order to find genes that are putatively metabolic in their function, but without specific functional assignment. These genes, and uncharacterised genes in close proximity to these, will be assessed by various bioinformatic techniques in order to determine more clearly their function. Genes annotated as a putative sorbose or mannose PTS system is determined by BLAST comparison with *Klebsiella pneumoniae* to be a putative L-sorbose PTS system and this will be experimentally tested.

1.6.2 Bacterial adaptation to heterologous metabolic genes

Metabolic genes encoding a carbon source uptake and utilisation operon from one strain of *E. coli* will be cloned into another *E. coli* using a plasmid vector in order to assess the ability of the *E. coli* to adapt to genes conferring novel metabolic capabilites on the strain, and the metabolic adaptation of this strain over a period of passage in defined media with that carbon source as sole carbon source.

1.6.3 A survey of shared metabolic capabilities of bacteria with completely sequenced genomes

A technique of whole genome gene-by-gene comparisons and synteny inferences based on sequence identity of neighbouring genes will be used to assess genes shared between MG1655 and 36 other bacteria of the family *Enterbacteriaceae*. This comparison will be used to compare *in silico* metabolic capabilities of these strains and to determine the applicability of gene essentiality characteristics in closely related bacteria.

1.6.4 Reliable inferences of bacterial metabolic models

One of the gene synteny inferences for an *E. coli* strain produced in this work will be used to reconstruct a metabolic model for that strain, using a subset of the reactions in model iAF1260 of Feist *et al* [1], and will be tested against data on growth and acetate production of that strain. Also a model of a derivative of wild-type *E. coli* which is commonly used in cloning, will be constructed and tested against experimental data from this work.

Table 1.1: Strains of genus *Escherichia* with genome sequences deposited in GenBank as of 1^{st} January 2009. Unless otherwise stated the strain is *Escherichia coli*, the RefSeq accession number is the number for the sequence in Genbank and the sequence status distinguishes complete chromosomal sequences ('C') from whole genome shotgun sequences ('W'). The length of 'W' sequences is the sum of the lengths of the contigs making up that sequence entry. Where there has been a publication reporting the completed genome sequence this has been indicated, however many of the genome sequences do not have associated publications so articles detailing the isolation of those strains have been included; these are marked with an asterisk. Where there is no published literature according to PubMed on a particular strain the sequencing centre has been indicated: JCVI (The J Craig Venter Institute) or JGI (The Joint Genome Institute).

Strain	Pathotype	RefSeq Accession Number	Sequence Status	Length [bp]	Reference
Escherichia albertii TW07627	EPEC	NZ_ABKX00000000	W	4698533	[36]*
101-1	EAEC	NZ_AAMK00000000	W	4979767	JCVI
536	UPEC	NC_008253	C	4938920	[37]
53638	EIEC	NZ_AAKB00000000	W	5071018	JCVI
55989	EAEC	NC_011748	C	5154862	[38]*
APEC O1	APEC	NC_008563	C	5082025	[39]
B171	EPEC	NZ_AAJX00000000	W	5426568	[40]*
B7A	ETEC	NZ_AAJT00000000	W	5300242	[41]*
ATCC 8739	Ethanologenic	NC_010468	C	4746218	JGI
CFT073	UPEC	NC_004431	C	5231428	[42]
E110019	EPEC	NZ_AAJW00000000	W	5376211	[43]*
E24377A	ETEC	NC_009801	C	4979619	JCVI
ED1a	Commensal	NC_011745	C	5209548	[44]*
F11	UPEC	NZ_AAJU00000000	W	5215961	[45]*
HS	Commensal	NC_009800	C	4643538	[46]*
IAI1	Commensal	NC_011741	C	4700560	[47]*
IAI39	UPEC	NC_011750	C	5132068	[47]*
str. K12 substr. DH10B	Commensal	NC_010473	C	4686137	[48]
str. K12 substr. MG1655	Commensal	NC_000913	C	4639675	[35]
str. K12 substr. W3110	Commensal	AC_000091	C	4646332	[49]
E22	EPEC	NZ_AAJV00000000	W	5528238	[45]*
O127:H6 str. E2348/69	EPEC	NC_011601	C	4965553	[28]
O157:H7 str. EC4024	EHEC	NZ_ABJT00000000	W	6199307	JCVI
O157:H7 str. EC4042	EHEC	NZ_ABHM00000000	W	5617728	JCVI
O157:H7 str. EC4045	EHEC	NZ_ABHL00000000	W	5634850	JCVI
O157:H7 str. EC4076	EHEC	NZ_ABHQ00000000	W	5705645	JCVI
O157:H7 str. EC4113	EHEC	NZ_ABHP00000000	W	5655847	JCVI
O157:H7 str. EC4115	EHEC	NC_011353	C	5572075	JCVI
O157:H7 str. EC4196	EHEC	NZ_ABHO00000000	W	5620606	JCVI
O157:H7 str. EC4206	EHEC	NZ_ABHK00000000	W	5629932	JCVI
O157:H7 str. EC4401	EHEC	NZ_ABHR00000000	W	5733133	JCVI
O157:H7 str. EC4486	EHEC	NZ_ABHS00000000	W	5933166	JCVI
O157:H7 str. EC4501	EHEC	NZ_ABHT00000000	W	5677181	JCVI
O157:H7 str. EC508	EHEC	NZ_ABHW00000000	W	5656666	JCVI
O157:H7 str. EC869	EHEC	NZ_ABHU00000000	W	5731065	JCVI
O157:H7 str. Sakai	EHEC	NC_002695	C	5498450	[50]
O157:H7 str. TW14588	EHEC	NZ_ABKY00000000	W	5670297	JCVI
O157:H7 EDL933	EHEC	NC_002655	C	5528445	[51]
S88	ECNM	NC_011742	C	5032268	[52]*
SE11	Commensal	NC_011415	C	4887515	[53]
SMS-3-5	Environmental	NC_010498	C	5068389	[54]
UMN026	UPEC	NC_011751	C	5202090	[55]*
UTI89	UPEC	NC_007946	C	5065741	[56]
Escherichia fergusonii ATCC 35469	Commesal	NC_011740	C	4588711	[57]*
Escherichia sp. 3_2_53FAA	Crohn's associated	NZ_ACAC00000000	W	5094952	Broad Institute

Chapter 2

Experimental Materials and Methods

2.1 Strains

Strains used throughout this project were stored in 15 % glycerol stocks at -80 °C between experiments. When experiments were undertaken a metal loop was used to scrape a little of the frozen stock and spread it on an agar plate with antibiotic where appropriate. An individual colony was then picked from this plate for use in each experiment. The following strains were used:

(i) *E. coli* strain CFT073 (American Type Culture Collection, Manassas, VA, US), ATCC 700928;

(ii) *E. coli* strain K-12 MG1655 (from the strain collection of JM Ward), ATCC 47076;

(iii) *E. coli* strain DH5α, ATCC 700790 (Invitrogen), genotype: F^{-} *fhuA2* Δ*(lacZ-argF)U169* ϕ*80*d*lacZ*Δ*M15 endA1 hsdR17 deoR nupG thi-1* *supE44 gyrA96 relA1 recAI phoA* λ^{-};

(iv) *E. coli* strain TOP10 (provided as One Shot® TOP10 *E. coli* from Invitrogen), genotype: F- *mcr*A Δ(*mrr-hsd*RMS-*mcr*BC) ϕ80*lac*Z ΔM15 Δ*lac*X74 *rec*A1 *ara*Δ139 Δ(*ara-leu*)7697 *gal*U *gal*K *rps*L (Str^{R}) *end*A1 *nup*G;

(v) *Klebsiella pneumoniae* producing *Kpn*I restriction enzyme (KP1; from the strain collection of JM Ward).

2.2 Buffers

The following items were made up from base chemicals in the lab:

(i) *Loading Buffer for agarose gel electrophoresis* (6x solution): 4 g of sucrose, 2 ml of 0.5 M EDTA (ethylenediaminetetraacetic acid) and 1 ml of 1.5 mg/ml BPB (bromophenol blue), topped up to 10 ml with purified water.

(ii) *Standard Reaction Buffer for restriction enzyme digestion* (10x solution): 500 mM Tris, 500 mM NaCl and 50 mM MgCl in purified water.

The following buffers were all provided by New England Biolabs, Hitchin, UK:

(i) *Antarctic Phosphatase Reaction Buffer* (10x solution): 500 mM Bis-Tris-Propane-HCl, 10 mM $MgCl_2$, 1 mM $ZnCl_2$ (pH 6.0 at 1x concentration at 25 °C);

(ii) *T4 DNA Ligase Reaction Buffer* (10x solution): 500 mM Tris-HCl, 100 mM $MgCl_2$, 10 mM ATP, 100 mM Dithiothreitol (pH 7.5 at 1x concentration at 25 °C);

(iii) *NEBuffer 1* (10x solution): 100 mM Bis-Tris-Propane-HCl, 100 mM MgCl2, 10 mM dithiothreitol (pH 7.0 at 1x concentration at 25 °C);

(iv) *NEBuffer 2* (10x solution): 100 mM Tris-HCl, 100 mM MgCl2, 500 mM NaCl, 10 mM dithiothreitol (pH 7.9 at 1x concentration at 25 °C);

(v) *NEBuffer 3* (10x solution): 500 mM Tris-HCl, 100 mM MgCl2, 1000 mM NaCl, 10 mM dithiothreitol (pH 7.9 at 1x concentration at 25 °C);

(vi) *NEBuffer 4* (10x solution): 200 mM Tris-Ac, 100 mM Mg-Ac, 500 mM K-Ac, 10 mM dithiothreitol (pH 7.9 at 1x concentration at 25 °C).

2.3 Molecular Biology Components

Enzymes are described as they are used throughout the methods section, including restriction enzymes. These enzymes are all provided by New England Biolabs, Hitchin, UK, unless otherwise stated.

2.3.1 Sterile deionised water

Sterile distilled water was used in many preparations during this project. This water was produced by an Elga Option 4 Water Purifier (Elga, Marlow, Buckinghamshire, UK) to 14 $M\Omega cm^{-1}$ and either autoclaved for 20 minutes at 121 °C or passed through a 0.22 μm sterile filter (Millipore, Billerica, MA). It will henceforth be referred to as purified water.

2.3.2 Media Components

Various media components were required for Molecular Biology and growth testing, and in all cases component chemicals were dissolved into purified water and either autoclaved at 121 °C for 20 minutes, or passed through a 0.22 μm sterile filter depending on the chemical.

Where rich media was required for growth, Nutrient Broth 2 (Oxoid, Hampshire, UK), henceforth referred to as NB2, was used. Other media required by some kit protocols are (made up to 1x): 'Lab-Lemco' Powder 10 g/L, Peptone 10 g/L, NaCl 5 g/L. LB medium was used in some cases (1x): tryptone 10 g/L, yeast extract 5 g/L and NaCl 10 g/L. SOC medium was also required for some molecular work (Invitrogen): tryptone 20 g/L, yeast extract 5 g/L, NaCl 10 mM, KCl 2.5 mM, $MgCl_2$ 10 mM, $MgSO_4$ 10 mM and glucose 20 mM.

For growth in defined minimal media a variant of M6 minimal medium (henceforth referred to as M6*), optimised for growth of *E. coli*, was used. It has the following composition: 5.20 gl^{-1} $(NH_4)_2SO_4$, 3.86 gl^{-1} $NaH_2PO_4.H_2O$, 4.03 gl^{-1} KCl, 4.16 gl^{-1} Citric Acid, 1.04 gl^{-1} $MgSO_4.7H_2O$, 0.25 gl^{-1} $CaCl_2.2H_2O$, 20.6 mgl^{-1} $ZnSO_4.7H_2O$, 27.2 mgl^{-1} $MnSO_4.4H_2O$, 8.1 mgl^{-1} $CuSO_4.5H_2O$, 4.2 mgl^{-1} $CoSO_4.7H_2O$, 100.6 mgl^{-1} $FeCl_3.6H_2O$, 0.3 mgl^{-1} H_3BO_3, 0.2 mgl^{-1} $Na_4MoO_4.2H_2O$ adjusted to pH 7.3 using 5M NaOH.

Antibiotics used in all experiments were provided by Sigma-Aldrich, Gillingham, Dorset, UK. Ampicillin was stored at 1000x concentration (100 mg/ml) at -20 °C. Kanamycin monosulfate was stored at -20 °C as a 50 mg/ml (1000 ×) stock.

Several carbon sources were used for growth tests, all were added to 1% w/v:

(i) D-Glucose (BDH Laboratory Supplies, Poole, UK),

(ii) L-Sorbose (Merck, Beeston, UK),

(iii) Creatinine (Sigma-Aldrich),

(iv) Urea (VWR International, Poole, UK) and

(v) DL-Malic Acid disodium salt (Sigma-Aldrich).

2.4 Agar plates for bacterial growth

Where plates were required for bacterial growth and selection, a standard mixture of 1.5 % w/v agar (Difco, Livonia, MI, US) and 25 g/L NB2 was added to purified water and autoclaved at 121 °C for 20 minutes. This was then left to cool to approximately 50 °C and poured immediately, or stored in at 50 °C until required. Plates poured without supplements will be referred to as NB2 agar plates and those with supplements referred to with the supplement(s) prefixed as an abbreviation, for instance kana NB2 agar plate for a plate containing kanamycin for selection.

The following standard final concentrations were used for each supplement (abbreviations are shown), unless otherwise stated in the text:

(i) Kanamycin 50 μg/ml (kana),

(ii) Ampicillin 100 μg/ml (amp),

(iii) X-Gal (Bioline, London, UK) 20 μg/ml (X-Gal),

(iv) Isopropyl β-D-1-thiogalactopyranoside (Melford, Ipswich, UK) 100 μM (IPTG).

Where other supplements are used their final concentrations are noted in the relevant section.

2.5 Growth Screening Tests (5 ml in universal tubes)

Where it was required to screen *E. coli* CFT073 and MG1655 for growth on various substrates as sole carbon sources (those listed in Section 2.3.2), M6* media was used supplemented with the carbon sources as required. Strains were taken directly from stocks and inoculated into 5 ml NB2 media in 20 ml universal tubes, then grown to stationary phase overnight at 37 °C shaken at 150 rpm in a KF-4 Chest Incubator (Infors HT, Hilden, Germany). The following morning, 25 μl of each strain was inoculated into 5 ml fresh NB2 media and grown in the same conditions for 6 hours until exponential phase growth was achieved.

25 μl of the resulting material was inoculated into separate 20 ml universal tubes containing 5 ml pre-warmed M6* containing a single carbon source at 1 % w/v. These universal tubes were then returned to the KF-4 and again grown overnight at 37 °C shaken at 150 rpm. These tubes were

observed at 24, 40 and 115 hours to determine whether there was growth of the strains using the relevant carbon source as a sole carbon source.

2.6 Growth Tests (200 ml in shake flasks)

Growth tests were conducted in 200 ml working culture volume in 2 l shake flasks to quantify uptake of substrates, growth and excretion of products of several bacteria at two temperatures using two different molecules, L-sorbose and D-glucose, as sole carbon sources. The two temperatures used were 30 °C and 37 °C and for each particular growth test the same temperature was used throughout the growth, from growing single colonies from stocks to the end of the shake flask observation. Ampicillin was used at a concentration of 100 μg/ml throughout all experiments involving *E. coli* containing plasmid pQR793 or a control plasmid, but this was not added for experiments involving *E. coli* CFT073.

To minimise stress on the bacteria before they were placed in the test medium care was taken not to expose them to stressful conditions, such as changes in temperature and starvation, while initial cultures were grown. This was achieved by keeping the bacteria at a constant temperature throughout growth, using rich media in starter cultures and ensuring that the starter cultures were in exponential growth phase at the point where they were transferred to shake flasks for growth observation.

Each strain was taken from a thawed 15 % glycerol stock and a loop placed into 5ml Nutrient Broth 2 (NB2; Oxoid) in a 20ml Universal, which was shaken in an incubator overnight at the appropriate temperature at 150rpm;

this was allowed to grow to stationary phase. 50 μl of this was then transferred to a starter culture of 20 ml of NB2 in a 50 ml falcon tube, again at the appropriate temperature. Preliminary testing determined approximate growth rates and final optical densities (used throughout as a surrogate for cell concentration) of the strain in these conditions and these data were used to monitor the growth of the starter culture. When the bacteria were in late exponential growth, defined as approximately three quarters of final cell density, 500 μl of this was transferred to 200 ml pre-warmed defined minimal medium in a 2 litre shake flask.

M6* medium was used for these experiments. 0.001 % thiamine supplement was added for all DH5α strains, which do not grow in the absence of thiamine. To this medium was added either 1 % w/v D-Glucose or 1 % w/v L-Sorbose, which each contain the same molar quantity of carbon.

Samples were taken at intervals through the growth, two 1 ml simultaneous samples at each time point. One sample was taken to test the optical density, which was measured at 600 nm using a CO8000 Cell Density Meter (WPA). For samples of optical density above 1, the sample was diluted to the range 0.2 to 0.7 and the actual optical density inferred by multiplying the reading by the dilution factor. The other sample was used to test the supernatant for glucose, L-sorbose and acetate. This sample was transferred to a microcentrifuge tube and spun for 10 minutes at 13,000 rpm in an accuSpin^{TM}microcentrifuge (Fisher Scientific, Loughborough, Leicestershire). The supernatant was then extracted by pipette and placed in a clean microcentrifuge tube and placed at -20 °C.

2.6.1 OD and biomass equivalence

Biomass was not directly measured in growth tests, but was taken to be proportional to OD. To determine this equivalence MG1655 was grown to OD 2 in 5 ml NB2 at 37 °C shaken at 150 rpm and trtiplicate 1 ml samples were taken. These samples were placed in preweighed 1.5 ml Eppendorf tubes that had been specially prepared by dessication for 48 hours. These samples were spun down in 1.5 ml Eppendorf tubes for 10 minutes at 13,000 rpm in the accuSpin$^{\text{TM}}$microcentrifuge, the supernatant decanted, then tissue was used to absorb excess moisture left on the pellets. These were dried at 100 °C in an oven for 48 hours and then difference in weight between the tubes before and after the addition of samples were calculated. The resultant equivalence was that 2 OD was equivalent to 0.6 ± 0.1 gDW, therefore mass during the project was calculated as 0.3 gDW/OD unit.

2.7 HPLC

High precision liquid chromatography was used for quantification of concentrations of glucose, L-sorbose and acetate in supernatant samples from growth tests. When all samples were collected from a single growth experiment, all 1 ml supernatant samples were removed and thawed on the bench for 20 minutes, then two aliquots of 260 μl from each sample were placed in wells in a 96-well 0.22um multiwell sterile filter plate (Pall Corporation, Port Washington, NY) and attached to the top of a 96-well plate (with a capacity of 300 μl per well) to collect flow-through. Samples were spun at 700 rpm in a 5804R centrifuge (Eppendorf, Hamburg, Germany) for 3 minutes.

These samples were then transferred to 2 ml sample containers for HPLC analysis.

If HPLC was immediately available samples were loaded immediately, and if not, they were stored at -20 °C, then thawed thoroughly and mixed using a Top Mix FB15024 (Fisher Scientific, Loughborough, UK) before loading. The HPLC equipment used was as follows: a P680 HPLC pump (Dionex, Sunnyvale, Ca), an ASI-100 Automated Sample Injector (Dionex, Sunnyvale, Ca), an STH 585 Column Oven (Dionex, Sunnyvale, Ca), a UVD170U ultraviolet detector (Dionex, Sunnyvale, Ca), an RI-101 refractive index detector (Shodex, Munich) and controlled by a UCI-100 Universal Chromatography Interface (Dionex, Sunnyvale, Ca). Software used was Chromeleon Version 6.4 SP8 Build 800 (Dionex, Sunnyvale, Ca).

The column used was an HPX-87H Aminex (Bio-Rad, Hemel-Hempstead, UK) ion exclusion and partition column, which is capable of separating both carbohydrates and organic acids. The HPLC analysis was run under the following conditions: eluent was 5 mM H_2SO_4, pH 2.2 (Fluka Chemika, Buchs, Switzerland) in purified water; column was kept at 60 °C throughout; flow rate was maintained at 0.6 ml/min; all sample injections were 20 μl; all wash cycles used 10 μl of 50% acetonitrile, 50% purified water. All water used in these samples was purified water, filtered through a 0.22 μm sterile filter. Sample standards were made up as follows: one vial of purified water, D-glucose dissolved in water at concentrations of 10, 8, 6, 4 and 2 g/L, L-sorbose dissolved in water at concentrations of 10, 8, 6, 4 and 2 g/L and sodium acetate dissolved in water to the concentrations of 30, 24, 18, 12 and 6 mM. All vials were sampled 3 times and each sample was run for

25 minutes through the column. UV absorbance was measured at 215 nm.

2.8 Protein Activity Assay

The protein activity assay described here was used to assess the temperature sensitivity of the activity of the glucitol-6-phosphate dehydrogenase in the L-sorbose degradation operon in CFT073 (gene locus c4986). Temperature sensitivity of a homologous glucitol-6-phosphate dehydrogenase in *Klebsiella pneumoniae* previously noted by Sprenger and Lengeler [118]. Sprenger and Lengeler observed impaired growth of *K. pneumoniae* at temperatures above 35 °C.

All growth media using cells containing constructs had ampicillin added to the concentration of 100 mg/ml. Cells were taken using a loop from frozen stocks, placed in 5 ml of prewarmed Nutrient Broth 2 (NB2;Oxoid, Basingstoke, Hampshire) and placed in either a KF-4 for 37 °C growth or an Amperetabelle Multitron 2 incubator for 30 °C growth, both at 150 rpm. This was left overnight to grow to stationary phase. Testing the heat sensitivity of the glucitol-6-phosphate dehydrogenase required that the cells be harvested at approximately mid-exponential phase, when the highest concentration of that enzyme per cell was presumed to be extant. This required that inoculum volumes were varied at this stage to ensure that overnight growth would result in mid-exponential phase growth in the carbon limited medium the following morning, thus that harvesting, processing and testing of the enzyme activity were all achieved in a single day. The relevant inoculum volume was therefore used for each set of conditions, depending on OD in the NB2 media at stationary phase and growth rate in the fresh minimal media at the relevant temperature. This volume was inoculated into 20 ml of M6* minimal media with 1 % w/v carbon source concentration in 50 ml falcon tubes. This was then placed in the relevant incubator for the correct temperature and left overnight. The following morning a single 1 ml sample was taken to check OD and if it was in the range 1.5 to 2.0 then the processing would begin. If the optical density was lower then the cells were left for the relevant period of time before being resampled and if then within the OD range required would be processed.

2.8.1 Protein extraction from bacterial cells

Crude extracts of intracellular material were used to assess the temperature dependence of the glucitol-6-phosphate dehydrogenase activity. For each sample the falcon tube was removed directly from the incubator into ice and all steps were done at 4 °C or on ice; all added chemicals were pre-cooled on ice before use in this method. The cells were then pelleted by centrifugation in a 5804R centrifuge at 5000 rpm for 10 minutes. The supernatant was decanted and the pellet was resuspended in 20ml ice cold 0.1 M Tris HCl (pH 9) to wash off the cells, then was spun again at 5000 rpm for 15 minutes to repellet. The supernatant was again decanted and the cell concentration was increased by resuspension in 10 ml 0.1 M Tris HCl (pH 9), which was transferred to a 15 ml falcon tube, to which was added 40 μl of 10 mg/ml lysozyme (Sigma-Aldrich). This was then transferred to a water bath at 37 °C for exactly 15 minutes, then removed to ice.

The suspension was then sonicated using a Soniprep 150 (MSE, Lower Sydenham, UK) at an amplitude of 9 microns, 10 seconds on and 12 seconds off, repeated 8 times, to break open the cells. The resultant material was transferred to a Sorvall tube (Sorvall, part of Thermo Scientific, Waltham, MA) and spun down in an RC6 Plus Superspeed Centrifuge (Sorvall) at RCF 30000 for 15 minutes to remove cell debris from the suspension. This suspension was then transferred to a universal tube on ice ready for the protein activity assay.

2.8.2 Assay of heat sensitivity of activity of Glucitol-6-Phosphate dehydrogenase

The protein activity assay method used for assessing the heat dependence of the activity of glucitol-6-phosphate dehydrogenase was adapted from Novotny *et al*'s [119] method, though in this investigation only crude lysate as per the procedure described in Section 2.8.1 was used, rather than purified protein. The assay was conducted in 0.1 M Tris HCl (pH 9) with the addition of glucitol-6-phosphate (Sigma) and NAD+ (Sigma). These concentrations were determined by testing various concentrations to find concentrations high enough to make the reaction insensitive to small variations in these concentrations, but not so high as to be inhibitory. The glucitol-6-phosphate and NAD+ were kept on ice before addition to the solution, as was the crude extract. The Tris HCl buffer was preheated to the relevant temperature before each sample was tested for activity, and the crude extract sample was preheated for 1 minute at the relevant temperature as well.

UV absorbance was determined using a U-1800 Spectrophotometer (Hitachi, Tokyo, Japan) with heated water provided by a GD120 waterbath and circulator (Grant Instruments Ltd, Cambridge, UK) to maintain the temperature required for each sample. The four components of the reaction sample were mixed together in a UV transparent cuvette (Fisher Scientific Ltd) in the following order: 920 μl 0.1 M Tris HCl (pH 9), NAD+ solution to 1 mM, glucitol-6-phosphate to 40 mM, then 50 μl crude cell extract. The resulting solution was mixed 3 times by pipetting. This was done as quickly as possible and then the lid of the spectrophotometer was closed and observation of the UV absorbance increase was observed. The time between mixing the crude extract into the buffer solution and beginning observation was about 4 seconds. The absorbance of the sample at 340 nm was observed for a minute. If the reaction worked, the first 10 seconds of observed absorbance change were used to determine an initial rate of reaction. This method was repeated twice at several temperatures between 30 °C and around 50 °C to produce a graph of protein activity as a function of temperature.

2.8.3 Glucitol-6-Phosphate dehydrogenase heat dependence assay

In order to determine the temperature dependence characteristics of the glucitol-6-phosphate dehydrogenase present in the L-sorbose uptake and utilisation operon in CFT073, the model of Lee *et al* [120] was used. The Equilibrium Model of enzyme activity temperature dependence takes into account inactivated but not denatured enzymes and takes into account the change in equilibrium coefficient between active and inactive enzyme due to changes in temperature. The method in Section 2.8 produces data in the form of initial activity values at particular temperatures for Glucitol-6-phosphate dehydrogenase, which can be modelled as followed. Using the Arrhenius equation and assuming Michaelis-Menten kinetics for the initial activity of an enzyme in excess substrate,

$$\begin{aligned} V_{init} &= k_{cat}. [E_a] & (2.1) \\ &= Ae^{-\frac{\Delta H^0}{RT}}. [E_a] & (2.2) \end{aligned}$$

where V_{init} is the initial reaction rate, k_{cat} is the rate constant for the reaction, $[E_a]$ is the concentration of active enzyme, A is the (Arrhenius) prefactor, ΔH^0 is the standard enthalpy of the reaction, R is the gas constant and T is the temperature at which the reaction is taking place.

From [120] the balance between active and heat inactivated (though not denatured) enzyme also follows Michaelis-Menten like kinetics with a process of activation/inactivation with an equilibrium constant K_{eq}. Although that work also treats denatured proteins, activity measurements in this work were taken with the absolute minimum of prior exposure to high temperatures and only initial activity readings were taken. Each set of activity measurements for a particular sample were done at the same time and readings of activity at lower temperatures both at the beginning and end of the measurement process were checked to ensure the activity was not significantly reduced by denaturation over the time period of the analysis. Further, although only the first minute of activity was measured for each sample, activity was regularly monitored for up to 10 minutes with no change in activity, so heat dependent denaturation was discounted as a significant process in this work. The chemical equation :

$$\mathrm{E_a} \xleftrightarrow{K_{eq}} \mathrm{E_i} \tag{2.3}$$

where K_{eq} is the equilibrium constant and E_i is the inactive form of the protein, the concentration of which can be written as total enzyme concentration $[E_t]$ minus active enzyme concentration $[E_a]$, so

$$\begin{aligned} \frac{[E_i]}{[E_a]} &= K_{eq} && (2.4) \\ \frac{[E_t]-[E_a]}{[E_a]} &= K_{eq} && (2.5) \\ [E_a] &= \frac{[E_t]}{1+K_{eq}}. && (2.6) \end{aligned}$$

Also from [120],

$$K_{eq} = e^{\frac{\Delta H_{eq}}{R}\left(\frac{1}{T_{eq}}-\frac{1}{T}\right)} \tag{2.7}$$

where R is the gas constant, ΔH_{eq} is the enthalpy at of the reversible equilibrium between active and inactive enzyme. Therefore

$$[E_a] = \frac{[E_t]}{1+e^{\frac{\Delta H_{eq}}{R}\left(\frac{1}{T_{eq}}-\frac{1}{T}\right)}} \tag{2.8}$$

and

$$\begin{aligned} V_{init} &= Ae^{-\frac{\Delta H^0}{RT}}.\frac{[E_t]}{1+e^{\frac{\Delta H_{eq}}{R}\left(\frac{1}{T_{eq}}-\frac{1}{T}\right)}} && (2.9) \\ &= \frac{Be^{-\frac{\Delta H^0}{RT}}}{1+e^{D\left(\frac{1}{T_{eq}}-\frac{1}{T}\right)}} && (2.10) \end{aligned}$$

where $B = A\,[E_t]$, $D = \frac{\Delta H^0}{R}$, ΔH^0 and T_{eq} are unknowns to be tuned to fit the available data. This model of initial activity as a function of temperature was then fitted to the observed data using a downhill simplex method (adapted from Nelder and Mead [121]) to minimise the sum of the residual squares of the differences between the observed values of activity and the model.

These parameters were determined independently for each assay and from them T_{max}, the temperature of maximum activity, was determined. Figure 5.8 shows the values of T_{max} determined for each of the strains tested. Strain *Klebsiella pneumoniae* KP1 was used as a control since it has been shown previously [118] that the glucitol-6-phosphate dehydrogenase of the L-sorbose operon of this strain is heat sensitive and it was the operon from which the CFT073 operon inferred its own function.

2.9 Cloning

Two putative operons in the genome of CFT073 were targeted for cloning, using one (in the case of the putative L-sorbose operon genes c4981-4987, see Section 4.3.3) or all (in the case of the putative 5- or 6-carbon sugar metabolism operon genes c3750-3756, see Section 4.3.4) of the below protocols for cloning.

2.9.1 Chemically competent cell preparation

Chemically competent cells were prepared in the following way:

(i) cells were grown overnight to stationary phase in 5 ml NB2 culture, at 37 °C shaken at 150 rpm,

(ii) this culture was added to 20 ml NB2 supplemented with 20 mM $MgCl_2$ and returned to the incubator for 1 hour,

(iii) it was removed to ice and spun down at 4 °C at 5000 rpm for 10 minutes in a 5804R centrifuge,

(iv) the supernatant was decanted and the cells were resuspended in 2 ml 75 mM $CaCl_2$ with 15 % w/v glycerol (all on ice) and

(v) 100 μl aliquots were transferred to eppendorf tubes and frozen at -80 °C.

2.9.2 Electrocompetent cell preparation

Where chemical transformation was inadequate for producing stable plasmid transformants, electroporation was used. Cells for this process required a different preparation, to reduce the ionic strength of the cell suspension to an absolute minimum:

(i) cells were grown overnight to stationary phase in 5 ml NB2 culture, at 37 °C shaken at 150 rpm,

(ii) the culture was added to 20 ml NB2 and returned to the incubator for 1 hour, (from this point all preparatory steps were done at 4 °C or on ice, with all components prechilled)

(iii) the cells were then transferred to ice and left to chill for 30 minutes,

(iv) they were then spun down at 4 °C at 5000 rpm for 10 minutes in a 5804R centrifuge,

(v) the supernatant was decanted and the cells were resuspended in 20 ml purified water,

(vi) the cells were again spun down at 4 °C at 5000 rpm for 10 minutes in a 5804R centrifuge,

(vii) steps vi) and vii) were repeated twice more,

(viii) the pellet was then resuspended in 20 ml 10 % w/v glycerol solution,

(ix) the cells were spun down a final time at 4 °C at 5000 rpm for 15 minutes,

(x) they were then resuspended in 2 ml 10 % w/v glycerol solution and

(xi) aliquots of 50 μl were placed in prechilled 1.5 ml Eppendorf tubes and either used immediately or frozen at -80 °C.

2.9.3 Primers

Primers were ordered from Operon (Ebersberg, Germany). The following primers were ordered:

(i) Putative operon including c4981 forward primer: 5'-GCCAGCGACATGCAGAGTTAAGTAGCGCGA-3, which has a melting temperature of 71.5C;

(ii) Putative operon including c4981 reverse primer: 5'-AAATCTCCTGTAAAACGCGGAATATACC-3, which has a melting temperature of 63C;

(iii) Putative operon including c3750 forward primer: 5'-TTTGCTTCC-AGAAATGGTAAAAAAATAATC-3, which has a melting temperature of 59C;

(iv) Putative operon including c3750 reverse primer: 5'-TAATGGACC-AATTCAATGCCCCACAGAGTG-3, which has a melting temperature of 67.5C.

2.9.4 Preparation of Genomic DNA

PCR of DNA fragments of CFT073 was done directly from the CFT073 genome. The method used to prepare the genomic DNA for PCR was by growing cells overnight to stationary phase in 10 ml 1 × NB2 at 37 °C at 150 rpm, then boiling at 100 °C for 20 minutes. The resultant material was either used immediately or stored at -20 °C until required.

2.9.5 Minipreps

Where plasmid DNA was recovered for analysis it was achieved by miniprep. Recovery was made from 1 ml samples of growth material. The standard protocol of the QIAprep Spin Miniprep Kit (Qiagen, Crawley, UK) was used for minipreps and resultant DNA was eluted into 50 μl EB Buffer for running in gels or freezing at -20 °C for future use. All buffers used for minipreps came from a Qiagen miniprep kit. The LyseBlue provided with the Qiagen kit was not used for any minipreps.

Where minipreps were done using a non-standard protocol the changes are noted in the relevant section below, but in all cases during the elution step, when Buffer EB was added to the QIAprep spin column it was left to stand for 2 minutes before centrifuging and was centrifuged for 2 minutes, rather than 1 minute, to achieve maximum DNA recovery.

2.9.6 DNA PCR amplification of putative L-sorbose operon (c4981-c4987)

PCR of DNA fragments of interest was conducted using a TC-512 gradient thermal cycler (Techne, Burlington, NJ, USA). The polymerase used was the Phusion High-Fidelity DNA Polymerase (Finnzyme), which is a *Pyrococcus*-like enzyme fused with a processivity-enhancing domain and has a very low error rate (base duplication errors per base duplicated) in Phusion HF Buffer of 4.4×10^{-7}. This polymerase produces blunt ended DNA products.

10 μl of genomic DNA prepared as in Section 2.9.4 was added to 990 μl of purified water, a 1/100 DNA solution and half of this was added to a further 500 μl of sterile distilled to produce a 1/200 DNA solution. Primers were diluted to 100 μM as per their datasheet from Operon, then 2.5 μl of each primer solution was diluted separately into 97.5 μl purified water.

Each reaction required 0.5 μl of 2 U/μl Phusion DNA polymerase (Finnzyme) and 10 μl of 5x Phusion$^{\text{TM}}$HF Buffer (Finnzyme), which was measured into a PCR tube. To this was added 5 μl of each primer solution and 1 μl of 10 mM dNTPs solution. The PCR annealing step was conducted at three temperatures: 53, 58 and 63 °C, using each of the two concentrations of DNA (1/100 and 1/200 × dilutions), of which 5 μl of the relevant solution was added to the PCR tubes. 23.5 μl sterile distilled was added to the PCR tubes to make the solution up to 50 μl. The PCR cycle was programmed thus:

(i) 3 minutes at 95 °C for initial denaturation,

(ii) 15 seconds at 95 °C for denaturation,

(iii) 15 seconds at annealing temperature for annealing,

(iv) 3.5 minutes at 72 °C for polymerisation and

(v) 10 minutes at 72 °C for final polymerisation.

Steps ii) to iv) were repeated 30 times in order before the final polymerisation step. The hot start option on the PCR machine was disabled for this programme and the 'heat lid' option was set at 105 °C.

2.9.7 DNA PCR amplification of putative 5- or 6-carbon operon (c3750-c3756)

PCR of the c3750-c3756 DNA fragment was achieved using the same procedure as in Section 2.9.6, with the following modifications: the PCR annealing step was conducted at four temperatures: 49, 52, 55 and 58 °C, due to the lower melting temperatures of the primers. The PCR cycle was programmed thus:

(i) 30 seconds at 98 °C for initial denaturation,

(ii) 10 seconds at 98 °C for denaturation,

(iii) 15 seconds at annealing temperature for annealing,

(iv) 240 seconds at 72 °C for polymerisation and

(v) 420 seconds at 72 °C for final polymerisation.

Steps ii) to iv) were repeated 30 times in order before the final polymerisation step.

2.9.8 Purification of PCR products

DNA purification of PCR products was achieved by gel extraction. 5 μl of PCR product solution was run in a gel with 1 μl loading buffer alongside a λ HindIII ladder (New England Biolabs), and the band at approximately 7 kb was cut out of the gel. DNA extraction from the gel was achieved using the standard protocol of the QIAquick Gel Extraction Kit using an accuSpin$^{\mathrm{TM}}$Micro microcentrifuge (Fisher Scientific, Loughborough, UK). Resulting DNA was eluted into 50 μl EB Buffer; 5 μl was run on a gel to check recovery and the rest was immediately frozen at -20 °C to be stored for later use.

2.9.9 Cloning kits and techniques

Cloning of PCR products into a multicopy plasmid vector was done in several ways, depending on the success of each cloning strategy. Two different cloning kits were used according to availability during the project, and when the kit failed to produce cloned products an alternative protocol using non-kit elements was attempted.

Advice in the relevant manuals concerning procedures for long PCR products (such as those cloned in this project which were both over 6 kb long) was followed. Both kits use blunt ended DNA fragments and vectors for ligation, and Phusion High-Fidelity DNA Polymerase was used (as described in Section 2.9.6) to produce these blunt ended long DNA fragments.

2.9.9.1 StrataClone Blunt PCR Cloning

In the first case the StrataClone Blunt PCR Cloning Kit (Stratagene, La Jolla, CA) was used with the blunt ended StrataClone$^{\text{TM}}$PCR Cloning Vector pSC-B (Stratagene, La Jolla, CA) as host vector. This system uses a topoisomerase and Cre recombinase to achieve ligation of two DNA arms with a heterologous DNA acceptor, then circularisation. All selection was then done using ampicillin. The protocol for ligation of long PCR products that came with the cloning kit was followed.

Transformation of vectors produced by cloning of the DNA fragment were achieved using the standard protocol of the Strataclone Blunt PCR Cloning Kit:

(i) 2 μl of cloned material was added to one tube of StrataClone SoloPack competent cells (provided with the kit) and mixed gently,

(ii) The transformation mixture was incubated on ice for 20 minutes,

(iii) The mixture was then heat-shocked at 42 °C for 45 seconds and returned to ice for 2 minutes,

(iv) 250 μl SOC medium (prewarmed to 42 °C) was then added and the mixture placed in an incubator at 37 °C shaken at 150 rpm for 1 hour.

Blue/white screening of transformants for recombinant vectors was achieved as described in Section 2.9.10, however since the *lacZ* gene is engineered to constitutively express β-galactosidase, IPTG was not required to induce expression on the screening plates.

2.9.9.2 Zero Blunt®TOPO®PCR Cloning

The second kit used was the Zero Blunt®TOPO®PCR Cloning Kit (Invitrogen, Eugene, OR, US) which used the blunt ended pCR®-Blunt II-TOPO®, which comes as a linearised vector with a *Vaccinia* virus DNA topoisomerase I covalently bound to each of the 3$'$ ends. This kit therefore also dispenses with the need for a DNA ligase enzyme. This vector also has the advantage that it does not require colour screening as it contains a fusion of the C-terminus of the LacZα fragment and lethal gene *ccd*B [122]. If recombination does not occur, cells containing the non-recombinant plasmid are killed on plating. This vector carries a kanamycin resistance gene, rather than ampicillin, so all plating was done on kana NB2 plates, and X-Gal and IPTG were not required as the vector is self-screening.

Chemically competent cells were used for transformation so the high salt cloning protocol provided in the kit manual was used. A single cloning step was required by the kit: 4 μl fresh PCR product (the maximum recommended amount was added to 1 μl salt solution (supplied in the kit) and 1 μl pCR®II-Blunt-TOPO®. This was incubated at 23 °C in a waterbath for 10 minutes then immediately transferred to ice.

One Shot®TOP10 Chemically Competent *E. coli* (Invitrogen) was used for transformations following the protocol set out in the manual:

(i) 2 μl of cloned material from the previous step was added to a vial of

One Shot®TOP10 *E. coli* and incubated on ice for 20 minutes,

(ii) The cells were heat shocked at 42 °C for 30 seconds and immediately transferred to ice,

(iii) 250 μl of room temperature SOC medium was added to the vial and this was mixed thoroughly and then incubated at 37 °C shaken at 150 rpm for 1 hour,

(iv) 10, 20, 40 and 50 μl aliquots of the transformation material was spread (with purified water to add up to 50 μl) onto pre-warmed kana NB2 plates and grown overnight at 37C.

2.9.9.3 Cloning of DNA fragments into pUC19 vector

Where pUC19 was used in cloning DNA fragments the following protocol was used. pUC19 (Fermentas, St. Leon-Rot, Germany) was cut using restriction enzyme SmaI (New England Biolabs) to produce linearised DNA: 10 μl of 1000 μg/ml pUC19 (Fermentas) was placed in solution with 1 μl SmaI, 2 μl Reaction Buffer 4 and 7 μl purified water in a 1.5 ml Eppendorf and placed in a waterbath at 37 °C for 90 minutes. 5 μl was then run on a gel to check digestion against a control sample of pUC19.

The vector was then dephosphorylated: 15 μl of the digest was added to 2.5 ml Antarctic Phosphatase Buffer (New England Biolabs) and 7.5 ml Antarctic Phosphatase (New England Biolabs), and this solution was left in a waterbath at 37 °C for 30 minutes. The phosphatase was then heat inactivated at 65 °C for 5 minutes. At the end of this step, the vector had been diluted to a concentration of 3 μg/μl.

A ligation step was then done which required all constituents and reaction vessels to be prechilled on ice. 1 μl of cut, dephosphorylated pUC19 solution was added to 5 μl of PCR product, 2 μl T4 DNA Ligase Reaction Buffer, 4 μl T4 DNA Ligase (New England Biolabs) and 8 μl purified water all in a 1.5 ml Eppendorf tube. The reaction mixture was briefly spun down and flicked gently to mix, then left at 4 °C overnight.

Chemically competent *E. coli* (both of strain MG1655 and DH5α) was used for transformation of the ligated vector. The following standard protocol was used for non-kit transformations:

(i) an aliquot of 50 μl or 100 μl of chemically competent cells and the ligated vector were thawed on ice,

(ii) 10-20 μl of vector was added to the competent cells, mixed gently and left for 45 minutes (on ice),

(iii) the cells were then heat shocked in a waterbath at 37 °C for exactly 5 minutes,

(iv) the cells were then transferred into 5 ml NB2 solution and incubated at 37 °C at 150 rpm for 2 hours.

2.9.9.4 Transformation of cells by electroporation

Where chemical transformation failed electroporation was used for electrotransformation of several strains of *E. coli* in an attempt to produce viable clones. A MicroPulser^{TM}(BioRad, Hercules, CA, US) was used to produce the electrical pulses required for the electroporation. An aliquot of 50 μl of electrocompetent cells prepared in the manner described in Section 2.9.2 were thawed on ice for the electroporation and the following procedure was used for the transformation:

(i) 2 μl of DNA fragment amplified by PCR was added to the cells, it was mixed well and left for 1 minute,

(ii) The MicroPulser was set to 'Ec1',

(iii) the material was transferred to a 0.1 cm electroporation cuvette (prechilled on ice) which was placed in the chamber and pulsed once,

(iv) the cuvette was removed immediately, 1 ml SOC medium was added and the cells were mixed thoroughly,

(v) the material was then transferred to a 20 ml universal tube and incubated at 37 °C shaken at 225 rpm for 1 hour.

The resultant transformed cells were then ready for selection and screening as with those cells transformed by chemical means.

2.9.10 Transformant Screening by amp/IPTG/X-Gal

Various amounts of transformation material from transformation reactions from 10 μl to 100 μl were then plated onto amp/X-Gal/IPTG NB2 agar plates for blue/white screening to detect cells containing recombinant vectors. The principle of blue/white screening depends on the fact that *E. coli* lacking the *lacZ* gene cannot metabolise X-Gal, the metabolism of which produces a blue chemical. Vectors, such as pUC19, have a *lacZ* gene encoded across their multiple cloning site (MCS, wherein the restriction site for *Sma*I is situated) which is disrupted on the successful ligation of a DNA fragment into the MCS. There are then up to three sets of cells in the consequent transformed material: those without a transformed vector, those with a non-recombinant (self-ligated) vector and those with a recombinant vector (although if DNA has not been cleaned up properly this is not a guarantee that any recombinant vectors contain the DNA fragment of interest).

The selection and screening is then achieved using amp/IPTG/X-Gal NB2 agar plates. Ampicillin causes those bacteria without the vector at all to lyse, but allows all vector-carrying cells to grow. IPTG is a non-metabolisable inducer for transcription of the *lacZ* gene, which produces β-galactosidase. β-galactosidase cleaves X-Gal into galactose and 5-bromo-4-chloro-3-hydroxyindole. 5-bromo-4-chloro-3-hydroxyindole then oxidises into 5,5'-dibromo-4,4'-dichloro-indigo, an insoluble blue product that turns a colony blue when produced. Thus only colonies that have retained their white colour during growth contain recombinant vectors, with a disrupted *lacZ* gene.

Inoculated plates were left overnight at 37 °C to screen for white colonies. When found these colonies were picked and inoculated into 5 ml NB2 medium containing ampicillin in a universal tube and grown overnight at 37 °C shaken at 150 rpm. 2 samples of 1 ml of material were then taken from the universal (to provide plenty of vector DNA for analysis and re-transformation) and miniprepped according to the standard QIAprep protocol, including the Buffer PB wash step, eluting the DNA into 50 μl of Buffer EB.

When MG1655 was used for transformation blue/white screening could not be used. MG1655 contains a wild-type *lacZ* gene on its chromosome which has expression inducible by IPTG, thus colonies of transformed MG1655 would always be blue in the presence of X-Gal, irrespective of whether the vector was recombinant or not. Where MG1655 was used for a large recombinant transformation efficiency was very low, so all colonies produced on an amp/IPTG/X-Gal NB2 plate were picked for growth, miniprepping and restriction analysis of the vector.

2.9.11 Restriction digests

All restriction digests were done according to a standard procedure, unless otherwise stated. The procedure was as follows: 10 μl DNA solution from miniprep (or other source if required) was added to 2 μl of 10x reaction buffer of the type appropriate for the restriction enzyme, 1 μl of restriction enzyme and 7 μl of purified water. This was then incubated for 90 minutes at 37 °C in a waterbath. For digestion with more than one enzyme the appropriate buffer was determined from the NEBuffer Activity Chart (New England Biolabs) table of buffers, and the amount of water added was adjusted to produce a 20 μl solution. Controls were produced in parallel, but replacing the reaction buffer and enzyme with purified water and freezing for the 90 minute period during which the samples were being digested before being thawed to run on a gel. After digestion 5 μl loading buffer was added to each solution and mixed gently; 5 μl of this was then transferred to a gel and run alongside an appropriate DNA ladder to determine fragment lengths.

2.9.12 Gels

All gels used for electrophoresis were made with electrophoresis grade agarose (Invitrogen), at a concentration of 1.5 % w/v, with 3 μl of 10,000x SYBR Safe DNA Gel Stain (Invitrogen, Eugene, OR, US) in each 30 ml gel slab. Although this agarose concentration is quite high for large DNA fragments no problems were encountered when using it, so agarose concentration was not changed for electrophoresis of large fragments. Electrophoresis was carried out using a Horizon 58 Gel Tank (Biometra, Goetingen, Germany) powered by a Model 200 powerpack (Biometra). Photographs of gels were taken by a Gene Genius Bio Imaging System (Syngene, Cambridge, UK) with a short wave UV light transilluminator, using GeneSnap software version 7.05 (Syngene, Cambridge, UK). Loading buffer as described in Section 2.2 was used for all electrophoresis experiments. 5 μl of sample and 1 μl of Loading Buffer were used in each lane, run against 6 μl of a DNA ladder appropriate for the sample. DNA markers used are listed below:

(i) 100 bp DNA ladder (New England Biolabs),

(ii) 1 kb DNA ladder (Promega),

(iii) λ DNA-HindIII Digest (New England Biolabs).

2.10 Errors

Unless otherwise stated in the data presented throughout the Results Chapters error bars are all standard errors derived from standard deviations of the samples taken of each measurement. Most measurements were taken in triplicate, so standard error on each value was calculated as standard deviation divided by the square root of 3. Expressed mathematically,

$$I_{\frac{1}{2}} = S_E = \frac{\sigma}{\sqrt{n}} \tag{2.11}$$

where $I_{\frac{1}{2}}$ is the difference between the mean value and the upper and lower bounds of the interval, S_E is the standard error, σ is the standard deviation of the sample values and n is the number of samples. All rates of bacterial growth are calculated as doubling rates as this can be related to other quantities more clearly than growth rate.

Chapter 3

Computational Methods

3.1 General computer programs and programming languages

Many computer programs were used in data analysis and research during this project. Where programs had specific roles in the project they are described in the relevant Section below, but several programs were used for multiple tasks and are described here. Due to the rate of change of computer programs various versions of each program were often used during the course of this project, however only the latest versions used during the project will be mentioned and where they are not commercial programs the author(s) and/or relevant website is noted.

Spreadsheets were used extensively both to do multiple parallel calculations and for conversion of data (e.g. to put in a database). Spreadsheets were all manipulated using the OpenOffice Calc program (version 3.0.1; Sun Microsystems Inc., Santa Clara, CA, US).

MATLAB (The MathWorks^{TM}, Cambridge, UK) was used extensively both in its own right for finding solutions to technical mathematical problems, such as sets of differential equations, and as an interface with various other programs, allowing automation of many tasks. These uses will be described in more detail in the relevant Sections below.

CLC Sequence Viewer (version 5.0.1; CLC bio, Hemyock, Devon, UK) was used for *in silico* restriction mapping of vectors, ClustalX (version 1.81; `www.clustal.org`) was used for multiple alignments of DNA and protein sequences of the order of kilobases in length, the GNU Image Manipulation Program (version 2.4.2; `www.gimp.org`) was used for image processing, TreeView (version 1.6.6; Roderic D. M. Page) produced phylogenetic trees from alignments outputted by ClustalX, Artemis Release 10 [123] was used for visualisation of genome sequences and selection of subsets of genome sequences for further analysis and SigmaPlot (version 11.0; Systat Software Inc., Chicago, IL, US) was used for graph creation.

Perl (ActivePerl version 5.10.0; ActiveState, Vancouver, BC, Canada) was used to interface with biological databases online through the BioPerl (version 1.5.2; `www.bioperl.org`) toolkit of Perl modules. It has also been used to perform some file manipulation and automation of BLAST searches. Python (version 2.5.1; `www.python.org`) has been used for its GenomeDiagram module (version 0.21; Leighton Pritchard, `http://-bioinf.scri.ac.uk/lp/programs.php`) which produces diagrams of genomes and sections of genomes. Both of these programming languages have been edited and run through Komodo Edit (version 4.0; Activestate).

Databases were produced and manipulated in the MySQL (version 5.0;

Sun Microsystems Inc.) database language using both OpenOffice Base (version 3.0.1; Sun Microsystems Inc.) and phpMyAdmin (version 2.11.8.1; `www.phpmyadmin.net`) which is an internet browser based MySQL interface. Databases were stored on an Apache HTTP server (version 2.2.9; The Apache Software Foundation, `www.apache.org`).

3.2 Genomic Comparisons

Many genome sequence comparisons were required during this project and they were standardised so that they would be reproducible and comparable. Between any set of more than two genomes multiple comparisons between each pair of genomes were possible, but in practice a single genome sequence was chosen as a reference genome and all other genome sequences were compared to that one, unless particular comparisons between two strains were required. One of two strains of *E. coli* were chosen to be reference strains depending on the requirements of each comparison, MG1655 because of the large body of knowledge about it and CFT073 as the first UPEC to have a complete published genome sequence [42].

The Basic Local Alignment Search Tool (BLAST [23]) has been used in this project both online and as a standalone program on a local computer. Where whole genome sequence comparisons were conducted the query and reference sequences were both downloaded from GenBank at the time of comparison and used the local BLAST program, as this was amenable to automation of BLAST runs and BLAST output.

In outline the procedure for comparing genome sequences was by using open reading frame (ORF) sequences from the query genome sequence and using BLASTp to find the closest related proteins in the reference organism. Early comparisons were done by manual inspection, replaced by DiagHunter (described below) later in the project, to determine whether sets of genes encoding these proteins were retained in similar places on each genome or whether they were retained in groups of adjacent genes, but in different places on the genomes of the two bacterial strains.

Genome sequences were all taken from Genbank [25], although due to the constant revision of Genbank files sometimes comparisons earlier in the project were rendered obsolete. Where possible comparisons were redone with the latest sequence files, if they were going to be used for further analysis.

3.3 Discovery of Putative Metabolic Genes and sets of co-linear genes adjacent to them in CFT073

3.3.1 Genomic comparisons for CFT073 metabolic comparisons

The initial genome sequence comparisons for the project were carried out by Preben Krabben (unpublished), and the putative metabolic genes were initially identified by him. All further work including bioinformatic analysis of the clusters and experimental verification of the identified genes were carried out by the author. These comparisons were used to look for uncharacterised metabolic genes that might have a role in CFT073 metabolism in the urinary tract. For this CFT073 was used as the reference genome and other genome sequences were compared to it.

The CFT073 genome [42] was compared gene-by-gene using TBLASTN [23] to get a sequence identity value. This percentage identity from TBLASTN was combined with the position and percentage identity of nearest neighbours and overall position in the genome to infer synteny conservation. The use of BLAST scores to determine gene conservation is well established (such as the BSR technique [124]) and this process was refined by adding a neighbour dependent analysis to determine gene synteny. For sets of two or more genes in a similar position on two genomes, retention of function of each gene was inferred by identity with a cut-off of 90% over the whole length of each gene. Single genes in similar positions on the two genomes were inferred as conserved only if their mutual identity was above 95% over their whole length.

3.3.2 Selection of genes for further analysis

Genes that were in CFT073 but not present in MG1655 according to the initial genomic comparison were selected for further analysis, since the MG1655 phenotype and genotype are very well characterised, and it does not colonise the urinary tract [125]. This produced an initial list of candidate genes that could be implicated in the metabolism of CFT073 in the urinary tract, but have not yet been characterised. This set of genes was inspected manually for those genes which appeared to have a metabolic function from their GenBank annotation [126], but without a definite, specific biological function, henceforth referred to as putatively metabolic genes (PMGs).

Where there was a PMG surrounded by uncharacterised genes, this region of the genome sequence was viewed using the NCBI's Sequence Viewer 2.1 and all adjacent genes transcribed in the same direction as the PMG and less than 100 base pairs separated were inferred to be part of a set of adjacent colinear (SAC) genes and included in the analysis as they could potentially be parts of an unidentified operon. The algorithm for inferring whether each set of genes was present in each of the *Escherichia* genomes considered was as follows: where the set was greater than 3 genes in length it was considered present if at most one of the genes was absent according to the synteny comparison; where there were 3 or fewer genes, all genes had to be present to conclude that the operon was present.

BLASTp was then used on each gene in the sets of genes against the full non-redundant database of Genbank [127] to try to find homologies to already annotated genes. Where homologs could not be found protein domain similarities were sought using Pfam [128] and SEED (http://theseed.uchicago.edu/FIG/index.cgi) in an endeavour to elucidate their function. Further, the NCBI's Conserved Domain Database [129] was searched for conserved domains.

```
85674276    16127996    0.0   0     820
85674276    16131778    1     090   810
85674276    16131850    2     040   449
85674277    16127997    1     177   310
85674277    145698255   0.37  0     1300
85674277    16128725    0.65  0     382
21321894    16127998    0.0   0     428
21321894    16128864    1.3   0     208
21321894    90111123    1.4   0     324
21321895    16127999    8     014   98
21321895    16130322    4     006   108
21321897    16128000    1     146   258
21321897    16129929    0.082 0     316
21321897    16128835    2.5   0     276
.           .            .     .     .
.           .            .     .     .
.           .            .     .     .
```

Figure 3.1: This is example output of an individual BLASTp comparison using the parameters described in the text. See text for description of the columns.

3.4 Full genome comparisons for metabolic network comparison

3.4.1 Gene-by-Gene Comparisons

For the genome synteny comparisons used in Chapter 6 individual comparisons between ORFs were required. BLASTp was used with the following non-default arguments: '-IT -m8 -b3', which produces output such as that seen in Figure 3.1. Although automation of the BLAST comparisons was used so multiple BLASTp comparisons were done together, each BLAST was conducted identically as if it had been run by itself. Results were then concatenated into a single file for further analysis, as described below.

ORFs were then taken one at a time and compared to the protein database of the reference strain. Non default parameter values were required for output from the program to be in the form of a list of at most three hits on the reference genome sequence per ORF (in the interests of computational efficiency) and in the form of a summary table without comments, which was conducive to automatic parsing for DiagHunter. Figure 3.1 shows sample output from the BLAST search conducted. The first column is the GI number of the query ORF and the second is the GI number of the ORF with a similarity in the reference protein database. The third and fourth columns are values X and Y respectively in the expression $Z = X \times 10^{-Y}$ where Z is the expectation that the two sequences have this level of similarity by chance. So for instance if the second line of Figure 3.1 is taken $Z = 1 \times 10^{-90}$ so it is very likely that the sequences have a common ancestor gene. Where both columns have the value 0 this represents an expectation value too close to zero to be expressible in the format illustrated above. The final column is a BLAST score value.

3.4.2 Gene position data and DiagHunter (the DH method)

DiagHunter [130] was used to find sets of genes transcribed in the same relative orientation and in similar relative positions (genes adjacent to each other) in two genomes. A standard method and set of parameters was used for gene synteny inferences as described below, apart from the 'lmin' parameter, which determined minimum numbers of adjacent genes to use for the inference of synteny by DiagHunter. This will be set on comparison with other methods of gene synteny inference in Section 6.4.1.3. This method of gene synteny inference will be termed the DH method for convenience through the course of this Thesis.

DiagHunter takes as input specially formatted BLASTp output, along with relative position and orientation data of genes in two compared bacterial strains, which is not part of the BLASTp output. This has required custom formatting that was achieved using MATLAB and a purpose built (MySQL) database of gene information for each strain. Each database contained a record for each inferred protein coding gene in the strain. The record contained the protein GI number which was used as a reference for finding the two appropriate records for each particular BLASTp result. The record also contained a counting index number indicating the order of all genes in a particular strain as well as the orientation of the gene, either on the Crick or Watson strand. For each BLASTp hit contained in the BLASTp output from Section 3.4.1 the GI numbers of the relevant genes (both query and reference) were looked up in the appropriate database and position and orientation data were collected and placed in a parsed file in the correct format for input into DiagHunter.

An example of the correct input format for DiagHunter can be seen in Figure 3.2, which is the output if the values in Figure 3.1 were parsed. The first and third columns are the GI numbers of the relevant proteins, but the significant addition to this file is the position and orientation data in the second and fourth columns. In each of these columns the numerical value is the number of genes counted from the origin of the genome to that gene, in the direction of transcription of the Thr operon leader peptide, which is encoded by gene 1 in this counting scheme. The letters in these columns represent the strand on which the protein is encoded, either the Watson (‘w’)

```
85674276    103w  16127996    2w    0     0     820
85674276    103w  16131778    3704w 1     90    810
85674276    103w  16131850    3775c 2     40    449
85674277    104w  16127997    3w    1     177   310
85674277    104w  145698255   1337w 0.37  0     1300
85674277    104w  16128725    695c  0.65  0     382
21321894    105w  16127998    4w    0     0     428
21321894    105w  16128864    835c  1.3   0     208
21321894    105w  90111123    342c  1.4   0     324
21321895    106w  16127999    5w    8     14    98
21321895    106w  16130322    2262w 4     6     108
21321897    107c  16128000    6c    1     146   258
21321897    107c  16129929    1879c 0.082 0     316
21321897    107c  16128835    806w  2.5   0     276
.           .     .           .     .     .     .
.           .     .           .     .     .     .
.           .     .           .     .     .     .
```

Figure 3.2: This is example input ready for DiagHunter, parsed from BLAST results such as those shown in Figure 3.1. For description of columns see text.

or the Crick (‘c’) strand. The final three columns are taken directly from the final three columns of the BLASTp output, indicating expectation value and BLAST score for each hit.

DiagHunter was then run using this parsed list of BLAST hits as its input. The performance settings used for DiagHunter were:

(i) Compress factor: 1, because gene order rather than absolute position was used,

(ii) Use orientation: 1, to use the orientation data available,

(iii) Score_cutoff: 20, to remove BLAST hits with poor BLAST scores,

(iv) Min_diag_len: variable, to assess the impact of this factor on gene inferences

(v) Min_diag_qual: 11, the running average sensitivity required for elongating a diagonal run of genes.

The outputs of DiagHunter are a diagram of gene synteny hits and a file listing sets of adjacent genes conserved between the two strains in the comparison. This list is then analysable itself looking at various comparisons and giving genes conserved across strains, but these lists of genes can also be used to infer metabolic models through knowledge of metabolic gene function.

3.5 Metabolic Models

There are many types of metabolic model taking into account various levels of knowledge about the physicochemical constraints within the system of interest, usually a cell. A summary of these models can be found in Section 1.3, but the models used in this project are stoichiometric whole-cell models, with very little by way of metabolic regulation constraining the model capabilities. Such models consist of several parts that when combined can be interrogated to find solutions that satisfy certain objective functions, by the minimisation or maximisation of such functions.

This solution in the case of stoichiometric models is a set of rates for every reaction that produces the desired objective. In the case of stoichiometric models covering the whole metabolism of a bacterium, it includes rates of metabolite uptake from the cell surroundings, rates of excretion of all waste products and, when included, the growth rate of the bacterium.

The first piece of the model is a matrix where each row represents a metabolite and each column represents a reaction; where a metabolite i is used up (or produced) in a reaction j the cell (j, i) has a value equal to minus (or plus) the amount used (or produced) in the reaction when the reaction is stoichiometrically balanced. For instance, in the reaction catalysed by acetolactate synthase two pyruvate molecules and one hydrogen ion react to produce one 2-acetolactate molecule and one carbon dioxide molecule. Thus in the column representing this reaction, in the cell representing pyruvate the value would be -2, in that for the hydrogen ion it would be -1 and for both 2-acetolactate and carbon dioxide the value would be 1. All other cells have values of 0.

Each reaction has limits to its rate: some reactions cannot run backwards (irreversible reactions) and all are limited by the capabilities of the enzyme catalysing the reaction. Further physicochemical conditions (such as pH, temperature and reactant and product concentration) have an effect on the rate of all reactions, but these are not taken into account for the purposes of whole genome metabolic models. The constraints on reaction rates are in the form of minimum and maximum rates for each reaction. Unless specifically constrained for a specific task at hand reaction limits are set to ∞ and $-\infty$ in the forward and reverse directions respectively except where a reaction is irreversible in which case the minimum rate is set to zero. Further, where external metabolites are taken up by the cell care is needed to limit the rate of uptake of these metabolites to realistic levels.

The last parts of the model required define the nature of the metabolite balancing achieved in the model. Solutions to models of this type rely on metabolite balancing, representing a steady state solution to the problem, where there is no net production or consumption of any individual metabolite. The right-hand side coefficients and senses (equal to, greater than or less than) of the equations (equality or inequality) produced from the dot multiplication of the stoichiometric matrix and the solution (vector of reaction rates) must be defined before a solution can be sought. Usually a steady state assumption is made about the cell, that no metabolites are net produced or net consumed in the cell during the period modelled. Thus all senses are set to equality and all right hand side coefficients are set to 0.

3.5.1 iAF1260 and related models

The model used as a basis of all the models in this project is that of Feist *et al* [1], called iAF1260. The aspects of the model described here are those which are important for the specific uses for which the model has been used, or modifications to the original model. It is the most complete stoichiometric model of *E. coli* metabolism publicly available and includes gene-protein-reaction (GPR) relationships for all the reactions for which these data are known. GPR relationships relate the genes present in an organism's genome (in this case that of *E. coli* K-12 MG1655) to the reactions which the bacterium is capable of catalysing. This link comes from the enzymes catalysing the reactions in question, which are coded for in the genes of the bacterium. So for instance the gene locus tag b0063, gene name araB encodes the enzyme L-ribulokinase which catalyses the reaction with EC number 2.7.1.16, the phosphorylation of ribulose. This represents information about other bacteria which also contain this gene, that they are in principle capable of catalysing the same reaction within themselves. Complications to GPR relationships occur in the form of multi-subunit enzymes and multiple enzymes catalysing the same reaction, but these are taken into

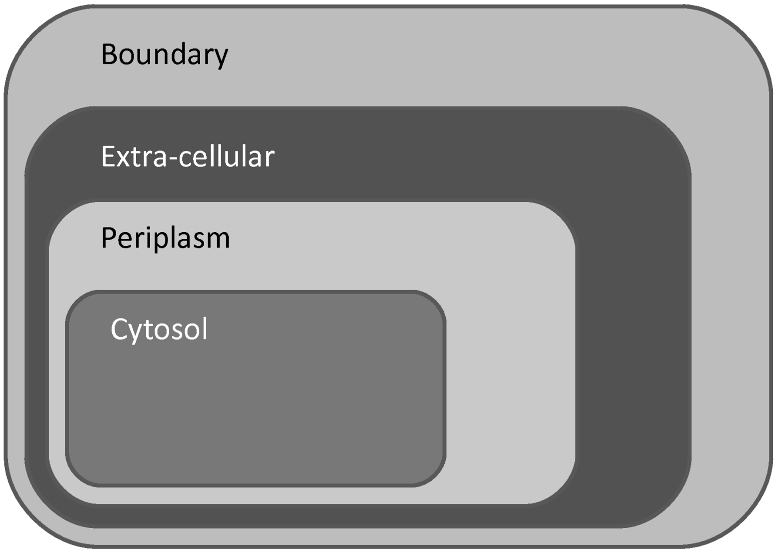


Figure 3.3: This diagram shows the compartments used in the stoichiometric models used in this project.

account in the iAF1260 GPR relationships.

An important aspect of the model iAF1260 is that it is partitioned into physically distinct compartments, that is, the cytosol, the periplasm and the extracellular environment. Metabolites present in any particular compartment are distinct from identical metabolites in other compartments as they are the only metabolites available for reactions within that compartment, and are not available for reactions in other compartments. The model therefore distinguishes metabolites in each different compartment. ‘Metabolites’ used in construction of the stoichiometric matrix are therefore defined not only by the metabolite itself, but by the position of the metabolite in the system. Transport across boundaries (membranes) between compartments (or reactions involving metabolites in more than one compartment) can occur by several processes, but each process must be defined by a reaction in the model, so diffusion across a membrane would be represented by a reaction like: 1 acetoacetate (extra-cellular) $\Leftrightarrow$ 1 acetoacetate (periplasm).

A further compartment has added to the model, that of the boundary, producing the layout seen in Figure 3.3. This compartment represents the region distant from the cell with which the cell cannot interact directly. This necessitates two sets of reactions, one set being diffusion between the extracellular compartment and the boundary compartment and another set representing the destruction or creation of ‘boundary’ metabolites in the boundary compartment. The boundary thus represents the medium in which the cells exist and from which the cells can draw nutrients - which can be depleted or added to in a modelling situation, as in a real situation. The destruction or creation of metabolites is represented in the model by one-sided reactions like the following: 1 water molecule (boundary) $\Leftrightarrow$.

3.5.2 Inferring metabolic models from synteny comparisons

The steps to creating an inferred metabolic model for a fully sequenced bacterium are fairly simple in principle. The genes inferred to be present by the method described in Section 3.2 are used to see which reactions could be catalysed by enzymes encoded by those genes according to the GPR relationships given with the iAF1260 model. Information about the GPR relationships in the model are in the form as shown in this example: ‘((b0936 and b0933 and b0934) or (b0365 and b0366 and b0367))’ where bXXXX is the Blattner number, as defined in GenBank, of the gene for a single reaction (in this case butanesulfonate transport via ABC system - periplasm). In these relationships ‘or’ means either one or the other or both and ‘and’ means that all of the genes connected in a row by ‘and’ must be present for an enzyme to be present. Thus either the enzyme complex

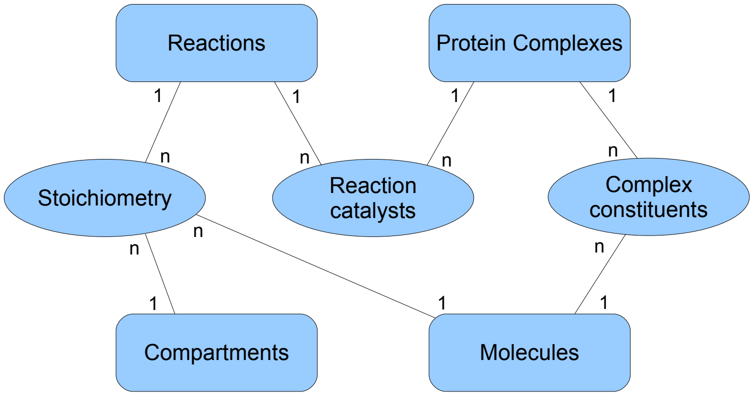


Figure 3.4: The database structure containing the information making up iAF1260. Rectangles with rounded corners represent entities and ovals represent links between those entities. Lines representing the relationships between the tables show the 1 to many relationships that exist in this database. Further details of the tables can be found in the text.

encoded by b0936, b0933 and b0934 together or the complex encoded by b0365, b0366 and b0367 together will catalyse the transport reaction. It does not matter for the purposes of this model if both sets of genes are present in the same genome.

These logical relationships between genes, proteins and reactions was represented for the purposes of building new metabolic models in a MySQL database with the structure detailed in Figure 3.4. The database consists of 7 tables split into entity tables and link tables. The entities are: reactions, compartments, molecules (including single atom species and gene products) and protein complexes (enzymes).

The first linking table represents the stoichiometry of the reactions. Each reaction has a set of records in the Stoichiometry table, one for each metabolite (molecule) in a specific compartment in that reaction. The record therefore consists of a reaction index number, a molecule index number, a compartment index number and the stoichiometry of that particular metabolite in the reaction. This table is for the purposes of populating the stoichiometric matrix when the reactions in a particular model are known. The records in ‘Stoichiometry’ associated with each present reaction are used to retrieve the correct stoichiometry for the model.

The second linking table is ‘Reaction catalysts’ that holds a record for each enzyme that catalyses each reaction. Each record therefore consists of the enzyme index number and the index number of a reaction catalysed by that enzyme. This table enables the presence or absence of available reactions in a model to be inferred from the complement of enzymes available in the model. The records in ‘Reaction catalysts’ associated with each present enzyme produce a list of reactions available.

The third linking table is ‘Complex constituents’, which describes the molecules required for the creation of each viable enzyme. Each record contains the index of a ‘Protein complex’ and the index of one of its ‘Molecule’ constituents. These constituents are the peptides encoded in the genes contained on the genome of the relevant organism.

This completes the necessary information to produce a metabolic model from the gene complement of an organism. The genes present determine the enzymes that have all their constituents and are functional; this list of enzymes constrains the list of reactions to those catalysable by the functional enzymes; the stoichiometric matrix can then be built according to the reactions available for the organism.

Several points are worth noting about this database. It contains quite a few reactions that are not catalysed by enzymes, or for which the catalysing genes have not been identified. The inference mechanism therefore checks each time it is run to see whether there are any enzymes (present or absent) that catalyse each reaction and where there are no potential enzymes the reaction is assumed to be present in the model.

There is also a limitation on the applicability of the model because using only the GPR relationships any reactions not present in MG1655 cannot be inferred to be present in the model. Where this was the case and further reactions were required these were added along with their GPR relations into the database and the presence of the required genes was determined manually.

An SQL dump of the database can be found in Supplementary File 'iaf1260.sql' and an Excel dump can be found in Supplementary File 'iaf1260.xls'.

3.6 Growth Models

Growth models were produced through the GPR relationships in iAF1260 for *E. coli* CFT073 for the simulations in Chapter 7. These were achieved with an additional table in the database consisting of a list of GI numbers of genes in MG1655 for which there were functional eqivalent genes in CFT073 (according to the genome comparisons carried out in Chapter 6), called 'gene_list'. The exact SQL query for production of the database can be seen in Supplementary File 'combined_presence.qry', but is described here for clarity. This is the algorithm used for producing new models:

(i) enzyme complexes with all their genes 'gene_list'were found,

(ii) all reactions without associated complexes (uncatalysed reactions) were found,

(iii) all reactions either in the uncatalysed list or catalysed by complexes found in step i) were combined to give a complete list of reactions,

(iv) all metabolites present in each of these present reactions were added to give a list of metabolites,

(v) the stoichiometries from the linking table ‘stoichiometry’ were then added to a list of all stoichiometries present.

Models were passed straight from the MySQL database to MATLAB and were used in MATLAB for linear optimisations and other model calculations. The simulations produced in Chapter 7 are described in that Chapter for convenience so that the simulation results and underlying calculations can better be compared.

Chapter 4

Isolating metabolic genes of interest: a bioinformatic and experimental approach

4.1 Preface

The complete genome sequence of *E. coli* CFT073 has opened further the possibility of investigating the metabolic basis (through metabolic genes) of bacterial survival during colonisation of the urinary tract. This is of great interest as a thorough knowledge of the metabolism of such UPEC would have the potential to offer new ways of treating urinary tract infections. The approach to functional genomics of using multiple genome sequences to infer commonalities and differences between bacterial strains living in different habitats has been used in the context of UPEC to an extent by others. Several strains' genome sequences have been compared to find common metabolic genes [62] and in this project the approach has been expanded to use all the available sequences of closely related strains (at the time of original comparison) and to investigate partially characterised metabolic genes and operons.

In this Chapter a complete gene synteny comparison has been carried out between nineteen fully sequenced strains from the genus *Escherichia* with the aim of finding as yet uncharacterised genes implicated in the metabolism of uropathogenic strains of *E. coli* (UPEC). Several sets of adjacent colinear genes have been identified which are present in all four UPEC included in this study (CFT073, F11, UTI89 and 536), annotated with putative metabolic functions, but are not found in any other strains considered.

An operon closely homologous to that encoding the L-sorbose degradation pathway in *Klebsiella pneumoniae* has been identified in *E. coli* CFT073; this operon is present in all of the UPEC considered, but only in 6 of the other 15 strains. The operon's function has been confirmed by cloning the genes into *E. coli* DH5α and testing for growth on L-sorbose. The functional genomic approach combining *in silico* and *in vitro* work presented here can be used as a basis for the discovery of other uncharacterised genes contributing to bacterial survival in specific environments.

4.2 Results and Discussion

4.3 Genome sequences

The genome sequences used in this Chapter can be found in Table 4.1, which indicates pathotype where known and indicates which have an inferred L-sorbose uptake and utilisation operon. The set of genome sequences used in this Chapter will be referred to as the EGSS (*Escherichia* genome sequence

Table 4.1: Strains of genus *Escherichia* used in this study, all with completely sequenced genomes or whole genome shotgun sequences freely available from GenBank. 'GI tract' is the gastrointestinal tract. Unless otherwise indicated they are *Escherichia coli*.

Strain	Type	Sorbose Operon	Source / Accession Number
CFT073	UPEC (uropathogenic)	+	AE014075
F11	UPEC	+	AAJU00000000
536	UPEC	+	CP000247
UTI89	UPEC	+	CP000243
042	EAEC (enteroaggregative)	–	Sanger Center
B7A	ETEC (enterotoxigenic)	–	AAJT00000000
E24377A	ETEC	–	CP000800
B171	EPEC (enteropathogenic)	–	AAJX00000000
E22	EPEC	–	AAJV00000000
E2348	EPEC	–	Sanger Center
E110019	EPEC	–	AAJW00000000
53638	EIEC (enteroinvasive)	–	AAKB00000000
MG1655	Commensal (GI tract)	–	U00096
HS	Commensal (GI tract)	–	CP000802
O157:H7 str. Sakai	EHEC (enterohaemorrhagic)	+	BA000007
O157:H7 EDL933	EHEC	+	AE005174
Shigella sonnei 53G	Bacillary Dysentery	+	Sanger Center
Shigella flexneri 2a str. 301	Bacillary Dysentery	+	AE005674
Shigella dysenteriae Sd197	Bacillary Dysentery	+	CP000034

set) for convenience, although in later Chapters the analysis is widened to include further bacteria, the sequences of which became available more recently than this analysis was done.

4.3.1 Multiple genome comparison

The results of the complete synteny comparison of the genome of CFT073 against the rest of the EGSS can be seen in Supplementary Table 1.

Overall 133 putatively metabolic genes (PMGs - as defined in Section 3.3.2) were inferred; they can be seen in Supplementary Table 2, along with the results of an NCBI Conserved Domain Search, and are summarised in Table 4.2. All of the individual genes marked with putative metabolic func-

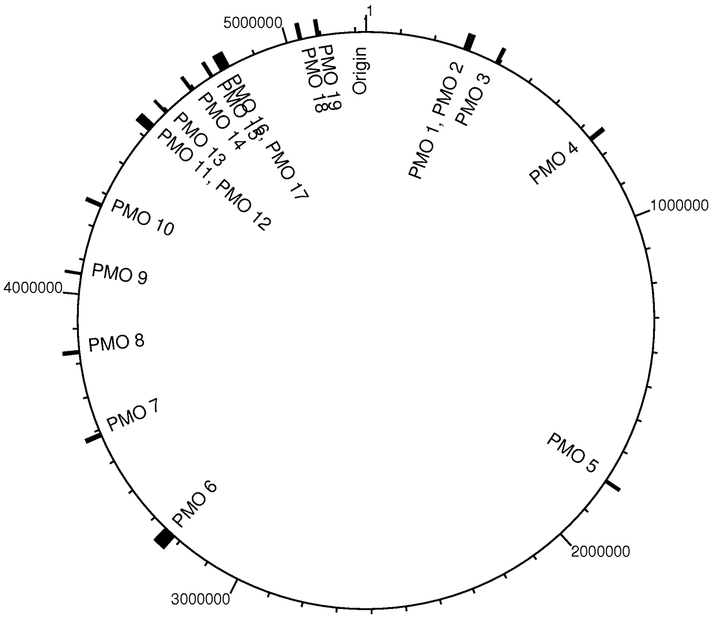


Figure 4.1: Positions of the SACs identified in the genome sequence of *E. coli* CFT073, as labelled in Table 4.2.

tions in the CFT073 genome sequence were positioned adjacent to genes transcribed in the same sense, so sets of adjacent genes colinear to these PMGs (SACs) were included in the investigation, a summary of which can be seen in Table 4.2, and positions of which in the CFT073 genome can be seen in Figure 4.1. It should be emphasised that the criterion for consideration of these genes was only that the genes marked with putative functions be absent from *E. coli* K-12 MG1655, without any further consideration of whether they would specifically be useful in the lifecycle of CFT073. 18 of these genes are present in all of and only in the UPEC and 49 genes are among those identified by Lloyd *et al* [101], which compared the gene content of 7 additional UPEC and 2 different faecal strains of *E. coli* by comparative genomic hybridisation against CFT073 to find those genes unique to uropathogens.

The synteny comparison shows several characteristics indicating a higher

Table 4.2: Synteny conservation of sets of adjacent colinear genes in 18 sequenced strains of genus *Escherichia*; all these sets of genes are present in *E. coli* CFT073. Those genes which are part of genomic islands, as identified in [101], are marked with an asterisk.

SAC no.	Gene c numbers	No. of genes	Putative function	*E. coli* F11	*E. coli* 536	*E. coli* UTI89	*E. coli* E2348	*E. coli* O42	*E. coli* B171	*E. coli* E24377A	*E. coli* E22	*E. coli* B7A	*E. coli* E110019	*E. coli* 53638	*S. sonnei* 53G	*E. coli* HS	*E. coli* K12 MG1655	*E. coli* O157H7	*E. coli* O157H7 EDL933	*S. flexneri* 2a str. 201	*S. dysenteriae* Sd197
1	c0317 c0323	7	Polysaccharide biosynthesis*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2	c0330 c0333	4	Fucose metabolism	+	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-
3	c0409 c0415	7	2,5-diketo-D-gluconic acid metabolism	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-
4	c0757 c0765	9	Chorismate biosynthesis	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-
5	c1955 c1960	6	PTS system, cellobiose specific	-	-	+	-	-	+	+	+	+	+	+	-	+	-	-	-	-	-
6	c3405 c3410	6	PTS system, maltose / glucose specific*	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-
7	c3750 c3756	7	5- or 6-carbon sugar metabolism	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8	c4013 c4018	6	Carbohydrate metabolism	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
9	c4276 c4280	5	PTS system, galactitol specific	-	-	+	-	+	+	-	+	-	-	+	-	-	-	+	+	-	-
10	c4481 c4488	8	PTS system, fructose specific	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-
11	c4756 c4759	4	PTS system, glucose specific	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-
12	c4760 c4780	21	Entner-Doudoroff pathway	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
13	c4828 c4830	3	Shikimate metabolism	+	+	+	+	-	-	-	-	-	+	-	+	+	-	-	-	-	-
14	c4924 c4926	3	Citrate metabolism	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-
15	c4981 c4987	7	L-sorbose metabolism	+	+	+	+	+	-	-	-	-	-	-	+	-	-	+	+	+	-
16	c5020 c5025	6	3-ketoacid metabolism	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
17	c5030 c5041	12	2-oxoglutarate metabolism	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-
18	c5298 c5303	6	3-ketoacid metabolism	+	+	+	+	-	+	+	+	+	+	+	-	+	-	-	-	-	-
19	c5346 c5351	5	Arginine metabolism	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

prevalence of the genes from CFT073 in the other UPEC than in the other strains. It is worth noting that the sets of adjacent colinear genes identified in this study are not in general parts of the large pathogenicity islands identified in CFT073 by others [101, 42], with the exception of those marked with an asterisk in Table 4.2.

SAC No. 6 (as labelled in Table 4.2) is within a large pathogenicity island, PAI-CFT073-*metV* (according to the nomenclature set out recently by Lloyd *et al* [101]), in which the SAC is restricted to the area c3405 to c3410. These genes are retained in all the UPEC and in *E. coli* E2348, but not in any other strain. Retention of the sorCDFBAME genes (SAC No. 15) in some of the EAEC and EPEC and all of the *Shigella* is consistent with the findings of Lehmacher and Bockemühl [131] who despite the negative phenotype showed that they retain the DNA for many of these genes.

It was found that 121 of the 133 PMGs identified are present in the same position in all four UPEC; these include the putative genes for L-sorbose degradation. The only SAC not present in any of the UPEC other than CFT073 is No. 1. Those found in the same place in all UPEC, but in none of the other strains, are 8, 12 and 19. The SACs identified here have a tendency to be present or absent as a whole, rather than on a gene-by-gene basis, which implies that their functions might be co-ordinated and co-dependent, eg. as enzymes for transport and metabolism of the same chemical.

An investigation of each of the PMGs present in all UPEC and absent from all others was conducted using BLAST, the SEED tool, Pfam and the NCBI Conserved Domain Database, in an attempt to identify putative functions for all the genes. SAC No. 8 (genes c4013 to c4018) has several genes annotated putatively already, c4015 to c4017 as part of a ribose ABC transporter, and c4018 as a tagatose 1,6-diphosphate aldolase. The hypothetical genes bear similarities to other sugar metabolism encoding genes: c4013 to a dehydrogenase and c4014 to a sugar kinase, possibly a fructokinase. These genes could encode enzymes for the uptake and catabolism of a 5- or 6-carbon sugar or sugar derivative.

SAC No. 19 (c5346 to c5351) has previously been identified as an arginine catabolic operon [62]. It contains three hypothetical proteins and the following annotated proteins: c5348 is annotated as an ‘ornithine carbamoyltransferase chain I’, c5349 as a ‘carbamate kinase’ and c5350 as an ‘arginine deiminase’. Gene c5346 resembles an ArgR arginine transcriptional repressor and c5347 bears a strong similarity to DcuC, a ‘C4-dicarboxylate anaerobic carrier’, though it is annotated in SEED as containing an ‘arginine/ornithine antiporter ArcD’. According to the MetaCyc [132] species comparison tool there are very few CFT073 genes annotated for arginine catabolism, but the synteny comparison carried out here implies that there are genes encoding the catabolism of L-arginine to succinate via agmatine, putrescine, 4-aminobutarol, 4-aminobutyrate and succinate semialdehyde. MG1655 contains an arginine ABC transporter, but lacks a verified L-arginine/L-ornithine antiporter, putatively present in CFT073. This leaves the pathway catalysed by this set of genes unknown; potentially according to SEED annotations, c5350 could produce citrulline from L-arginine and c5348 could produce L-ornithine from citrulline, bypassing L-arginino-succinate in the urea cycle. The third reaction (c5349), putatively producing carbamoyl-phosphate from ammonia, would seem to be necessary only if fine control of this reaction needed to be coupled to the other reactions encoded by this set of genes, as there is a gene of the same function elsewhere in the genome (c0635).

Since there is a limited number of carbon sources present in urine - predominantly urea, uric acid and creatinine (and L-sorbose in small amounts) - it might seem plausible that the sets of genes identified here may encode proteins for the utilisation of these chemicals. However, an analysis of the genes investigated here has so far failed to find any good matches to known metabolic pathways for these three carbon containing compounds. The putative arginine utilisation operon described above encodes enzymes that catalyse reactions whose metabolic functions are close to those of urea catabolism though the only product of urea possible in the reactions annotated in MetaCyc is ammonia. UPEC are not confined to using the carbon sources in urine; they adhere primarily to the epithelial cells in the urinary tract so potentially the mucus produced by these cells could be used to fulfil the metabolic needs of UPEC.

The benefit of D-serine utilisation genes for UPEC has been suggested [62], so these genes were investigated in the UPEC included in this study to assess their relative prevalence. It was noted that according to the initial synteny comparison (Supplementary Table 1) these genes are not present in *E. coli* F11 or 536. However there are D-serine utilisation genes in both F11 and 536 at an alternative position in the genome identified by Brzuszkiewicz *et al.* Moreover, UTI89 also has D-serine utilisation genes in this alternative position on its genome as well as those found in this study. None of the non-UPEC in the EGSS had this alternative operon, which is characterised by a particularly large intergenic region ($\sim$1 kb) adjacent to it, conserved between F11, 536 and UTI89. This D-serine utilisation operon is therefore unique within these strains to the UPEC, indicating that it was present in a common ancestor of only these strains. The two D-serine operons present in UTI89 have a mutual identity along their entire length of 96 %, which indicates that there may have been a duplication event at some point in its history. This also indicates that the differing positions of the D-serine operons are not due to movement of the genes, but by duplication and subsequent loss of the first operon.

4.3.2 Putative L-sorbose utilisation operon *in silico* analysis

SAC 15 is noted in Table 4.2 as a putative L-sorbose metabolism operon. This was based on the annotations of the CFT073 genome sequence of the genes c4981-c4987, which indicated a sorbose, mannose or related compound PTS system. L-sorbose [133] can be present in urine (and in the gastrointestinal tract). It can be seen from Table 4.2 that it is predominantly the UPEC and *Shigella* strains which contain the operon enabling use of L-sorbose as a carbon source.

The putative L-sorbose operon in CFT073 was compared to that of *Klebsiella pneumoniae* using BLASTp. Both of these strains are in the family *Enterobacteriaceae* and the function of *Klebsiella*'s operon has been experimentally verified [118]. The results of this comparison are shown in Table 4.3, which shows identity above 90% for all but one of the genes.

A phylogenetic comparison between the putative L-sorbose operons from those sequenced bacteria of the genus *Escherichia*, using *Klebsiella pneu-*

Table 4.3: Comparison of SAC No. 15 in CFT073 to the *Klebsiella pneumoniae* L-sorbose degradation operon.

CFT073 gene ID	CFT073 name	*Klebsiella* name	*Klebsiella* annotation	Percentage Identity
c4981	Putative oxidoreductase	L-sorbose 1-phosphate reductase	SorE	91
c4982	PTS system, mannose-specific IID component	second subunit of EII-Sor	SorM	97
c4983	PTS system, mannose-specific IIC component	first subunit of EII-Sor	SorA	95
c4984	Putative sorbose PTS component	EIII-B Sor PTS	SorB	92
c4985	Putative sorbose PTS component	EIII-F Sor PTS	SorF	82
c4986	sorbitol-6-phosphate 2-dehydrogenase	D-glucitol-6-P -Dehydrogenase	SorD	92
c4987	Putative transcriptional regulator of sorbose uptake and utilisation genes	Sor regulator	SorC	92

moniae (GI: 150953431) and *Klebsiella oxytoca* (see acknowledgments for reference) as an outgroup, was conducted using ClustalX [134]. BLAST was used to extract the putative L-sorbose operons for most of the bacteria, but Artemis [123] was required to extract the relevant parts of both the *Klebsiella* strains and F11, which did not at the time have complete single contig genome sequences.

This ClustalX comparison of the putative sorbose operons of 14 strains of bacteria produced the phylogenetic tree shown in Figure 4.2. These bacteria are all of genus *Escherichia* (except the *Klebsiella* strains) and include several *Shigella* strains, *Shigella boydii* Sb227 (Accession Number: CP000036.1) and *Shigella sonnei* Ss046 (Accession Number: CP000038.1), not used in the full genome analysis because their genomes are not completely sequenced. The operons were located in the genome sequence thus far generated for them and extracted to compare to the others in the EGSS. Also included is the inferred L-sorbose operon from the recently published

Avian Pathogenic *Escherichia coli* O1:K1:H7 (APEC 01) [39] (Accession Number: CP000468.1), which clusters with the UPEC, separate from the enterohaemorrhagic *E. coli* (EHEC) and *Shigella* strains. The EHEC and *Shigella* L-sorbose operons are grouped together. Two of the operons are split (those of O157:H7 strain Sakai and *Shigella dysenteriae* Sd197) by DNA insertions, but for the purposes of the ClustalX computation the insertions were removed and the homologous parts of the respective genes were allowed to line up independently.

4.3.3 Experimental verification of function of genes c4981 to c4987 from *Escherichia coli* CFT073

The genes c4981 to c4987 were successfully cloned into the pSC-B plasmid for experimental verification of function according to the method in Section 2.9 and confirmed by sequencing as being oriented as shown in Figure 4.3. The growth of the three strains in minimal media containing either glucose or L-sorbose as sole carbon source can be seen in Figure 4.4. The negative control, DH5α with pUC19, was unable to utilise this carbon source. DH5α containing the pQR793 plasmid (called DH5αL from now for convenience) grew using L-sorbose as the sole carbon source.

The functions of genes c4981 to c4987 from *Escherichia coli* CFT073 have therefore been confirmed as those encoding a pathway for the utilisation of L-sorbose as a sole carbon source. The very high similarity between this operon (along its entire length) with the operon identified and experimentally characterised in *Klebsiella pneumoniae* [135] implies that not only do the two operons both utilise L-sorbose, but that the same pathway

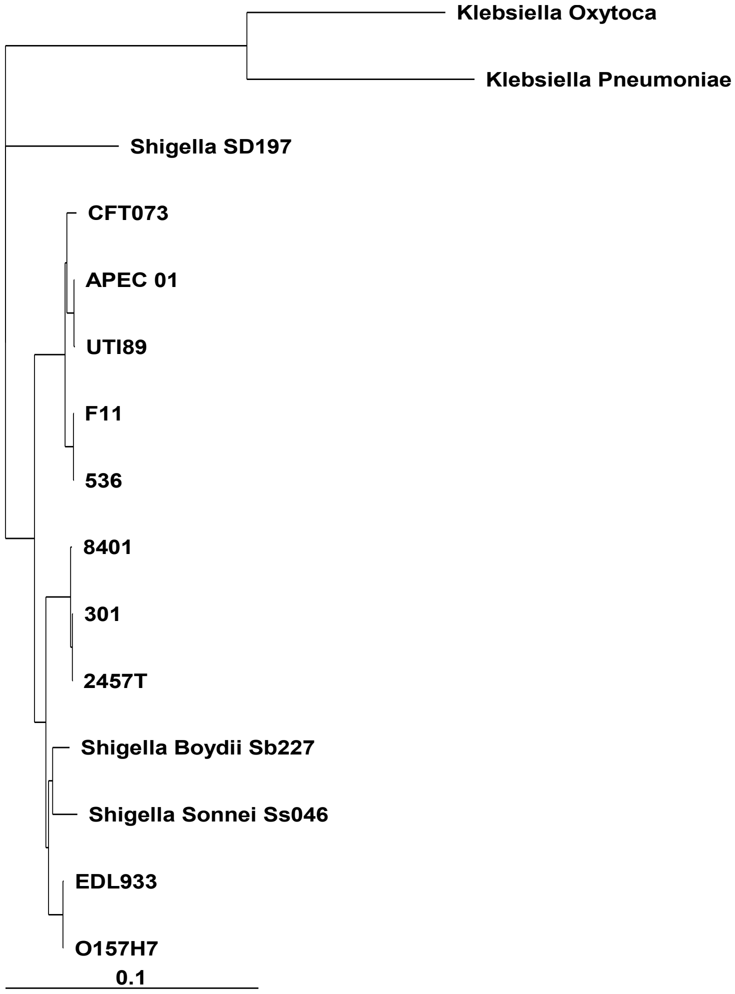


Figure 4.2: Phylogenetic tree of L-sorbose operons (both confirmed and putative) in the genus *Escherichia* and in two *Klebsiella* strains. The scale is in units of substitutions per site. Unless otherwise stated the strain is *Escherichia coli*.

is used by both. The phylogenetic analysis of the L-sorbose operon conducted as part of this research shows how this operon fits in the phylogeny of L-sorbose operons in the genus *Escherichia*. The Glucitol-6-phosphate dehydrogenase found in *Klebsiella pneumoniae* has been shown to be temperature sensitive [118], which might account for the long lag phase and slow growth of DH5α containing the plasmid.

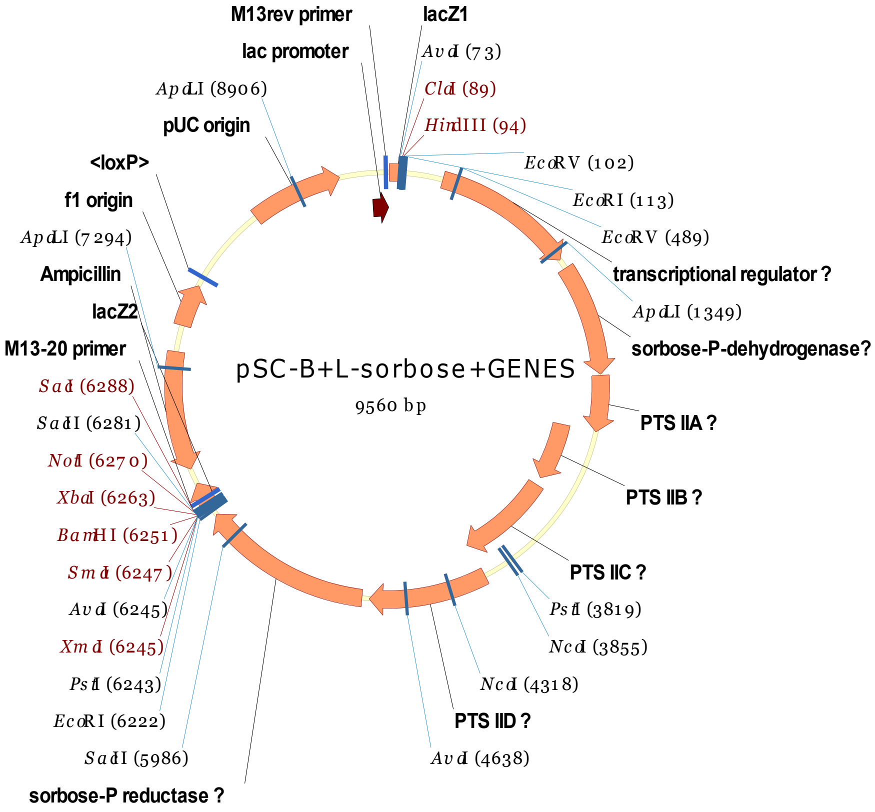


Figure 4.3: Structure of the pSC-B plasmid with the putative L-sorbose operon insert from CFT073.

4.3.3.1 The L-sorbose pathway and its link with central metabolism in *E.coli*

The cloning of the L-sorbose operon from CFT073 into DH5α resulted in a positive phenotypic response of DH5α in L-sorbose minimal medium, and close similarity between this operon and the L-sorbose utilisation operon in *Klebsiella pneumoniae* imply that the reactions catalysed by the products of these genes, as shown in Figure 4.5, were added to the metabolism of DH5α, linking to it where D-glucitol 6-phosphate is converted to D-fructose 6-phosphate.

This Figure also shows in overview how these reactions are linked to both biomass production (driven by energy transfer conducted through the TCA cycle) and to acetate production, a nuisance metabolite in many cir-

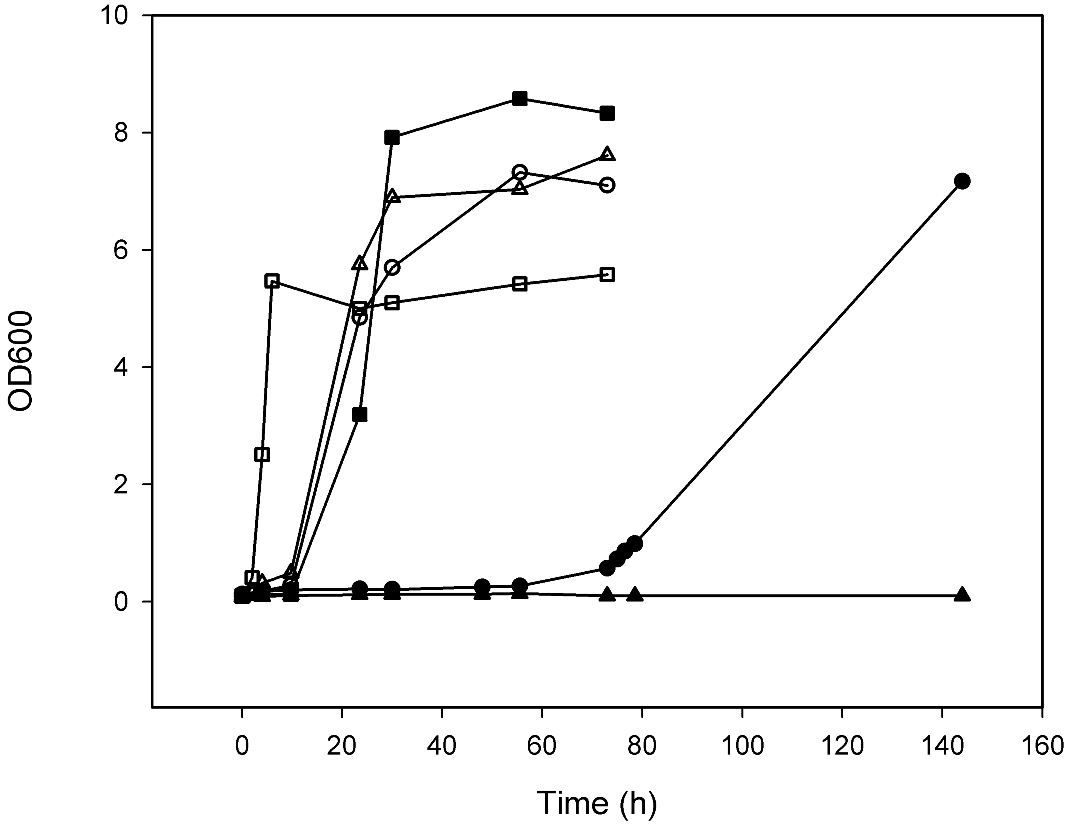


Figure 4.4: Growth curves for DH5α containing plasmid pQR793, compared to CFT073 and DH5α with an empty pUC-19 plasmid. DH5α with pQR793 is represented by ○ and •, DH5α with pUC19 by △ and ▲ and CFT073 by □ and ■ where empty symbols represent growth on glucose and filled symbols represent growth on L-sorbose. Where duplicate samples were taken readings varied by less than 0.01 OD_{600} units using a CO8000 Cell Density Meter (WPA).

cumstances and the reduction of which in engineering contexts has been the subject of much work (for instance [136, 137]). Changes in acetate production in response to a novel carbon source for a bacterium, and changes due to serial passage have both been investigated in the course of this project, so it has been included in this Figure.

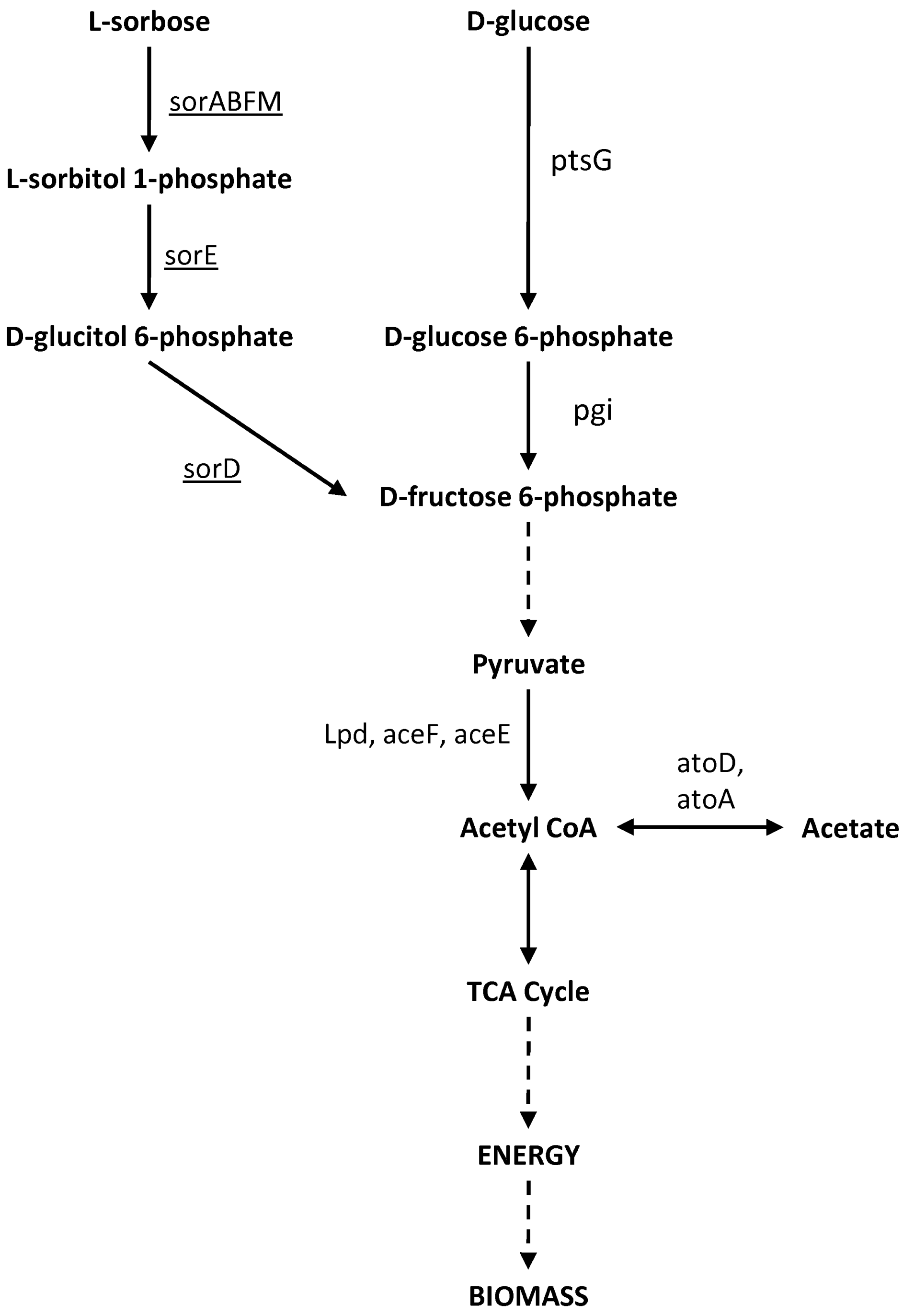


Figure 4.5: Showing in overview the link between L-sorbose degradation and central metabolism, entering the glycolysis I pathway at D-fructose 6 phosphate. Solid lines represent single reactions linking metabolites and dashed lines represent pathways and more general transitions. Where single reactions are shown, relevant genes have been indicated, underlined if they are not present in MG1655.

Table 4.4: Comparison of SAC No. 7 in CFT073 - putatively involved in the metabolism of a 5- or 6-carbon sugar - with the NCBI Conserved Domain Database. Current annotation can be seen alongside the domain inferences.

GI number	Locus tag	Annotation	Conserved Domain Database entry
26249581	c3750	Putative regulator	FadR, Transcriptional regulators
26249582	c3751	Hypothetical oxidoreductase ydfI	MtlD, Mannitol-1-phosphate /altronate dehydrogenases
26249583	c3752	Hypothetical zinc-type alcohol dehydrogenase-like protein yjjN	Tdh, Threonine dehydrogenase and related Zn-dependent dehydrogenases
26249584	c3753	Ureidoglycolate dehydrogenase	Ldh_2, Malate/L-lactate dehydrogenase
26249585	c3754	Putative c4-dicarboxylate transport system binding protein	SBP_bac_7, Bacterial extracellular solute-binding protein, family 7
26249586	c3755	hypothetical protein c3755	DctQ, Tripartite ATP-independent periplasmic transporters, DctQ component.
26249587	c3756	c4-dicarboxylate permease	ArsB_NhaD_permease, Anion permease ArsB/NhaD

4.3.4 Bioinformatic analysis of genes c3750-c3756

Current annotations of genes c3750-c3756 are quite general, and the NCBI Conserved Domains Database provides more specific putative functions - as shown in Table 4.4. There are several potential metabolic functions, such as mannitol, malate and L-lactate degradation, as well as a potential regulator gene (c3750) and transporter genes (c3755 and c3756). It was therefore selected for cloning and further characterisation.

4.3.4.1 Cloning of genes c3750-c3756 into *E. coli*

Cloning of this cluster of genes was attempted using the methods in Section 2.9. Cloning was not achieved using any of the methods described in the methods so biochemical characterisation was not possible for these genes.

4.3.5 A putative type VI protein secretion system

Adjacent to SAC No. 6 was a set of genes of mostly unknown function that was initially considered part of that SAC for this analysis. However, when the PTS system genes were investigated it showed that this set of genes was not a single set of genes. The genes that were not part of the putative PTS operon contained two poorly characterised genes, c3391 and c3392, that had annotations, 'Secreted protein Hcp' and 'ClpB protein' respectively. Hcp is a haemolysin coregulated protein and ClpB is a protein disaggregation chaperone.

It therefore looked like a candidate for some sort of protein secretion system, rather than an operon encoding a metabolic function. The genes comprising this set, c3385 to c3403, were BLASTed against the GenBank protein database and it was found that many of them are related to the Type VI protein secretion system (T6SS) recently elucidated by Pukatzki *et al* [138]. This could be a novel Type VI protein secretion system, but it was beyond the scope of the work presented here to investigate it further at this point.

4.4 Conclusions

In this Chapter multiple genome sequence analysis has been used to identify several sets of adjacent colinear genes in *E. coli* CFT073 that are not present in *E. coli* MG1655 and might be implicated in metabolism in the human urinary tract. Specifically a previously incompletely annotated operon encoding proteins involved in L-sorbose catabolism has been identified and experimentally confirmed encompassing genes c4981 to c4987 of the genome of uropathogenic *Escherichia coli* strain CFT073.

The sets of genes from CFT073 found solely in UPEC include an arginine metabolic operon and one that is potentially a 5- or 6-carbon sugar catabolic operon, though the substrate specificity for any of the encoded enzymes is currently unknown and awaits investigation. The use of such sets of genomic data will become increasingly important as the rate of sequencing increases while experimental verification of gene function lags considerably behind. Although elucidation of novel gene function cannot be done purely through comparative genomics it can, as shown here, aid searches for important genes, not necessarily previously characterised in other species or strains.

Chapter 5

Bacterial response to novel metabolic genes: a case study

5.1 Preface

Increasingly metabolic engineering of bacteria involves the addition of heterologous genes to construct novel pathways and to add existing pathways to add important capabilities to engineered bacteria. This is combined with controlled change of expression of various genes to channel metabolites towards target products and away from waste products(for instance in [139]). Because of the complexity of problems in understanding the behaviour of large scale metabolic networks many of these control mechanisms are only semi-rational, that is, there is a certain amount of trial and improvement involved in finding optimal reaction fluxes and the highest level of production of a target product is not necessarily known, so incremental improvements do not necessarily have a benchmark. This problem is solved to an extent by whole genome stoichiometric metabolic models since the stoichiometric limits set by such models are a minimal set of constraints on a metabolic network and can potentially give the maximum yield of a target product, when the models are adapted to take into account the changes in the metabolic network achieved by the addition and removal of metabolic genes undertaken in engineering the bacterium.

Bacteria have been adapting to novel conditions and gene complements for billions of years and as such are very efficient evolvers. Their evolutionary goals have been different to what is required of them for biotechnological applications, but nonetheless understanding the relative effects of various techniques of adaptation bacteria apply and how they integrate novel metabolic capabilities into their metabolism might shed light on current and future attempts to engineer bacteria for such applications. In this Chapter a model system, that of the metabolic operon encoding proteins for the uptake and catabolism of L-sorbose (from CFT073) cloned into a DH5α host, has been investigated in an attempt to elucidate how these metabolic genes are integrated into the host metabolism. This is of interest partly because of the very poor growth seen in the previous chapter of DH5αL initially in M6L containing L-sorbose. This analysis has been conducted using a single medium (M6L), which contains L-sorbose as a sole carbon source, to apply a selection pressure on the bacterium in a medium optimised for fast *E. coli* growth.

Those aspects of growth most easily quantified have been used to follow the adaptation of a bacterial strain (in DH5α) containing the pSC793, from now on referred to as DH5αL, to a strain that has a high, steady growth rate in the defined medium, referred to as DH5αM. These have been compared to CFT073, which is used as a reference strain. Growth rate, carbon source uptake rate and acetate production have all been measured at the beginning and end of adaptation over approximately 100 generations of DH5αL to characterise the adaptation of this strain to its novel gene complement.

5.2 Results

5.2.1 Passage

To produce the new strain DH5αM, DH5αL was subjected to selective conditions in serial passage through a series of identical shakes flasks, all grown in the same conditions as those described in Section 2.6, in minimal medium with 1 % L-sorbose as the sole carbon source at 37 °C. During each shake flask experiment doubling rate was measured in order to monitor the change in doubling rate during adaptation. The doubling rates of DH5αL during passage can be seen in Figure 5.1.

This Figure uses effective generations as the independent variable. This has been calculated in a simple way: at the end of each growth, just as the inoculum from the culture was about to be transferred to fresh media, the optical density at 600 nm of the old culture was taken and the factor by which the inoculum was diluted was calculated for the fresh media. Because the media was clear ($\text{OD}_{600} = 0.00$ compared to water) this gave an effective optical density of the culture at the beginning of growth in the fresh media. Optical density at the end of growth in a particular culture was then divided by the optical density at the beginning of that culture to give a multiplicative factor by which the culture had grown. The logarithm to the base 2 of that factor was then equal to the number of times the culture had doubled in

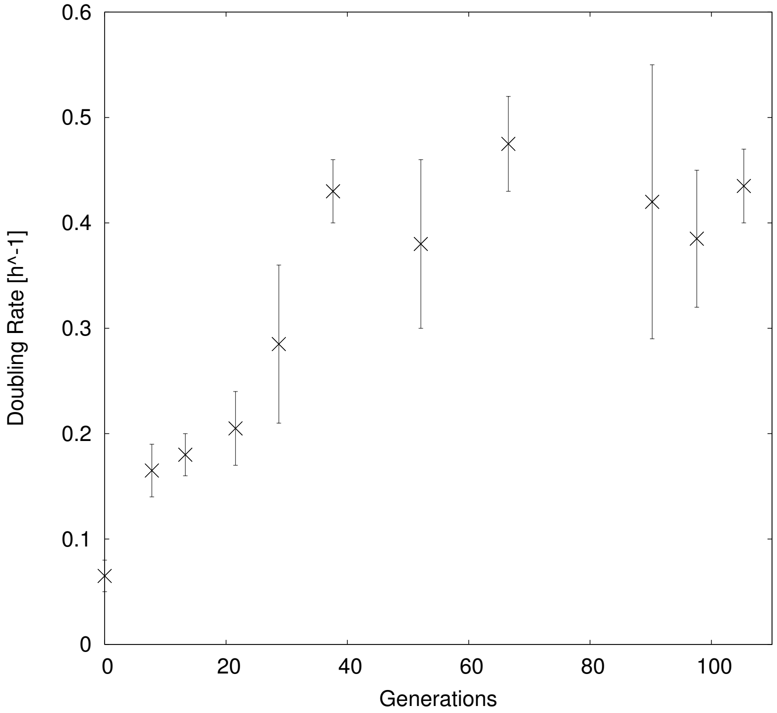


Figure 5.1: Doubling rates of *E. coli* DH5αL passaged over approximately 100 generations. The media used was M6L supplemented with 1 % L-sorbose and growth was carried out at 37 °C, identical conditions to those used to produce the results shown in Figure 5.4. Doubling rates were determined by testing OD during growth in each shake flask as the passage was taking place. Error bars indicate a 95% confidence interval (2 standard errors).

amount during the course of growth in that shake flask. Overall number of generations was then calculated by summing the number of generations grown in all previous cultures and the current culture.

5.2.2 Strain growth in defined media

The three strains considered here, *E. coli* CFT073, DH5αL and DH5αM, were grown in four different conditions, of varying temperature and using different carbon sources. Cultures were grown according to the method described in Section 2.6 and samples for quantification of glucose, L-sorbose and acetate were analysed by the method described in Section 2.7. The reason for growing the strains at two different temperatures was that it has

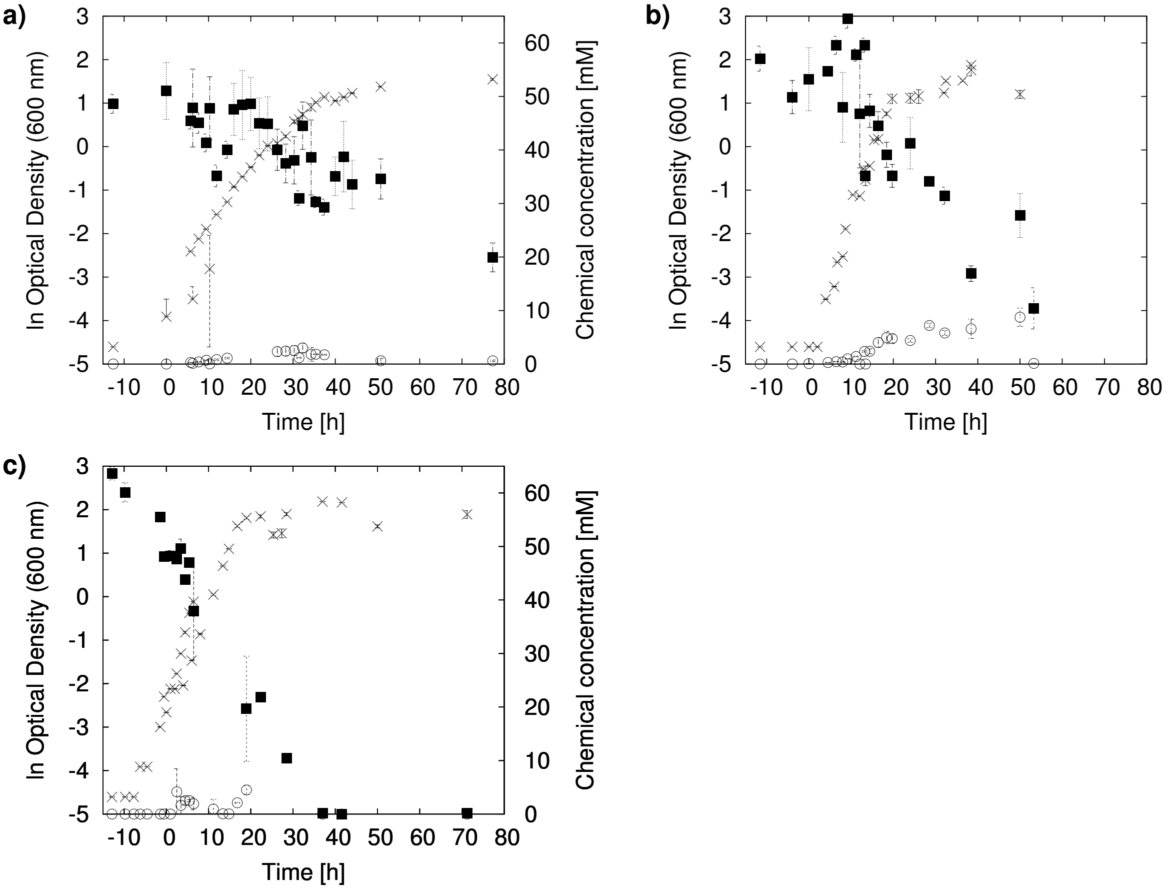


Figure 5.2: This figure shows the growth characteristics of each of three strains of *E. coli* - DH5αL (a), DH5αM (b), CFT073 (c) - in M6L medium supplemented with 1% L-sorbose, grown at 30 °C. Each graph shows three biological replicates of the conditions, sampled at different points through their growth to get data on a wide spread of the growth period. Optical density measurements are labelled by 'x', extracellular concentration of L-sorbose by '■', and extracellular concentration of acetate by '○'. Measurements of each point were made in triplicate and error bars represent one standard error on these triplicate measurements. All errors are plotted (1 standard error), though some are so small that they only cover a single pixel on the graphs, so are not visible. The three growth curves of each strain from the three biological replicates are lined up (as described in the text) to clarify their patterns. Optical density is plotted on a logarithmic axis to show the exponential nature of the bacterial growth.

been reported that the glucitol-6-phosphate dehydrogenase encoded by the L-sorbose operon *Klebsiella pneumoniae*, from which the function of the CFT073 operon was inferred, was heat sensitive and initially it was thought that the orignal difference between growth rates of DH5αL and CFT073 might have something to do with this temperature dependence.

For repeated experiments, because different inoculum sizes were used to

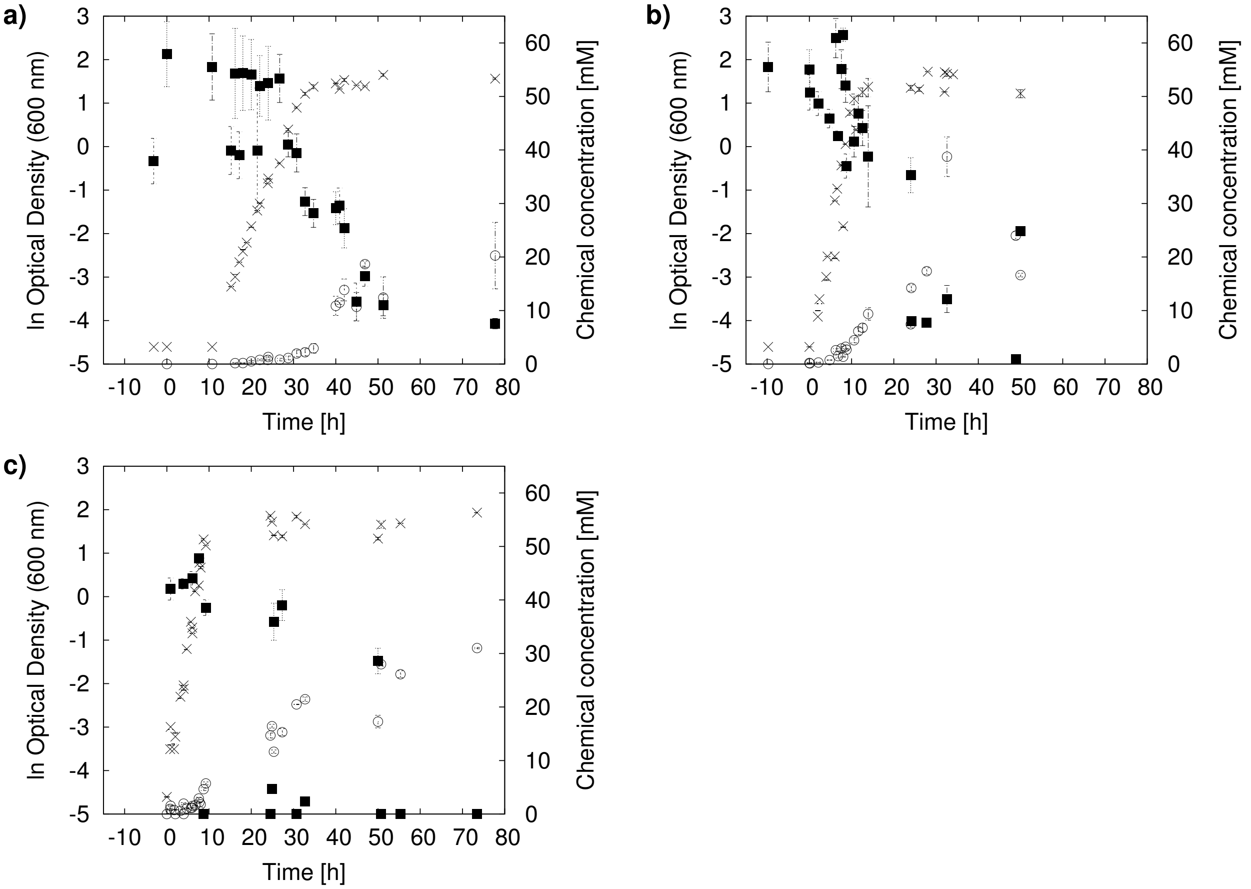


Figure 5.3: This figure shows the growth characteristics of each of three strains of *E. coli* - DH5αL (a), DH5αM (b), CFT073 (c) - in M6L medium supplemented with 1% Glucose, grown at 30 °C. Each graph shows three biological replicates of the conditions, sampled at different points through their growth to get data on a wide spread of the growth period. Optical density measurements are labelled by 'x', extracellular concentration of Glucose by '■', and extracellular concentration of acetate by '○'. Measurements of each point were made in triplicate and error bars represent one standard error on these triplicate measurements. All errors are plotted (1 standard error), though some are so small that they only cover a single pixel on the graphs, so are not visible. The three growth curves of each strain from the three biological replicates are lined up (as described in the text) to clarify their patterns. Optical density is plotted on a logarithmic axis to show the exponential nature of the bacterial growth.

view different parts of the growth curve, the growth patterns do not line up if only time from inoculation is used. Therefore the growth curves were lined up using their optical density measurements during exponential growth. The slopes of ln(OD) were roughly the same for each strain in each condition so vertical offset of the exponential lines at time = 0 hours and the mean slope between the lines was used to calculate a horizontal offset. All experiments

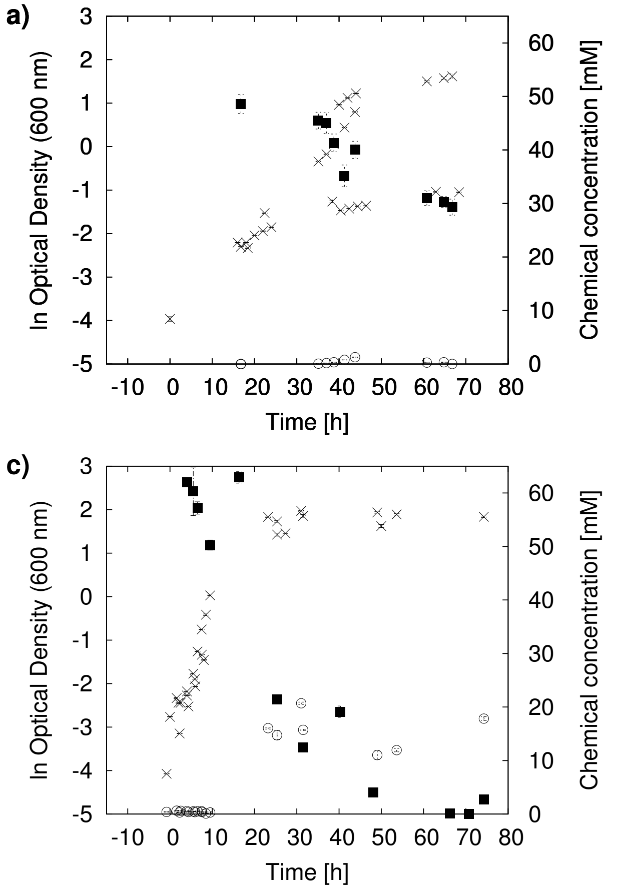


Figure 5.4: This figure shows the growth characteristics of each of three strains of *E. coli* - DH5αL (a), DH5αM (b), CFT073 (c) - in M6L medium supplemented with 1% L-sorbose, grown at 37 °C. Each graph shows three biological replicates of the conditions, sampled at different points through their growth to get data on a wide spread of the growth period. Optical density measurements are labelled by 'x', extracellular concentration of L-sorbose by '■', and extracellular concentration of acetate by '○'. Measurements of each point were made in triplicate and error bars represent one standard error on these triplicate measurements. All errors are plotted (1 standard error), though some are so small that they only cover a single pixel on the graphs, so are not visible. The three growth curves of each strain from the three biological replicates are lined up (as described in the text) to clarify their patterns. Optical density is plotted on a logarithmic axis to show the exponential nature of the bacterial growth.

with all strain and condition combinations were repeated three times and the times for the first one of these repetitions were used as reference times. The times for the other two repetitions of each carbon source and temperature combination were modified by their individually calculated horizontal offset to better compare the other data obtained during growth.

Figures 5.2, 5.3, 5.4 and 5.5 show the results of the growths split by

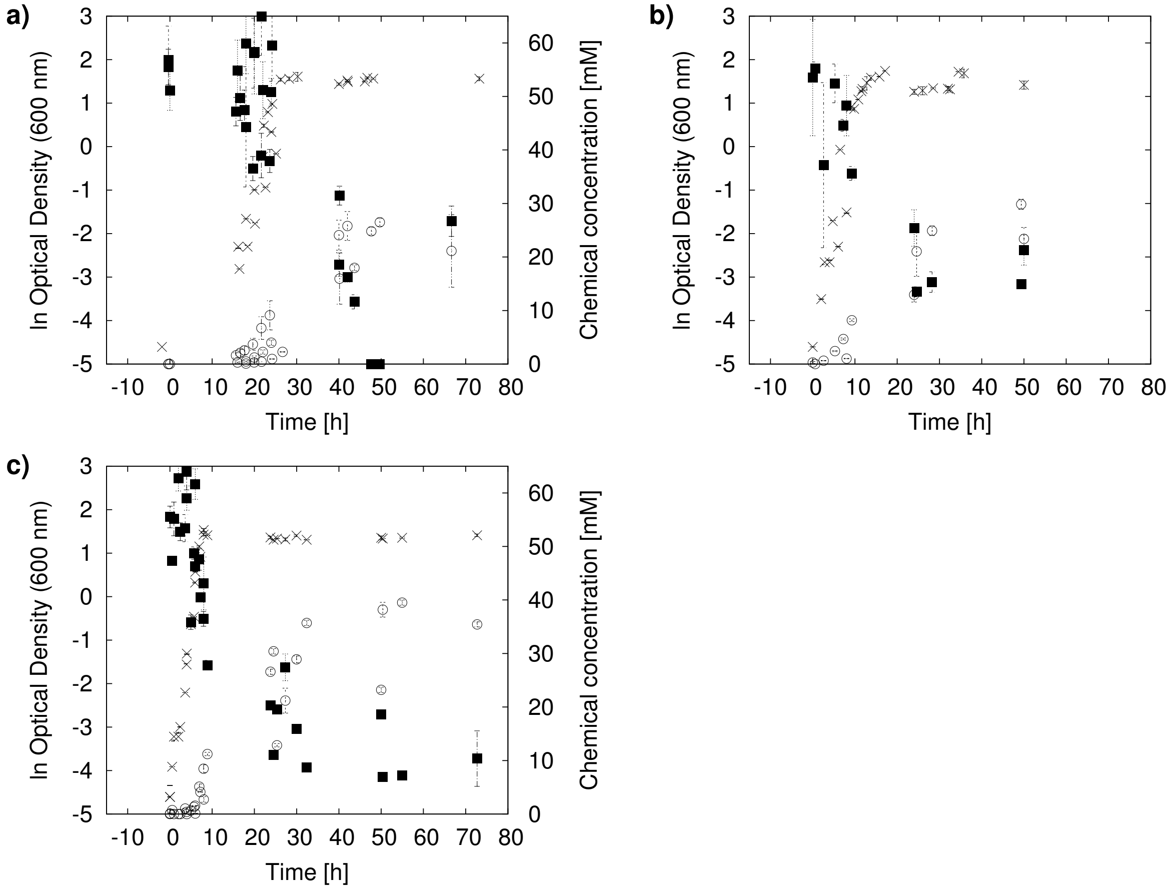


Figure 5.5: This figure shows the growth characteristics of each of three strains of *E. coli* - DH5αL (a), DH5αM (b), CFT073 (c) - in M6L medium supplemented with 1% Glucose, grown at 37 °C. Each graph shows three biological replicates of the conditions, sampled at different points through their growth to get data on a wide spread of the growth period. Optical density measurements are labelled by ’x’, extracellular concentration of Glucose by ’■’, and extracellular concentration of acetate by ’○’. Measurements of each point were made in triplicate and error bars represent one standard error on these triplicate measurements. All errors are plotted (1 standard error), though some are so small that they only cover a single pixel on the graphs, so are not visible. The three growth curves of each strain from the three biological replicates are lined up (as described in the text) to clarify their patterns. Optical density is plotted on a logarithmic axis to show the exponential nature of the bacterial growth.

growth conditions into L-sorbose at 30 °C, Glucose at 30 °C, L-sorbose at 37 °C and Glucose at 37 °C respectively. Table 5.1 shows the doubling rates calculated from each of the conditions and each of the different strains considered and Table 5.2 shows t values for the differences between each strain in each of the different conditions; the t-statistic value for each of these comparisons of 2.776. (This value is the same for all of the comparisons

Table 5.1: Doubling rates of the three strains grown at 30 and 37 °C on L-sorbose and Glucose. Growths were done in triplicate, and the values in brackets indicate the uncertainty in the last digit of the quoted value (one standard error).

Temperature	DH5αL [h^{-1}]	DH5αM [h^{-1}]	CFT073 [h^{-1}]
Glucose			
30 ^{o}C	0.39(1)	0.63(8)	0.94(3)
37 ^{o}C	0.49(2)	0.71(10)	1.12(4)
Ratio	**1.26**	**1.13**	**1.19**
L-sorbose			
30 ^{o}C	0.16(1)	0.44(1)	0.40(6)
37 ^{o}C	0.24(2)	0.42(6)	0.50(5)
Ratio	**1.5**	**0.95**	**1.25**

Table 5.2: T-statistics comparing each of the strains' doubling rates to the other strains in the same conditions, taken from Table 5.1. Since each growth rate was calculated from 3 biological replicates, for the t-test there were 4 degrees of freedom, using a 2-tailed 95 % confidence test, the value of the test statistic is 2.776. Values shown in this table that are greater than 2.776 therefore represent a statistically significant difference between the doubling rates of the two strains compared in the conditions stated, at the 5 % level.

Conditions	**DH5αL vs. DH5αM**	**DH5αL vs. CFT073**	**DH5αM vs. CFT073**
L-sorbose, 30 °C	19.799	3.946	0.658
Glucose, 30 °C	2.977	17.393	3.628
L-sorbose, 37 °C	2.846	4.828	1.024
Glucose, 37 °C	2.157	14.087	3.807

because the number of degrees of freedom and significance levels are the same.) All datapoints have associated errors, but some errors calculated are so small that they fall within the datapoints themselves and are not visible.

Biological replicates of each growth where plotted on the same axes; although they observed different parts of the growths all had a sufficient set of data points taken during the exponential growth phase to infer a reliable doubling rate. Table 5.3 shows the optical density achieved at the end of the exponential growth phase for each of the conditions and strains - with errors obtained in the same way as for the growth rates in Table 5.1.

Table 5.3: Optical density of cells at the beginning of stationary phase of the three strains grown at 30 and 37 °C on L-sorbose and Glucose. Growths were done in Biological triplicate, and the values in brackets indicate the uncertainty in the last digit of the quoted value (one standard error).

Temperature	DH5αL	DH5αM	CFT073
Glucose			
30 ^{o}C	4.6(3)	3.92(4)	5.3(7)
37 ^{o}C	4.70(7)	4.3(7)	4.1(3)
Ratio	1.0	1.1	0.8
L-sorbose			
30 ^{o}C	3.5(2)	3.7(5)	5.9(5)
37 ^{o}C	3.0(10)	4.4(5)	5.9(4)
Ratio	0.78	1.2	0.99

5.2.3 Re-transformation of the passaged plasmid into naive DH5α

Retransformation of the plasmid from DH5αM was achieved into naive DH5α in order to examine whether changes in growth rate were related to changes in the plasmid sequence or the chromosome of the passaged DH5α. The new strain, DH5αLM, was tested for growth in M6* minimal medium supplemented with 1 % L-sorbose (the same as the passage conditions) and it was found that the doubling rate for this strain was 0.26(2) h^{-1} at 37 °C. This strain was serially passaged over 80 generations in the same conditions as the original passage and within that time achieved a maximum doubling rate of 0.47(4) h^{-1} at 37 °C.

5.2.4 Acetate production

Acetate production was evaluated by measuring its concentration in the supernatant at each time point where OD was measured during growth. These measurements can be seen in Figures 5.2 - 5.5. Where there was no peak visible in the HPLC traces at the appropriate time acetate concentration was

Table 5.4: Showing total extracellular acetate concentration measured at the end of growth (in mM/L). Errors are taken from biological replicates and represent uncertainty in the final digit (one standard error).'ND' (not detected) is indicated where no acetate peak was detected in the HPLC traces taken for those samples, so it is inferred that no net acetate production occured in these conditions.

Conditions	DH5αL	DH5αM	CFT073
L-sorbose, 30 °C	0.5(3)	7.5(6)	ND
Glucose, 30 °C	15(2)	26(7)	21(3)
L-sorbose, 37 °C	ND	12(1)	18(2)
Glucose, 37 °C	24(2)	27(3)	31(5)

assumed below the limit of detection of the HPLC, so was plotted as zero concentration. Table 5.4 shows the acetate concentrations at the beginning of the stationary phase for each strain and condition combination, where these data could be taken from the acetate data from the HPLC. This represents net acetate production during growth and is compared with a model of overflow metabolism in Chapter 7.

Korz *et al* [140] have shown that below a certain growth rate threshold in a glucose medium acetate is not excreted from E. coli; this threshold is around a doubling rate of $0.2 - 0.25h^{-1}$, which fits well with the observations of DH5αL shown above, where at and below this threshold (in L-sorbose) little or no acetate is excreted. There is some indication from the data above that there is a linear relationship between growth rate of each strain and acetate production above a certain threshold, independent of growth conditions and even strain. A graph of net acetate production as a function of growth rate for the growths studied is presented here as Figure 5.6.

Acetate production as a result of overflow metabolism during carbon uptake has been observed in *E. coli* in continuous culture [141], related to

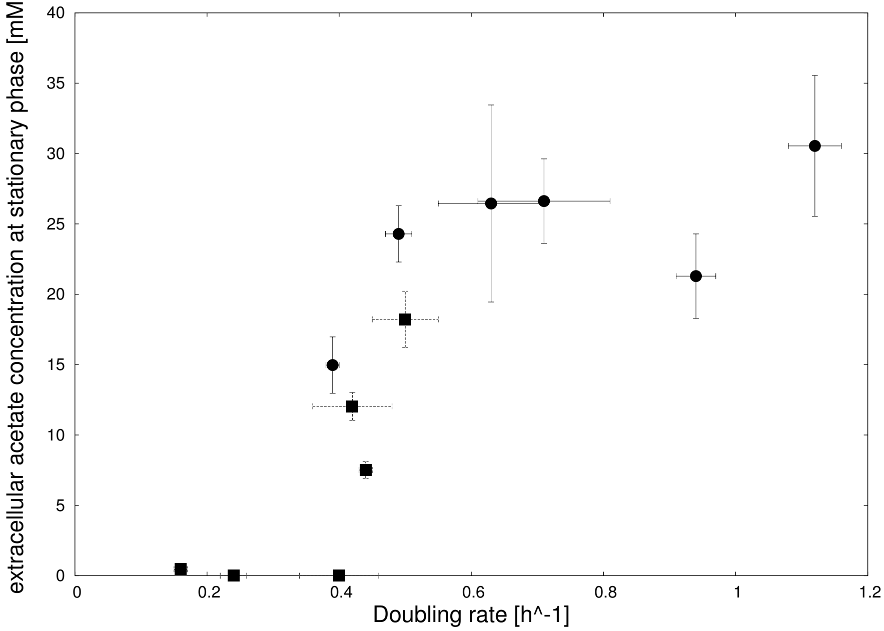


Figure 5.6: Final extracellular acetate concentrations as a function of doubling rate for all combinations of strains and conditions tested. L-sorbose growths are marked as '■', and glucose growths by '•'. Errors are plotted for both growth rate and acetate concentration (1 standard error).

the specific rate of carbon uptake, rather than growth rate, but the glucose uptake data presented here are not sufficiently precise to determine this relationship for these growth experiments. Further analysis of acetate production as a result of carbon overflow metabolism accompanies the modelling of growth in these conditions in Chapter 7.

5.2.5 Glucitol-6-phosphate dehydrogenase activity

The change of growth characteristics of strain DH5aL over the 100 generation passage period could have been achieved in several ways. One of these is a change in the characteristics of the enzyme presumed to be the bottleneck in the pathway for L-sorbose utilisation. Since the activity of the close homolog of the glucitol-6-phosphate dehydrogenase in *Klebsiella pneumoniae* has been reported to be temperature dependent, the corresponding en-

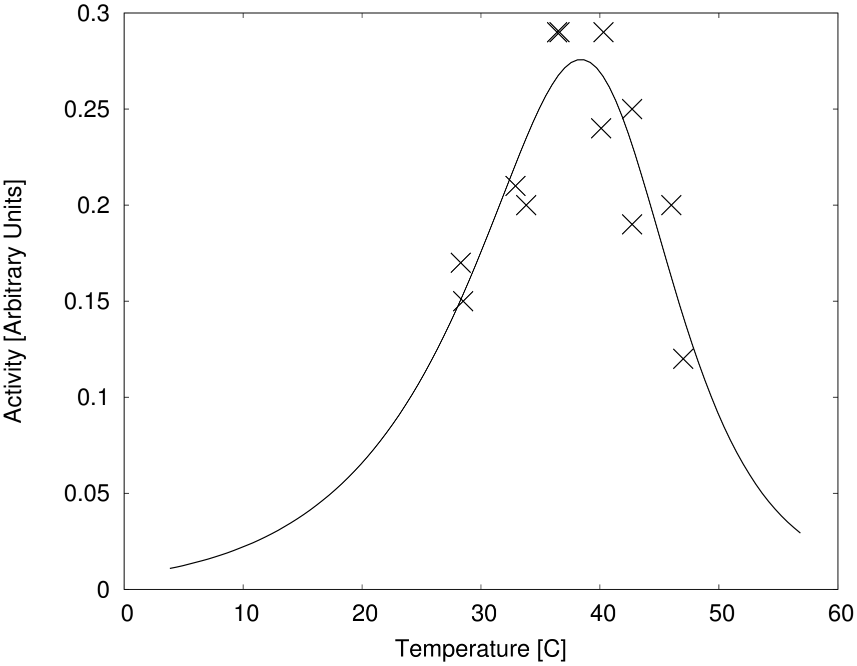


Figure 5.7: An example of the results of activity tests on glucitol-6-phosphate dehydrogenase found in the crude lysate of cells after growth on a medium with L-sorbose as the sole carbon source. The activity was tested at various temperatures around 37 °C to determine the dependence of the activity of this enzyme on temperature. This example shows results from DH5αL grown at 37 °C plotted as crosses. The line shows a line of best fit of the model of Lee *et al* [120]. In this case T_{eq} is 40.95C.

zyme in the CFT073 L-sorbose operon was tested for heat sensitivity both before and after passage. To test the change in glucitol-6-phosphate activity temperature dependency over the period of 100 generations of passage the glucitol-6-phosphate activity of crude lysate from freshly grown cells was measured using the method described in Section 2.8.

Figure 5.7 shows an example of one of the datasets produce to calculate the heat sensitivity of the Glucitol-6-phosphate dehydrogenase. Crosses show measurements and the solid line shows the model derived from Lee *et al* [120] with parameters fitted using the downhill simplex method of Nelder and Mead [121].

The fitted model gave a temperature of maximum activity, T_{max}. Fig-

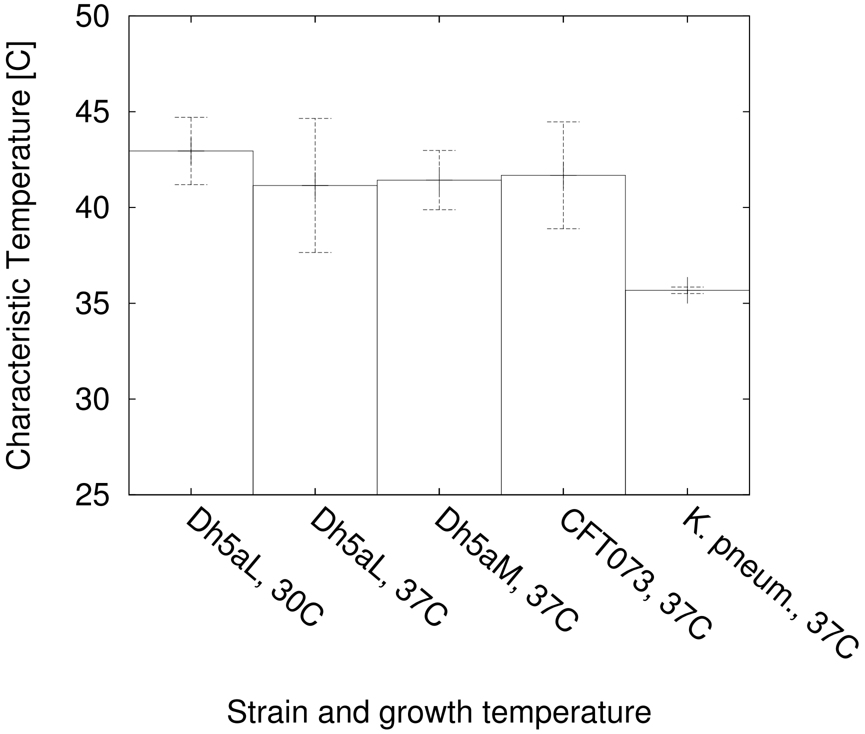


Figure 5.8: The results of all of the glucitol-6-phosphate dehydrogenase experiments carried out to test heat sensitivity of this enzyme in unpassaged and passaged strains. Growth temperatures in the x labels show what temperature the cells were grown at before crude extracts were taken. DH5αL was grown initially at 30 °C to ensure that no adaptive pressure towards the higher temperature was placed on the bacterium before the activity test was carried out. Errors represent 95 % confidence intervals of T_{eq} values according to the Lee model of enzyme heat dependence used in data analysis. All values of T_{eq} were inferred from line fitting (of which Figure 5.7 is an example) and errors were obtained from biological replicates for each strain.

ure 5.8 shows the values of T_{max} determined for each of the strains tested. Strain *Klebsiella pneumoniae* KP1 was used as a control since it has been shown previously [118] that the glucitol-6-phosphate dehydrogenase of the L-sorbose operon of this strain is heat sensitive and it was the operon from which the function of the CFT073 operon was inferred.

5.3 Discussion

5.3.1 Passage

The first notable result in this Chapter is that DH5α did indeed change its behaviour over the 100 generations as shown in Figure 5.1 increasing its doubling rate in the face of consistent selective conditions - using L-sorbose as a sole carbon source growing at 37 °C. The doubling rate increased almost fourfold during the course of adaptation and the Figure shows that doubling rate changed mostly over the first 40 or so generations to its higher level. This change appears to have happened over a period of 40 generations, rather than in few discrete jumps in doubling rate.

Growth curves for these passages (not shown) did not show any discontinuities in doubling rate which might indicate abrupt changes in capability, although if considering the mechanism of change (some genetic alteration which would initially occur in just a single cell) the effects on overall doubling rate of a change conferring a markedly greater growth rate in that cell would be massively diluted by the large number of unadapted cells already growing in the culture. It is therefore not possible to eliminate the possibility that just a few key adaptations were the cause of the large change in doubling rate, especially when considering the error bounds marked on the Figure. What is perhaps more clear from the Figure is that once the adaptation to a doubling rate of $0.4h^{-1}$ there was a plateau in doubling rate capability (constant within the experimental error) with no further significant increase.

Modelling of the growth of this bacterial strain will not be discussed until

Chapter 7, where theoretical limits of growth rate will be calculated using a stoichiometric model of this strain. In work by Ibarra *et al* [108] 500 generations of *E. coli* adaptation were observed in chemostat culture using a sole carbon source with the express goal of determining how close growth rates on that carbon source could come to theroetical maxima according to stoichiometric models. In that study doubling rate on malate was shown to increase 21 %, on glucose by 18 %, on succinate by 17 % and on acetate by 20 %. The work presented here shows a far larger increase in doubling rate of around 150 % (if the initial growth at 0 generations on Figure 5.1 is ignored since there may have been residual effects from the cloning procedure affecting its growth).

Fong *et al* [142] have similarly investigated the relationship between the adaptation and the approach to optimal flux distributions in genetically engineered bacteria. Their passage took place over 1000 generations and involved knockout mutants, rather than the addition of novel metabolic genes, but similar selective pressure would have been placed on the strains during growth as they used M9 media supplemented with glucose, similar to that used in the work in this Chapter which used M6* minimal media supplemented with L-sorbose. Interestingly increases in growth rate observed by Fong *et al* of up to 133 % were of the same order as those seen above, even though adaptation in that Chapter was only observed over 100 generations. Changes in the adapted strains observed by Fong *et al* over the first 100 generations were far less pronounced than the data shown here. This difference could be because of the nature of the changes in the metabolic network - since the L-sorbose uptake and utilisation genes were novel to DH5α its flux distribution to begin with was far further from optimal for growth rate than the gene deletion mutants used by Fong *et al.*

One factor that may have influenced this large rate increase is the experimental conditions. Although every effort was made to keep consistent conditions during each growth there is a chance that media and cell treatment were improved unwittingly during the course of the passage. Evidence for this is in the form of the growth tests performed later on DH5αL, during which the mean doubling rate was 0.24 h^{-1}, rather than 0.16 h^{-1}; so taking this into account, DH5αM may only be growing 70 % faster than DH5αL in the same conditions, but this is still considerably greater a jump than that observed by Ibarra *et al.* Two further explanations for the large difference in scale of these changes seem feasible: firstly that different mechanisms might be at work in the adaptation of the bacteria, or secondly that the same mechanisms are at work, but that there is more room for evolution to occur via these mechanisms in DH5αL. The second of these two explanations seems more plausible, as bacteria will use any and all available strategies for adaptation so even if there are several mechanisms whereby faster growth could be achieved, they would be explored both in continuous culture and in serial passage.

Ferenci [143] has recently reviewed the mechanisms of evolution in chemostat cultures in bacteria and concluded that both regulatory and enzymatic changes are important in adaptation in these conditions. Although continuous chemostat culture has not been used in this work, the principles of evolution elucidated in experiments of this sort should be valid at least to an extent. In continuous culture cells experience a constant environment with a low carbon source concentration, whereas in this work the cells alternate between a carbon-rich and a carbon-poor environment, so selection pressures will be somewhat different. However, the mechanisms used to adapt to these different conditions should be the same. Without further investigations however, such as expression profiles for the strains involved to determine how regulation has changed to cope with the culture conditions there is no way to know to what extent regulation rather than enzyme characteristics have changed to produce the observed change in growth rate.

Two evolutionary mechanisms could have occurred to produce the observed change in growth rate: the integration of the genes on the pQR793 plasmid into the regulatory architecture of the cell or a change in the characteristics of one of the proteins involved in L-sorbose uptake and degradation. The purpose of the glucitol-6-phosphate dehydrogenase activity assay (discussed below) was to determine whether this enzyme (presumed to be the most significant bottleneck in growth at least in the early stages of passage) had changed its behaviour as a function of temperature or not, to elucidate how much this might have affected growth rate during the course of the passage. Observations of glucose uptake and build-up of extracellular acetate were made in order to elucidate more of the metabolic changes brought about by the passage.

Another feature of Figure 5.1 is the plateau between about 40 and 100 generations where growth rate doesn't change (within the error bounds). Ibarra *et al*'s observations of *E. coli* showed a steady climb in growth rate for almost the entire 500 generation period of adaptation towards the point of theoretical maximum growth rate (as determined by a genome scale stoichiometric model). It could be that the regular change of conditions from starvation to excess and back selected against ever faster growth rates to balance with ability to cope in the nutrient poor conditions experienced by the bacteria during serial passage.

5.3.2 Adaptation

Adaptation is an ongoing process, so there is a chance that growth rate would continue to increase after the 100 or so generations observed here (to a maximum constrained by the physicochemical limitations on the strain), but it can be seen that the doubling rate appears to have stabilised (in the conditions used for the passage) to around $0.4h^{-1}$ and remained there over 60 generations (see Figure 5.1). At 100 generations the passaged strain, named DH5αM, was sampled and used for growth tests as seen above.

Figures 5.2 to 5.5 show the results of these growth tests and provide many interesting glimpses of the changes that have taken place over the 100 generations. Table 5.2 shows that there is a significant difference between DH5αL and CFT073 doubling rates, even when grown on Glucose (where there is no metabolic dependence on the genes encoded on pQR793). The L-sorbose genes on this plasmid are not artificially inducible, but the operon was cloned with its upstream region and all the regulatory region controlling the expression of the L-sorbose uptake and utilisation enzymes in this operon. The regulator, SorC, has no close homologues within DH5α, so it would not be expected that there would be any regulatory cross-talk interfering with the growth of DH5αL. There is no genotypic reason why DH5αL would grow any slower on glucose and metabolic burden of plasmids is small when there is not very high expression of genes encoded on a plasmid, so the reason for this poor growth in all conditions is unclear.

When DH5αL and DH5αM are compared it is clear that passage on L-sorbose has had a significant effect on L-sorbose growth and to some extent in glucose (the difference is significant at 30 °C, but not at 37 °C). This represents evidence that DH5αM grows better in conditions not limited to those conditions used during passage. This indicates that as a growth strain DH5α (albeit containing a cloned plasmid) is not well adapted. Yau *et al* [144] compared growth of various strains of bacteria containing a plasmid and showed that DH5α grows about 10 % slower than a wild-type strain (W3110), probably somewhat similar in metabolic character to CFT073. The difference in growth rates in DH5αM compared to CFT073 are not significantly different in L-sorbose at the 5 % level, but CFT073 still grows faster on glucose than DH5αM, clearly indicating that there is some aspect of DH5α metabolism holding it back as yet unidentified.

It can be seen from Table 5.1 that both DH5αL and CFT073 increase their doubling rate at higher temperature when L-sorbose is the sole carbon source, whereas DH5αM does not significantly change its rate. *E. coli* grows better at 37 °C than at 30 °C, all other things being equal [145] - so there must be some temperature dependent aspect of the L-sorbose operon that remains, or some way in which the bacteria have adapted when faced with 37 °C passage for 100 generations that renders growth at 30 °C far improved, more so than at 37 °C (to make up for the suboptimal growth rate due to temperature).

Doubling rate for the adapted strain in this work reached a plateau (see

Figure 5.1) and did not reach the growth rate observed in the L-sorbose native strain CFT073. The rate observed in CFT073 is not necessarily the maximum doubling rate achievable using L-sorbose as sole carbon source in these growth conditions. Therefore it is known that DH5αM did not reach the theoretical maximum doubling rate. Observed consistently between the work of Fong *et al* and the work presented here is that adaptation occurred gradually and evenly, rather than in jumps of doubling rate, indicating that changes in doubling rate were due to many small modifications (presumably in regulation rather than by changes in the activity of the enzymes responsible for L-sorbose uptake and utilisation, although this has not been experimentally verified), rather than a few large changes.

The data presented above show clearly that adaptation occurs in bacteria when presented with some particular set of conditions which is not necessarily specific to those conditions but spills over into similar situations. Adaptation in this case is beneficial (in terms of growth rate and final OD) for the passage conditions but may be even more beneficial for other conditions, as shown above in the high growth rate of DH5αM in L-sorbose medium at 30 °C compared to 37 °C. However, comparing the established strain, CFT073, with the strain passaged for improved performance, DH5αM, the rate of growth and final OD are not the same even after 100 generations of adaptation. There is therefore some aspect of genetics that eludes DH5αM either in terms of a lack of certain genes (a comparison of gene complements will be done in a later chapter), or in terms of regulatory adaptation to make full use of the L-sorbose genes provided on the plasmid pQR793. Growth comparisons on glucose show that DH5αM does not grow as well as CFT073 on this carbon source either, but the conditions of the passage mean that it is unclear whether DH5αL could have adapted towards a growth ability comparable to CFT073 if the passage conditions had been different.

The question of where the adaptation took place, whether it was on the plasmid or on the chromosome, was addressed by re-transformation of the plasmid from the serially passaged strain DH5αM into a naive DH5α. The growth rates observed for this new strain, before and after passage, are not statistically different to DH5αL and DH5αM respectively, indicating that a large part of the adaptation may have occurred on the chromosome of DH5α during passage, rather than on the plasmid.

5.3.3 Glucose and L-sorbose uptake

It is clear from Figures 5.2 to 5.5 that glucose and L-sorbose (the carbon sources) in the supernatant was not very well measured by the HPLC in the conditions used for these experiments, however clear patterns emerge as to the changes that have taken place in the course of the passage. Two features of many of the glucose traces shown are striking: the first obvious - that the faster growing bacteria take up the carbon sources faster than slower growing ones; the second is that there seem to be several cultures in which the carbon source is not completely consumed, but remains in the supernatant unused by the cells, often for tens of hours after the cells have reached stationary phase. This is most often the case in DH5αL (seen most obviously in Figure 5.2 ’a)’) and seems to be more pronounced in the cases of L-sorbose uptake, rather than glucose.

Although DH5αM consumes a larger proportion of the carbon source, in general, than DH5αL it still leaves some unused in the growth medium. Some of this failure to take up the carbon source might be to do with regulation of the L-sorbose metabolic genes, which might turn off prematurely in both strains, though later in the case of DH5αM. Further, there might be some other metabolic limiting factor such as Thiamine, which DH5α requires for growth (one of the genotypic differences between DH5α and CFT073), so although DH5αM is better adapted there is still a limiting factor other than carbon source availability preventing further growth (and carbon source uptake). However, whether or not glucose represses acetate assimilation during growth, it is consistently seen that there is net acetate production in many *E. coli* strains (not just those studied here) in glucose media, as discussed by Eiteman *et al* [146]. It looks in general as if there is less glucose left than L-sorbose in otherwise the same conditions for all the strains where carbon source is left in the media.

The fact that CFT073 has left some carbon source in similar conditions indicates that Thiamine may not be the limiting growth factor, but some other vital media component. Growth limiting factors will be discussed further in Chapter 7 where a stoichiometric model will be used to assess how much of each media component is required for growth. The data further show that the failure to take up all the carbon source in a medium is reversed in CFT073. In this case glucose seems to be left unused where it is sole carbon source whereas L-sorbose is fully assimilated. A possible explanation of this is that growth efficiency is sacrificed for growth rate in the case of Glucose (CFT073's doubling rate on glucose is twice that on L-sorbose), but further conclusions about this assertion require more experimentation.

The carbon source data in Figures 5.2 to 5.5 also show the variability of uptake of these carbon sources. The filled squares represent three lots of data from three biological replicates which have been grouped for clarity in these figures, but where there is a spread of carbon source results this is mostly inter-replicate variability rather than inconsistency in a single replicate that has been the cause. Where there were inconsistent results between replicates HPLC traces were reanalysed manually to check that the inferred peaks were indeed measurements of the carbon source, and in the case where no carbon source was seen that there had not been otherwise unidentified peaks that did in fact indicate the presence of the carbon source. Also, several runs were repeated to check that this was not an error in the sample analysis - which in all cases it was not. This inconsistency is apparent in Figures 5.3 'b)' and 'c)' and 5.5 'a)', all with glucose as sole carbon source.

The final optical densities achieved by cells show some variability over both strain and growth conditions. The most striking result illustrated in Table 5.3 is the low final optical density of CFT073 compared to the other strains in conditions *a priori* seen as the best conditions for growth - in glucose medium at 37 °C. This result fits well with the observation of incomplete use of carbon source by CFT073, which indicates that perhaps there is some growth inhibition independent of glucose, for instance acetate levels in the medium.

5.3.4 Acetate production

The variation between net production rates of acetate in the strains and conditions presented here is stark. Some strain and condition combinations produce no acetate at all, such as DH5αL grown at 37 °C with L-sorbose, others producing as much as 39 mM net extracellular concentration over the course of growth, such as CFT073 at 30 °C with glucose.

CFT073 produces somewhat more acetate during growth on glucose than on L-sorbose, a trend maintained in DH5αL but if not reversed, then at least reduced in DH5αM. There is a pronounced change in initial acetate production rates from DH5αL to DH5αM when grown in the passage conditions, L-sorbose at 37 °C, whereby acetate production is increased. CFT073 data for these conditions show very low initial acetate production and DH5αM production is rather high. Although this increased production of acetate is associated with an increase in the final OD for DH5αL, this higher final OD is still well short of that of CFT073, only 75 % of that value.

It seems that the adaptation observed here has prioritised growth rate over final OD and this has come with an associated cost in terms of production of acetate. For many strains of *E.coli* high growth rate on a medium of glucose as sole carbon source is associated with high acetate production and low yield and introduction of acetate into a medium inhibits both growth rate and final yield [147], so this assertion seems plausible.

Final extracellular acetate concentrations measured at or near the beginning of stationary phase are shown in Table 5.4. The size of the errors on some of these values is not necessarily because of difficulty in measuring acetate concentration (which was mostly highly accurate), but probably because the measurements were not necessarily taken at the same time in each biological replicate (for the reasons discussed above).

The first notable result in this Table is the lack of net acetate production of DH5αL when using L-sorbose as a carbon source at 37 °C. With acetate production of the same order of magnitude in CFT073 and DH5αL when grown on glucose, it is apparent that both strains do produce acetate and in similar amounts in some circumstances, but this is not the case for the L-sorbose. Net acetate production is also drastically reduced on growth in L-sorbose at 30 °C. The reactions catalysed by the enzymes from the L-sorbose operon link with *E. coli* central metabolism at fructose-6-phosphate, which should not in itself reduce acetate production. This presents the possibility that when the bacterium, due to difficulty in growth because L-sorbose is difficult to metabolise when the relevant genes are not well regulated, has a reduced growth rate (as has been seen by others [147]) and acetate production is not upregulated.

For all of the strains and conditions in this study Figure 5.6 shows acetate production as a function of growth rate. This Figure indicates minimal acetate production at low growth rates and a steep curve up from doubling rates of about $0.3h^{-1}$ to $0.6h^{-1}$, then fairly steady production of acetate above that doubling rate. Growth on L-sorbose and growth on glucose are shown as separate data series in this Figure, to compare the two carbon sources. The production amounts do not appear to be much different between the two carbon sources where they overlap, but this overlap is quite small $(0.3 - 0.5h^{-1})$ so only a trend can be seen and no firm conclusions can be made about the dependence of acetate production on carbon source at a given growth rate. Although there are not enough datapoints here to come to quantitative conclusions about the relationship between acetate production and growth rate this Figure shows strikingly the qualitative invariance of response between the *E. coli* strains and between different conditions.

An alternative hypothesis that L-sorbose fails to suppress acetate reassimilation use as a carbon source is supported by the low but measureable acetate concentrations in DH5αL growths in L-sorbose during the exponential phase of growth, which then drop to zero shortly thereafter. The increased production of acetate in DH5αM over DH5αL in all conditions, faster growth rates and higher final ODs in L-sorbose indicate that something regulated in parallel with the acetate production is worth the cost in terms of energy, carbon and growth inhibition at high concentrations, although as mentioned above there is no evidence that acetate production is disproportionately greater in DH5αM, when the change in growth rate is taken into account. This adaptation comes at a price though, in the form of lower final ODs in glucose, the specialisation of DH5αM to the L-sorbose carbon source; whereas higher acetate production rates are seen in CFT073, final ODs are higher than in either DH5αL or DH5αM.

5.3.5 Adaptation and flux changes in DH5αL

The above observations clearly show that flux distributions are quite different between DH5αL and DH5αM, even in very similar environmental conditions. The use of linear optimisation to model this change of flux distribution has been done in Chapter 7.

5.3.6 Glucitol-6-phosphate dehydrogenase activity

Figure 5.7 shows the activity profile of crude extract from DH5αL grown at 37 °C in L-sorbose as a function of temperature. The model from Lee *et al* [120] shown as the solid line on this figure fits well with the parameters determined by minimisation of sum least squares differences between the data and the model. With four parameters it could be possible to fit many implausible shapes, but the shape of all the fitted lines against the datapoints was peaked at consistent and reasonable values, and the tails of all the lines go to zero at high and low temperatures. These factors provide evidence that the model is plausible as a representation of the underlying phenomena of enzymatic action and the actual characteristics of the enzymes tested.

The data presented in Figure 5.8 show clearly that there is little difference in the temperature sensitivity of the glucitol-6-phosphate dehydrogenase before and after passage. This indicates that the changes during passage are not due to change in this enzyme, but due to some other factor or factors. The *Klebsiella pneumoniae* as a positive control shows that differences can be distinguished between temperature sensitivities of different enzymes. All of the other strains are not significantly different, but *Klebsiella pneumoniae* is statistically significantly different at the 5 % level. Sprenger and Lengeler [118] noted heat sensitivity of this enzyme and determined heat inactivation of it after 10 minutes at particular temperatures, indicating this sensitivity. Although those results are not directly comparable to the results presented here, due to the differing methods of enzyme extraction and type of heat sensitivity tested, the two results appear to corroborate one another.

5.4 Conclusions

In this Chapter it has been shown that a bacterium presented with novel metabolic genes will adapt to these genes over a period of tens of generations when passaged in selective media. However it has also shown that in this case it does not achieve the efficiency of use achieved by the wild-type strain over 100 generations, even in the particular conditions in which the passage was undertaken.

This adaptation does not appear to have affected the metabolic relationship between carbon source uptake and acetate production that comes from overflow metabolism. One reason that might account for the change, the heat sensitivity of the glucitol-6-phosphate dehydrogenase in the L-sorbose operon of CT073, has been ruled out by an assay of this heat sensitivity around the temperatures at which the experiments were undertaken.

The mechanisms behind the adaptation of DH5αL over the period of passage to DH5αM are yet to be elucidated. Remaining possibilities to explain the differences in growth rates and final ODs achieved are: changes in some other genes - such as those encoding transport proteins for L-sorbose (changes in transport proteins have been observed during continuous culture as shown by Tsen *et al* [148]) - or regulatory changes. Whether any part of the adaptation is specific to the L-sorbose genes, rather than a general improvement in metabolic capability (observed above by improvement in growth characteristics on glucose) is yet to be elucidated.

Chapter 6

The use of genome sequence comparisons for inferring metabolic models in bacteria

6.1 Preface

Bacterial genomes are being sequenced and published at an exponentially increasing rate. Figure 6.1 shows published genomes of bacteria according to GOLD (Genomes OnLine Database v2.0) as of June 8, 2009 [149] and clearly shows the exponential increase in this number, though it should be borne in mind that increasingly genomes are sequenced without subsequent publication. The current rate of publication is around 200 per year and the slope of the best fit line calculated from Figure 6.1 ‘b)’ indicates that this number is doubling every two years, implying that by 2010 there will be at least one newly sequenced bacterial genome per day being published. This rate of sequencing presents amazing opportunities for probing many fundamental principles of bacterial evolution, including the evolution of their

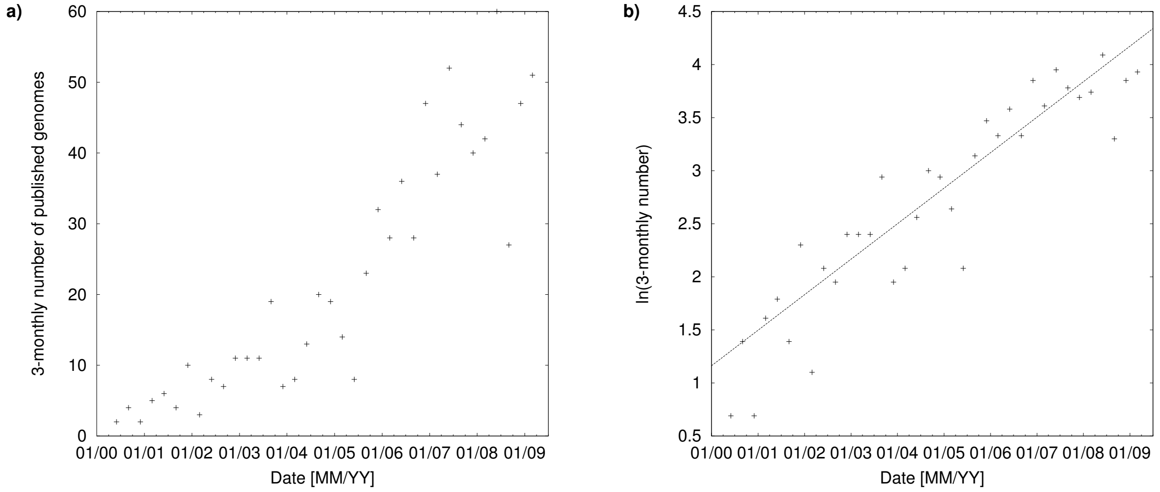


Figure 6.1: Showing the number of bacterial genome sequences published per quarter since 2000. 'a)' shows actual values and 'b)' shows them on a natural logarithm scale, showing a clear exponential pattern. The line of best fit for these data assuming an exponential increase has been included on plot 'b)'.

metabolic networks.

Reliable inferences of gene function are crucial to establishing the metabolic networks of newly sequenced bacteria and much effort has been channelled into establishing reliable annotation strategies for certain key reference genomes (such as using gene ontologies [150]), although BLASTp against the non-redundant protein database (nr, NCBI) is still generally used to annotate newly sequenced *E. coli* (for instance [28, 53]). Genome comparisons are usually done without reference to annotation, but rather by using the BLAST score ratio (BSR) method of Rasko *et al* [151], which has been used to survey many of *E. coli* with complete genome sequences to look at the pangenome of *E. coli* [124].

MetaCyc's Pathologic program [82] produces gene complement comparisons and metabolic networks by comparison of annotation, but if annotation is based only on BLASTp comparisons against NCBI's non-redundant database, then it is vulnerable to the problem of successive inferred annotation. This problem can potentially be overcome by limiting comparisons to only experimentally characterised genes, although identification of these can be difficult as there is no field in GenBank entries to indicate the source of a gene annotation. Feature notes are often used to indicate close orthologs to other genes, but do not in themselves determine which gene's function was experimentally elucidated. One potential way of overcoming this problem when attempting to reconstruct a metabolic model is to find a well characterised model organism on which a great amount of work has been done to experimentally verify gene functions (such as *E. coli* MG1655 in the case of *E. coli*) and take inferences of gene function from just that strain.

The amount of information gathered about MG1655 has allowed whole genome-scale models of metabolism in this organism to be established based on these biochemical characteristics, for instance model iAF1260 [1] which takes into account 1260 ORFs (out of 4149 protein coding genes in MG1655) to produce a stoichiometric metabolic model of the metabolism of MG1655. Using the data relating these ORFs to chemical reactions within the cell, and assuming that in general reactions in closely related strains of bacteria are associated with the same ORFs, it should be possible to reconstruct the subset of the reactions found in MG1655 that are also present in any other bacterium if the set of genes common to the two strains is known.

The work presented in this chapter shows the application of a gene function inference strategy to establishing common genes between a set of closely related bacteria and MG1655, thus through the GPR relationships compiled by Feist *et al* for iAF1260 [1] inferring the subset of the metabolic network of MG1655 that is present in each of these bacteria. This procedure has been automated so that a new comparison can be achieved through knowledge only of the GI number of the nucleotide entry in GenBank of a new bacterial genome sequence. All but one of the programs used to achieve this are freely available for download from the Internet, the exception being MATLAB, which has been used to combine the steps in the inference process.

The tools required for the process are achieved using freely available computer applications and languages: MySQL, Perl, DiagHunter, a local BLAST server and several of the Bioinformatics-specific modules available for Perl, which are named in the following way: ‘Bio::Perl’, which contains bioinformatic tools for Perl and ‘Bio::DB::GenBank’ which contains tools for communicating with the GenBank website for downloading the genome sequences required for this work. This process of inferring a metabolic model takes of the order of 10 minutes per organism and thus a large number of genome comparisons and therefore metabolic networks and models can be compared quickly.

MetaCyc [78] has been used to validate the comparisons done in this study, as it is the most comprehensive publicly available resource of gene complement comparisons with MG1655 for the purposes of metabolic research. Lists of essential genes from various sources have also been used to validate these comparisons. The model iAF1260 [1] has been used as the basis of all the models produced in this work.

6.2 Data

6.2.1 Bacterial Genome Sequences

The strains used for the metabolic inference are listed in Table 6.1 which gives strain names, their pathotype (where available), the GI numbers of the GenBank entries for their genome sequences and whether they have been analysed by Pathologic and posted on the MetaCyc website.

Phylogenetic relationships between the strains studied here (according to 16S rRNA comparisons) are shown in Figure 6.2, calculated on the Ribosomal Database Project website [152, 153].

6.3 Results

6.3.1 Genomic comparison data

Genomic comparisons were conducted according to the method detailed in Section 3.4, using DiagHunter to combine positional and identity data for a more robust inference than BLAST data alone. Variations of the minimum diagonal length for inference of gene synteny were used in each comparison in DiagHunter in order to assess the effect of this criterion on gene synteny inferences. Minimum diagonal lengths of 1, 2 and 3 were used for these comparisons. Data for these three different comparisons can be seen in Supplementary Table 4.

Tables 6.2 and 6.3 give an overview of the data contained in Supplementary Table 4, the complete set of inferences carried out in this study. For comparison details of the equivalent comparison held on the MetaCyc web-

Table 6.1: All bacterial strains the complete genome sequences of which have been compared in this study. Where the strain is *E. coli* this species label has been omitted for clarity. Pathotype has been included where possible. Accession number and GI number refer to the Genbank nucleotide entries which include complete sequences and annotated genes for these strains. The indication of existence of an entry for each strain in MetaCyc has been included to show the coverage of these across the species included in this study.

Strain Name	Type	Accession Number	GI number	In MetaCyc?
ATCC 8739	Commensal	CP000946	170018061	-
BW2952	Commensal	CP001396	238899406	-
IAI1	Commensal	CU928160	218552585	-
K-12 DH10B	Commensal	CP000948	169887498	-
K-12 HS	Commensal (Gastrointestinal)	CP000802	157159467	-
K-12 MG1655	Commensal (Gastrointestinal)	U00096	49175990	+
SE11	Commensal (Gastrointestinal)	AP009240	209917191	-
SMS-3-5	Commensal (environmental)	CP000970	170679574	+
ED1a	Avirulent O81	CU928162	218687878	-
CFT073	UPEC (uropathogenic)	AE014075	26111730	+
536	UPEC	CP000247	110341805	+
IAI39	UPEC	CU928164	218698419	-
UMN026	UPEC	CU928163	218430358	-
UTI89	UPEC	CP000243	91209055	+
E2348	EPEC	FM180568	215485161	-
E24377A	ETEC	CP000800	157154711	-
55989	EAggEC	CU928145	218693476	-
O157:H7 str. Sakai	EHEC (enterohaemorrhagic)	BA000007	15829254	+
O157:H7 EDL933	EHEC	AE005174	56384585	+
O157:H7 str. EC4115	EHEC	CP001164	209395693	-
S88	ECNM (neonatal meningitis)	CU928161	218556939	-
APEC 01	Avian pathogenic	CP000468	117622295	-
Shigella boydii CDC 3083-94	Bacillary Dysentery	CP001063	187730020	-
Shigella boydii Sb227	Bacillary Dysentery	CP000036	82542618	-
Shigella dysenteriae Sd197	Bacillary Dysentery	CP000034	81239530	-
Shigella flexneri 2a str. 2457T	Bacillary Dysentery	AE014073	30061571	+
Shigella flexneri 2a str. 301	Bacillary Dysentery	AE005674	24080789	+
Shigella flexneri 5 str. 8401	Bacillary Dysentery	CP000266	110804074	+
Shigella sonnei Ss046	Bacillary Dysentery	CP000038	74310614	+
Edwardsiella ictaluri 93-146	Enteric septicaemia (Catfish)	CP001600	238917983	-
Escherichia fergusonii ATCC 35469	UTI and wound, bacteraemia	CU928158	218547440	-
Klebsiella pneumoniae subsp. *pneumoniae* MGH 78578	UTI and Pneumonia	CP000647	150953431	+
Legionella pneumophila str. Corby	Pneumonia-like	CP000675	148358139	-
Proteus mirabilis str. HI4320	UTI and others (opportunistic)	AM942759	172046403	-
Salmonella typhimurium LT2	Gastroenteritis	AE006468	16445344	+
Yersinia pestis KIM	Bubonic Plague	AE009952	22002119	+

site [82], which will be discussed below, have been included in these Tables. Table 6.3 shows a breakdown of inference-by-inference differences between MetaCyc and the DiagHunter (DH) method used in this study.

Table 6.4 shows an overview of the results obtained by the DH method separated by category of gene. Categories were taken according to the MultiFun categorisation in GenProtEC [154] as they cover all types of gene

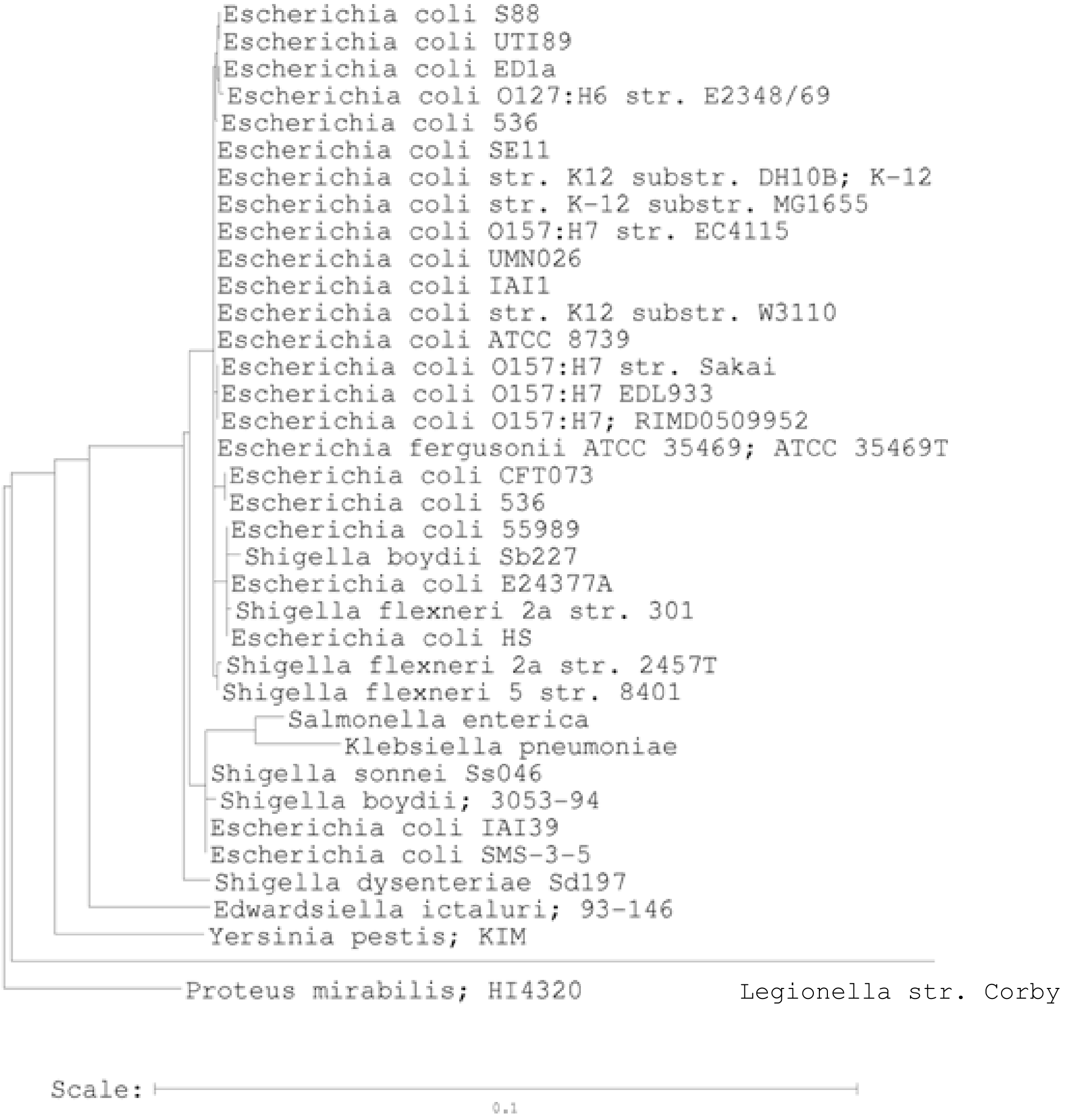

Escherichia coli S88
Escherichia coli UTI89
Escherichia coli ED1a
Escherichia coli O127:H6 str. E2348/69
Escherichia coli 536
Escherichia coli SE11
Escherichia coli str. K12 substr. DH10B; K-12
Escherichia coli str. K-12 substr. MG1655
Escherichia coli O157:H7 str. EC4115
Escherichia coli UMN026
Escherichia coli IAI1
Escherichia coli str. K12 substr. W3110
Escherichia coli ATCC 8739
Escherichia coli O157:H7 str. Sakai
Escherichia coli O157:H7 EDL933
Escherichia coli O157:H7; RIMD0509952
Escherichia fergusonii ATCC 35469; ATCC 35469T
Escherichia coli CFT073
Escherichia coli 536
Escherichia coli 55989
Shigella boydii Sb227
Escherichia coli E24377A
Shigella flexneri 2a str. 301
Escherichia coli HS
Shigella flexneri 2a str. 2457T
Shigella flexneri 5 str. 8401
Salmonella enterica
Klebsiella pneumoniae
Shigella sonnei Ss046
Shigella boydii; 3053-94
Escherichia coli IAI39
Escherichia coli SMS-3-5
Shigella dysenteriae Sd197
Edwardsiella ictaluri; 93-146
Yersinia pestis; KIM
Proteus mirabilis; HI4320
Legionella str. Corby
Scale:
0.1

Figure 6.2: A phylogenetic tree of the strains analysed in this Chapter inferred by 16S rRNA sequence comparisons [152, 153], with outgroup defined as *Yersinia pestis*.

function.

6.3.2 iAF1260 gene synteny

The subset of genes linked through the GPR relationships in iAF1260 to metabolism was then compared, and a summary of the results for those

Table 6.2: Summary statistics of the genome comparisons produced in this study and comparative set of MetaCyc comparisons. This gives an idea of the number of individual gene comparisons required for this work. Assuming each of the 36 strains used in this comparison were compared with each other, and each has approximately 4000 genes, there would be 2.7 million comparisons, therefore comparing with a single reference genome to begin with reduces computational burden by the order of tenfold for this number of genome sequences.

Property	Amount
Total no. of genomes used	36
Total no. of genes compared	4095
Total no. of synteny inferences	143,325
Total no. of positive inferences	114,677
No. of these genomes in MetaCyc	16
No. of genes positively identified in MetaCyc	3907
No. of gene inferences in MetaCyc	58,605
No. of positive gene inferences in MetaCyc	47,805

Table 6.3: A summary of the differences in inferences between this study and MetaCyc. These data have been constructed those inferences for which there is a MetaCyc inference to compare with the study data.

Property	Minimum diagonal length = 1	Minimum diagonal length = 2	Minimum diagonal length = 3
No. positive inferences	49,975	47,755	47,090
% inferred by DH, but not MetaCyc	5.2	2.5	2.0
% inferred by MetaCyc, but not DH	1.5	2.6	3.2
% of all inferences differing between DH and MetaCyc	6.7	5.0	5.3

genes can be seen in Table 6.5. The genes included in this comparison are indicated in Supplementary Table 4, Worksheet ‘Minimum diagonal length 2’, Column AO. Organisms were separated by relatedness to see how different parts of metabolism were affected differently over different phylogenetic distances. Table 6.6 shows a summary of the essential genes according to the iAF1260 model.

Gene complement is important in establishing metabolic models, as it gives an indication of which reactions are catalysed by gene products in a particular organism. The GPR relationships in iAF1260 give the reactions in

Table 6.4: An overview of gene synteny comparisons by gene category (as defined by the MultiFun catalogue of gene roles used in GenProtEC [154]). Where genes have been classified as unknown or not been classified in GemProtEC they have been added to the 'Unclass./Unknown/Other' category as have genes from three categories: 'location of gene products', 'DNA sites', and 'cryptic genes' which did not contain enough genes to provide reliable data for the table. % conserved values have been evaluated by taking the mean retention percentage of all genes in a particular category, and standard deviations ('Std dev') have been included to indicate the spread of synteny conservation within these categories. Standard errors on the mean retentions indicate how good the estimation of gene retention is by category. Separate data for essential genes (see Section 6.4.2.1 and Supplementary Table 3 for details of these genes) have been included, broken down by category.

Category	All genes				Essential genes		
	Number	% conserved	Std error	Std dev	Number	% conserved	Std dev
Metabolism	1597	85.74	0.43	17.08	85	96.99	2.82
Information	513	87.48	0.77	17.52	72	97.38	4.19
Regulation	59	81.31	2.26	17.33	0	-	-
Transport	371	83.2	0.89	17.18	13	98.93	1.41
Cell processes	126	84.68	1.58	17.76	14	98.02	1.3
Cell structure	228	81.74	1.22	18.43	3	100	0
Extrachromosomal	214	30.75	1.74	25.4	0	-	-
Unclass./Unknown /Other	987	77.69	0.67	21.03	4	86.11	25.96

an organism, given the genes available. Table 6.7 shows the reactions conserved across the species included in this study, broken down by category and organism relationship to MG1655, as in Tables 6.5 and 6.6.

6.4 Discussion

6.4.1 Synteny comparisons

6.4.1.1 Reliability of comparisons using the DH method

DiagHunter has previously been used for inference of plant and human gene synteny, but to our knowledge it has never been applied to a bacterial system. This use of gene context data should improve reliability of gene synteny inferences and these give significantly different inferences of gene complements to other databases of such data, such as MetaCyc. It is therefore

Table 6.5: A summary of how many of the genes included in iAF1260 are synteny conserved over the bacteria in this study. Subsystems are defined in iAF1260, and where genes are in multiple subsystems they have been placed in that row of the table. The *Shigella* column contains several non-*Shigella* strains: the *Escherichia fergusonii*, the *Salmonella* and the *Klebsiella*, as their 16S rRNA indicates that they are rather more closely related to *Shigella* than the four more distantly related organisms. Of these four, three are also in the order of *Enterobacteriales* (family of *Enterobacteriaceae*), whereas *Legionella pneumophila* str. Corby is a member of a different order of *Gammaproteobacteria*, *Legionellales*. Results were therefore separated to reflect previously inferred relatedness for ready comparison of metabolic conservation.

Subsystem	No. genes (iAF1260)	Conserved overall %	Conserved in *E. coli* %	Conserved in *Shigella* %	Conserved in other Enterobacteriaceae %	Conserved in *Legionella* %
Number of strains		**36**	**22**	**10**	**3**	**1**
Alanine and Aspartate Metabolism	9	91.7	98.0	94.4	77.8	0.0
Alternate Carbon Metabolism	134	73.9	88.2	64.3	41.5	1.5
Anaplerotic Reactions	9	89.2	98.5	85.6	74.1	0.0
Arginine and Proline Metabolism	33	84.9	94.9	82.7	60.6	9.1
Cell Envelope Biosynthesis	48	89.1	94.1	90.8	75.7	22.9
Citric Acid Cycle	20	91.9	98.0	93.0	81.7	20.0
Cofactor and Prosthetic Group Biosynthesis	125	94.1	99.3	95.5	83.2	17.6
Cysteine Metabolism	10	93.3	99.1	97.0	76.7	0.0
Folate Metabolism	6	96.8	99.2	98.3	88.9	50.0
Glutamate Metabolism	8	85.8	96.6	85.0	45.8	0.0
Glycerophospholipid Metabolism	17	94.1	98.9	95.9	88.2	11.8
Glycine and Serine Metabolism	8	92.4	99.4	96.3	66.7	0.0
Glycolysis/Gluconeogenesis	27	92.3	98.0	94.1	84.0	11.1
Glyoxylate Metabolism	2	70.8	97.7	40.0	0.0	0.0
Histidine Metabolism	9	99.7	100.0	100.0	100.0	88.9
Ion Transport and Metabolism	65	91.6	98.0	93.8	76.9	0.0
LPS Biosynthesis / Recycling	45	75.7	80.7	76.7	60.0	8.9
Membrane Lipid Metabolism	12	93.1	97.7	91.7	83.3	41.7
Methionine Metabolism	10	85.0	90.0	87.0	76.7	0.0
Methylglyoxal Metabolism	5	91.1	97.3	96.0	66.7	0.0
Multiple Subsystems	94	91.2	98.2	92.0	73.8	14.9
Murein Biosynthesis	8	92.4	98.3	87.5	83.3	37.5
Murein Recycling	26	92.7	97.9	97.3	82.1	3.8
Nitrogen Metabolism	16	78.3	92.3	70.0	35.4	0.0
Nucleotide Salvage Pathway	50	89.3	96.6	86.8	72.0	14.0
Oxidative Phosphorylation	86	92.5	97.7	91.5	77.9	31.4
Pentose Phosphate Pathway	11	84.8	92.1	84.5	63.6	9.1
Purine and Pyrimidine Biosynthesis	21	96.6	99.8	99.0	96.8	4.8
Pyruvate Metabolism	20	87.6	97.0	83.5	61.7	15.0
Threonine and Lysine Metabolism	13	93.6	100.0	93.1	89.7	0.0
Transport, Inner Membrane	222	84.8	94.4	81.9	65.0	8.6
Transport, Outer Membrane	17	79.4	92.2	70.6	47.1	0.0
tRNA Charging	24	94.3	97.9	96.3	86.1	33.3
Tyr, Trp and Phe Metabolism	19	94.4	98.8	96.3	84.2	36.8
Unassigned	20	83.5	95.0	75.5	55.0	5.0
Val, Leu, and Ile Metabolism	11	92.4	96.7	98.2	81.8	0.0

in principle testable whether using this additional contextual information is more reliable than BLAST comparisons from which most annotations are taken. Also, this potential increase of reliability comes at a very small com-

Table 6.6: A summary of the essential genes (those without which the iAF1260 model does not function) divided by category and relationship to MG1655. The order has been preserved as in Table 6.5; those functional categories that do not contain any essential genes are indicated on the Table with dashes.

Subsystem	No. essential Genes (iAF1260)	Conserved overall %	Conserved in *E. coli %*	Conserved in *Shigella %*	Conserved in other Enterobacter–iaceae %	Conserved in *Legionella %*
Number of strains		**36**	**22**	**10**	**3**	**1**
Alanine and Aspartate Metabolism	0	-	-	-	-	-
Alternate Carbon Metabolism	1	100	100.0	100.0	100.0	100.0
Anaplerotic Reactions	0	-	-	-	-	-
Arginine and Proline Metabolism	9	94.14	100.0	100.0	63.0	0.0
Cell Envelope Biosynthesis	18	91.98	92.7	97.8	83.3	44.4
Citric Acid Cycle	2	93.06	93.2	100.0	100.0	0.0
Cofactor and Prosthetic Group Biosynthesis	75	95.52	99.5	98.1	83.1	18.7
Cysteine Metabolism	5	95	100.0	100.0	73.3	0.0
Folate Metabolism	1	88.89	100.0	90.0	33.3	0.0
Glutamate Metabolism	0	-	-	-	-	-
Glycerophospholipid Metabolism	6	96.76	98.5	100.0	100.0	16.7
Glycine and Serine Metabolism	0	-	-	-	-	-
Glycolysis/Gluconeogenesis	2	91.67	97.7	95.0	66.7	0.0
Glyoxylate Metabolism	0	-	-	-	-	-
Histidine Metabolism	8	100	100.0	100.0	100.0	100.0
Ion Transport and Metabolism	2	94.44	100.0	100.0	66.7	0.0
LPS Biosynthesis / Recycling	28	81.15	85.2	83.2	66.7	14.3
Membrane Lipid Metabolism	5	98.89	99.1	100.0	100.0	80.0
Methionine Metabolism	4	94.44	100.0	97.5	75.0	0.0
Methylglyoxal Metabolism	0	-	-	-	-	-
Multiple Subsystems	16	95.49	98.9	96.9	89.6	25.0
Murein Biosynthesis	0	-	-	-	-	-
Murein Recycling	0	-	-	-	-	-
Nitrogen Metabolism	0	-	-	-	-	-
Nucleotide Salvage Pathway	3	100	100.0	100.0	100.0	100.0
Oxidative Phosphorylation	0	-	-	-	-	-
Pentose Phosphate Pathway	0	-	-	-	-	-
Purine and Pyrimidine Biosynthesis	18	96.6	99.7	98.9	96.3	5.6
Pyruvate Metabolism	0	-	-	-	-	-
Threonine and Lysine Metabolism	8	95.83	100.0	100.0	83.3	0.0
Transport, Inner Membrane	3	94.44	97.0	96.7	100.0	0.0
Transport, Outer Membrane	4	94.44	100.0	95.0	83.3	0.0
tRNA Charging	0	-	-	-	-	-
Tyr, Trp and Phe Metabolism	11	96.97	99.2	100.0	84.8	54.5
Unassigned	0	-	-	-	-	-
Val, Leu, and Ile Metabolism	5	93.89	94.5	100.0	100.0	0.0

putational cost, a single run of DiagHunter comparing two bacteria on an Intel Pentium®M 1.7 GHz computer with 1 Gb RAM takes approximately 4 seconds.

Table 6.7: A summary of reactions conserved in the iAF1260 model according to DH method. Categories are as in Table 6.5. The reactions included in the table do not represent all reactions in iAF1260, there are 437 reactions that do not have a gene complex associated with them in the model and have been omitted from the comparison as they are inferred to be present irrespective of gene complement.

Subsystem	No. of reactions	Conserved overall %	Conserved in *E. coli %*	Conserved in *Shigella %*	Conserved in other Enterobacter–iaceae %	Conserved in *Legionella %*
Number of strains		**36**	**22**	**10**	**3**	**1**
Alanine and Aspartate Metabolism	9	94.14	99.49	97.78	74.07	0.00
Alternate Carbon Metabolism	173	77.78	90.62	70.92	31.98	1.16
Anaplerotic Reactions	8	94.44	100.00	96.25	70.83	25.00
Arginine and Proline Metabolism	37	89.19	98.65	87.84	50.45	10.81
Cell Envelope Biosynthesis	134	93.45	97.05	96.49	77.86	30.60
Citric Acid Cycle	13	91.88	97.55	91.54	76.92	15.38
Cofactor and Prosthetic Group Biosynthesis	148	94.18	99.26	95.07	79.28	18.24
Cysteine Metabolism	12	93.52	98.86	97.50	69.44	8.33
Folate Metabolism	6	92.13	99.24	90.00	72.22	16.67
Glutamate Metabolism	6	89.81	97.73	93.33	50.00	0.00
Glycerophospholipid Metabolism	218	94.85	99.19	96.10	87.31	9.63
Glycine and Serine Metabolism	14	92.06	98.38	97.86	54.76	7.14
Glycolysis/Gluconeogenesis	22	94.19	98.76	98.18	75.76	9.09
Glyoxylate Metabolism	4	82.64	98.86	70.00	33.33	0.00
Histidine Metabolism	10	99.72	100.00	100.00	100.00	90.00
Ion Transport and Metabolism	73	92.16	99.88	94.38	58.90	0.00
LPS Biosynthesis / Recycling	57	75.49	80.38	76.14	60.23	7.02
Membrane Lipid Metabolism	42	92.79	99.13	93.81	70.63	9.52
Methionine Metabolism	14	82.14	86.69	84.29	69.05	0.00
Methylglyoxal Metabolism	5	92.78	99.09	94.00	73.33	0.00
Murein Biosynthesis	10	99.17	100.00	100.00	100.00	70.00
Murein Recycling	34	93.95	98.26	97.94	78.43	5.88
Nitrogen Metabolism	13	66.88	80.77	56.92	20.51	0.00
Nucleotide Salvage Pathway	130	94.25	98.88	95.46	79.49	24.62
Oxidative Phosphorylation	52	90.49	98.43	88.27	66.67	9.62
Pentose Phosphate Pathway	10	93.89	98.18	97.00	76.67	20.00
Purine and Pyrimidine Biosynthesis	24	96.06	99.24	99.17	93.06	4.17
Pyruvate Metabolism	9	91.05	97.47	87.78	85.19	0.00
Threonine and Lysine Metabolism	18	94.29	99.24	98.33	75.93	0.00
Transport, Inner Membrane	277	86.33	94.96	85.13	55.23	1.81
Transport, Outer Membrane	34	87.58	97.99	85.59	47.06	0.00
Transport, Outer Membrane Porin	247	91.67	100.00	100.00	33.33	0.00
tRNA Charging	22	95.71	97.93	99.09	89.39	31.82
Tyr, Trp, and Phe Metabolism	22	94.19	98.76	94.55	77.27	40.91
Unassigned	22	89.14	97.31	86.36	65.15	9.09
Val, Leu and Ile Metabolism	15	95.56	97.27	100.00	100.00	0.00

6.4.1.2 Lists of genes in GenBank

Although every effort was made to keep up-to-date genome sequences for comparisons this was not possible in some circumstances. For instance, iAF1260 was based on a GenBank list of genes from 2008 so this list was used where necessary. The changes in the GenBank list of protein coding genes between that time and 04/06/09 amount to the removal of 12 gene annotations and the addition of 28 more, less than 1 % of the total number of genes annotated in the MG1655 genome sequence. None of these changes affect any genes used in the iAF1260 gene-protein-reaction relationships that determine the metabolic network structure of MG1655 (discussed in the next chapter). Apart from genes that have been newly annotated between the time of the production of iAF1260 and June 2009 (b4586 - b4689, 52 genes), there are only two genes in the list of genes from 2009 that have not been used in the comparisons done in this study: b1500 (a two component system connector membrane protein, EvgSA to PhoQP) and b4543 (an uncharacterised gene predicted to encode a protein). These two genes have been omitted for the purposes of this study and 4095 of the 4149 currently annotated protein coding genes in MG1655 have been used for this comparison.

6.4.1.3 MetaCyc comparisons and DiagHunter parameter selection

MetaCyc's Pathway Tools have been used to infer metabolic pathways conserved between species through their shared gene complement. These comparisons are freely available online and will be used here to validate and calibrate the genomic comparison implemented above. According to the MetaCyc website the gene comparisons are taken from the annotations of sequenced genomes [82], which are usually inferred by the sequencers of the genome sequence in question by mutual best BLAST hits from the BLAST database. Data from MetaCyc have been downloaded through the 'Comparative Analysis and Statistics' page (`http://metacyc.org-/comp-genomics`) comparison of orthologs. These comparisons provide locus tags for identification and through these it is possible to assign Blattner (b) numbers [35] to the vast majority of them. Those genes for which b numbers could be assigned unambiguously for MG1655 were compared with those strains for which data has been posted on the MetaCyc website, using the comparison tool on that website, and the results can be seen in Worksheet 4 of Supplementary Table 4.

It can be seen in Table 6.2 that approximately half of the genomes have entries in MetaCyc as do the vast majority of genes, which gives a large overlap for the purposes of comparative validation. Table 6.3 shows the disparities between the data obtained by this analysis and MetaCyc. With the minimum diagonal length parameter of DiagHunter set at a value of '2' the result is most consistent with MetaCyc, also it is the least biased in terms of positive disagreements and negative disagreements. This genomic comparison will therefore be used in the rest of this Chapter and references to the DH method refer to the inferences based on that parameter selection unless otherwise stated.

Setting of the minimum diagonal length to '2' has been done with reference to the MetaCyc data. The physical interpretation of this parameter is that it has eliminated all candidate genes for functional conservation that are displaced with respect to the gene order in MG1655 without corresponding movement of any adjacent genes. It therefore seems that genes do not tend to move around the genome individually, according to MetaCyc. The privileged position of MG1655's gene order is not arbitrary; it is one of the most intensively studied organisms and has much of its biochemical function (and relationship to its genome sequence) experimentally verified so its genes are known to function in the order in which they are positioned on the genome. This is not true of the vast majority of other bacteria which, as in this study, must have their biochemical functions inferred from computational comparisons with the reference organism.

For the majority of strains the results that are inconsistent between this study and MetaCyc are split evenly between those inferred present in MetaCyc and not by the DH method and those inferred present by the DH method but not by MetaCyc. However, *Klebsiella pneumoniae* and *Yersinia pestis* each have considerably more genes inferred present by MetaCyc. The inferences are more consistent the closer the strains are in the phylogeny. Whereas the *E. coli* are inconsistent in 3 % of cases, the disparity between MetaCyc and the DH method goes up to around 11 % for *Legionella*. These observations are consistent with the differing methods of inference between Pathologic and the DH method. Since the DH method is free from any successive inferred annotation it is more conservative in inferring shared function between distantly related organisms, so is less consistent with MetaCyc as phylogenetic distance between the strains increases.

6.4.1.4 Applicability of synteny comparison

In general of the strains included in this study, a large number of genes are conserved over a large majority of the strains and inferences overall are closely consistent with those in MetaCyc. The overall positive synteny inference is a little over 80 % and for genes included in iAF1260 it is around

88 %, indicating that in general the metabolic genes in MG1655 are more frequently conserved in other species than are other types of gene.

The phylogenetic distance between MG1655 and *Legionella pneumophila* str. Corby has rendered the inferred gene synteny comparison very sparse. The absence of inferred conserved genes in *Legionella pneumophila* str. Corby is not necessarily indicative of lack of functional conservation, but more that even orthologs with the same function are too dissimilar to be confident of a positive inference. Corby is inferred to contain just 13 % of the genes in iAF1260 and $<$ 25 % of the essential model genes. These observations do not necessarily indicate that functions are not conserved between the two bacteria, but that bioinformatic evidence of synteny conservation between the bacteria is insufficient to warrant the inference of shared functions for a majority of genes.

It is apparent from these observations that inferring a meaningful metabolic model from a single model organism is not possible bioinformatically across an entire taxonomic class, but is perhaps limited to a single order or even a smaller set of bacteria. The DH method was used on two other *gammaproteobacteria* in different orders to test this assertion. The GenBank genome sequences of *Pseudomonas aeruginosa* PAO1 (GI 110227054) and *Haemophilus influenzae* Rd KW20 (GI 6626252) were compared to MG1655 and conserved metabolic genes amounted to just 30 % and 21 % of the total number of iAF1260 genes respectively and 43 % and 31 % of the iAF1260 essential genes, so even if metabolic networks were reconstructed from these data, working stoichiometric models for these bacteria would require a huge amount of further work.

If Corby is removed from the comparisons, 83 % of all inferences are positive and 90 % of iAF1260 genes are inferred to be conserved. It should be stressed that this does not necessarily indicate typical gene overlaps within or between orders of organisms, since the organisms here are not necessarily a representative sample of all *Enterobacteriaceae*, but increasing numbers of sequenced genomes in the near future will continue to fill the gaps and determine the extent to which a single reference model (such as iAF1260) can be used to infer other metabolic models.

6.4.1.5 False negative synteny inferences from MG1655 against self comparison

A test of MG1655 against itself was conducted to assess the number of false-negatives produced using the DH method and 4068 out of 4095 genes were inferred as synteny conserved (99.34 %), thus the false negative rate for an identical genome sequence is less than 0.5 %. Table 6.8 shows the genes not inferred using the DH method. The majority of these are insertion sequence (IS) or phage genes, and the failure of the DH method in these cases is explainable. IS and phage elements are often present in multiple places of the genome of an individual strain, but in the interests of computational expediency the DH method is limited to 3 BLAST hit results per gene. Thus where there are more than 3 identical copies of a particular gene (for instance there are 5 InsE genes in MG1655) only three have been inferred to be conserved. The genes highlighted in grey in Table 6.8 are not IS or phage elements, so there is no obvious reason for their exclusion by this method of inference.

Although a thorough examination of the genes highlighted here would be required for further refinement of the DH method, here only the gene in

Table 6.8: Genes wrongly inferred to be absent (false negatives) according to the DH method of gene inference.

B number	In MetaCyc?	GI protein	Function
b0259	-	16128244	IS5 transposase and trans-activator
b0298	+	16128283	IS3 element protein InsE
b0299	+	16128284	IS3 element protein InsF
b0360	+	145698224	KpLE2 phage-like element; IS2 element repressor InsA
b0361	+	16128346	KpLE2 phage-like element; IS2 element transposase InsAB'
b0372	-	16128357	IS3 element protein InsF
b0373	-	16128358	IS3 element protein InsE
b0552	-	16128535	IS5 transposase and trans-activator
b0656	-	16128639	IS5 transposase and trans-activator
b1331	-	16129292	IS5 transposase and trans-activator
b1402	-	16129363	KpLE2 phage-like element; IS2 element transposase InsAB'
b1403	-	145698254	KpLE2 phage-like element; IS2 element repressor InsA
b1573	-	16129532	Qin prophage; predicted protein
b1588	+	145698266	oxidoreductase subunit
b1597	-	90111305	acid shock-inducible periplasmic protein
b1616	+	16129574	glucuronide transporter
b1832	+	49176157	conserved protein
b1996	-	16129937	KpLE2 phage-like element; IS2 element transposase InsAB'
b1997	-	145698282	KpLE2 phage-like element; IS2 element repressor InsA
b2030	-	16129971	IS5 transposase and trans-activator
b2192	-	16130129	IS5 transposase and trans-activator
b2558	+	16130483	predicted transglycosylase
b2598	-	16130519	pheA gene leader peptide
b2635	-	90111471	CP4-57 prophage; predicted inner membrane protein
b3643	+	157783152	defective ribonuclease PH
b3906	+	16131746	Transcriptional activator, L-rhamnose-binding
b4568	-	145698350	predicted protein

this list which is also in the iAF1260 model, b1588, is considered in further detail. On close examination, although the BLASTp part of the analysis worked correctly, an inconsistency in GI numbers for the protein encoded by gene b1588 due to an out-of-date entry in the MySQL database used to establish the gene order in MG1655 means that this particular gene is inferred as absent. In all but *Legionella* genes to either side of this gene are overwhelmingly conserved.

b1588 is part of an enzyme complex (along with b1587, b1589 and b1590) that catalyses two reactions: dimethyl sulfoxide reduction (DM-

SOR1) and Trimethylamine N-oxide reduction (TMAOR1). There is however an enzyme complex consisting of the protein products of b0894, b0895 and b0896 that also catalyses both these reactions. These three genes are inferred to be present in all of the organisms in this study, apart from *Shigella boydii* Sb227, *Shigella dysenteriae* Sd197. What is striking about the absence in the *Shigella* strains of genes required for this enzyme complex is that even ignoring the absence of b1588 both have genes missing from both complexes, so neither strain can catalyse these reactions. None of the other strains are missing any of the other genes than b1588.

6.4.1.6 Specific strains against MG1655 synteny comparison

Some strains compared to MG1655 by the DH method reconstruct interesting features of the relationships between the bacterial genome sequences. Many of these features were noted in the relevant sequencing paper, and these validate the method to an extent. Figure 6.3 is a map of gene synteny between MG1655 and CFT073. This figure shows clearly the net addition of genes to CFT073 (indicated by horizontal gaps in the map) many of which are virulence related [42], as seen by Rasko *et al* [124] in a similar comparison of the two strains MG1655 and CFT073. There is also an absence of rearrangements with respect to MG1655 - these are clearly closely related strains despite large differences in gene complements. The gene complement of CFT073 with respect to that of MG1655 will be discussed further in Chapter 7 where a stoichiometric model of CFT073 has been constructed.

Figure 6.4 shows some of the other maps of synteny inference obtained using DiagHunter. They show how synteny conserved genes are distributed

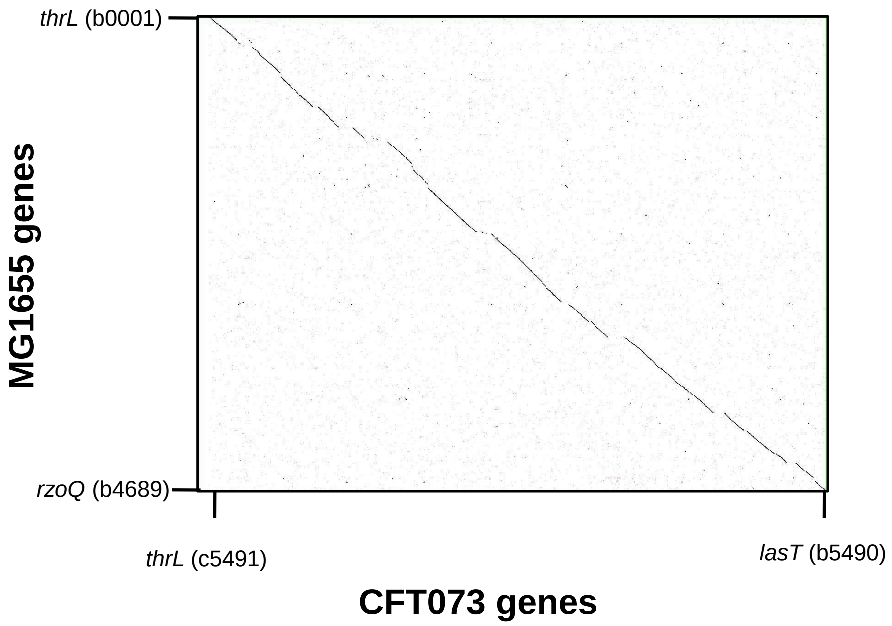


Figure 6.3: DiagHunter output map for *E. coli* CFT073 versus *E. coli* K12 MG1655. MG1655 protein coding genes are each represented by a row (starting from the top at *thrL*) on the maps and query sequence genes are represented by columns (starting from the left, also at the *thrL* gene for that organism). Where a gene is conserved between the two strains there is a black dot in the corresponding row and column. There are 4149 rows, each corresponding to one of the MG1655 genes. Due to technical issues with DiagHunter numbering of non-MG1655 genes had to be started at 101 rather than 1, so the first 100 columns of this map is empty - an artifact of the computational method and not a genuine feature of the comparison.

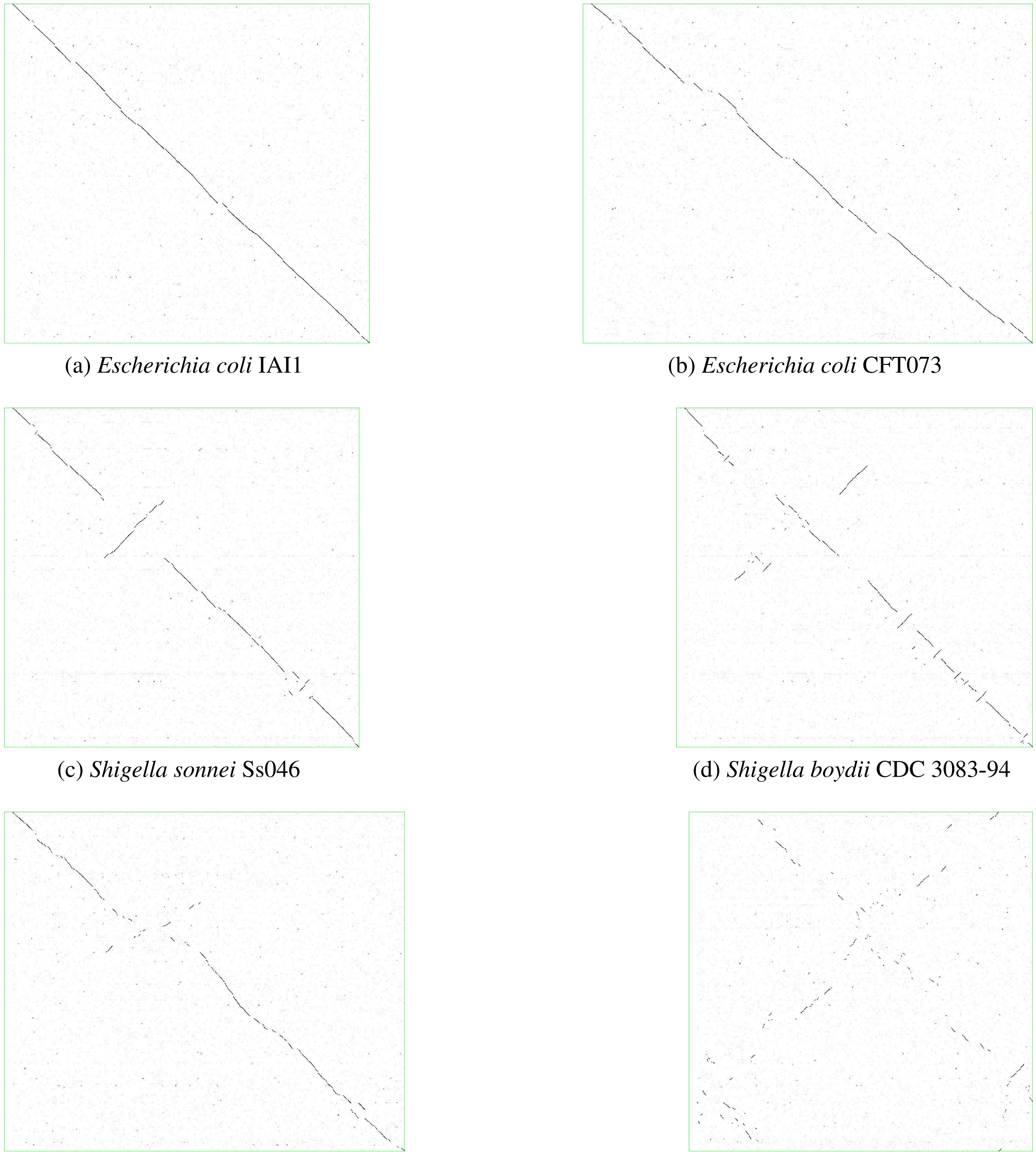
between the query strains and *E. coli* K12 MG1655. A clear backbone of conserved genes can be seen in strains as evolutionarily distant as *Klebsiella* even *Yersinia pestis* shows vestiges of the backbone. It should be borne in mind that these maps do not indicate which differences between the bacterial genomes have occurred in which of the two strains, merely that there are dfferences. Changes since the last common ancestor could have occured in the consequent ancestors of either MG1655 or the query strain.

Figures 6.4 (c) - (d) both show the very close relationship between *Shigella* and *Escherichia* though it can be seen that different parts of these genomes have been affected by genomic rearrangements, rather than there being common rearrangements to the *Shigella* strains.

Another feature of maps (d) - (f) in this Figure are the apparent linkage of pairs of genomic rearrangements. This can most clearly be seen in map (d) where the two sets of genes that fall distant from the backbone of genes appear to be two sets of genes that have swapped positions in the genome between MG1655 and *Shigella boydii* CDC 3083-94. The most likely explanation for these linked differences is that they are the result of two separate events: one *in situ* reorientation of a large region of DNA, then *in situ* reorientation of a subsection of this DNA back to its original orientation. Although this has not been observed before in this strain, comparison of other strains, including other *Shigella* [155] and *Y. pestis* [156], have observed a similar phenomenon.

Figure 6.4 (f) shows that despite its disparate synteny conservation with MG1655 the conserved genes in *Yersinia pestis* appear to fall primarily along two crossed lines of synteny. This indicates that the rearrangements between it and MG1655 are primarily due to large-scale DNA inversions, rather than translocation in any form and confirms the observations of Deng *et al* [156], who did a similar comparison, of two synteny conserved 'backbones' corresponding to the two replichores of *Y. pestis*.

There are many other observations on a genome scale to be made here, indicating mechanisms of genome adaptation and relationships between various bacteria, but much of this analysis has been covered by others in original genome sequence publications such as Yang *et al* [155] and Deng et al [156]. Also the BLAST score ratio method of Rasko *et al* [151] has recently been used to produce similar comparisons at the genome level [124], so the rest of the focus of this Chapter will be on conservation of metabolic genes



(a) *Escherichia coli* IAI1

(b) *Escherichia coli* CFT073

(c) *Shigella sonnei* Ss046

(d) *Shigella boydii* CDC 3083-94

(e) *Klebsiella pneumoniae* subsp. pneumoniae MGH 78578

(f) *Yersinia pestis* KIM

Figure 6.4: DiagHunter output maps of synteny conserved genes for selected genome sequences used in this study, compared with *Escherichia coli* K12 MG11655. For an explanation of the plots see the caption for Figure 6.3.

included in the iAF1260 model.

6.4.2 Gene and reaction retention in the *Enterobacteriaceae*

6.4.2.1 Essential genes

It is difficult to determine with bioinformatic comparisons of genome sequences whether numbers of inferred genes are over- or under-estimated. Currently experimental validation is required for any definitive evidence about whether genes have conserved function. However, there are a set of genes that have been experimentally determined to be required for survival of *E. coli K12* and its very close relatives (including MG1655). Due to the complexity of the interactions between the many parts of the *E. coli* and the apparent high gene retention between strains it seems plausible that these essential genes will be essential across closely related strains, so they have been used here to validate the comparisons undertaken. The list of essential genes was compiled from several sources as described in Section 1.4 [88, 89, 90] and the list of the 191 genes determined by all three sources to be essential can be seen in Supplementary Table 3.

The essential genes present in all the *Enterobacteriaceae* amount to 146 of the 191 experimentally determined essential genes. 45 essential genes are therefore not present in all of the strains. One of these genes, b1085, is absent in the majority of strains and one is absent in a quarter, b1718. b1085 is annotated as a predicted protein, and has an unknown function. However b1718 is protein chain initiation factor IF-3 *inf*C, adjacent to a 50S ribosomal subunit downstream and threonyl-tRNA synthetase upstream on the MG1655 genome. Its absence in *E. coli* DH10B is notable as this strain is a derivative of the same wild-type *E. coli* as MG1655. Further, the only strains lacking this gene are other *E. coli* and *Escherichia Fergusonii*. This raises the question of whether the gene is indeed essential in *E. coli* and if not why has it been identified by all of the experimental approaches used to produce the list of essential genes.

In terms of using experimentally derived lists of essential genes to determine the reliability of the DH method in inferring gene complements, the large number of highly conserved essential genes (especially across closely related strains) indicates the effectiveness of this method at inferring positive gene synteny. Further, both the DH method and MetaCyc inferences provide very similar inferences for all of the essential genes identified across the strains analysed here. Essential genes cannot aid in determination of whether gene synteny comparisons are too sensitive and incorrectly infer genes as present (false positives), so it remains an open question the best way to determine this bioinformatically.

6.4.2.2 Gene types conserved

It can be seen from Table 6.4 that retention of genes across the sample of strains used in this analysis is influenced in a small but significant way by the category of function of the genes.

The standard deviations of each of the known categories (except extra-chromosomal), indicating the spread of conservation of genes is consistent across categories. This indicates that the rate at which new genes are accommodated in genomes and at which (non-essential) genes are eliminated from them is not dependent on the broad functional categories in which the genes fall, but more probably on some general gene mixing mechanism.

This implies that metabolic and transport genes are no more likely to be permanently removed or added than genes encoding cell structure components. This result is surprising given the heterogeneity in metabolic capabilities (even between different strains of *E. coli*). This evidence of functional category independent gene mixing is further supported by the similar values of conserved genes and spread of conservation in the Unclassified/Unknown/Other category to the other categories.

The extrachromosomal category of genes represents those genes that are insertion element, prophage, plasmid and colicin related and it can be seen that these are far less consistently conserved, which is expected as they are associated with mobile genetic elements. As expected, The essential genes as identified in Section 6.4.2.1 are to a very great extent present irrespective of category. The low percentage conserved in the 'unclassified' category is due to the lack of gene b1085 in many strains, but the other three unclassified genes are very highly conserved.

Genes from iAF1260 were further broken down into the subsystems identified by Feist *et al* [1] and summaries of these genes can be seen in Tables 6.5 and 6.6. Although this included a large number of genes, it does not include all 1597 genes in the metabolism category according to GenProtEC, due to insufficient experimental data to relate the extra genes to specific metabolic functions.

Table 6.5 shows several interesting features. The least well conserved genes are those implicated in alternate carbon metabolism, glyoxylate metabolism and lipopolysaccharide biosynthesis / recycling. Alternate carbon metabolism includes many pathways specific to particular carbon containing compounds, without being essential to core metabolic processes, so heterogeneity in carbon source uptake and utilisation enzymes is to be expected. Glyoxylate metabolism consists of only two genes, the products of which catalyse two of the reactions converting glyoxylate to 3-phosphoglycerate, which is part of the Embden-Meyerhof pathway. This can also be attributed to the heterogeneity of carbon source utilisation mechanisms required in various environments. The poor conservation of lipopolysaccharide genes is because of differing lipopolysaccharide coats produced in even closely related bacteria for the purposes of virulence as has been seen in *Shigella* [157] and in *E. coli*, sufficient that a database of structures has been created [158].

These Tables also show that in many subsystems *Shigellae* contain as many common genes with MG1655 as other *E. coli*, with the advent of sequencing it has been possible to group *E. coli* and *Shigella* than their genera might indicate [159]. The genes of one subsystem, that of valine, leucine and isoleucine metabolism is in general better conserved in the *Shigella* strains than the *E. coli* strains. This information by itself does not have an obvious explanation, but rather by looking at the reactions and pathways that are missing due to these genes gives a better idea of how well these inferences will translate into new metabolic models.

6.4.3 Metabolic model construction

There is currently no absolute method of determining gene synteny - except through experimental comparisons. In the case of sets of organisms, or even single organisms, thousands of experiments would have to be carried out to validate the bioinformatic gene comparisons presented here and elsewhere (e.g. MetaCyc). However, the use of another source of synteny comparisons (MetaCyc) and experimentally elucidated essential genes have both provided evidence that the synteny inferences made using DiagHunter are of comparable reliability to those of Pathologic.

6.4.3.1 Essential metabolic genes conserved and models

For models of growth it is necessary that all genes essential to the model are present, or some alternative is found for each one, or that growth media are changed to accommodate the strain being modelled (for instance DH10B is auxotrophic for leucine). There are 234 essential metabolic genes according to iAF1260 (as determined with model medium used in the next Chapter). Their retention across various subsystems can be seen in Table 6.6. The most striking result that comes out of this Table is that according to this analysis, essential genes relating to cell envelope biosynthesis are less well conserved in other *E. coli* than in *Shigella* even though more genes in this subsystem are conserved overall in *E. coli*.

These values are due almost entirely to the absence of two genes from most *E. coli*, that are conserved across most of the *Shigella*, b2038 (*rfbC*) and b2040 (*rfbD*) encoding a dTDP-4-deoxyrhamnose-3,5-epimerase and a dTDP-4-dehydrorhamnose reductase subunit of dTDP-L-rhamnose synthase respectively. Both genes are essential in dTDP-rhamnose biosynthesis [160] which is part of the O-antigen production pathway in gram-negative bacteria. Although the structure of the O-antigen of *E. coli* K12 (and therefore MG1655) has been reported [161], it was found by reconstruction of the *rfb* operons of K12, since K12 does not express O-antigens due to two deleterious mutations in the *rfb* operons presumably accumulated during early experiments after it was originally isolated. Neither of these genes is essential according to the list of essential genes compiled from the sources described in Section 1.4. It is therefore not clear why the reactions catalysed by the products of genes b2038 and b2040 are present in the model of MG1655 and why these reactions are essential for biomass accumulation in the model.

The other relatively poorly conserved essential genes are mostly due to individual strain attributes. For instance, DH10B's relationship as a laboratory adapted strain from the same original K-12 as MG1655 explains the low leucine biosynthesis subsystem retention. DH10B is auxotrophic for leucine, due to deletion of its leucine biosynthetic genes (b0071 to b0074), so these are in fact essential for growth on glucose minimal medium, though not on rich media.

Since rich media do not have fully defined constituents it is not possible to determine which metabolic genes according to iAF1260 are absolutely essential for growth. The subset of genes in iAF1260 that are essential for growth in the model has been determined to be consistent with experimental determinations of essential genes, as was noted when the model was published [1]. There are 63 iAF1260 genes that are essential according to all of the sources mentioned in Section 1.4, and are also essential according to the iAF1260 model. According to this analysis, 16 of the 22 *E. coli* strains contain all 63 of these, as do 7 out of the 10 *Shigella* strains. Even the strain from these two groups with the lowest number of conserved essential genes, CFT073, still has 60 out of 63 of them. This makes the production of working metabolic models feasible with the comparisons presented here, since only a very few anomalous results (resulting in models that do not predict growth in any circumstances) need be addressed. This is also the case with the other *Enterobacteriaceae*, which each contain at least 57 of the 60 essential genes, but is not the case for *Legionella* as it is inferred to have only 15 of the 63 genes in its genome.

6.4.3.2 Reactions conserved and gene redundancy in derived strains

The aim of this gene synteny comparison is to produce metabolic models for newly sequenced bacteria, purely from metabolic model iAF1260 and the genome sequences of MG1655 and the query strain. Therefore the gene synteny comparisons carried out here were used through the Gene-Protein-Reaction relationships set out in iAF1260 to infer metabolic networks and models, based on those of MG1655. The reactions conserved can be seen in Table 6.7 broken down by subsystem (as in Feist *et al* [1]).

This Table of reactions excludes all those reactions that do not yet have an assigned gene dependency. For those without a gene dependency there are several reasons. In some cases the relevant genes have not been characterised, in others the reactions are spontaneous and do not require catalysis. These second type include diffusion across cell membranes and exchange reactions which interface the model of the cell to a model environment. These are therefore not considered in the calculations for Table 6.7.

Possibly the largest problem with the models presented in this Chapter is the limitation to reactions present in MG1655, rather than all known pathways, as is attempted by MetaCyc. According to MetaCyc MG1655 contains 1793 reactions out of a total of 6483 in the database (as of 2007 [78]). According to the comparative analysis on that page (calculated by Pathologic), comparing the available annotations of the bacteria in this study, there are on average 400 reactions in each strain that are not in MG1655, ranging from 0 extra reactions in *Yersinia Pestis* KIM to 809 in *Klebsiella pneumoniae* subsp. pneumoniae MG78578. Although these additional reactions ostensibly amount to 30 % of metabolism in the worst case, inference of these extra reactions rely on annotation, so might be over-estimations of extra reaction content due to successive inferred annotation.

6.5 Conclusions

In this Chapter a quick method of reliably inferring bacterial metabolic models from a reference model (iAF1260, a model of MG1655 metabolism) and the genome sequence of a novel bacterium has been analysed to determine to what extent it can produce a working model for the novel bacterium. This has been determined by analysis of genes essential for biomass production in the reference model and essential genes according to other sources. More than 93 % of the iAF1260 genes considered essential were recovered over the set of bacteria analysed and almost 97 % of metabolic genes considered essential by all sources considered were retained.

This analysis shows that even very closely related bacteria to the reference strain (for instance, DH10B) do not necessarily contain all the genes required for a working model, indicating that this could provide evidence of non-essentiality in some cases. The method of model construction was used on bacteria of varying phylogenetic distance to MG1655 and it was established that although it works well within a bacterial family this method is not effective between orders of bacteria. Its applicability between families in the same order was not tested as there are no genome sequences outside the family of *Enterobacteriaceae* within the order of *Enterobacteriales*.

Different subsystems of the model iAF1260 have been retained differently depending on strain according to the comparison presented. The limitations of the inference to known reactions in MG1655 has been discussed, but since it relies on direct comparison between newly sequenced genes and the large number of experimentally characterised genes in MG1655, its inferences should be very reliable. For the family *Enterobacteriaceae* this quick reliable method of model inference recovers a large proportion of the metabolic networks of these bacteria.

Chapter 7

Stoichiometric modelling of *E. coli* DH5α and CFT073 to investigate metabolic adaptation to heterologous metabolic genes

7.1 Preface

The gene complements inferred for all of the bacteria discussed in Chapter 6 have not been tested in simulations of growth. This is because many of these models would not run because after initial synteny comparisons some genes essential for biomass growth in minimal media using glucose as sole carbon source in the model were not present. The absence of any one of these genes renders the model inoperative. Further, there are several parameters that are set from experimental observation in model iAF1260 (such as the non-growth-associated energy drain on the system due to cell maintenance) which may not have the same value either between bacterial strains or in various experimental conditions.

The model uropathogen *E. coli* CFT073 was originally chosen in this project as a template for seeking novel genetic contributions to uropathogenic metabolism and with the data on growth of CFT073 and DH5α with added metabolic genes in defined media shown in Chapter 5 it was a natural choice for selection to test the inference of its metabolism by the DH method (see Section 3.4.2). The genes investigated for function in Chapter 5 were confirmed to encode an L-sorbose uptake and utilisation operon, so experiments have concentrated on L-sorbose metabolism, using glucose as a control.

These models are purely stoichiometric, they do not include any regulatory elements, as regulation of bacterial metabolism is so complex and to a great extent unknown. If known regulation were added to the model it would take infeasible computing power to run simulations so it is hoped that with a few key parameters limiting reaction rates (constraints) will be enough to produce qualitatively correct predictions.

The aims in this Chapter are: to test whether the model of CFT073 produced here can reproduce growth observed in minimal media, to test whether constraints alone can allow simulations to reproduce observed growth patterns and to determine whether serial passage of DH5αL has allowed this strain to attain its theoretical growth rate limit, according to the stoichiometry of its metabolic network. All reaction abbreviations used in this Chapter are those used in the iAF1260 of Feist *et al* [1].

Table 7.1: Showing how the genotype of DH5α is predicted to affect its metabolic network compared to MG1655. Each of the genes rendered inoperative by disruption or deletion that are included in iAF1260 are shown. ‘Reactions lost’ indicates which reactions from the stoichiomeric model are removed by the gene deletions, giving their iAF1260 designations.

Genotype	Blattner number	Subsystem	Reactions lost
fhuA2	b0150	Transport, Outer Membrane	FE3HOXtonex, FECRMtonex, FEOXAMtonex
Δ(lacZYA-argF)U169	b0273	Arginine and Proline Metabolism	–
	b0312	Unassigned	OCBT, BETALDHx, BETALDHy
	b0314	Transport, Inner Membrane	–
	b0323	Unassigned	–
	b0331	Alternate Carbon Metabolism	MCITL2
	b0333	Alternate Carbon Metabolism	MCITS
	b0334	Alternate Carbon Metabolism	MCITD
	b0335	Alternate Carbon Metabolism	ACCOAL
	b0336	Transport, Inner Membrane	CSNt2pp
	b0337	Nucleotide Salvage Pathway	CSND
	b0339	Unassigned	–
	b0340	Nitrogen Metabolism	CYNTAH
	b0341	Transport, Inner Membrane	CYNTt2pp
	b0343	Transport, Inner Membrane	LCTStpp
	b0344	Alternate Carbon Metabolism	LACZ
phoA	b0383	Unassigned	PPTHpp
relA1	b2784	Cofactor Biosynthesis	GTPDPK
nupG	b2964	Transport, Inner Membrane	INSt2pp, DGSNt2pp, DINSt2pp, GSNt2pp, INSt2pp

7.2 Results

7.2.1 DH5α genotype

The genotype of DH5α is as follows: F$^-$ *fhuA2 Δ(lacZYA-argF)U169 ϕ80* d*lacZΔM15 endA1 hsdR17 deoR nupG thi-1 supE44 gyrA96 relA1 recAI phoA* λ^-. Many of these changes from the wild-type *E. coli* (from which MG1655 was derived) are metabolic in nature and the model used for the analysis in this Chapter reflects these changes. Table 7.1 indicates which aspects of the genotype have an effect on the model and which reactions have been removed due to the genotype.

7.2.2 CFT073 model construction

In Chapter 6 the genes conserved between *E. coli* MG1655 and CFT073 were inferred. This set of conserved genes has been used to construct the part of CFT073 metabolism shared with MG1655.

Table 7.2 shows a list of genes that are missing from CFT073, but that are essential according to iAF1260 in glucose minimal medium. Also shown in this Table are the reactions that could be removed due to these missing genes. All but one of the missing genes is related to lipopolysaccharide (LPS) biosynthesis. Most are explicitly so and b0173 (dxr) which is part of the 'Cofactor and Prosthetic Group Biosynthesis' subsystem is involved in isoprenoid production, which is used in LPS and enterobacterial common antigen biosynthesis. Of note however, dxr is also vital for menaquinone and ubiquinone biosynthesis, which are important in electron transport at the bacterial membrane. It is therefore not clear why this does not appear to be present in the analysis presented here.

Without a full reconstruction of the LPS biosynthetic pathways of CFT073 available the assumption has been made that although the exact reactions may be incorrect, the contribution of LPS and cell envelope material to overall biomass formation is approximately equal between CFT073 and MG1655. Therefore reactions considered essential for this aspect of cell growth have been included in the CFT073 model.

The only essential gene that is not related directly to LPS biosynthesis is b1262 (*trpC*), which encodes an enzyme that catalyses two steps in *de novo* tryptophan biosynthesis. The work presented in Chapter 4 shows that CFT073 is not auxotrophic for tryptophan, since it grows in minimal

Table 7.2: Showing the iAF1260 genes that are not present in CFT073 and are essential for the model to run. Reaction details can be found in supplementary material accompanying the publication of Feist *et al* [1].

Gene name	Blattner number	Subsystem	Reactions affected
dxr	b0173	Cofactor and Prosthetic Group Biosynthesis	DXPRIi
lpxK	b0915	LPS Biosynthesis / Recycling	TDSK
fabG	b1093	Cell Envelope Biosynthesis	3OAR100, 3OAR120, 3OAR121, 3OAR140, 3OAR141, 3OAR160, 3OAR161, 3OAR180, 3OAR181, 3OAR40, 3OAR60, 3OAR80
trpC	b1262	Tyr, Trp, and Phe Metabolism	IPGS, PRAIi
rfbC	b2038	Cell Envelope Biosynthesis	TDPDRE
rfbD	b2040	Cell Envelope Biosynthesis	TDPDRR
rfaC	b3621	LPS Biosynthesis / Recycling	HEPT1
waaU	b3623	LPS Biosynthesis / Recycling	HEPT4
rfaZ	b3624	LPS Biosynthesis / Recycling	MOAT3C
rfaB	b3628	LPS Biosynthesis / Recycling	GALT1
rfaS	b3629	LPS Biosynthesis / Recycling	RHAT1

medium with glucose as sole carbon source. Due to the retention of genes encoding other enzymes in the tryptophan biosynthesis pathway, and the fact that these commit metabolites to the tryptophan biosynthesis pathway, it seems feasible that an alternative but as yet unidentified enzyme (or enzymes) does catalyse these steps in CFT073. The two reactions IGPS and PRAIi will therefore be retained in the model of CFT073.

7.2.3 Addition of L-sorbose operon to models

E. coli MG1655 (and therefore DH5α) is unable to metabolise L-sorbose. The L-sorbose operon in CFT073, the function of which has been confirmed in this work, contains genes that encode an L-sorbose uptake and utilisation pathway (as originally elucidated in *Klebsiella penumoniae* [118]): extracellular L-sorbose → cytosolic L-sorbose-1-phosphate → cytosolic D-glucitol-6-phosphate → cytosolic D-fructose-6-phosphate. Table 7.3 shows

Table 7.3: Showing the details of L-sorbose exchange, uptake and utilisation reactions added to models for CFT073 and DH5αM. Metabolite short names are taken from iAF1260 where the metabolite is part of that model. The reactions are the following: SORpts is L-sorbose transport across the cell membrane by phosphotransferase transport system with concomitant phosphorylation, SPR and SPRalt are L-sorbose-1-phosphate reduction to D-glucitol-6-phosphate using NADH or NADPH respectively as proton donor and EX_sor-L is the exchange reaction for L-sorbose. All metabolites taking part in these reactions are in the cytosol apart from those marked with letters 'b' or 'e', which are in the boundary and extracellular compartments respectively.

Reaction name	Substrate name	Stoichiometry	Product name	Stoichiometry
SORpts	pep	1	pyr	1
	L-sorbose (e)	1	L-sorbose-1-phosphate	1
SPR	h	1	D-glucitol-6-phosphate	1
	NADH	1	NAD	1
	L-sorbose-1-phosphate	1		
SPRalt	h	1	D-glucitol-6-phosphate	1
	NADPH	1	NADP	1
	L-sorbose-1-phosphate	1		
EX_sor-L	L-sorbose (e)	1	L-sorbose (b)	1

the stoichiometries of the added reactions. The final reaction step, glucitol-6-phosphate to fructose-6-phosphate, is already included in the model as it is part of the D-sorbitol utilisation pathway, catalysed by the product of gene srlD (b2705).

7.2.4 Growth simulation calculations

The model iAF1260 is stoichiometric and linear optimisation maximising the bacterial growth rate was used to establish how efficiently carbon source uptake was converted into biomass accumulation. The models derived from this, of both DH5α and CFT073, were the same and were considered in the following way. Timescales over which bacterial regulation operate are fast compared to the times over which the environment in which the bacteria resided changed, therefore it was assumed that steady state assumptions allowed the use of linear optimisation with flux balance.

Known quantities from experiments in Chapter 5 were biomass, extra-cellular carbon source and extra-cellular acetate concentrations. Therefore the simulations were designed to observe these quantities. Energy (and carbon) from the carbon source was divided by bacteria in the experiments between biomass accumulation and acetate production, and there is also an energy cost associated with non-growth-associated maintenance of cells in the culture which was unobservable in these experiments, but that was included in simulations as it is well established (it requires a flux of 8.39 mmol/gDW/h through reaction 'ATPM' in model iAF1260).

Specific c-source uptake rate was assumed to be proportional to the c-source concentration in the media up to a maximum value determined by Fischer *et al* [162] experimentally to be around 10 mM/gDW/h, depending on growth conditions. Therefore uptake rate was as follows:

$$R = \begin{cases} k_1 \times \rho & \text{if } R < 10 \text{ mM/gDW/h} \\ 10 & \text{otherwise} \end{cases} \tag{7.1}$$

where k_1 is an affinity parameter representing how well bacteria can take up c-source when it is available and ρ is the millimolar concentration of c-source in the simulated medium.

Acetate production was modelled using a simple relationship established from Figure 5.6, but modified in light of Vemuri *et al* [141], relating specific c-source uptake rate (R) to specific acetate production rate (A) due to overflow metabolism. This relationship is as follows:

$$A = \begin{cases} 0 & \text{if } R < R_0 \\ (R - R_0)y_1 & \text{if } R_0 \leq R < R_m \\ (R_m - R_0)y_1 & \text{if } R \geq R_m \end{cases} \tag{7.2}$$

where R_0 is the threshold c-source uptake rate for overflow to acetate production, R_m is the uptake rate of c-source corresponding to the maximal acetate production rate where acetate production saturates and y_1 is the slope of the line corresponding to acetate production due to overflow between these two values. The two threshold values were calculated by minimising glucose uptake rates at the doubling rates corresponding to the thresholds for acetate production identified in Figure 5.6. These were $R_0 = 3.36$ and $R_M = 6.66$ mM/gDW/h.

7.2.4.1 Simulation algorithm

The simulations were run as follows:

(i) at timepoint i, time from beginning of simulation was $t = i \times \Delta t$,

(ii) $R(t)$ was calculated from Equation 7.1 from $\rho(t - \Delta t)$,

(iii) $A(t)$ was calculated from Equation 7.2 using $R(t)$,

(iv) $B(t)$ was calculated by maximising the biomass equation subject to the constraints on $R(t - \Delta t)$ and $A(t - \Delta t)$,

and the following values were calculated:

$$M(t) = M(t-\Delta t) + \Delta t \times B(t), \tag{7.3}$$

$$\rho(t) = \rho(t-\Delta t) - \Delta t \times R(t) \times M(t-\Delta t), \tag{7.4}$$

$$\epsilon(t) = \epsilon(t-\Delta t) + \Delta t \times A(t) \times M(t-\Delta t), \tag{7.5}$$

where $M(t)$ is biomass in gDW, $\rho(t)$ is c-source concentration in the medium and $\epsilon(t)$ is acetate concentration in the medium, both of these in mM.

7.2.5 Simulation results

Simulations compared to experimental data over the period of bacterial growth (the growth was only simulated during this period) are shown in Figures 7.1, 7.2 and 7.3 for CFT073, DH5αL and DH5αM. Simulations were run from different times, depending on the offsets of data used for comparison (from Chapter 5), and were ended at the point where achieving biomass rate ≥ 0 was not possible by linear optimisation.

Simulation parameters M_0 and k_1 were adjusted to give an accurate biomass trace during the exponential phase of growth. ρ_0 was calculated as a mean of the first three c-source experimental points after the beginning time of the simulation. The parameter determining acetate production as a function of c-source uptake rate, y_1 was determined such that it fit acetate production in the early phase of growth (when the assumption of steady state was the most reasonable). Simulation parameters for each of these simulations are shown in Table 7.4.

Simulation results for growth rates, final biomass concentrations and acetate concentration at the end of the exponential phase of growth are shown

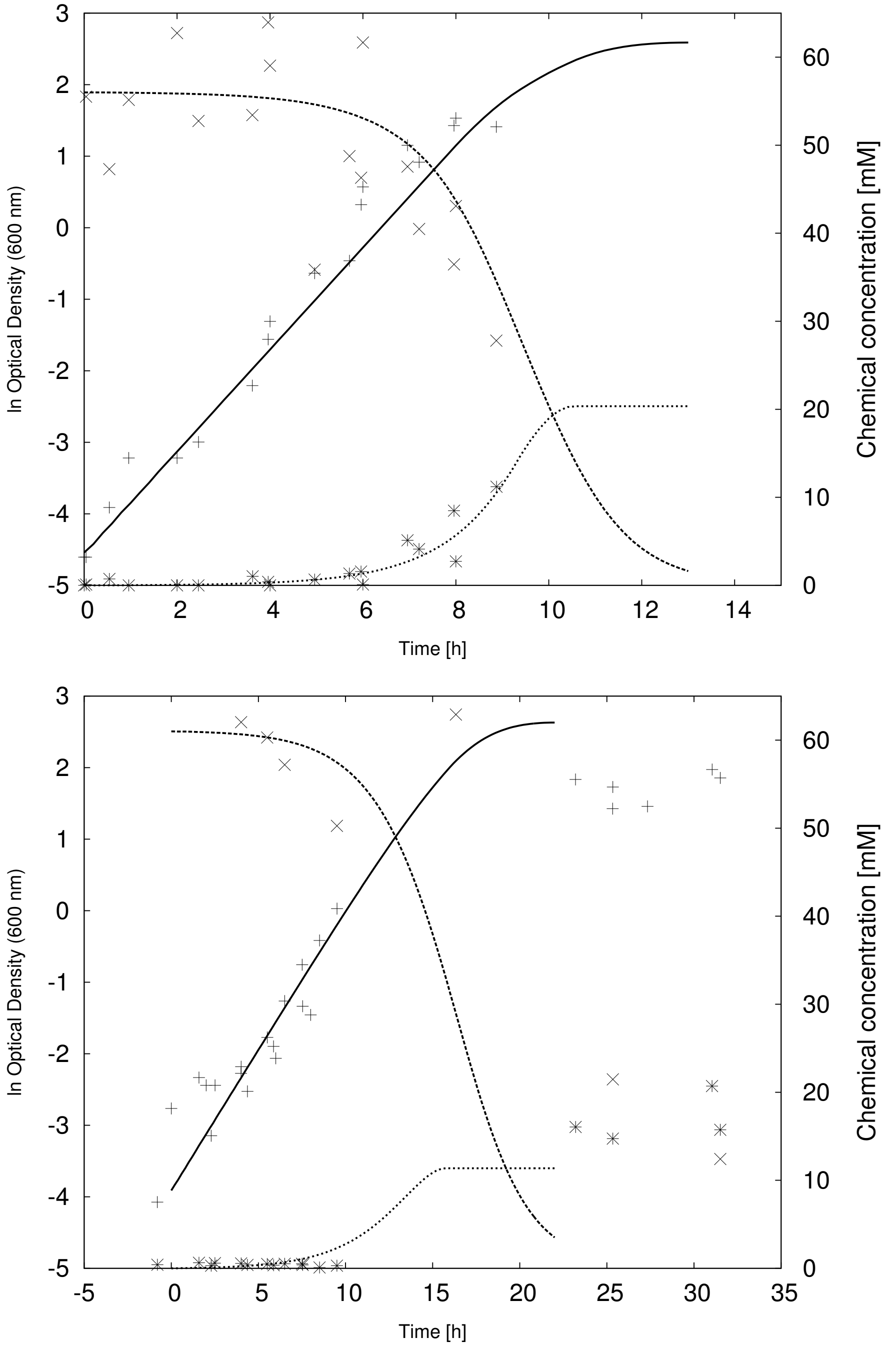


Figure 7.1: Simulations compared to experimental data for growth of CFT073 in M6 minimal media with supplemented with different c-sources. The experimental data are as follows: optical density '+', c-source concentration '×' and acetate concentration '∗'. Simulation results are as follows: optical density - solid line, c-source concentration - dashed line and acetate concentration - dotted line. Top: using glucose as carbon source. Bottom: using L-sorbose as carbon source.

in Table 7.5 along with the corresponding values from the experimental data (from Tables 5.1, 5.3 and 5.4 in Chapter 5).

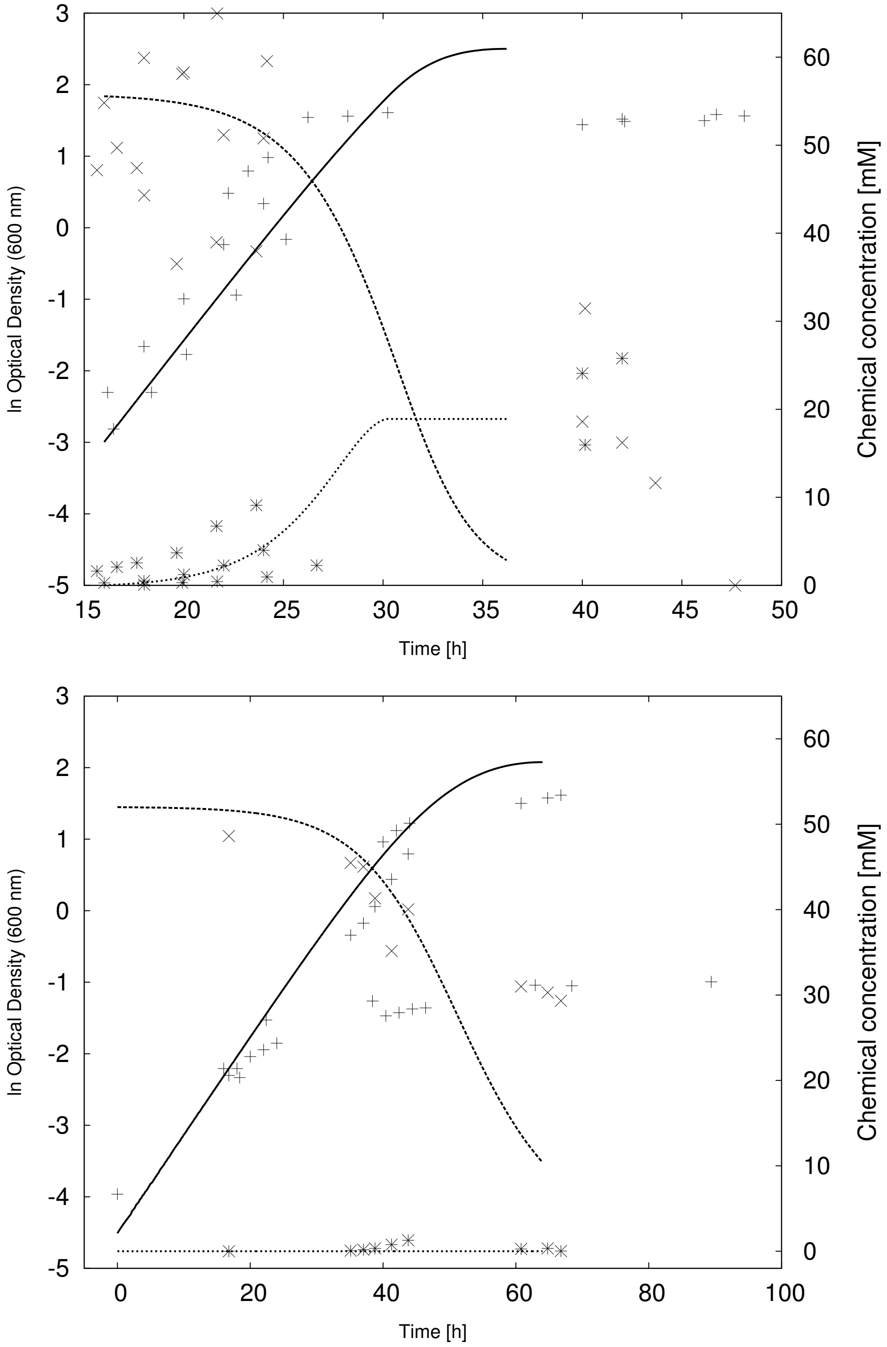


Figure 7.2: Simulations compared to experimental data for growth of DH5αL in M6 minimal media with supplemented with different c-sources. The experimental data are as follows: optical density '+', c-source concentration '$\times$' and acetate concentration '$*$'. Simulation results are as follows: optical density - solid line, c-source concentration - dashed line and acetate concentration - dotted line. Top: using glucose as carbon source. Bottom: using L-sorbose as carbon source.

7.3 Discussion

7.3.1 Model inference

The model of DH5α metabolism shows interesting characteristics when compared to the original model of MG1655 metabolism (iAF1260). Due to

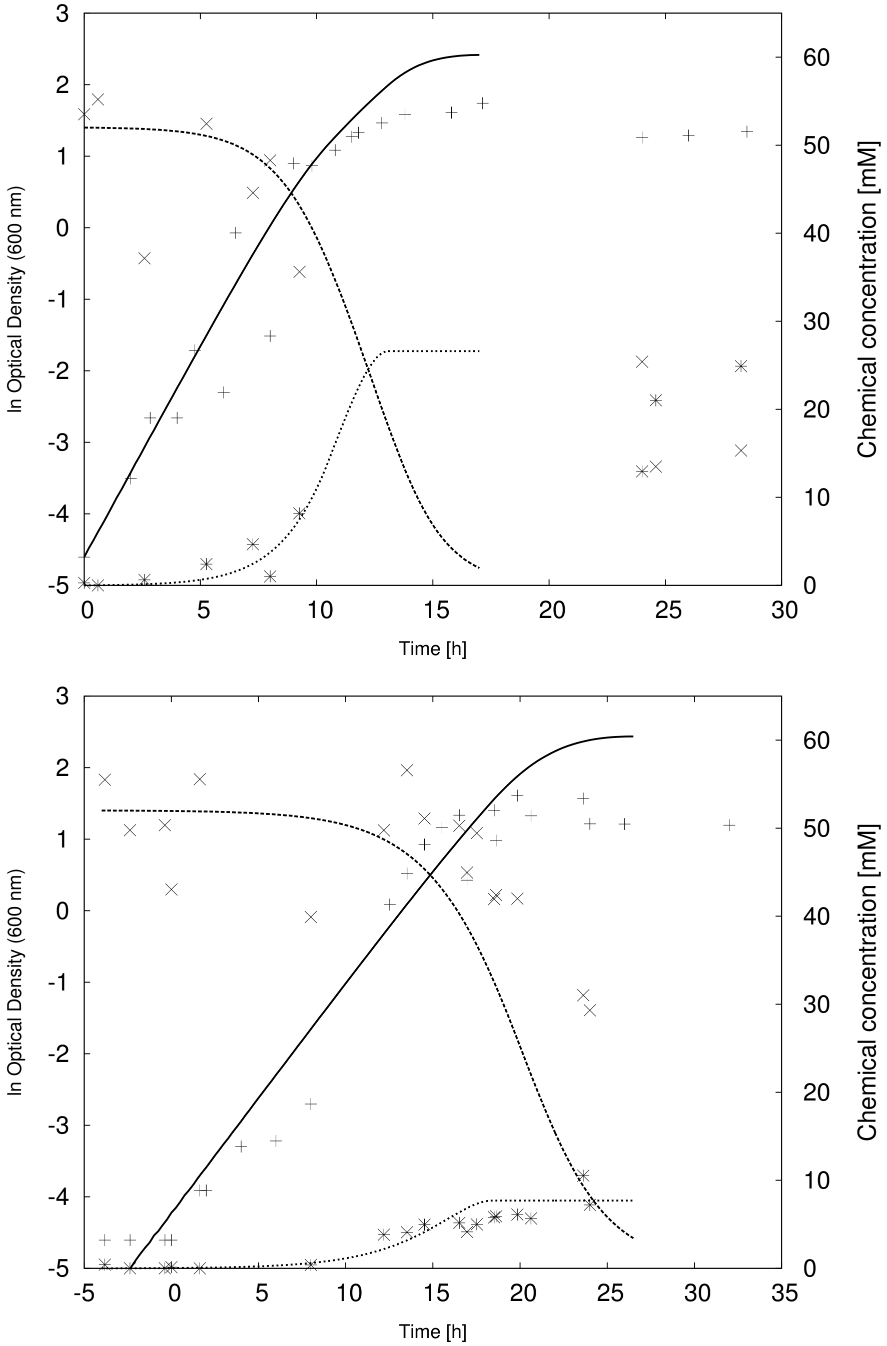


Figure 7.3: Simulations compared to experimental data for growth of DH5αM in M6 minimal media with supplemented with different c-sources. The experimental data are as follows: optical density '+', c-source concentration '$\times$' and acetate concentration '$*$'. Simulation results are as follows: optical density - solid line, c-source concentration - dashed line and acetate concentration - dotted line. Top: using glucose as carbon source. Bottom: using L-sorbose as carbon source.

the derivation of DH5α from the same wild-type strain as MG1655, the only reactions it is missing are those either as-yet uncharacterised in MG1655 or

Table 7.4: Showing the simulation parameters calculated for the simulations shown in Figures 7.1, 7.2 and 7.3. The '-' for y_1 when DH5αL is grown on L-sorbose is because for the other model values L-sorbose uptake rate never increases over the threshold value for carbon overflow. The inferred maximum specific c-source uptake rate for each of the strains, at the initial c-source concentration, is indicated in the final column.

	ρ_0 [mM]	M_0 [gDW/l $\times 10^{-3}$]	k_1	y_1	Max specific c-source uptake [mM/gDW/h]
Glucose					
DH5αL	56	15.00	0.120	2.2	6.7
DH5αM	52	3.00	0.180	2.2	9.4
CFT073	56	3.15	0.220	1.4	12.3
L-sorbose					
DH5αL	52	3.30	0.035	-	1.8
DH5αM	52	1.20	0.100	2.0	5.2
CFT073	61	6.00	0.100	1.5	6.1

those removed by genomic disruption, as laid out in Table 7.1, 23 reactions in total. What is surprising is that this has an impact on the maximum efficiency with which the DH5α could grow.

On comparison of the results of parallel linear optimisations it can be seen that the prevention of two reactions which have a flux in iAF1260 in the conditions above have potentially impacted on DH5α performance, ACCOAL and INSt2pp, which are the acetate-CoA ligase (ADP-forming) reaction and inosine transport in via proton symport (periplasm) reactions respectively. The rate for INSt2pp is an artifact of the linear solution found in this case, the alternative reverse periplasmic transport reaction, INSt2rpp, has an equal but negative flux. So the difference inefficiencies is due to ACCOAL, which produces acetyl-CoA from acetate.

The model of CFT073 metabolism is not so complete. Analysis in MetaCyc indicates that there are 269 reactions present in CFT073 that are not present in MG1655, which represents about 13 % of the total number of reactions in that strain if correct. The problem of successive inferred an-

Table 7.5: Showing the simulation predictions of doubling rate, final biomass concentration and extracellular acetate concentration at the end of exponential growth, along with the experimentally derived values of these from Chapter 5.

			Doubling Rate [h^{-1}]	Final Biomass [OD]	Acetate conc. [mM]
CFT073	*Glucose*	Data	1.12	4.1	31
		Model	1.08	13.97	14.56
	L-sorbose	Data	0.5	5.9	18
		Model	0.56	13.88	11.37
DH5αM	*Glucose*	Data	0.71	4.3	27
		Model	0.85	11.2	26.6
	L-sorbose	Data	0.42	4.4	12
		Model	0.47	11.43	7.7
DH5αL	*Glucose*	Data	0.49	4.7	24
		Model	0.51	12.2	18.9
	L-sorbose	Data	0.24	3	0
		Model	0.2	7.98	0

notation was discussed in Chapter 6, so it will only be mentioned here that these reactions have not been experimentally confirmed in CFT073, inferences of these reactions are not necessarily based on comparisons with experimentally characterised genes, so it seems plausible that this number of other reactions is an overestimate. Another important characteristic of these reactions if they are present in CFT073 is that since iAF1260 represents a real bacterium it contains all reactions essential for bacterial survival and the vast majority of central pathway reactions, so inferences of a CFT073 model will not be missing these reactions unless they are genuinely absent.

The efficiencies in acetate production and ATP production (shown by ATPM) in CFT073 appear to be somewhat compromised by the removal of reactions from the original model. Two reactions missing from the CFT073 model that have positive fluxes are PDH and R15BPK, the pyruvate dehydrogenase and ribose-1,5-bisphosphokinase reactions. There is an alternative pathway for production of 5-Phospho-alpha-D-ribose 1-diphosphate catalysed by the enzyme encoded by gene b1207 (ribose-phosphate diphosphokinase, reaction PRPPS), thus R15BPK flux does not affect these efficiencies.

The loss of PDH requires that an alternative way of producing acetyl-CoA, an essential intermediate in central metabolism. An analysis of flux distributions using iAF1260 with and without PDH shows increased flux through PFL (catalysed by pyruvate formate lyase) which also produces acetyl-CoA, while converting pyruvate to formate. Both reactions use pyruvate to form acetyl-CoA, but there is an efficiency cost in dealing with formate as a by-product, rather than NADH (as in reaction PDH).

Part of MetaCyc's algorithm for finding reactions in new species tries to fill gaps in metabolism of such new species where there appear to be catalysed reactions on either side of reactions in a metabolic pathway, but no gene encoding an enzyme for that intermediate reaction. It is assumed that this is why MetaCyc reports CFT073 as being able to catalyse the pyruvate dehydrogenase reaction. However, this work shows that there is no reason in principle why CFT073 would require this reaction. Using BLASTx it was determined that there is definitely a truncated *ace*F gene in the position corresponding to the actual *ace*F gene in MG1655, so this indicates that the gene is no longer functional. While the idea of pathway closing was used above to infer that reactions in the tryptophan pathway were present, the difference here is that there are metabolic alternatives to the reaction inferred as missing in this case. This means that no deadend reactions are produced by the elimination of this reaction and it seems less likely in this case that the pyruvate dehydrogenase reaction does actually occur in CFT073.

7.3.2 Simulations

The simulations shown in Figures 7.1, 7.2 and 7.3 show good qualitative agreement with the data collected previously on the growth of these strains. The simple models of carbon overflow into acetate production effectively reproduce acetate excretion during growth and for DH5αL and DH5αM show reasonably good agreement with total acetate produced at the end of growth. The failure of DH5αL to produce acetate in L-sorbose medium is predicted by the model due to the low L-sorbose uptake rate in the simulation. Acetate traces tend to increase, then become flat towards the end of growth and this is because growth rate drops below the threshold value for acetate production in the model. This plateau can be seen in most of the data in these Figures, so it would seem to support the theory that acetate is only produced as some overflow when specific c-source uptake is high.

Maximum specific uptake rate of c-source, as shown in Table 7.4, shows a clear difference between the three strains considered here. It can be seen that passage improves DH5αL's ability to take up c-source in both of the media used here to compare the strains. However the most drastic improvement is with the uptake of L-sorbose (the conditions of the passage). Indeed the maximum specific uptake rate approaches that of CFT073, although even after 100 generations of adaptation it is still 15 % lower than the CFT073 value. Measured maximum specific uptake rates in *E. coli* MG1655 and W3110, where they have been measured in continuous culture, are 10 and 10.5 mM/gDW/h [1, 163], and according to the stoichiometries and the point at which the L-sorbose pathway enters central metabolism (see Figure 4.5) there is no metabolic reason why L-sorbose could not be taken up at the same rate as glucose and growth occur at the same rate. Clearly this does not happen, so there may well be constraints exerted by the L-sorbose pathway itself in L-sorbose uptake.

There are several parameters in the iAF1260 model that have been taken as constant for the purposes of the simulations in Chapter 7. Firstly the non-growth dependent ATP utilisation for maintenance was taken to be 8.39 mM/gDW/h. This value has been shown to be 7.6 mM/gDW/h in W3110 [163], so how consistent this value is between other strains (MG1655 and W3110 being so closely related) remains open to debate. It is a very important value because it represents an intrinsic drain on energy and carbon during the growth of these bacteria. It is not a simple reaction, as represented in the model, but a wide variety of different activities for cell maintenance, so finding an *E. coli* with a low maintenance rate might be of interest to metabolic engineers looking to maximise growth or product yields in fermentation. Maximum glucose uptake rate has been shown to be 10.5 mM/gDW/h in W3110 [163], which is consistent with the value of 10 mM/gDW/h used here.

These simulations break down in terms of biomass accumulation and c-source uptake usually towards the point where there becomes a lack of carbon source. Indeed in Table 7.5 it can clearly be seen that biomass accumulation is predicted to be almost three times that seen in reality. There are several potential explanations for this, but most likely is that *E. coli* have adapted to curb their growth even in the presence of some carbon source, as profligate growth would entirely deplete potential energy sources required for maintenance and to store for potential regulatory shifts when the external environment changes. None of this control of metabolism is represented in the simulations presented here.

Indeed the data presented in Chapter 5 show that DH5αL fails to take up large amounts of carbon source available to it in the media. This is particularly true when DH5αL is presented with L-sorbose, a novel carbon source for it and it must make use of its newly acquired metabolic genes. This could represent an inability to make full use of the L-sorbose, but it does not then explain how the DH5αL grows in the first place. It therefore suggests that these bacteria with metabolic genes as yet poorly integrated into their regulatory network adopt a conservative strategy when growing in this carbon source that is taken up at a very limited rate. This is actually shown to an extent in Figure 7.2 bottom panel, though not to the extent that L-sorbose is left in the experimental case.

The other unexplained aspect of these simulations is the predicted low net acetate excretion by CFT073. The actual final acetate concentrations in media are higher than in the simulations and this shows that although qualitatively this overflow mechanism is correct perhaps it has slightly different R_0 and R_M values to DH5α, perhaps related to its loss of pyruvate dehydrogenase activity. The threshold for acetate production calculated from the data presented in Chapter 5 and the models calculated here is similar to that for acetate production in chemostat experiments conducted by Vemuri *et al* [141] whose calculated lower limit was $R_0 = 4.4mM/gDW/h$.

Although some interesting results come out of the simulations presented here it should be borne in mind that they are greatly simplified representations of the metabolism of the underlying organisms, and acetate production and c-source uptake are very much more complicated in reality than in the models presented here. For instance, uptake rate for carbon source should be subject to some saturation effect at high c-source availability and could be modelled by Michaelis-Menten kinetics with a single substrate (extracellular c-source) with the transporter as the enzyme catalysing the reaction.

Further, there are several parameters that, if more sophisticated (and realistic) models were used, could be estimated from experimental data rather than, as here, fitting to the data during simulation.

7.3.3 Model construction and validity

It can be seen from the simulations presented here that the models inferred from the original iAF1260 model are adequate for representing qualitative results about growth in simple circumstances. Due to the direct comparison between the genomes of CFT073 and MG1655 conducted in Chapter 6 this is a reliable indication of the core metabolic functions of CFT073 and much of the peripheral metabolism shared by the two bacteria since they are so closely related. Since CFT073 is one of the most distantly related of the *E. coli* compared in Chapter 6 to MG1655 it seems plausible that the direct inference of models from iAF1260 for any *E. coli* is posible and represents a quick and easy way of beginning the reconstruction of the metabolism of these bacteria.

Quantitative results used to constrain the model for MG1655 may not be accurate for other bacteria, especially those not closely related to it - Feist *et al*'s model uses a flux through a general ATP hydrolysis equation to account for non-growth associated energy use for cell maintenance amounting to

8.39 mmol ATP gDW^{-1} H^{-1}, but this may vary in other bacteria and might have to be experimentally obtained on a strain-by-strain basis, though this is beyond the scope of the work presented here.

While the problem of successive inferred annotation and the limit of comparisons due to proteins with high homology but different functions (e.g. [164]) this method of inference appears to be a very reliable system. Despite the initial limitation of comparison to pathways in MG1655 models can easily be expanded due to their simple stoichiometric nature, as shown above with the addition of L-sorbose uptake and utilisation genes.

It should be noted that by limiting this model reconstruction of CFT073 to the iAF1260 model and the L-sorbose operon with experimentally confirmed function in Chapter 4 a lot of reaction pathways will be missing that could potentially be added by careful comparison of the CFT073 genome with the set of genes from non-*E. coli* organisms that also have experimentally verified functions. The difficulty with inferring these reactions is in collecting the specific genes from a source which connects the genes to reactions (such as MetaCyc) and collecting the DNA sequences of these genes for comparison with CFT073.

7.4 Conclusions

It has been shown in this Chapter that models inferred from the genome comparisons presented in Chapter 6, based on stoichiometric model iAF1260, represent feasible, though incomplete, models of bacterial metabolism. Since the organism used to compare with MG1655 is CFT073, one of the *E. coli* with the fewest mutually conserved genes with MG1655, it shows that this method of model inference is probably valid for any *E. coli* strain, but validity beyond this, for instance for *Shigella* strains, has not yet been established.

It has also been shown that simulations based on one of these inferred models, using a simple model of overflow metabolism and c-source uptake, qualitatively predict observed growth characteristics of the strains in batch growth in shake flask experiments. However the models fail to predict bacterial strategies for coping with low c-source concentrations and would require further development to predict this aspect of behaviour.

The modelling of acetate production during growth shows qualitative agreement with experimental data and also reasonable quantitative final extracellular acetate concentrations for the strains related to MG1655 (DH5αL and DH5αM), however prediction of final extracellular acetate for CFT073 is below that seen experimentally and implies that CFT073 may have different threshold values for acetate overflow - perhaps due to its lack of pyruvate dehydrogenase activity, the only enzyme missing from the central metabolism of CFT073 compared to MG1655.

Chapter 8

Conclusions and further work

8.1 The complete process of metabolic model generation for newly sequenced bacteria

The results presented in this Thesis represent bioinformatic and experimental approaches to reconstruction of stoichiometric metabolic models for bacteria with newly sequenced genomes. The initial aims of this project were to assess the feasibility of reconstruction of such metabolic models and the use they would be in metabolic engineering contexts. Newly sequenced bacteria often have initial annotations, but they can be subject to mis-annotation and the problem of successive inferred annotation (as described in Chapter 1). It was therefore decided that basing a model on a previous well characterised model (initially iJR904 [165], then iAF1260 [1]) and use a single inference step directly from open reading frames of the query bacterium against the genome of the reference bacterium (in this case MG1655). Then ways of bridging the gaps between the sequence and the reality of the metabolism of the strain would be investigated to try to produce a more complete model by bioinformatic and experimental approaches. Even in such a homologous group of bacteria as *E. coli* there are vast differences in metabolic capabilities of strains, so this experimental verification of reactions would be vital for most reconstructions. As an example strain CFT073 was selected, since its genome had recently been sequenced and its differences in habitat (in the urinary tract) to MG1655 were known. The principle of metabolic reconstruction could have been used on any sequenced strain of *E. coli*, but it was thought that this difference in habitat would aid in the characterisation of previously uncharacterised metabolic genes in CFT073 to enlarge and improve the metabolic model.

In comparing the two strains, CFT073 and MG1655, metabolic genes in CFT073 were effectively separated into three categories, characterised genes present in both strains, genes present in CFT073 only (but with an annotated function) and uncharacterised genes in CFT073 only. The core metabolic model (present in both) was used as the basis of the metabolic model used in Chapter 7 and characterisation of the uncharacterised genes in CFT073 only was attempted to reconstruct as much of the CFT073 model as possible, as shown in Chapter 4. In this Chapter a bioinformatic approach was taken to try to elucidate the function of these uncharacterised genes (the results of which can be seen in Supplementary Table 2), but only one of the sets of genes analysed, that of L-sorbose uptake and utilisation, was sufficiently well characterised to merit inclusion in the model presented in Chapter 7. This shows the scale of the task of reconstructing complete metabolic networks for newly sequenced bacteria as many such uncharacterised sets of genes are present in each newly sequenced bacterium. For CFT073 a purely bioinformatic approach was not enough to provide enough information to complete the metabolic model, so the construction of such a model will often require biochemical characterisation. One large limitation of the model presented here is that it is incomplete, but it was sufficiently complete to simulate growth in limited conditions (with L-sorbose or glucose) as carbon source, but further investigation of the model would require the addition of the missing reactions.

In looking for these extra uncharacterised genes that could be of use to CFT073 in the urinary tract there are problems in terms of the timescales of adaptation and evolution in nature. Although some abilities distinct from those of MG1655 are required for CFT073's ability to colonise the urinary tract, there might well be a lot of genetic material that is not subject to selective pressure, so might be retained just because it has not been selected against.

As discussed in Chapter 7 reliable inference of the function of the well annotated genes in CFT073, but not in MG1655, also potentially contributing to metabolism in the urinary tract, was not done. This would be one of the first steps in improving the incomplete model of CFT073 that was used in Chapter 7, but would require a lot of work linking the huge number of gene-protein-reaction relationships (as presented in, say, MetaCyc) to gene sequences (as presented in GenBank).

Metabolic regulation massively influences the phenotypic response of bacteria to a particular environment, be it soil, the human gut or a shake flask culture. Therefore the genes confirmed as encoding an L-sorbose uptake and utilisation pathway were used as a model for the addition of heterologous metabolic pathways to an *E. coli* strain. It was hoped that this would elucidate some of the factors involved in such addition of metabolic genes in metabolic engineering. The results of this can be seen in Chapter 5. Initial transformation of the genes resulted in a very slow growing strain, which potentially has implications for use of such heterologous pathways in metabolic engineering. Strain capabilities approaching the 'native' wild-type capabilities of CFT073 were only achieved through passage - in this case over about 40 generations. Even then the growth rate and biomass concentrations were below those of the wild-type - and beyond 40 generations these capabilities did not improve. This shows that strains engineered by addition of such genes might well have to take into account poor integration of such genes into the metabolism of the host bacterium.

One potential reason for the poor growth of this initial strain DH5αL could just have been low expression of the genes on the plasmid, since the genes were not put under the control of a constitutive or tuneable promoter. This might well produce faster growing strains, but it would be informative to look at growth rate as a function of expression of the genes, since the rest of metabolism would have to adapt to changes in flux through that pathway, and an easier way to tune the metabolism might just be by selection of fast growing mutants of DH5αL (as was done here through passage), without the need for a tuneable promoter.

The use of metabolic models to guide metabolic engineering and improve strain characteristics for the production of heterologous gene products has been done. For instance, Luo *et al* [166] have used flux balance analysis in a model of central *E. coli* metabolism to improve production of recombinant human-like collagen in *E. coli*, so expansion of these models to include the details of the entire metabolism of the host cell should improve analyses like this further, by taking into account all the possible flux distributions potentially available to the *E. coli*.

One of the features of the data collected about acetate excretion in the strains tested in Chapter 5 is that passage over 100 generations failed to abolish the relationship between carbon source uptake and acetate excretion. Changing regulatory control in central metabolism by genetic engineering has modified strain MG1655 to retain growth rate and total biomass accumulation in minimal media with glucose as sole carbon source [136]. This result indicates that just reducing the excretion of acetate will not necessarily increase growth rate or biomass yield in conditions of minimal media with glucose as sole carbon source. Some *E. coli* strains produce acetate in far lower quantities in similar conditions than MG1655 and closely related strains, for instance strain BL21 [167]. Differing acetate excretion levels are attributed to differing gene expression to MG1655, rather than different unique pathways in such strains [168], so it is not clear why regulation in MG1655 and related bacteria (such as DH5α) produces high acetate excretion rates at high specific glucose uptake rates.

One observation from the work presented in Chapter 5 is that L-sorbose seems to induce the same acetate production at a given growth as glucose, so introduction of this pathway linking to central metabolism does not seem to disrupt the regulation of the overall central metabolism of the cells. It is not yet clear why the passage presented here has not reduced acetate production or, given the observations of Veit *et al* [136], whether any length of passage without careful choice of selection conditions would produce a reduction of acetate production.

Concerning the mechanism of adaptation of DH5αL during passage, the plasmid from DH5αM was re-transformed into a naive DH5α, grown and repassaged (Chapter 5). The results of this indicated that changes had certainly occurred on the chromosome, and it was possible that none of the changes on the plasmid (if there were any) had any effect on the metabolism of the passaged strains. This result is of interest because it indicates that if the changes are based on regulation of the L-sorbose operon then some regulatory mechanism on the chromosome has perhaps adapted to interact with the L-sorbose operon and upregulate expression.

The model of CFT073 presented here is incomplete as it has ignored genes with an annotated function that are not present in MG1655. According to MetaCyc there are 269 reactions that exist in CFT073 that are not in MG1655, although due to the problem of successive inferred annotation and potentially wrongly inferred reactions without a gene dependence (such as for reaction ‘PDH’ in iAF1260 discussed in Chapter 7), this number may be somewhat smaller. The method of gene synteny inference presented in Chapter 6 is conservative in that only the genes (and consequent reactions) in MG1655 have been considered. There are of course many other metabolic genes with biochemically confirmed functions which could be used to assess whether the 269 reactions mentioned above would indeed be part of the CFT073 metabolic network by comparison with CFT073 genes. The difficulty in such an analysis would be extraction of the relevant biochemically characterised genes for BLAST comparison, since although such genes are clearly labelled in MetaCyc such genes are not labelled in GenBank.

Overall this work shows an example of the reconstruction of the metabolic model of a newly sequenced genome and the associated work, both bioinformatic and experimental, that might be required for each new bacterial strain sequenced to produce a complete stoichiometric model of metabolism. One such model has been used to simulate growth of CFT073 in defined media, and models have been used to investigate the adaptation of a bacterial strain with heterologous carbon source uptake genes. The enlargement and development of such models will be increasingly important for analyses of such engineered bacteria, and also in directing engineering for increasing target product yields.

8.2 Functional Genomics and *E. coli*

The availability of sequence data for bacteria far outstrips the capability of current research to analyse and understand those data. For example *Escherichia coli*, despite 1.76×10^6 publications (according to Google Scholar `http://scholar.google.co.uk/scholar`) that mention it and tens of closely related genome sequences, is not fully characterised. The most well characterised strain *E. coli* MG1655 contains 475 genes encoding either ‘conserved’, ‘hypothetical’ or ‘expressed’ proteins, none with a known function.

The functional genomic approach to gene characterisation taken in Chapter 4 takes the approach that data relating to bacteria that survive in particular environments will have patterns, i.e. that genes more frequently observed in the genome sequences of bacteria from a particular environmental niche than of bacteria residing outside that niche might have some beneficial function in that niche. This is an assumption that underlies much other work on *E. coli*, most comprehensively by Rasko *et al* [124] for metabolism but also focusing on bacterial pathogenesis (such as [62]).

This approach has been successfully used here in the context of UPEC, where genes of unconfirmed function have been identified for targeted characterisation and many other candidates for characterisation have been identified.

8.3 *E. coli* response to the introduction of heterologous metabolic genes

The response of bacteria to gene deletion and amplification have been observed in many instances of metabolic engineering (such as recently [169, 170] and many others) and addition of genes through cloning of plasmids is well established as a way of changing the effective genotype of a bacterial strain (for instance [171]). Heterologous biosynthetic pathways have successfully been cloned into *E. coli*. For instance an echinomycin biosynthetic pathway has been heterologously expressed in *E. coli* by Watanabe *et al* [172] and a mevalonate pathway has been engineered for the production of terpenoids also in *E. coli* [173]. Investigation of how bacteria respond to this sort of engineering, by way of a metabolic pathway for uptake of a novel carbon source, was decided for this project.

The response of DH5α over approximately 100 generations of growth in defined media to the introduction of L-sorbose uptake and utilisation genes can be seen in Chapter 5, where it could be seen that growth rate and final biomass concentration were both increased for the conditions in which the adaptation took place. However this adaptation did not bring the adapted strain DH5αM up to the yields achieved by the 'native' strain CFT073, which apart from the L-sorbose uptake operon has quite a similar metabolic network (as seen in Chapter 7). A plateau in growth rate was observed and this indicates that there are mechanisms that might require further analysis for a system such as this to achieve optimal performance.

It is clear that there is a difference between the growth characteristics of the three strains tested in Chapter 5. The observation of heat sensitivity of the *Klebsiella pneumoniae* glucitol-6-phosphate dehydrogenase by Sprenger and Lengeler [118] gave a clue as to a potential reason for differences in the growth profiles of these strains, however Figure 5.8 clearly demonstrates that it is not this enzyme that constitutes the difference between the strains.

8.4 Gene and reaction complement comparisons using genome sequences

The inclusion of a context-based system for determining gene synteny in bacteria, using DiagHunter, and the removal of the possibility for successive inferred annotation, were both achieved in the work presented in Chapter 6. This comparison has given a reliable basis of gene inference for model building for bacteria of the order of *Enterobacteriaceae*, which can be supplemented by functional genomics approaches as described in Chapter 4.

The relationship between gene retention and reaction retention for the *Enterobacteriaceae* has been analysed in Chapter 6 and it has been shown that the DH method returns approximately 83 % of the reactions inferred by MetaCyc to be present in these bacteria. The validation of many of these reactions could potentially be done by extraction of the DNA of the relevant genes from GenBank and use of BLAST on just these relevant genes, to ascertain reliable inferences of gene function conservation.

8.5 Whole genome stoichiometric models

Whole genome stoichiometric models of two strains of *E. coli* have been produced and simulations of growth in defined conditions have been run, as shown in Chapter 7. qualitative predictions of growth have been good and a simple model of acetate production as an overflow metabolite has predicted the production or non-production of acetate during the course of growth.

The analysis of the retention of essential genes for stoichiometric models is of great importance as without these genes models will not run. Chapter 7 shows that in some cases experimental evidence can be used to plausibly argue the existence of reactions inferred absent by gene comparisons, such as the reactions encoded by *trpC*, and this has been directly incorporated in the CFT073 model.

Although these models are only subsets of the possible models for the bacteria it is felt that they are the best basis for building further models due to their reliability. The addition of further reactions that are experimentally validated (e.g. the L-sorbose uptake and utilisation operon) retains the close link between model building and experimental verification of gene function. The cost of such a strategy is the amount of functional information known about many genes that are not present in MG1655, which will not be present in these new models, but this is mitigated by the reliability of the inferences made by the DH method. Further, development of a tool for direct comparison of query genes to genes with experimentally validated metabolic functions would greatly aid in the reconstructive effort.

8.6 Potential further work

8.6.1 Functional genomics and CFT073 metabolism

The bioinformatic data collected in Chapter 4 potentially holds much information about the metabolism of CFT073 and functional characterisation of the sets of metabolic genes found solely in UPEC, and in other pathotypes, could potentially benefit from an approach based on genomic comparisons and expanded from single genes by identification of sets of conserved genes in such pathogens. As the number of sequenced bacteria grows and grows, incorporation of these new bacteria would make methods of functional genomics such as this more and more powerful.

8.6.2 Further investigation of the adaptation of engineered *E. coli*

Further characterisation of the metabolic flux changes during passage of the strain DH5αL by way of radioactive carbon labelling and measurement could provide more insight into the adaptation of this strain to the addition of heterologous metabolic genes.

To benchmark the adaptation observed here (and for similar adaptive tests with heterologous genes from closely related donor strains of bacteria) passage of CFT073 over 100 generations in the same conditions to determine its characteristics after the same selective pressure would provide more information about the potential flux distributions accessible to *E. coli* when growing on L-sorbose. This could potentially shed light on the reason why adaptation did not recover the rates of growth seen in CFT073 in the passage conditions.

Since the beginning of this project fast genome sequencing, such as by Solexa [174], have made feasible the prospect of sequencing single bacteria in a matter of days, and with the MG1655 reference genome, DH5αL and DH5αM could be sequenced to look for single nucleotide polymorphisms (SNPs) and other changes both on the chromosome and the plasmid to try to explain the adaptation seen here. If CFT073 were also passaged then another set of SNPs could be identified by sequencing of passaged and unpassaged CFT073, as a control. These SNPs could then potentially each be studied individually in naive DH5α containing naive pQR793 L-sorbose plasmid, to determine whether any of the changes were individually responsible for changes in L-sorbose uptake and growth in this system, in order to identify how exactly the integration of new metabolic pathways might occur naturally in bacterial systems.

8.6.3 Validation of the DH method of gene synteny inference

The application of the DH method to finding synteny conserved genes gives differing inferences of gene conservation to MetaCyc. Validation of this method would require further bioinformatic and experimental validation of the specific genes on which there are disagreements between the two sets of inferences. This could potentially feed back into a refined version of the DH method for gene inference. Further, combining DiagHunter with the BSR method of gene homology measurement of Rasko *et al* [151] could further refine the technique presented in this work.

8.6.4 Using further bioinformatic evidence for building new metabolic models

There is a great deal of biochemical knowledge about Gene-Protein-Reaction relationships that are not from *E. coli* MG1655, many of which are in the MetaCyc database of reaction pathways. Further reliable addition of pathways from this database could be achieved by direct comparison of query strain genes (and sets of genes) with the genes experimentally characterised and deposited in MetaCyc. One way of doing this would be by developing a tool for extracting the genes and their GPR relationships from MetaCyc, extracting the amino acid sequences of those genes from Genbank and using BLASTp to compare them to the query genes.

Further construction of the CFT073 model in particular will require further experimental verification of the putative metabolic genes presented in this work and development of a system for reliable inference of reactions from genes not present in MG1655, combining sources like MetaCyc and GenBank to produce an automatic system for direct inference of function from experimentally characterised genes from other organisms.

8.7 Final conclusions

The application of context based gene synteny comparisons to model building and addition of experimentally verified pathways to these has been successfully carried out. One such initial model, that of CFT073, has been experimentally validated in some limited conditions. The DH method is an additional tool in the reliable inference of metabolic models and could be incorporated into any existing methods of inference based on BLAST to potentially improve bacterial synteny comparisons. The addition of a metabolic pathway from CFT073 to DH5α has been used to investigate the metabolic changes that occur on addition of such heterologous pathways and how they might integrate into the metabolism of a host bacterium. It is hoped that the work presented here gives an indication of the type of work that will be required in building complete models of metabolism for bacteria and modifying them by addition of reactions from heterologous sources, a combination of bioinformatic and experimental approaches.

Bibliography

[1] A. Feist, C. Henry, J. Reed, M. Krummenacker, A. Joyce, P. Karp, L. Broadbelt, V. Hatzimanikatis, and B. Palsson, “A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information,” *Mol. Syst. Biol.*, vol. 3, p. 121, 2007.

[2] J. G. Mendel, “Versuche ber Plflanzenhybriden,” *Verhandlungen des naturforschenden Vereines*, vol. 4, pp. 3–47, 1865.

[3] G. Beadle and E. Tatum, “Genetic control of biochemical reactions in neurospora,” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 27, pp. 499–506, Nov 1941.

[4] F. Jacob and J. Monod, “Genetic regulatory mechanisms in the synthesis of proteins,” *J. Mol. Biol.*, vol. 3, pp. 318–356, Jun 1961.

[5] V. Sanchez-Torres, T. Maeda, and T. K. Wood, “Protein engineering of the transcriptional activator FhlA To enhance hydrogen production in *Escherichia coli*,” *Appl. Environ. Microbiol.*, vol. 75, pp. 5639–5646, Sep 2009.

[6] A. A. Aristidou, K. Y. San, and G. N. Bennett, “Metabolic engineering of *Escherichia coli* to enhance recombinant protein production through acetate reduction,” *Biotechnol. Prog.*, vol. 11, pp. 475–478, 1995.

[7] J. D. Watson and F. H. Crick, “Molecular structure of nucleic acids;

a structure for deoxyribose nucleic acid," *Nature*, vol. 171, pp. 737–738, Apr 1953.

[8] P. A. Levene, "The Structure of Yeast Nucleic Acid. IV. Ammonia Hydrolysis," *J. Biol. Chem.*, vol. 40, p. 415, Nov 1919.

[9] F. Sanger and A. R. Coulson, "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase," *J. Mol. Biol.*, vol. 94, pp. 441–448, May 1975.

[10] F. Sanger, S. Nicklen, and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 74, pp. 5463–5467, Dec 1977.

[11] A. M. Maxam and W. Gilbert, "A new method for sequencing DNA," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 74, pp. 560–564, Feb 1977.

[12] L. M. Smith, J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. Heiner, S. B. Kent, and L. E. Hood, "Fluorescence detection in automated DNA sequence analysis," *Nature*, vol. 321, pp. 674–679, 1986.

[13] M. D. Adams, A. R. Kerlavage, J. M. Kelley, J. D. Gocayne, C. Fields, C. M. Fraser, and J. C. Venter, "A model for high-throughput automated DNA sequencing and analysis core facilities," *Nature*, vol. 368, pp. 474–475, Mar 1994.

[14] R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, and J. M. Merrick, "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd," *Science*, vol. 269, pp. 496–512, Jul 1995.

[15] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, *et al.*, "The

sequence of the human genome," *Science*, vol. 291, pp. 1304–1351, Feb 2001.

[16] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, *et al.*, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, pp. 860–921, Feb 2001.

[17] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg, "Genome sequencing in microfabricated high-density picolitre reactors," *Nature*, vol. 437, pp. 376–380, Sep 2005.

[18] M. Ronaghi, S. Karamohamed, B. Pettersson, M. Uhln, and P. Nyrn, "Real-time DNA sequencing using detection of pyrophosphate release," *Anal. Biochem.*, vol. 242, pp. 84–89, Nov 1996.

[19] D. G. Hert, C. P. Fredlake, and A. E. Barron, "Advantages and limitations of next-generation sequencing technologies: a comparison of electrophoresis and non-electrophoresis methods," *Electrophoresis*, vol. 29, pp. 4618–4626, Dec 2008.

[20] E. Y. Chan, "Advances in sequencing technology," *Mutat. Res.*, vol. 573, pp. 13–40, Jun 2005.

[21] P. K. Gupta, “Single-molecule DNA sequencing technologies for future genomics research,” *Trends Biotechnol.*, vol. 26, pp. 602–611, Nov 2008.

[22] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool,” *J. Mol. Biol.*, vol. 215, pp. 403–410, Oct 1990.

[23] S. Altschul, T. Madden, A. Schffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman, “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs,” *Nucleic Acids Res.*, vol. 25, pp. 3389–3402, Sep 1997.

[24] G. D. Schuler, J. A. Epstein, H. Ohkawa, and J. A. Kans, “Entrez: molecular biology database and retrieval system,” *Meth. Enzymol.*, vol. 266, pp. 141–162, 1996.

[25] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, “GenBank,” *Nucleic Acids Res.*, vol. 37, pp. 26–31, Jan 2009.

[26] K. Liolios, N. Tavernarakis, P. Hugenholtz, and N. C. Kyrpides, “The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide,” *Nucleic Acids Res.*, vol. 34, pp. D332–334, Jan 2006.

[27] M. Cren, A. Kondorosi, and E. Kondorosi, “An insertional point mutation inactivates NolR repressor in Rhizobium meliloti 1021,” *J. Bacteriol.*, vol. 176, pp. 518–519, Jan 1994.

[28] A. Iguchi, N. Thomson, Y. Ogura, D. Saunders, T. Ooka, I. Henderson, D. Harris, M. Asadulghani, K. Kurokawa, P. Dean, B. Kenny, M. Quail, S. Thurston, G. Dougan, T. Hayashi, J. Parkhill, and G. Frankel, “Complete genome sequence and comparative genome analysis of enteropathogenic *Escherichia coli* O127:H6 strain E2348/69,” *J. Bacteriol.*, vol. 191, pp. 347–354, Jan 2009.

[29] A. Marchler-Bauer, J. B. Anderson, M. K. Derbyshire, C. DeWeese-Scott, N. R. Gonzales, M. Gwadz, L. Hao, S. He, D. I. Hurwitz, J. D. Jackson, Z. Ke, D. Krylov, C. J. Lanczycki, C. A. Liebert, C. Liu, F. Lu, S. Lu, G. H. Marchler, M. Mullokandov, J. S. Song, N. Thanki, R. A. Yamashita, J. J. Yin, D. Zhang, and S. H. Bryant, "CDD: a conserved domain database for interactive domain family analysis," *Nucleic Acids Res.*, vol. 35, pp. D237–240, Jan 2007.

[30] O. Man, T. Atarot, A. Sadot, T. Olender, and D. Lancet, "From subgenome analysis to protein structure," *Curr. Opin. Struct. Biol.*, vol. 13, pp. 353–358, Jun 2003.

[31] M. Punta and Y. Ofran, "The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function," *PLoS Comput. Biol.*, vol. 4, p. e1000160, Oct 2008.

[32] D. Pal and D. Eisenberg, "Inference of protein function from protein structure," *Structure*, vol. 13, pp. 121–130, Jan 2005.

[33] R. A. Laskowski, J. D. Watson, and J. M. Thornton, "ProFunc: a server for predicting protein function from 3D structure," *Nucleic Acids Res.*, vol. 33, pp. 89–93, Jul 2005.

[34] F. Sanger, G. Air, B. Barrell, N. Brown, A. Coulson, C. Fiddes, C. Hutchison, P. Slocombe, and M. Smith, "Nucleotide sequence of bacteriophage phi X174 DNA," *Nature*, vol. 265, pp. 687–695, Feb 1977.

[35] F. R. Blattner, G. r. Plunkett, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao, "The complete genome sequence of *Escherichia coli* K-12," *Science*, vol. 277, pp. 1453–1474, Sep 1997.

[36] G. Huys, M. Cnockaert, J. Janda, and J. Swings, “Escherichia albertii sp. nov., a diarrhoeagenic species isolated from stool specimens of Bangladeshi children,” *Int. J. Syst. Evol. Microbiol.*, vol. 53, pp. 807–810, May 2003.

[37] B. Hochhut, C. Wilde, G. Balling, B. Middendorf, U. Dobrindt, E. Brzuszkiewicz, G. Gottschalk, E. Carniel, and J. Hacker, “Role of pathogenicity island-associated integrases in the genome plasticity of uropathogenic *Escherichia coli* strain 536,” *Mol. Microbiol.*, vol. 61, pp. 584–595, Aug 2006.

[38] C. Bernier, P. Gounon, and C. Le Bougunec, “Identification of an aggregative adhesion fimbria (AAF) type III-encoding operon in enteroaggregative *Escherichia coli* as a sensitive probe for detecting the AAF-encoding operon family,” *Infect. Immun.*, vol. 70, pp. 4302–4311, Aug 2002.

[39] T. Johnson, S. Kariyawasam, Y. Wannemuehler, P. Mangiamele, S. Johnson, C. Doetkott, J. Skyberg, A. Lynne, J. Johnson, and L. Nolan, “Genome Sequence of Avian Pathogenic *Escherichia coli* Strain O1:K1:H7 Shares Strong Similarities with Human ExPEC Genomes,” *J Bacteriol*, vol. 189, pp. 3228–3236, Apr 2007.

[40] L. Paulozzi, K. Johnson, L. Kamahele, C. Clausen, L. Riley, and S. Helgerson, “Diarrhea associated with adherent enteropathogenic *Escherichia coli* in an infant and toddler center, Seattle, Washington,” *Pediatrics*, vol. 77, pp. 296–300, Mar 1986.

[41] H. DuPont, S. Formal, R. Hornick, M. Snyder, J. Libonati, D. Sheahan, E. LaBrec, and J. Kalas, “Pathogenesis of *Escherichia coli* diarrhea,” *N. Engl. J. Med.*, vol. 285, pp. 1–9, Jul 1971.

[42] R. A. Welch, V. Burland, G. Plunkett 3rd, P. Redford, P. Roesch, D. Rasko, E. L. Buckles, S.-R. Liou, A. Boutin, and other authors,

"Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*," *Proc Natl Acad Sci U S A*, vol. 99, pp. 17020–17024, Dec 2002.

[43] M. Viljanen, T. Peltola, S. Junnila, L. Olkkonen, H. Jrvinen, M. Kuistila, and P. Huovinen, "Outbreak of diarrhoea due to *Escherichia coli* O111:B4 in schoolchildren and adults: association of Vi antigen-like reactivity," *Lancet*, vol. 336, pp. 831–834, Oct 1990.

[44] P. Escobar-Pramo, K. Grenet, A. Le Menac'h, L. Rode, E. Salgado, C. Amorin, S. Gouriou, B. Picard, M. Rahimy, A. Andremont, E. Denamur, and R. Ruimy, "Large-scale population structure of human commensal *Escherichia coli* isolates," *Appl. Environ. Microbiol.*, vol. 70, pp. 5698–5700, Sep 2004.

[45] A. Stapleton, S. Moseley, and W. E. Stamm, "Urovirulence determinants in *Escherichia coli* isolates causing first-episode and recurrent cystitis in women," *J. Infect. Dis.*, vol. 163, pp. 773–779, Apr 1991.

[46] M. Levine, E. Bergquist, D. Nalin, D. Waterman, R. Hornick, C. Young, and S. Sotman, "Escherichia coli strains that cause diarrhoea but do not produce heat-labile or heat-stable enterotoxins and are non-invasive," *Lancet*, vol. 1, pp. 1119–1122, May 1978.

[47] B. Picard, J. Garcia, S. Gouriou, P. Duriez, N. Brahimi, E. Bingen, J. Elion, and E. Denamur, "The link between phylogeny and virulence in *Escherichia coli* extraintestinal infection," *Infect. Immun.*, vol. 67, pp. 546–553, Feb 1999.

[48] T. Durfee, R. Nelson, S. Baldwin, G. Plunkett, V. Burland, B. Mau, J. Petrosino, X. Qin, D. Muzny, M. Ayele, R. Gibbs, B. Csrgo, G. Psfai, G. Weinstock, and F. Blattner, "The complete genome sequence of *Escherichia coli* DH10B: insights into the biology of a laboratory workhorse," *J. Bacteriol.*, vol. 190, pp. 2597–2606, Apr 2008.

[49] K. Hayashi, N. Morooka, Y. Yamamoto, K. Fujita, K. Isono, S. Choi, E. Ohtsubo, T. Baba, B. Wanner, H. Mori, and T. Horiuchi, “Highly accurate genome sequences of *Escherichia coli* K-12 strains MG1655 and W3110,” *Mol. Syst. Biol.*, vol. 2, 2006.

[50] T. Hayashi, K. Makino, M. Ohnishi, K. Kurokawa, K. Ishii, K. Yokoyama, C. G. Han, E. Ohtsubo, K. Nakayama, T. Murata, M. Tanaka, T. Tobe, T. Iida, H. Takami, T. Honda, C. Sasakawa, N. Ogasawara, T. Yasunaga, S. Kuhara, T. Shiba, M. Hattori, and H. Shinagawa, “Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12,” *DNA Res*, vol. 8, pp. 11–22, Feb 2001.

[51] N. Perna, G. Plunkett, V. Burland, B. Mau, J. Glasner, D. Rose, G. Mayhew, P. Evans, J. Gregor, H. Kirkpatrick, G. Psfai, J. Hackett, S. Klink, A. Boutin, Y. Shao, L. Miller, E. Grotbeck, N. Davis, A. Lim, E. Dimalanta, K. Potamousis, J. Apodaca, T. Anantharaman, J. Lin, G. Yen, D. Schwartz, R. Welch, and F. Blattner, “Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7,” *Nature*, vol. 409, pp. 529–533, Jan 2001.

[52] S. Bonacorsi, O. Clermont, V. Houdouin, C. Cordevant, N. Brahimi, A. Marecat, C. Tinsley, X. Nassif, M. Lange, and E. Bingen, “Molecular analysis and experimental virulence of French and North American *Escherichia coli* neonatal meningitis isolates: identification of a new virulent clone,” *J. Infect. Dis.*, vol. 187, pp. 1895–1906, Jun 2003.

[53] K. Oshima, H. Toh, Y. Ogura, H. Sasamoto, H. Morita, S. Park, T. Ooka, S. Iyoda, T. Taylor, T. Hayashi, K. Itoh, and M. Hattori, “Complete genome sequence and comparative analysis of the wild-type commensal *Escherichia coli* strain SE11 isolated from a healthy adult,” *DNA Res.*, vol. 15, pp. 375–386, Dec 2008.

[54] W. Fricke, M. Wright, A. Lindell, D. Harkins, C. Baker-Austin, J. Ravel, and R. Stepanauskas, “Insights into the environmental resistance gene pool from the genome sequence of the multidrug-resistant environmental isolate *Escherichia coli* SMS-3-5,” *J. Bacteriol.*, vol. 190, pp. 6779–6794, Oct 2008.

[55] A. Manges, J. Johnson, B. Foxman, T. O’Bryan, K. Fullerton, and L. Riley, “Widespread distribution of urinary tract infections caused by a multidrug-resistant *Escherichia coli* clonal group,” *N. Engl. J. Med.*, vol. 345, pp. 1007–1013, Oct 2001.

[56] S. Chen, C. Hung, J. Xu, C. Reigstad, V. Magrini, A. Sabo, D. Blasiar, T. Bieri, R. Meyer, P. Ozersky, J. Armstrong, R. Fulton, J. Latreille, J. Spieth, T. Hooton, E. Mardis, S. Hultgren, and J. Gordon, “Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: a comparative genomics approach,” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 103, pp. 5977–5982, Apr 2006.

[57] J. Farmer, G. Fanning, B. Davis, C. O’Hara, C. Riddle, F. Hickman-Brenner, M. Asbury, V. Lowery, and D. Brenner, “Escherichia fergusonii and Enterobacter taylorae, two new species of Enterobacteriaceae isolated from clinical specimens,” *J. Clin. Microbiol.*, vol. 21, pp. 77–81, Jan 1985.

[58] C. Hill and B. Harnish, “Inversions between ribosomal RNA genes of Escherichia coli,” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 78, pp. 7069–7072, Nov 1981.

[59] S. Gama-Castro, V. Jimnez-Jacinto, M. Peralta-Gil, A. Santos-Zavaleta, M. Pealoza-Spinola, B. Contreras-Moreira, J. Segura-Salazar, L. Muiz-Rascado, I. Martnez-Flores, H. Salgado, C. Bonavides-Martnez, C. Abreu-Goodger, C. Rodrguez-Penagos, J. Miranda-Ros, E. Morett, E. Merino, A. Huerta, L. Trevio-Quintanilla, and J. Collado-Vides, “RegulonDB (version 6.0): gene

regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation," *Nucleic Acids Res.*, vol. 36, pp. D120–124, Jan 2008.

[60] C. Fraser, J. Gocayne, O. White, M. Adams, R. Clayton, R. Fleischmann, C. Bult, A. Kerlavage, G. Sutton, J. Kelley, R. Fritchman, J. Weidman, K. Small, M. Sandusky, J. Fuhrmann, D. Nguyen, T. Utterback, D. Saudek, C. Phillips, J. Merrick, J. Tomb, B. Dougherty, K. Bott, P. Hu, T. Lucier, S. Peterson, H. Smith, C. Hutchison, and J. Venter, "The minimal gene complement of Mycoplasma genitalium," *Science*, vol. 270, pp. 397–403, Oct 1995.

[61] R. Fani, M. Brilli, and P. Lio, "The origin and evolution of operons: the piecewise building of the proteobacterial histidine operon," *J Mol Evol*, vol. 60, pp. 378–390, Mar 2005.

[62] E. Brzuszkiewicz, H. Bruggemann, H. Liesegang, M. Emmerth, T. Olschlager, G. Nagy, K. Albermann, C. Wagner, C. Buchrieser, and other authors, "How to become a uropathogen: comparative genomic analysis of extraintestinal pathogenic *Escherichia coli* strains," *Proc Natl Acad Sci U S A*, vol. 103, pp. 12879–12884, Aug 2006.

[63] A. R. Mushegian and E. V. Koonin, "A minimal gene set for cellular life derived by comparison of complete bacterial genomes," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 93, pp. 10268–10273, Sep 1996.

[64] J. G. Lawrence and H. Ochman, "Molecular archaeology of the *Escherichia coli* genome," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 95, pp. 9413–9417, Aug 1998.

[65] E. V. Koonin, A. R. Mushegian, M. Y. Galperin, and D. R. Walker, "Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea," *Mol. Microbiol.*, vol. 25, pp. 619–637, Aug 1997.

[66] J. L. Siefert, K. A. Martin, F. Abdi, W. R. Widger, G. E. Fox, and G. E. Fox, “Conserved gene clusters in bacterial genomes provide further support for the primacy of RNA,” *J. Mol. Evol.*, vol. 45, pp. 467–472, Nov 1997.

[67] T. Tsuru and I. Kobayashi, “Multiple genome comparison within a bacterial species reveals a unit of evolution spanning two adjacent genes in a tandem paralog cluster,” *Mol. Biol. Evol.*, vol. 25, pp. 2457–2473, Nov 2008.

[68] S. M. Turner, R. R. Chaudhuri, Z.-D. Jiang, H. DuPont, C. Gyles, C. W. Penn, M. J. Pallen, and I. R. Henderson, “Phylogenetic comparisons reveal multiple acquisitions of the toxin genes by enterotoxigenic *Escherichia coli* strains of different evolutionary lineages,” *J Clin Microbiol*, vol. 44, pp. 4528–4536, Dec 2006.

[69] C. L. Barrett and B. O. Palsson, “Iterative reconstruction of transcriptional regulatory networks: an algorithmic approach,” *PLoS Comput Biol*, vol. 2, p. 52, May 2006.

[70] Q. Yang and S. Sze, “Large-scale analysis of gene clustering in bacteria,” *Genome Res.*, vol. 18, pp. 949–956, Jun 2008.

[71] M. W. Covert, C. H. Schilling, and B. Palsson, “Regulation of gene expression in flux balance models of metabolism,” *J Theor Biol*, vol. 213, pp. 73–88, Nov 2001.

[72] M. W. Covert, N. Xiao, T. J. Chen, and J. R. Karr, “Integrating metabolic, transcriptional regulatory and signal transduction models in *Escherichia coli*,” *Bioinformatics*, vol. 24, pp. 2044–2050, Sep 2008.

[73] T. Handorf, O. Ebenhh, and R. Heinrich, “Expanding metabolic networks: scopes of compounds, robustness, and evolution,” *J. Mol. Evol.*, vol. 61, pp. 498–512, Oct 2005.

[74] H. W. Ma and A. P. Zeng, "The connectivity structure, giant strong component and centrality of metabolic networks," *Bioinformatics*, vol. 19, pp. 1423–1430, Jul 2003.

[75] T. Ochiai, J. C. Nacher, and T. Akutsu, "A constructive approach to gene expression dynamics," *Physical Letters A*, vol. 30, pp. 313–321, Sep 2004.

[76] J. C. Nacher, T. Ochiai, T. Yamada, M. Kanehisa, and T. Akutsu, "The role of log-normal dynamics in the evolution of biochemical pathways," *BioSystems*, vol. 83, pp. 26–37, Jan 2006.

[77] M. DeJongh, K. Formsma, P. Boillot, J. Gould, M. Rycenga, and A. Best, "Toward the automated generation of genome-scale metabolic networks in the SEED," *BMC Bioinformatics*, vol. 8, p. 139, 2007.

[78] R. Caspi, H. Foerster, C. Fulcher, P. Kaipa, M. Krummenacker, M. Latendresse, S. Paley, S. Rhee, A. Shearer, C. Tissier, T. Walk, P. Zhang, and P. Karp, "The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases," *Nucleic Acids Res.*, vol. 36, pp. D623–631, Jan 2008.

[79] G. M. Oddone, D. A. Mills, and D. E. Block, "A dynamic, genome-scale flux model of Lactococcus lactis to increase specific recombinant protein expression," *Metab. Eng.*, Aug 2009.

[80] A. P. Oliveira, J. Nielsen, and J. Frster, "Modeling Lactococcus lactis using a genome-scale flux model," *BMC Microbiol.*, vol. 5, p. 39, 2005.

[81] Q. Wang, X. Chen, Y. Yang, and X. Zhao, "Genome-scale in silico aided metabolic analysis and flux comparisons of Escherichia coli to improve succinate production," *Appl. Microbiol. Biotechnol.*, vol. 73, pp. 887–894, Dec 2006.

[82] P. D. Karp, S. Paley, and P. Romero, “The Pathway Tools software,” *Bioinformatics*, vol. 18 Suppl 1, pp. S225–232, 2002.

[83] R. A. Notebaart, F. H. van Enckevort, C. Francke, R. J. Siezen, and B. Teusink, “Accelerating the reconstruction of genome-scale metabolic networks,” *BMC Bioinformatics*, vol. 7, p. 296, 2006.

[84] C. Mdigue and I. Moszer, “Annotation, comparison and databases for hundreds of bacterial genomes,” *Res. Microbiol.*, vol. 158, pp. 724–736, Dec 2007.

[85] M. Durot, P. Y. Bourguignon, and V. Schachter, “Genome-scale models of bacterial metabolism: reconstruction and applications,” *FEMS Microbiol. Rev.*, vol. 33, pp. 164–190, Jan 2009.

[86] A. L. Knorr, R. Jain, and R. Srivastava, “Bayesian-based selection of metabolic objective functions,” *Bioinformatics*, vol. 23, pp. 351–357, Feb 2007.

[87] S. S. Fong, J. Y. Marciniak, and B. . Palsson, “Description and interpretation of adaptive evolution of *Escherichia coli* K-12 MG1655 by using a genome-scale in silico metabolic model,” *J. Bacteriol.*, vol. 185, pp. 6400–6408, Nov 2003.

[88] T. Baba, T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner, and H. Mori, “Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection,” *Mol. Syst. Biol.*, vol. 2, 2006.

[89] S. Y. Gerdes, M. D. Scholle, J. W. Campbell, G. Balzsi, E. Ravasz, M. D. Daugherty, A. L. Somera, N. C. Kyrpides, I. Anderson, M. S. Gelfand, A. Bhattacharya, V. Kapatral, M. D’Souza, M. V. Baev, Y. Grechkin, F. Mseeh, M. Y. Fonstein, R. Overbeek, A. L. Barabsi, Z. N. Oltvai, and A. L. Osterman, “Experimental determination and

system level analysis of essential genes in *Escherichia coli* MG1655," *J. Bacteriol.*, vol. 185, pp. 5673–5684, Oct 2003.

[90] Y. Yamazaki, H. Niki, and J. Kato, "Profiling of *Escherichia coli* Chromosome database," *Methods Mol. Biol.*, vol. 416, pp. 385–389, 2008.

[91] G. Bertani, "Studies on lysogenesis. I. The mode of phage liberation by lysogenic *Escherichia coli*," *J. Bacteriol.*, vol. 62, pp. 293–300, Sep 1951.

[92] A. Ronald, "The etiology of urinary tract infection: traditional and emerging pathogens," *Dis Mon*, vol. 49, pp. 71–82, Feb 2003.

[93] H. L. Mobley, D. M. Green, A. L. Trifillis, D. E. Johnson, G. R. Chippendale, C. V. Lockatell, B. D. Jones, and J. W. Warren, "Pyelonephritogenic *Escherichia coli* and killing of cultured human renal proximal tubular epithelial cells: role of hemolysin in some strains," *Infect Immun*, vol. 58, pp. 1281–1289, May 1990.

[94] D. E. Johnson, C. V. Lockatell, R. G. Russell, J. R. Hebel, M. D. Island, A. Stapleton, W. E. Stamm, and J. W. Warren, "Comparison of *Escherichia coli* strains recovered from human cystitis and pyelonephritis infections in transurethrally challenged mice," *Infect. Immun.*, vol. 66, pp. 3059–3065, Jul 1998.

[95] G. Blum, M. Ott, A. Lischewski, A. Ritter, H. Imrich, H. Tschpe, and J. Hacker, "Excision of large DNA regions termed pathogenicity islands from tRNA-specific loci in the chromosome of an *Escherichia coli* wild-type pathogen," *Infect. Immun.*, vol. 62, pp. 606–614, Feb 1994.

[96] J. Hacker, G. Blum-Oehler, I. Mhldorfer, and H. Tschpe, "Pathogenicity islands of virulent bacteria: structure, function and

impact on microbial evolution," *Mol. Microbiol.*, vol. 23, pp. 1089–1097, Mar 1997.

[97] D. A. Rasko, J. A. Phillips, X. Li, and H. L. Mobley, "Identification of DNA sequences from a second pathogenicity island of uropathogenic *Escherichia coli* CFT073: probes specific for uropathogenic populations," *J. Infect. Dis.*, vol. 184, pp. 1041–1049, Oct 2001.

[98] D. M. Guyer, J. S. Kao, and H. L. Mobley, "Genomic analysis of a pathogenicity island in uropathogenic *Escherichia coli* CFT073: distribution of homologous sequences among isolates from patients with pyelonephritis, cystitis, and Catheter-associated bacteriuria and from fecal samples," *Infect. Immun.*, vol. 66, pp. 4411–4417, Sep 1998.

[99] T. J. Wiles, R. R. Kulesus, and M. A. Mulvey, "Origins and virulence mechanisms of uropathogenic *Escherichia coli*," *Exp. Mol. Pathol.*, vol. 85, pp. 11–19, Aug 2008.

[100] D. M. Guyer, I. R. Henderson, J. P. Nataro, and H. L. Mobley, "Identification of sat, an autotransporter toxin produced by uropathogenic *Escherichia coli*," *Mol. Microbiol.*, vol. 38, pp. 53–66, Oct 2000.

[101] A. Lloyd, D. Rasko, and H. Mobley, "Defining Genomic Islands and Uropathogen-Specific Genes in Uropathogenic Escherichia coli," *J Bacteriol*, vol. 189, pp. 3532–3546, May 2007.

[102] R. Steadman and N. Topley, *Urinary Tract Infections*, ch. 3. Chapman & Hall, 1998.

[103] U. Lindberg, L. A. Hanson, U. Jodal, G. Lidin-Janson, K. Lincoln, and S. Olling, "Asymptomatic bacteriuria in schoolgirls. II. Differences in escherichia coli causing asymptomatic bacteriuria," *Acta Paediatr Scand*, vol. 64, pp. 432–436, May 1975.

[104] V. Roos and P. Klemm, "Global gene expression profiling of the

asymptomatic bacteriuria *Escherichia coli* strain 83972 in the human urinary tract," *Infect. Immun.*, vol. 74, pp. 3565–3575, Jun 2006.

[105] S. J. Hultgren, W. R. Schwan, A. J. Schaeffer, and J. L. Duncan, "Regulation of production of type 1 pili among urinary tract isolates of *Escherichia coli*," *Infect. Immun.*, vol. 54, pp. 613–620, Dec 1986.

[106] V. Roos, G. C. Ulett, M. A. Schembri, and P. Klemm, "The asymptomatic bacteriuria *Escherichia coli* strain 83972 outcompetes uropathogenic E. coli strains in human urine," *Infect Immun*, vol. 74, pp. 615–624, Jan 2006.

[107] R. O. Darouiche, W. H. Donovan, M. Del Terzo, J. I. Thornby, D. C. Rudy, and R. A. Hull, "Pilot trial of bacterial interference for preventing urinary tract infection," *Urology*, vol. 58, pp. 339–344, Sep 2001.

[108] R. U. Ibarra, J. S. Edwards, and B. O. Palsson, "*Escherichia coli* K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth," *Nature*, vol. 420, pp. 186–189, Nov 2002.

[109] L. Emődy, M. Kerenyi, and G. Nagy, "Virulence factors of uropathogenic *Escherichia coli*," *Int J Antimicrob Agents*, vol. 22 Suppl 2, pp. 29–33, Oct 2003.

[110] D. M. Gordon and M. A. Riley, "A theoretical and experimental analysis of bacterial growth in the bladder," *Mol Microbiol*, vol. 6, pp. 555–562, Feb 1992.

[111] D. Benson, I. Karsch-Mizrachi, D. Lipman, J. Ostell, and D. Wheeler, "GenBank," *Nucleic Acids Res.*, vol. 36, pp. 25–30, Jan 2008.

[112] T. A. Russo, S. T. Jodush, J. J. Brown, and J. R. Johnson, "Identification of two previously unrecognized genes (guaA and argC) important for uropathogenesis," *Mol. Microbiol.*, vol. 22, pp. 217–229, Oct 1996.

[113] P. L. Roesch, P. Redford, S. Batchelet, R. L. Moritz, S. Pellett, B. J. Haugen, F. R. Blattner, and R. A. Welch, “Uropathogenic *Escherichia coli* use d-serine deaminase to modulate infection of the murine urinary tract,” *Mol. Microbiol.*, vol. 49, pp. 55–67, Jul 2003.

[114] A. T. Anfora and R. A. Welch, “DsdX is the second D-serine transporter in uropathogenic *Escherichia coli* clinical isolate CFT073,” *J. Bacteriol.*, vol. 188, pp. 6622–6628, Sep 2006.

[115] A. T. Anfora, B. J. Haugen, P. Roesch, P. Redford, and R. A. Welch, “Roles of serine accumulation and catabolism in the colonization of the murine urinary tract by *Escherichia coli* CFT073,” *Infect. Immun.*, vol. 75, pp. 5298–5304, Nov 2007.

[116] A. T. Anfora, D. K. Halladin, B. J. Haugen, and R. A. Welch, “Uropathogenic *Escherichia coli* CFT073 is adapted to acetatogenic growth but does not require acetate during murine urinary tract infection,” *Infect. Immun.*, vol. 76, pp. 5760–5767, Dec 2008.

[117] M. Sabri, S. Houle, and C. M. Dozois, “Roles of the extraintestinal pathogenic *Escherichia coli* ZnuACB and ZupT zinc transporters during urinary tract infection,” *Infect. Immun.*, vol. 77, pp. 1155–1164, Mar 2009.

[118] G. A. Sprenger and J. W. Lengeler, “L-Sorbose metabolism in Klebsiella pneumoniae and Sor+ derivatives of *Escherichia coli* K-12 and chemotaxis toward sorbose,” *J Bacteriol*, vol. 157, pp. 39–45, Jan 1984.

[119] M. J. Novotny, J. Reizer, F. Esch, and M. H. Saier, “Purification and properties of D-mannitol-1-phosphate dehydrogenase and D-glucitol-6-phosphate dehydrogenase from *Escherichia coli*,” *J. Bacteriol.*, vol. 159, pp. 986–990, Sep 1984.

[120] C. K. Lee, R. M. Daniel, C. Shepherd, D. Saul, S. C. Cary, M. J. Danson, R. Eisenthal, and M. E. Peterson, “Eurythermalism and the temperature dependence of enzyme activity,” *FASEB J.*, vol. 21, pp. 1934–1941, Jun 2007.

[121] J. A. Nelder and R. Mead, “A simplex method for function minimisation,” *The Computer Journal*, vol. 7, pp. 308–313, 1965.

[122] P. Bernard, P. Gabant, E. M. Bahassi, and M. Couturier, “Positive-selection vectors using the F plasmid ccdB killer gene,” *Gene*, vol. 148, pp. 71–74, Oct 1994.

[123] K. Rutherford, J. Parkhill, J. Crook, T. Horsnell, P. Rice, M. A. Rajandream, and B. Barrell, “Artemis: sequence visualization and annotation,” *Bioinformatics*, vol. 16, pp. 944–945, Oct 2000.

[124] D. A. Rasko, M. J. Rosovitz, G. S. Myers, E. F. Mongodin, W. F. Fricke, P. Gajer, J. Crabtree, M. Sebaihia, N. R. Thomson, R. Chaudhuri, I. R. Henderson, V. Sperandio, and J. Ravel, “The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates,” *J. Bacteriol.*, vol. 190, pp. 6881–6893, Oct 2008.

[125] J. Johnson, T. Berggren, D. Newburg, R. McCluer, and J. Manivel, “Detailed histopathological examination contributes to the assessment of *Escherichia coli* urovirulence,” *J. Urol.*, vol. 147, pp. 1160–1166, Apr 1992.

[126] National Institute of Health, “GenBank,” *http://www.ncbi.nlm.nih.gov/Genbank/index.html*, 2006.

[127] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler, “GenBank,” *Nucleic Acids Res*, vol. 35, pp. 21–25, Jan 2007.

[128] A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, and other authors, "The Pfam protein families database," *Nucleic Acids Res*, vol. 32, pp. 138–141, Jan 2004.

[129] A. Marchler-Bauer, J. B. Anderson, P. F. Cherukuri, C. DeWeese-Scott, L. Y. Geer, M. Gwadz, S. He, D. I. Hurwitz, J. D. Jackson, Z. Ke, C. J. Lanczycki, C. A. Liebert, C. Liu, F. Lu, G. H. Marchler, M. Mullokandov, B. A. Shoemaker, V. Simonyan, J. S. Song, P. A. Thiessen, R. A. Yamashita, J. J. Yin, D. Zhang, and S. H. Bryant, "CDD: a Conserved Domain Database for protein classification," *Nucleic Acids Res.*, vol. 33, pp. D192–196, Jan 2005.

[130] S. B. Cannon, A. Kozik, B. Chan, R. Michelmore, and N. D. Young, "DiagHunter and GenoPix2D: programs for genomic comparisons, large-scale homology discovery and visualization," *Genome Biol.*, vol. 4, p. R68, 2003.

[131] A. Lehmacher and J. Bockemhl, "L-Sorbose utilization by virulent *Escherichia coli* and *Shigella*: different metabolic adaptation of pathotypes," *Int. J. Med. Microbiol.*, vol. 297, pp. 245–254, Jul 2007.

[132] R. Caspi, H. Foerster, C. A. Fulcher, R. Hopkinson, J. Ingraham, P. Kaipa, M. Krummenacker, S. Paley, J. Pick, and other authors, "MetaCyc: a multiorganism database of metabolic pathways and enzymes," *Nucleic Acids Res*, vol. 34, pp. 511–516, Jan 2006.

[133] D. S. Wishart, D. Tzur, C. Knox, R. Eisner, A. C. Guo, N. Young, D. Cheng, K. Jewell, D. Arndt, and other authors, "HMDB: the Human Metabolome Database," *Nucleic Acids Res*, vol. 35, pp. 521–526, Jan 2007.

[134] J. D. Thompson, T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins, "The CLUSTAL_X windows interface: flexible strategies

for multiple sequence alignment aided by quality analysis tools," *Nucleic Acids Res*, vol. 25, pp. 4876–4882, Dec 1997.

[135] G. A. Sprenger and J. W. Lengeler, "Mapping of the sor genes for L-sorbose degradation in the chromosome of Klebsiella pneumoniae," *Mol. Gen. Genet.*, vol. 209, pp. 352–359, Sep 1987.

[136] A. Veit, T. Polen, and V. F. Wendisch, "Global gene expression analysis of glucose overflow metabolism in *Escherichia coli* and reduction of aerobic acetate formation," *Appl. Microbiol. Biotechnol.*, vol. 74, pp. 406–421, Feb 2007.

[137] M. S. Wong, S. Wu, T. B. Causey, G. N. Bennett, and K. Y. San, "Reduction of acetate accumulation in *Escherichia coli* cultures for increased recombinant protein production," *Metab. Eng.*, vol. 10, pp. 97–108, Mar 2008.

[138] S. Pukatzki, A. T. Ma, D. Sturtevant, B. Krastins, D. Sarracino, W. C. Nelson, J. F. Heidelberg, and J. J. Mekalanos, "Identification of a conserved bacterial protein secretion system in Vibrio cholerae using the Dictyostelium host model system," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 103, pp. 1528–1533, Jan 2006.

[139] K. H. Lee, J. H. Park, T. Y. Kim, H. U. Kim, and S. Y. Lee, "Systems metabolic engineering of *Escherichia coli* for L-threonine production," *Mol. Syst. Biol.*, vol. 3, p. 149, 2007.

[140] D. J. Korz, U. Rinas, K. Hellmuth, E. A. Sanders, and W. D. Deckwer, "Simple fed-batch technique for high cell density cultivation of *Escherichia coli*," *J. Biotechnol.*, vol. 39, pp. 59–65, Feb 1995.

[141] G. N. Vemuri, E. Altman, D. P. Sangurdekar, A. B. Khodursky, and M. A. Eiteman, "Overflow metabolism in *Escherichia coli* during steady-state growth: transcriptional regulation and effect of the re-

dox ratio," *Appl. Environ. Microbiol.*, vol. 72, pp. 3653–3661, May 2006.

[142] S. S. Fong, A. P. Burgard, C. D. Herring, E. M. Knight, F. R. Blattner, C. D. Maranas, and B. O. Palsson, "In silico design and adaptive evolution of *Escherichia coli* for production of lactic acid," *Biotechnol. Bioeng.*, vol. 91, pp. 643–648, Sep 2005.

[143] T. Ferenci, "Bacterial physiology, regulation and mutational adaptation in a chemostat environment," *Adv. Microb. Physiol.*, vol. 53, pp. 169–229, 2008.

[144] S. Y. Yau, E. Keshavarz-Moore, and J. Ward, "Host strain influences on supercoiled plasmid DNA production in *Escherichia coli*: Implications for efficient design of large-scale processes," *Biotechnol. Bioeng.*, vol. 101, pp. 529–544, Oct 2008.

[145] S. L. Herendeen, R. A. VanBogelen, and F. C. Neidhardt, "Levels of major proteins of *Escherichia coli* during growth at different temperatures," *J. Bacteriol.*, vol. 139, pp. 185–194, Jul 1979.

[146] M. A. Eiteman and E. Altman, "Overcoming acetate in *Escherichia coli* recombinant protein fermentations," *Trends Biotechnol.*, vol. 24, pp. 530–536, Nov 2006.

[147] G. W. Luli and W. R. Strohl, "Comparison of growth, acetate production, and acetate inhibition of *Escherichia coli* strains in batch and fed-batch fermentations," *Appl. Environ. Microbiol.*, vol. 56, pp. 1004–1011, Apr 1990.

[148] S. D. Tsen, S. C. Lai, C. P. Pang, J. I. Lee, and T. H. Wilson, "Chemostat selection of an *Escherichia coli* mutant containing permease with enhanced lactose affinity," *Biochem. Biophys. Res. Commun.*, vol. 224, pp. 351–357, Jul 1996.

[149] K. Liolios, K. Mavromatis, N. Tavernarakis, and N. C. Kyrpides, "The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata," *Nucleic Acids Res.*, vol. 36, pp. D475–479, Jan 2008.

[150] D. Barrell, E. Dimmer, R. P. Huntley, D. Binns, C. O'Donovan, and R. Apweiler, "The GOA database in 2009–an integrated Gene Ontology Annotation resource," *Nucleic Acids Res.*, vol. 37, pp. 396–403, Jan 2009.

[151] D. A. Rasko, G. S. Myers, and J. Ravel, "Visualization of comparative genomic analyses by BLAST score ratio," *BMC Bioinformatics*, vol. 6, p. 2, 2005.

[152] J. R. Cole, Q. Wang, E. Cardenas, J. Fish, B. Chai, R. J. Farris, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, T. Marsh, G. M. Garrity, and J. M. Tiedje, "The Ribosomal Database Project: improved alignments and new tools for rRNA analysis," *Nucleic Acids Res.*, vol. 37, pp. D141–145, Jan 2009.

[153] J. R. Cole, B. Chai, R. J. Farris, Q. Wang, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, A. M. Bandela, E. Cardenas, G. M. Garrity, and J. M. Tiedje, "The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data," *Nucleic Acids Res.*, vol. 35, pp. D169–172, Jan 2007.

[154] M. H. Serres, S. Goswami, and M. Riley, "GenProtEC: an updated and improved analysis of functions of *Escherichia coli* K-12 proteins," *Nucleic Acids Res.*, vol. 32, pp. D300–302, Jan 2004.

[155] F. Yang, J. Yang, X. Zhang, L. Chen, Y. Jiang, Y. Yan, X. Tang, J. Wang, Z. Xiong, J. Dong, Y. Xue, Y. Zhu, X. Xu, L. Sun, S. Chen, H. Nie, J. Peng, J. Xu, Y. Wang, Z. Yuan, Y. Wen, Z. Yao, Y. Shen, B. Qiang, Y. Hou, J. Yu, and Q. Jin, "Genome dynamics and diver-

sity of *Shigella* species, the etiologic agents of bacillary dysentery," *Nucleic Acids Res.*, vol. 33, pp. 6445–6458, 2005.

[156] W. Deng, V. Burland, G. Plunkett, A. Boutin, G. F. Mayhew, P. Liss, N. T. Perna, D. J. Rose, B. Mau, S. Zhou, D. C. Schwartz, J. D. Fetherston, L. E. Lindler, R. R. Brubaker, G. V. Plano, S. C. Straley, K. A. McDonough, M. L. Nilles, J. S. Matson, F. R. Blattner, and R. D. Perry, "Genome sequence of Yersinia pestis KIM," *J. Bacteriol.*, vol. 184, pp. 4601–4611, Aug 2002.

[157] B. Liu, Y. A. Knirel, L. Feng, A. V. Perepelov, S. N. Senchenkova, Q. Wang, P. R. Reeves, and L. Wang, "Structure and genetics of *Shigella* O antigens," *FEMS Microbiol. Rev.*, vol. 32, pp. 627–653, Jul 2008.

[158] R. Stenutz, A. Weintraub, and G. Widmalm, "The structures of *Escherichia coli* O-polysaccharide antigens," *FEMS Microbiol. Rev.*, vol. 30, pp. 382–403, May 2006.

[159] G. M. Pupo, R. Lan, and P. R. Reeves, "Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 97, pp. 10567–10572, Sep 2000.

[160] D. F. Macpherson, P. A. Manning, and R. Morona, "Characterization of the dTDP-rhamnose biosynthetic genes encoded in the rfb locus of *Shigella flexneri*," *Mol. Microbiol.*, vol. 11, pp. 281–292, Jan 1994.

[161] G. Stevenson, B. Neal, D. Liu, M. Hobbs, N. H. Packer, M. Batley, J. W. Redmond, L. Lindquist, and P. Reeves, "Structure of the O antigen of *Escherichia coli* K-12 and the sequence of its rfb gene cluster," *J. Bacteriol.*, vol. 176, pp. 4144–4156, Jul 1994.

[162] E. Fischer, N. Zamboni, and U. Sauer, "High-throughput metabolic flux analysis based on gas chromatography-mass spectrometry de-

rived 13C constraints," *Anal. Biochem.*, vol. 325, pp. 308–316, Feb 2004.

[163] A. Varma and B. O. Palsson, "Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110," *Appl Environ Microbiol*, vol. 60, pp. 3724–3731, Oct 1994.

[164] A. E. Kister and J. C. Phillips, "A stringent test for hydrophobicity scales: two proteins with 88 % sequence identity but different structure and function," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 105, pp. 9233–9237, Jul 2008.

[165] J. L. Reed, T. D. Vo, C. H. Schilling, and B. O. Palsson, "An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR)," *Genome Biol*, vol. 4, no. 9, p. R54, 2003.

[166] Y. E. Luo, D. D. Fan, L. A. Shang, H. J. Shi, X. X. Ma, Y. Mi, and G. F. Zhao, "Analysis of metabolic flux in Escherichia coli expressing human-like collagen in fed-batch culture," *Biotechnol. Lett.*, vol. 30, pp. 637–643, Apr 2008.

[167] J. Shiloach, J. Kaufman, A. S. Guillard, and R. Fass, "Effect of glucose supply strategy on acetate accumulation, growth, and recombinant protein production by Escherichia coli BL21 (lambdaDE3) and Escherichia coli JM109," *Biotechnol. Bioeng.*, vol. 49, pp. 421–428, Feb 1996.

[168] J. N. Phue, S. B. Noronha, R. Hattacharyya, A. J. Wolfe, and J. Shiloach, "Glucose metabolism at high density growth of E. coli B and E. coli K: differences in metabolic pathways are responsible for efficient glucose utilization in E. coli B as determined by microarrays and Northern blot analyses," *Biotechnol. Bioeng.*, vol. 90, pp. 805–820, Jun 2005.

[169] Z. G. Qian, X. X. Xia, and S. Y. Lee, “Metabolic engineering of *Escherichia coli* for the production of putrescine, a four carbon diamine,” *Biotechnol. Bioeng.*, Aug 2009.

[170] Z. L. Fowler, W. W. Gikandi, and M. A. Koffas, “Increasing malonyl-CoA biosynthesis by tuning the *Escherichia coli* metabolic network and its application to flavanone production,” *Appl. Environ. Microbiol.*, Jul 2009.

[171] V. E. Balderas-Hernndez, A. Sabido-Ramos, P. Silva, N. Cabrera-Valladares, G. Hernndez-Chvez, J. L. Bez-Viveros, A. Martnez, F. Bolvar, and G. Gosset, “Metabolic engineering for improving anthranilate synthesis from glucose in *Escherichia coli*,” *Microb. Cell Fact.*, vol. 8, p. 19, 2009.

[172] K. Watanabe, H. Oguri, and H. Oikawa, “Diversification of echinomycin molecular structure by way of chemoenzymatic synthesis and heterologous expression of the engineered echinomycin biosynthetic pathway,” *Curr Opin Chem Biol*, vol. 13, pp. 189–196, Apr 2009.

[173] V. J. Martin, D. J. Pitera, S. T. Withers, J. D. Newman, and J. D. Keasling, “Engineering a mevalonate pathway in Escherichia coli for production of terpenoids,” *Nat. Biotechnol.*, vol. 21, pp. 796–802, Jul 2003.

[174] S. Bennett, “Solexa Ltd,” *Pharmacogenomics*, vol. 5, pp. 433–438, Jun 2004.

Publication List

Publications

W. A. Bryant, P. Krabben, F. Baganz, Y. Zhou and J. M. Ward "Multiple genome comparison of genus Escherichia and its application to the discovery of uncharacterised metabolic genes in uropathogenic *Escherichia coli*" - under submission

Oral presentations

W. A. Bryant, P. Krabben, F. Baganz, Y. Zhou and J. M. Ward "The search for uncharacterised metabolic genes in *E. coli* CFT073 using a multiple genome synteny comparison." Exploiting Genomics: *Escherichia coli* user group final meeting, University of Birmingham, 28th November 2007

Posters

W. A. Bryant, P. Krabben, F. Baganz, Y. Zhou and J. M. Ward "Stoichiometric models and their application to growth in defined conditions." Inaugural BESG Young Researchers Meeting (IChemE), University College London, 4th January 2008

W. A. Bryant, P. Krabben, F. Baganz, Y. Zhou and J. M. Ward "The genome as insight into the behaviour of cells in defined conditions." Genomes to Systems 2008, Manchester Central Convention Complex, 17th - 19th March 2008

Acknowledgements

This work was supported by the Biotechnology and Biological Sciences Research Council (UK). The sequence data for strains *Escherichia coli* O42, E2348 and *Shigella sonnei* 53G were produced by the Pathogen Sequencing Group at the The Wellcome Trust Sanger Institute, were used with permission, and can be obtained from `ftp://ftp.sanger.ac.uk/pub/pathogens/Escherichia_Shigella`. *Klebsiella oxytoca*'s L-sorbose operon sequence was produced by the Genome Sequencing Center at Washington University School of Medicine in St. Louis and can be obtained from `ftp://genome.wustl.edu/pub/organism/Microbes/Enteric_Bacteria/Klebsiella_oxytoca/assembly/Klebsiella_oxytoca-1.0/output/`