A Systems Biology approach for the characterization of metabolic bottlenecks in recombinant protein production processes

Sónia Madalena Azevedo Carneiro

UMinho|2010

Universidade do Minho
Escola de Engenharia

Sónia Madalena Azevedo Carneiro

A Systems Biology approach for the characterization of metabolic bottlenecks in recombinant protein production processes

Setembro de 2010

Universidade do Minho
Escola de Engenharia

Sónia Madalena Azevedo Carneiro

A Systems Biology approach for the characterization of metabolic bottlenecks in recombinant protein production processes

Tese de Doutoramento
Doutoramento em Engenharia Química e Biológica

Trabalho efectuado sob a orientação do
Doutor Eugénio Manuel de Faria Campos Ferreira
e da
Doutora Isabel Cristina de Almeida Pereira da Rocha

Setembro de 2010

Autor

Sónia Madalena Azevedo Carneiro

Email: soniacarneiro@deb.uminho.pt

Telefone: +351 253 604 400

BI: 11563890

Título da tese

A Systems Biology approach for the characterization of metabolic bottlenecks in recombinant protein production processes

Orientadores

Doutor Eugénio Manuel de Faria Campos Ferreira

Doutora Isabel Cristina de Almeida Pereira da Rocha

Ano de conclusão 2010

Doutoramento em Engenharia Química e Biológica

Universidade do Minho, Setembro de 2010

AGRADECIMENTOS/ACKNOWLEDGMENTS

“Não há nada como o sonho para criar o futuro” (Victor Hugo)

Embora as palavras sejam exíguas para expressar toda a minha gratidão, reservo este espaço para agradecer a todos aqueles que de alguma forma contribuíram para a realização deste trabalho.

Não poderia deixar de começar por agradecer aos meus orientadores, Doutor Eugénio Ferreira e Doutora Isabel Rocha, pela confiança, disponibilidade, orientação e estímulo científico, mas acima de tudo pela possibilidade que me foi dada para “criar o meu futuro” que outrora foi um sonho.

Ao Doutor Silas Vilas-Bôas, por me ter recebido no seu laboratório e pela partilha de conhecimentos que se revelaram valiosos no decorrer dos últimos tempos.

À instituição de acolhimento, Centro de Engenharia Biológica da Universidade do Minho pela disponibilidade do espaço e equipamentos indispensáveis no decorrer do meu trabalho científico e à Fundação para a Ciência e Tecnologia pela atribuição da bolsa de doutoramento (SFRH/BD/22863/2005).

Aos meus colegas de laboratório pela partilha de ideias, conversas e algumas noites a “fermentar” bactérias! Agradeço em especial ao Rafael Costa pela ajuda prestada no laboratório e ainda à Doutora Ana Cristina Veloso, que muito me ensinou nos primeiros tempos.

Aos colegas do grupo de investigação, pela partilha de ideais e cooperação em alguns trabalhos desenvolvidos durante estes últimos anos. Em especial à Doutora Anália Lourenço, com quem aprendi imenso sobre o mundo da mineração de texto e bases de dados, e também ao Rafael Carreira e José Pedro Pinto pela cooperação em alguns dos trabalhos.

Aos colegas e amigos que encontrei nas antípodas, e com os quais também partilhei algumas noites a “fermentar”! I especially want to thank Xavier and Raphael, truthful friends, for the great talks while tasting great New Zealand’s wine!

Aos amigos, que sempre me apoiaram.

À minha família por toda a luta e fé! Sem o vosso apoio não teria sido possível!

Ao Jorge…não há palavras que definam a importância do seu apoio, dedicação e carinho!

ABSTRACT

A Systems Biology approach for the characterization of metabolic bottlenecks in recombinant protein production processes

The main purpose of this thesis is to investigate the influence of recombinant protein production in the reorganization of the metabolic activities and the resulting stress-induced responses in the bacterium *Escherichia coli*. More specifically, the focus is on the RelA-mediated stringent response, a stress response that is triggered by the sudden lack of intracellular amino acids and that has been associated with the metabolic burden imposed by recombinant processes.

To identify the main metabolic bottlenecks in recombinant biosynthetic processes, which include maintenance of recombinant DNA and expression of heterologous genes, a systematic modelling approach is proposed, capable of predicting the amino acid shortages caused by recombinant processes and the consequent activation of the RelA-dependent guanosine pentaphosphate (ppGpp) synthesis.

The view of ppGpp as a primarily regulator of gene transcription has been expanded and it is now clear that the response controlled by ppGpp is crucial for cell survival during the adaptation to stressful conditions. Major advances have been achieved to understand this regulatory system governing gene expression in response to environmental growth perturbations, but so far mainly transcriptome and proteome analyses that have been applied to elucidate the stringent control mediated by ppGpp. Metabolomics analysis can provide substantial information on the impact of this stress response at the biochemical level, in particular during recombinant bioprocesses. Therefore, two metabolomics-based approaches were applied: metabolic profiling to evaluate the intracellular metabolic profiles and metabolic footprinting to estimate the profiles of extracellular metabolites.

In these metabolomics studies two *E. coli* strains (*E. coli* W3110 and the isogenic Δ*rel*A mutant) were used to investigate the influence of recombinant processes on the host cells' metabolism, as well as the main metabolic activities affected by the RelA activity under different growth conditions. The mutant strain presented a “relaxed” phenotype that characterized this bacterial system by an acute delay in most metabolic adaptations to transient growth conditions. Most importantly, it was shown that these mutant cells lack metabolic adjustments that are often observed after metabolic burden phenomena. Nevertheless, this cellular system presented major advantages in terms of biomass yield and productivity, which imply a remarkable improvement in recombinant bioprocesses. Thus, alleviating stress responses can be beneficial if they impair the desired quality and quantity of the recombinant product. However, it must be pointed out that this may be an alternative as long as recombinant bioprocesses are designed to achieve a finer balance between strain improvement strategies and culturing conditions.

RESUMO

Caracterização das limitações metabólicas durante a produção de proteína recombinante usando abordagens da Biologia de Sistemas

O trabalho realizado no âmbito desta tese teve como principal finalidade a avaliação das alterações metabólicas relacionadas com a produção de proteínas recombinantes em células bacterianas de *Escherichia coli* e a consequente activação de respostas de *stress*. Foi evidenciada a resposta restringente promovida pela actividade da enzima RelA, dado ser uma das principais respostas de *stress* induzidas pelo decréscimo da quantidade de aminoácidos disponíveis no meio intracelular como consequência da expressão de proteínas recombinantes. As diferenças na composição em aminoácidos entre as proteínas da biomassa e recombinantes, têm sido apontadas como principais causas para o desequilíbrio metabólico que conduz à exaustão de alguns metabolitos, nomeadamente de aminoácidos.

De modo a explorar estes fenómenos e avaliar o impacto dos processos recombinantes no metabolismo das células hospedeiras, foi proposto um modelo matemático capaz de identificar pontos de estrangulamento na rede metabólica. Estes locais correspondem a vias metabólicas que apresentam limitações na capacidade catalítica e que serão essenciais para compensar o consumo desproporcionado de aminoácidos levado a cabo pela síntese de proteínas recombinantes. Associado a este fenómeno foi considerada a descrição da síntese de nucleótidos guanosina pentafosfato (ppGpp) induzida pela escassez de aminoácidos no meio intracelular.

O reconhecimento deste nucleótido como um regulador fundamental na transcrição da informação genética tem sido amplamente descrito e tornou-se evidente que as respostas celulares controladas pelo ppGpp são determinantes para a sobrevivência e adaptação dos organismos a condições adversas. Neste sentido, vários estudos foram elaborados para elucidar o papel do ppGpp no controlo destas respostas de *stress* e nas alterações fisiológicas que advêm destes processos, nomeadamente ao nível do metabolismo. A análise do metaboloma, em comparação com o transcriptoma ou o proteoma, é capaz de capturar de forma mais directa a relação entre as actividades metabólicas e a fisiologia dos organismos, designadamente em sistema recombinantes.

Neste trabalho foram elaborados alguns estudos em que se aplicaram duas abordagens de análise metabolómica distintas: *profiling* metabólico, que se refere à análise do perfil de metabolitos intracelulares; e *footprinting* metabólico, que se refere à análise do perfil de metabolitos extracelulares. Nestes estudos foram usadas duas estirpes de *E. coli* (W3110 e a estirpe isogénica com mutação no gene *rel*A) clonadas com um vector de expressão pTRC-His-AcGP1 que codifica a proteína verde fluorescente AcGFP1, derivada da proteína AcGFP da *Aequorea coerulescens*. Foram avaliadas as principais alterações metabólicas provocadas pela indução da produção de proteína recombinante e pela actividade catalítica da enzima RelA em diversas condições de crescimento. Comparando os perfis metabólicos das duas estirpes, foram estimadas várias diferenças significativas que se podem revelar críticas durante processos recombinantes. A estirpe mutante revelou um comportamento típico de um fenótipo "relaxado", que é caracterizado por um retardamento significativo na adaptação do metabolismo a alterações nas condições de crescimento. Não obstante, a estirpe mutante exibiu melhores resultados em termos de rendimento em biomassa e produtividade, o que representa uma vantagem notável para a aplicação destes sistemas bacterianos recombinantes ao nível industrial. Em resumo, a restrição de respostas de stress pode trazer benefícios se a qualidade e quantidade do produto estiverem em causa, mas deve salientar-se que não é uma solução absoluta, sendo que as condições de processamento devem ser também levadas em consideração na implementação destes bioprocessos.

LIST OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

OUTLINE OF THE THESIS

The key contributions of this thesis are to the elucidation of the impact of the stringent response in the *E. coli* metabolism during recombinant bioprocesses and, in broader terms, to the importance of metabolomics-driven approaches to understand the metabolic behaviour of these recombinant cells.

To assist the study of the stringent response phenomenon in *E. coli* cells, a literature mining exercise was first developed to compile, process, and analyse information that has been reported in the last four decades. Fundamental details on the mechanisms underlying the activation of this stress response and the pleiotropic effects of the ppGpp on cellular processes are given in Chapter 2.

In Chapter 3, a modelling approach is proposed to describe the effects of recombinant protein production in the host cells' metabolism and the consequent induction of the activity of RelA, a ribosome-bounded enzyme that synthesizes the key regulator of the stringent response, ppGpp. This approach combines a genome-scale model for the *E. coli* metabolism and kinetic descriptions for biomass, recombinant protein and ppGpp formation, as well as plasmid maintenance.

To investigate how the inactivation of the *rel*A gene is associated with changes in *E. coli* metabolism, metabolomics techniques were used In Chapter 4, *E. coli* W3110 and the isogenic mutant Δ*rel*A strains were grown under steady-state conditions at different dilution rates. The intracellular metabolite levels of amino and non-amino organic acids were measured by GC/MS and metabolite profiles were evaluated to discriminate between metabolic states of *E. coli* cells. In Chapter 5, the Δ*rel*A mutant strain, cloned with the vector pTRC-His-AcGFP1, was used to assess the impact of recombinant processes on the *E. coli* metabolism, which includes plasmid maintenance and protein formation. Because the stringent factor RelA is expected to impair the production of recombinant proteins, this work aimed at understanding if the deletion of the *rel*A gene might be a potential strategy to enhance the performance of recombinant cells.

An alternative metabolomics approach was applied to inspect the metabolic behaviour of *E. coli* cells during the production of recombinant proteins. Chapter 6 presents the implementation of a metabolic footprinting approach to measure the extracellular metabolite levels during recombinant *E. coli* fed-batch cultures, the common industrial mode of operation of these processes. The physiological and metabolic changes associated to the induction of protein expression and nutritional shifts during the fed-batch process were inspected.

In Chapter 7, conclusions and final remarks are devised.

Objectives of the thesis

Systems engineering strategies to enhance the productivity of recombinant proteins are still reliant on the *E. coli* cellular capabilities to cope with stressful conditions elicited by recombinant processes. To gain fundamental insight into the molecular mechanisms governing these events, particularly the stringent response, and to use this information to identify strategies for minimizing the impact of this phenomenon, the following objectives were considered in this thesis:

- To present a comprehensive analysis of the literature on the *E. coli* stringent response, where key players and molecular mechanisms involved in this response are detailed;
- To uncover new molecular players, which functional roles have not been directly associated with the *E. coli* stringent response in previous studies, but are relevant entities or participate in cellular processes closely related with this stress response;
- To systematically analyse and characterize the stimulus of the stringent response during recombinant processes, as a consequence of the amino acids deprivation caused by the additional drainage of biosynthetic precursors for the production of recombinant proteins;
- To identify key metabolic bottlenecks in the *E. coli* metabolism during recombinant processes, which indicate network points at which metabolic fluxes are to a certain extent restricted due to the unbalanced withdrawn of biosynthetic precursors;
- To assess the ability of metabolomics approaches, in particular metabolic profiling and metabolic footprinting, to discriminate between cells in different metabolic states, as well as with different genetic backgrounds (e.g. *E. coli* W3110 and the isogenic mutant Δ*rel*A).
- To characterize the RelA-mediated stringent response by comparing the metabolic behaviour of an *E. coli* W3110 strain with its isogenic mutant (Δ*rel*A);
- To evaluate the impact of recombinant processes, including plasmid maintenance and expression of a recombinant protein (AcGFP1, a green fluorescent protein derived from

Aequorea coerulescens), in the metabolism of *E. coli* cells, and evaluate the effect of the *rel*A mutation in this response;

- To verify if *E. coli* Δ*rel*A mutant strains can enhance the productivity of recombinant proteins, since the *rel*A gene mutation would limit the ppGpp-induced stringent response during recombinant processes;
- To determine the metabolic differences when recombinant bioprocesses are performed at different growth rates.

CHAPTER 1

GENERAL INTRODUCTION

"...a new understanding of life emerged at the forefront of science."

Fritjof Capra in *Complexity and Life*,

Theory Culture Society 2005; 22; 33

1.1 ABSTRACT

This chapter aims to introduce the main aspects involving the application of system-level approaches in the understanding of cellular processes during recombinant bioprocesses. The successful use of genetically engineered bacteria, like *Escherichia coli*, to overexpress recombinant proteins and the main problems associated with these bioprocesses, in particular the metabolic burden and physiological stress responses, with major focus on the stringent response, are reviewed. To fully characterize the impact of recombinant processes in the metabolism of *E. coli* and the consequent induction of the stringent response, modelling approaches are suggested. Finally, metabolomics approaches are introduced as relevant tools for studying recombinant *E. coli* bioprocesses. The main procedures to perform metabolome analysis are viewed in detail, as well as the last developments in bioinformatics tools for storing, analysing and interpreting metabolome data. Several examples are provided to illustrate the strengths and still existing limitations of this analytical approach. Topics on the main questions studied throughout this thesis are referred in the end of this chapter.

1.2 *E. COLI*: A MICROBIAL SYSTEM FOR RECOMBINANT PROTEIN EXPRESSION

In the last decades, microbial cells have been exploited for the production of a variety of products. Recombinant DNA technologies, developed in the late 70's, offered a very powerful tool for the economical and large-scale production of recombinant proteins. These products are diverse and find their main applications in industries such as pharmaceutical, food, health care and environment. For instance, the human insulin was the first recombinant protein approved by the Food and Drug Administration (FDA) to enter in the market and is produced in the bacterium *E. coli*. Since then, many other recombinant proteins with industrial interest have been produced in several microbial systems such as bacteria, being *E. coli* the most common system, or yeasts, such as *Saccharomyces cerevisiae* and *Pichia pastoris*. In Figure 1.1 some of the therapeutic products that have been produced in recombinant *E. coli* systems within the last years are summarized.

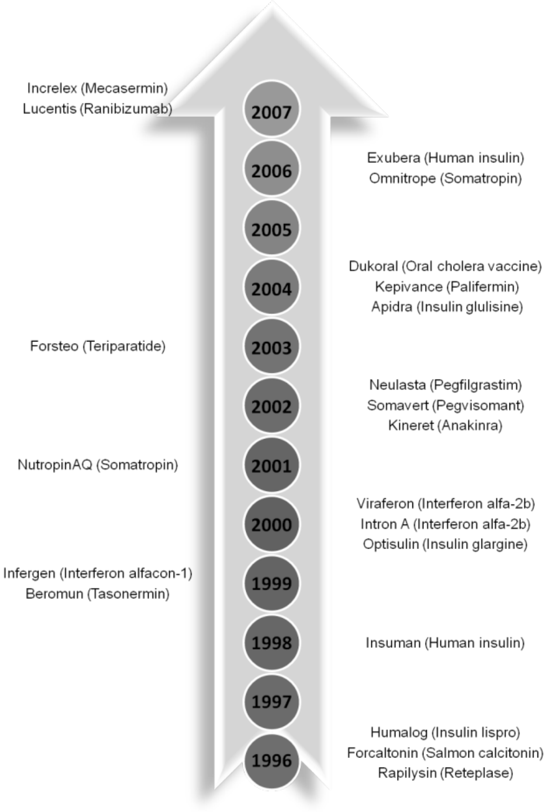


Figure 1.1. Biopharmaceutical products approved by the European Union that are synthesised via recombinant *E. coli* systems (Ferrer-Miralles N *et al.*, 2009).

E. coli is by far the first choice for the production of recombinant proteins, and has been quite important in the development of certain molecular procedures, such as cloning, genetic modification and also for the small-scale production for research purposes. Besides its ability to grow rapidly and achieve high cellular densities on inexpensive substrates, it is one of the best characterized biological systems. However, several obstacles to the production of high quality proteins limit its application as a microbial factory. Post-translational modifications are often limited in *E. coli* systems and only recombinant proteins that are naturally non-glycosylated or are also active without glycosylation can be expressed with this microbial system (Demain AL and Vaishnav P, 2009; Ferrer-Miralles N *et al.*, 2009; Schmidt FR, 2004). Moreover, the frequency of codons occurring in eukaryotic genes is different from that appearing in *E. coli* genes, which basically determines the low abundance of specific tRNAs, often leading to the premature termination of protein synthesis or to amino acid *mis*-incorporation, reducing the efficiency of protein expression (Gustafsson C *et al.*, 2004; Kane JF, 1995).

Yet, recent progress in the fundamental understanding of cellular processes in *E. coli*, together with the availability of improved expression vector systems, is making this bacterium more adequate for the expression of complex eukaryotic proteins. Several strategies to enhance recombinant protein production have become available: (i) reduction of plasmid copy number (Jones KL *et al.*, 2000); (ii) use of different promoters to regulate expression (Peti W and Page R, 2007); (iii) secretion of proteins into the periplasmic space or into the medium (Mergulhao FJ *et al.*, 2005); (iv) changing the growth medium (Sahdev S *et al.*, 2008); (v) lowering of temperature (Sahdev S *et al.*, 2008); (vi) chromosomal insertion of the foreign genes (Srinivasan S *et al.*, 2003); and (vii) coexpression/knockout of certain key genes (Kim SY *et al.*, 2009; Wong MS *et al.*, 2008). This last approach is probably one of the most effective ones and is related with the metabolic engineering field. Stephanopoulos and co-authors (Stephanopoulos G *et al.*, 1998) defined it as "*the directed improvement of product formation or cellular properties through the modification of specific biochemical reactions or introduction of new ones with the use of recombinant DNA technology*". The metabolic engineering field proposes the directed modification of the genetic background of organisms to implement desirable metabolic capabilities in living cells. For example, Flores and co-workers (Flores S *et al.*, 2004) developed a strategy to overcome the reduction of growth rate due to foreign protein expression, by modifying the pentose phosphate (PP) pathway. By increasing the carbon flux through the oxidative branch of the PP pathway, it was possible to partially recover growth capacities of the recombinant *E. coli* cells.

These approaches have resulted in some high-quality results by identifying some metabolic activities that could be manipulated to improve cell physiology during recombinant protein production, though other cellular phenomena capable to hinder recombinant bioprocesses are still unresolved. For example, the metabolic burden is associated with cellular processes related to plasmid DNA replication, plasmid-encoded mRNA transcription and the corresponding recombinant protein synthesis. These processes require the additional drain of biosynthetic precursors, energy and other cellular resources that are shared with the analogous host metabolic processes. The competition for a limited pool of cellular resources, like deoxyribonucleotides (dATP, dGTP, TTP, dCTP), ribonucleotides (ATP, GTP, UTP, CTP), amino acids and high-energy molecules, like NADH and NADPH, provokes serious perturbations in the cellular metabolism (Glick BR, 1995). Typically, in recombinant cultures this promotes the reduction of the cellular growth rate and final biomass yields. As metabolic precursors and high-energy molecules are being consumed in the recombinant process, the host metabolic processes engaged in the cellular growth are severely burdened, unbalancing the host metabolism.

Moreover, as many amino acid precursors are drained from the TCA cycle, the fluxes over this pathway are lowered and the carbon flux through glycolysis exceeds the capacity of the TCA cycle to assimilate the surplus of acetyl-CoA, which is directed to the production of acetate (Majewski RA and Domach MM, 1990). The accumulation of toxic levels of acetate is undesirable, mostly because it represents a diversion of carbon that might otherwise have generated biomass or the recombinant product and it ends by retarding cellular growth (Luli GW and Strohl WR, 1990; Ponce E, 1999; Suarez DC and Kilikian BV, 2000; Turner C *et al.*, 1994) and inhibiting protein formation (Shiloach J *et al.*, 1996; Suarez DC and Kilikian BV, 2000; Turner C *et al.*, 1994). Also, since most cellular processes are closely regulated by growth rate conditions, a decrease in the biomass formation indicates reduced amounts of components for the protein producing system, including ribosomal proteins or translation elongation factors, and of most catabolic enzyme levels. Thereby, the translational activity and the supply for most anabolic precursors become limiting factors as the recombinant protein synthesis rate increases.

Additionally, it has been shown that during the expression of recombinant proteins in *E. coli*, a variety of cellular responses can be elicited (Duerrschmid K *et al.*, 2008; Seo JH *et al.*, 2003): heat-shock-like responses, due to the accumulation of misfolded proteins (Allen SP *et al.*, 1992; Thomas JG and Baneyx F, 1996); SOS responses if cells are exposed to agents that cause damage to DNA or interfere with DNA replication (Gill RT *et al.*, 2000); and starvation responses caused by the excessive drainage of metabolic precursors and energy (Bonomo J and Gill RT, 2005; Sanden AM *et al.*, 2003). It has been reported (Dedhia N *et al.*, 1997; Sanden AM *et al.*, 2003) that the expression of recombinant proteins may lead to an increase in intracellular concentration of guanosine tetraphosphate or ppGpp, which was first identified as a key regulator involved in the cellular response to amino acid starvation, the so-called stringent response (Cochran JW and Byrne RW, 1974; Haseltin WA and Block R, 1973; Magnusson LU *et al.*, 2005; Stephens JC *et al.*, 1975) (detailed in the next subsection and Chapter 2). Lately, this molecule was also associated with other nutrient starvation responses (Lin HY *et al.*, 2004; Traxler MF *et al.*, 2006). For example, the *E. coli* response to glucose starvation was linked to the induction of both the stringent and the general stress responses (Schweder T *et al.*, 2002). The concentration of the corresponding regulators, i.e. ppGpp and the alternative RpoS sigma factor (also called σ^S) were shown to be tightly regulated by the cell in situations when a nutrient shift occurs. Indeed, many stress responsive genes that are regulated by the σ^S have been observed to be expressed during recombinant protein production (Duerrschmid K *et al.*, 2008), namely *dna*K and *ibp*A that encode proteins involved in the heat shock like response (Han MJ *et al.*, 2004; Jurgen B *et al.*, 2000). Such typical stress response is commonly elicited to prevent the aggregation of unfolded proteins into inclusion bodies and has been explored to improve the productivity of specific recombinant systems (Endo S *et al.*, 2006; Kohda J *et al.*, 2002; Kwon MJ *et al.*, 2002; Yokoyama K *et al.*, 1998).

In general, different stresses might be elicited simultaneously, working together as a complex regulatory system with a multitude of molecular components to ensure a coordinated and an effective answer. For example, it has been shown (Gill RT *et al.*, 2000), by monitoring transcriptome and proteome profiles, that the production of recombinant proteins induces the expression of genes belonging to the heat-shock (e.g. *dna*K and *ibp*A), SOS (e.g. *rec*A), and starvation response regulons (e.g. *rpo*S), albeit to different levels and with different time profiles.

Thus, the interplay of several stress proteins may be critical to protect bacterial cells when exposed to environmental insults. As suggested in Chapter 2, some of the biological entities involved in the stringent response may as well participate in other stress responses. Proteins that are involved in responses to starvation, DNA damage, osmotic, oxidative or SOS stresses were highlighted in the bibliome analysis of the *E. coli* stringent response, which suggests that ppGpp might be involved in many cellular responses, other than the stringent response. Indeed, some links between the stringent response and others stress responses were evidenced. For example, the response to DNA damage stimulus was assigned by the RecA, RecG and Mfd proteins that intervene in the early dissociation of the elongation complex stalled by ppGpp (Trautinger BW *et al.*, 2005) and the RecA regulator and the UvrABC nucleotide excision repair complex have been implicated in the DNA repair process and SOS response (Bichara M *et al.*, 2007).

1.2.1 STRINGENT RESPONSE

From such a variety of stress responses, stringent response has been perhaps the less studied in recombinant systems. Currently, the regulatory mechanisms of the ppGpp activity during recombinant processes and the impact of this stress response in the metabolism are not entirely clear. It is acknowledged that under conditions of stress, cells accumulate high levels of ppGpp initiating a global change in the cellular physiology. However, most studies (Durfee T *et al.*, 2008; Haddadin FT *et al.*, 2009; Haddadin FT and Harcum SW, 2005) have just uncovered the main changes observed at the proteome or transcriptome levels. To investigate this cellular response at the metabolic level (i.e. the closest indicator of the physiological state of cells), the basis of the ppGpp mechanisms are first reviewed.

The discovery that this unusual guanosine nucleotide is accumulated in response to starvation was followed by extensive studies on the downstream regulatory circuits of the stringent response, which suggested that the accumulation of ppGpp is an important link between nutritional stress and bacterial adaptation. It was first proposed by (Haseltin WA and Block R, 1973) that the ratio of aminoacylated tRNA to free tRNA is one of the critical parameters that regulate the synthesis of ppGpp. When free tRNA is encountered at the A-site of the 50S ribosome, protein synthesis is delayed, resulting in an idling reaction in which ribosome-bound

RelA is activated to synthesize ppGpp (initially pppGpp is produced and is then converted to ppGpp). Thus, an increase in the population of free tRNA during starvation leads to an accumulation of ppGpp (Chatterji D and Ojha AK, 2001) (Figure 1.2). Since the induction of recombinant proteins can cause the exhaustion of the intracellular amino acid pools resulting in a change in the aminoacylated-tRNA to uncharged tRNA ratio, the accumulation of ppGpp is presumably associated with the induction of the classical stringent response.

Less is known about the mechanism behind SpoT-dependent production of ppGpp and how SpoT senses starvation conditions. Nevertheless, it appears that SpoT is primarily responsible for the accumulation of ppGpp in response to most stresses and nutrient limitations apart from amino acid starvation (Murray KD and Bremer H, 1996; Vinella D *et al.*, 2005). In addition, SpoT exhibits dual functions and is also responsible for hydrolyzing ppGpp. A strain lacking both RelA and SpoT is completely unable to produce ppGpp and the response to starvation of such a strain is called the relaxed response (Xiao H *et al.*, 1991).

One of the first effects of high-level intracellular ppGpp to be discovered was a sudden decrease in the transcriptional rate of ribosomal RNA (Baracchini E and Bremer H, 1988). Direct negative effects of ppGpp on promoters have been detected *in vitro* (Barker MM *et al.*, 2001b; Jores L and Wagner R, 2003; Kajitani M and Ishihama A, 1984; Raghavan A and Chatterji D, 1998) and several mechanisms for direct negative regulation by ppGpp have been suggested (Barker MM *et al.*, 2001b; Barker MM *et al.*, 2001a; Gralla JD, 2005; Jores L and Wagner R, 2003; Roberts JW, 2009; Srivatsan A and Wang JD, 2008). One of the effects of ppGpp appears to be a destabilization of the RNAP–promoter open complex. The rRNA promoters form intrinsically unstable open complexes with RNAP and are therefore thought to be specifically sensitive to further destabilization (Magnusson LU *et al.*, 2005; Oshima T *et al.*, 2002; Paul BJ *et al.*, 2004b). It has recently been shown that the negative effects of ppGpp on transcription in vitro are amplified by the presence of the protein DksA. DksA binds to RNAP by protruding into the secondary channel of RNAP (Perederina A *et al.*, 2004) and decreases open complex stability, which accentuates the negative effects of ppGpp on rRNA promoters. DksA has also been suggested to contribute to the positive effects of ppGpp (Paul BJ *et al.*, 2004a).

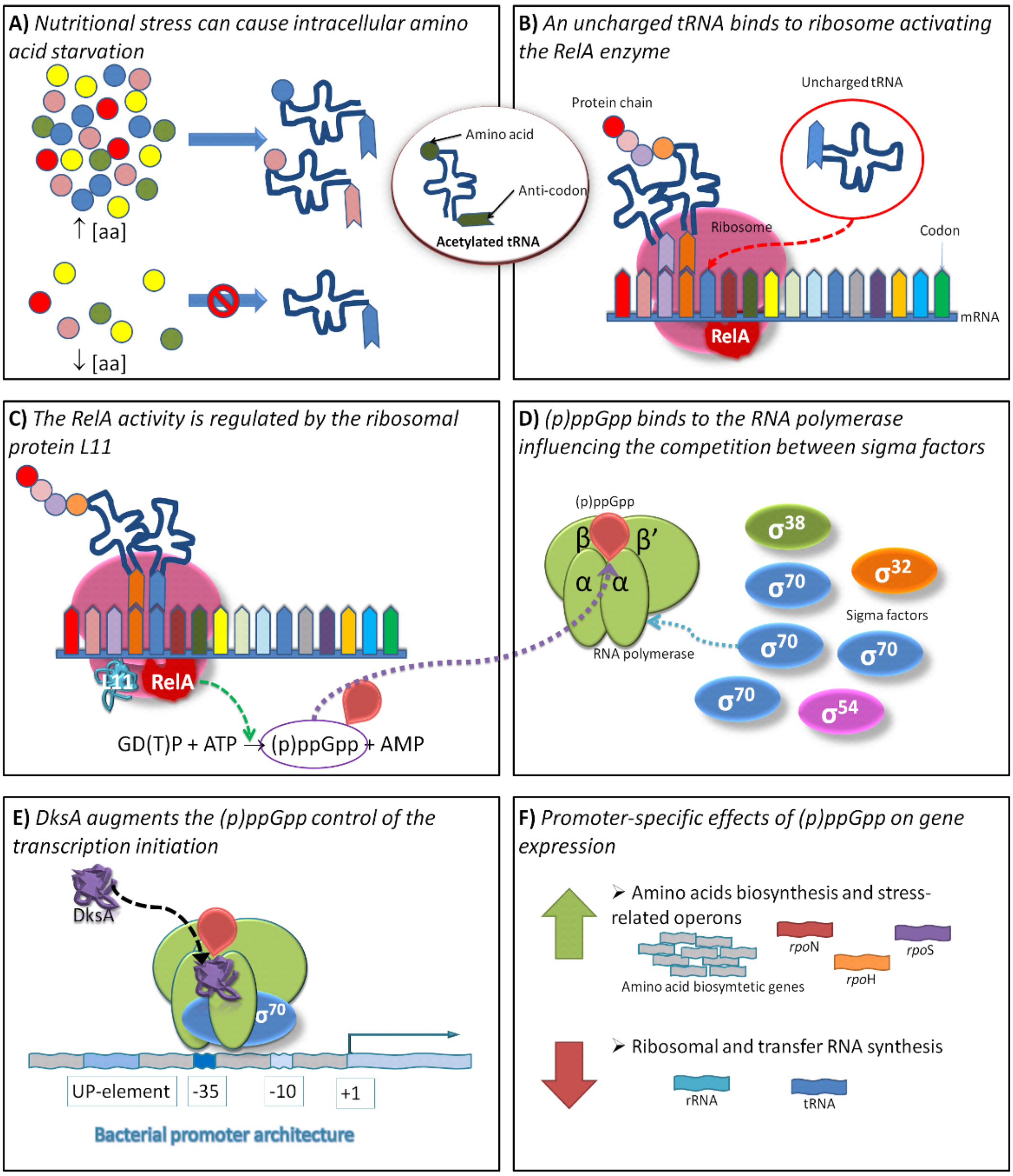

A) Nutritional stress can cause intracellular amino acid starvation
↑ [aa]
↓ [aa]
Amino acid
Anti-codon
Acetylated tRNA
B) An uncharged tRNA binds to ribosome activating the RelA enzyme
Protein chain
Uncharged tRNA
Ribosome
Codon
mRNA
RelA
C) The RelA activity is regulated by the ribosomal protein L11
L11
RelA
GD(T)P + ATP → (p)ppGpp + AMP
D) (p)ppGpp binds to the RNA polymerase influencing the competition between sigma factors
(p)ppGpp
β
β'
α
α
RNA polymerase
σ38
σ70
σ32
Sigma factors
σ70
σ70
σ70
σ54
E) DksA augments the (p)ppGpp control of the transcription initiation
DksA
σ70
UP-element
-35
-10
+1
Bacterial promoter architecture
F) Promoter-specific effects of (p)ppGpp on gene expression
Amino acids biosynthesis and stress-related operons
rpoN
rpoS
rpoH
Amino acid biosymtetic genes
Ribosomal and transfer RNA synthesis
rRNA
tRNA

Figure 1.2. The (p)ppGpp-mediated stringent response (Figure 2.1 in Chapter 2).

(A) Low amino-acid concentrations lead to decreased charging of the corresponding tRNAs. (B) The translational machinery depends on the translocation along the mRNA whereby a new acetylated-tRNA is positioned in the ribosome. Whenever an uncharged tRNA binds to the ribosome, the elongation of the polypeptide chain is stalled. (C) The stringent factor RelA is then activated in the presence of the ribosomal protein L11, catalyzing the synthesis of (p)ppGpp nucleotides. (D) These nucleotides bind directly to the RNA polymerase and affect the binding abilities of sigma factors to the core RNA polymerase. (E) The co-factor DksA also binds to the RNA polymerase and augments the (p)ppGpp regulation of the transcription initiation at certain σ^{70}-dependent promoters, functioning both as negative and positive regulators. (F) These regulators change the gene expression: (i) decreasing the transcription activity of genes involved in the translational activity; (ii) and increasing the transcription of stress-related operons and genes encoding for enzymes needed for the synthesis and the transport of amino acids.

Among the positively regulated promoters are many that are dependent on alternative sigma factors (e.g. σ^S and σ^{54}) (Brown L *et al.*, 2002) and also those that are dependent on the housekeeping sigma factor, σ^{70} (e.g. promoters controlling genes encoding the universal stress proteins (Kvint K *et al.*, 2003) and proteins involved in amino acid biosynthesis and uptake). In general, σ^{70}-dependent genes involved in cell proliferation and growth are negatively regulated by ppGpp, whereas the σ^{70}-dependent genes implicated in maintenance and stress defence are positively regulated by the alarmone (Nystrom T, 2004). Other models suggest that ppGpp acts through changes in the availability of RNAP, rather than decreasing the open complex stability. An increased availability of free RNAP during growth arrest has been suggested to be a consequence of RNAP falling off stable RNA promoters as a result of decreased open complex stability (Barker MM *et al.*, 2001a; Zhou YN and Jin DJ, 1998). The positively regulated promoters are then induced because they are argued to be relatively poor at recruiting RNAP and are sub-saturated during normal growth. Jensen and Pedersen (Jensen KF and Pedersen S, 1990) have argued that 'stringent' promoters, for example stable RNA promoters, require high concentrations of RNAP to be transcribed at their maximal rate and are therefore especially sensitive to a diminished availability of RNAP.

Despite the many existing hypothesis, there is a general consensus that stringent response influences a cascade of events, starting with the over-accumulation of ppGpp molecules that later mediate the regulation of basic cellular processes, like the metabolism, translational activities or stress responses. The discovery of ppGpp accumulation during bacterial recombinant processes promoted even more the interest of the industry and researchers on this stress response. Limitations imposed by this response during recombinant bioprocesses represent significant losses in the quality and quantity of recombinant products. Moreover, the discovery of ppGpp as a key regulator of various cellular processes is another significant achievement that suggests that adaption to starvation is more complex than it was initially understood. And so, four decades after its discovery, stringent response still has an enormous scope for study.

1.3 MODELLING THE STRINGENT RESPONSE IN RECOMBINANT SYSTEMS

All biological systems are characterized by a multitude of functional units that support cellular growth, reproduction and survival. The description of these systems is often complex, not only because it increases with the amount of functional units enclosed in the system, but also with the intricacy of their relationships. With the recent developments in omics technologies, combined with computational analysis, the identification of most biological molecules and the multi-level interactions among them needed to carry out cellular functions, was made possible (Figure 1.3). Most of these research studies address the system-level understanding of the organization and dynamics of cellular processes (e.g. transcriptional regulation, signal transduction pathways, etc).

With the increasing interest on the development of new strategies to improve the efficiency of recombinant bioprocesses, several modelling approaches considering this systems-level perspective have been implemented (Chou CP, 2007; Gnoth S *et al.*, 2008). Process optimization for recombinant protein production has been traditionally focused on genetic-based solutions for high-level gene expression (Baneyx F, 1999; Hannig G and Makrides S, 1998; Sorensen HP and Mortensen KK, 2005). However, with the application of these novel approaches, aspects like the host metabolism or its genetic background can now be engineered, improving considerably the performance of recombinant bioprocesses. A successful example was demonstrated by Wong and co-workers (Wong MS *et al.*, 2008) that reduced ten-fold acetate accumulation in recombinant *E. coli* culture by disabling the phosphoenolpyruvate:sugar phosphotransferase system (PEP-PTS) through the deletion of the *pts*HI operon. It is acknowledged that the production of acetate retards cell growth and inhibits protein formation, and these were significantly attenuated when using this mutated host strain and the final biomass concentration and volumetric productivity of recombinant proteins were increased.

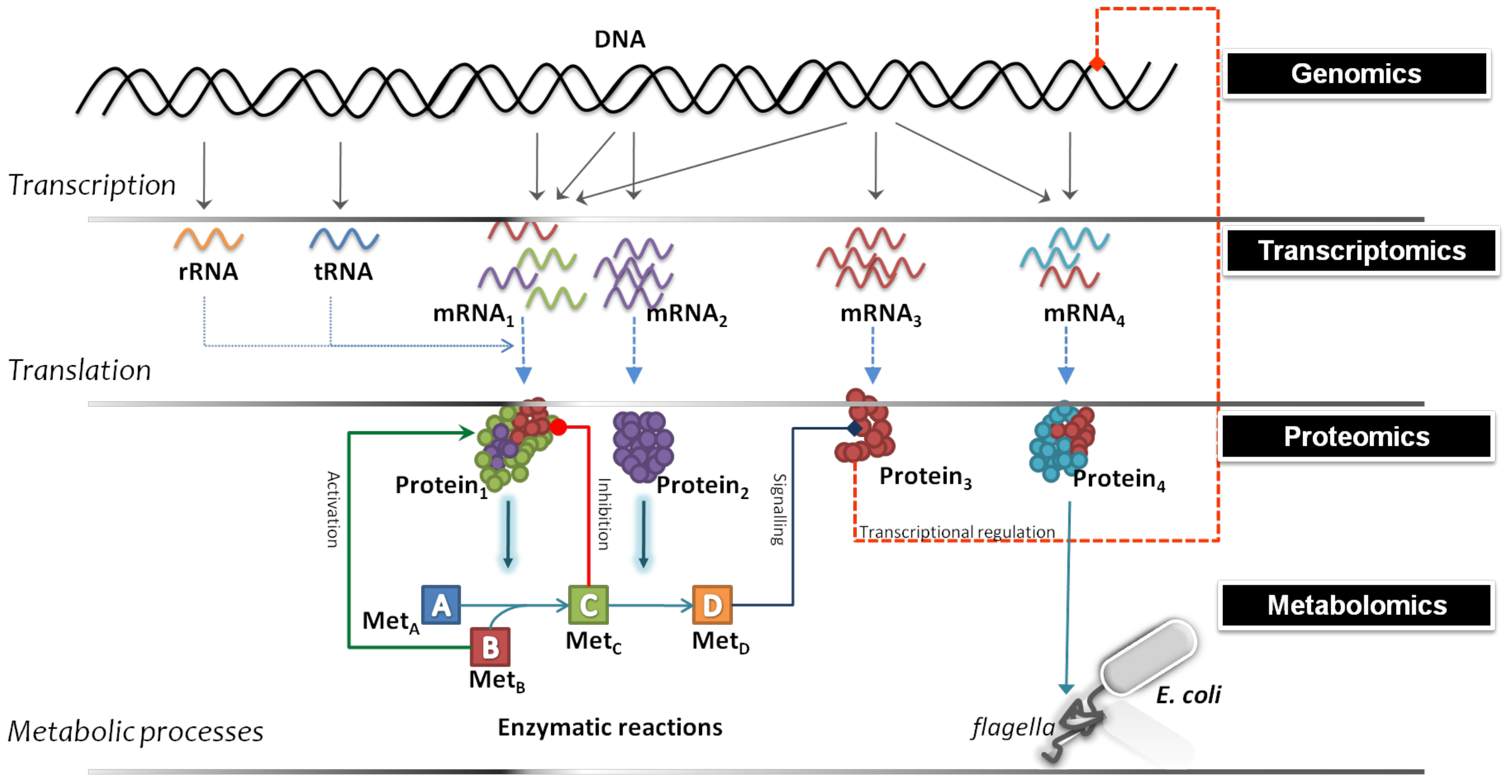


Figure 1.3. Schematic representation of the cellular organization at a systems-level perspective.

Biological systems consist in numerous functional units that assist cellular processes by engendering diverse interactions between them. Genes, proteins, and metabolites are basic components of the system that carry information for transcriptional, translational or metabolic processes. Interactions between these components might be multi-level. For example, metabolites involved in enzymatic reactions and gene-encoding enzymes represent two different levels of interaction that, ultimately control the cellular metabolism. These interactions can be even more complex when the transcription of genetic information is also controlled by the availability of reactants or the excess of end-products. Typically, there are biological molecules that act as regulators constraining the information processing (signalling and enzymatic inhibition/activation). With the advent of omics technologies, screening all these levels of information is now possible and new insights into the modelling, analysis and control of systems will be available.

Similar strategies to overcome the consequences of metabolic burden or physiological stresses, triggered during the overproduction of recombinant proteins, have been reported (Chou CP, 2007; Gnoth S *et al.*, 2008). For instance, the manipulation of stress-responsive genes (e.g. *ibp*AB, *rel*A) or metabolic factors, can improve cell physiology during recombinant bioprocesses. As an example, the co-expression of the small heat-shock proteins (IbpA and IbpB) has proven to reduce the physiological stress associated with protein misfolding (Lethanh H *et al.*, 2005). Moreover, it has been reported that by increasing the availability of NADH, a common energy carrier and cofactor involved in various biosynthesis pathways, cell physiology and recombinant protein production can be improved, which was achieved through the overexpression of the *pnc*B gene in *E. coli* (San KY *et al.*, 2002). The impact of the stringent response in recombinant protein production was also verified and it was found that ppGpp-less strains were able to produce recombinant protein 20-fold higher than the wild-type cells (Dedhia N *et al.*, 1997). Although significant improvements in the production of recombinant proteins were obtained with these strategies, rational selection of the proper gene(s) to be coexpressed, overexpressed or deleted is still challenging. Systematic approaches for identifying potential targets and develop fine-tuned strategies to overexpress the recombinant product and suppress the natural responses to physiological stresses, becomes critical to improve recombinant bioprocesses.

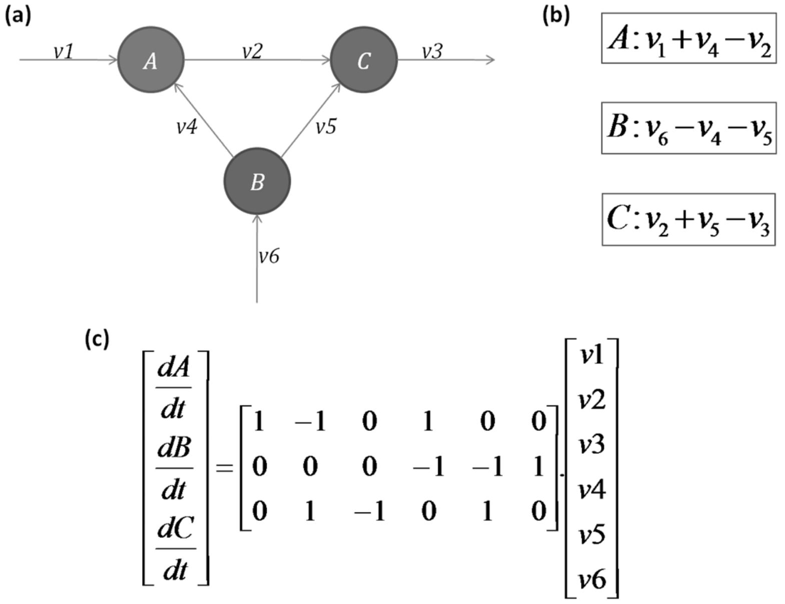
Mathematical modelling emerged as an important tool to examine central problems in the biological sciences, ranging from the organizational principles of individual cells to the dynamics of particular cellular processes. Although the mathematical representation of biological systems is perhaps one of the most complicated problems in systems biology, it is also one of the most practical ways to describe the properties of a system. Thus, to understand the whole scenario that leads to most of the limitations in recombinant bioprocesses, it would be important to explore the dynamics of the main cellular activities involved in the synthesis of recombinant protein, and in particular cellular responses to the metabolic burden and the ppGpp-induced response.

To achieve a systems-level understanding of the impact of recombinant protein production in the host metabolism, it is crucial to mathematically represent the entire metabolic network of the organism. There are several modelling methodologies capable of describing the cellular metabolism, but there is, however, one modelling and simulation approach that has shown a surprising ability to simulate and predict the metabolic behaviour of living cells - Flux Balance Analysis (FBA) (Edwards JS *et al.*, 2001; Edwards JS *et al.*, 2002; Pramanik J and Keasling JD, 1997; Price ND *et al.*, 2003; Reed JL and Palsson BO, 2003; Saner U *et al.*, 1992; Schilling CH *et al.*, 1999; Urbanczik R and Wagner C, 2005; Varma A and Palsson BO, 1993; Varner JD, 2000) that assumes a steady-state condition for the internal fluxes of the system and optimal flux distribution. Many organisms have been studied by using these so-called stoichiometric models (Borodina I *et al.*, 2005; Chung BK *et al.*, 2010; Duarte NC *et al.*, 2004; Feist AM *et al.*, 2006; Feist AM *et al.*, 2007; Navid A and Almaas E, 2009; Nogales J *et al.*, 2008; Oliveira AP *et al.*, 2005; Reed JL *et al.*, 2003; Schilling CH *et al.*, 2002; Sheikh K *et al.*, 2005; Varma A and Palsson BO, 1994), and *E. coli* metabolism was one of the first to be modelled using this approach (Varma A and Palsson BO, 1994), where the stoichiometry of metabolic pathways and maximal growth principles were used to predict cellular growth and the flux distribution through the metabolic network.

More generally, metabolic models usually based on the fundamental law of mass conservation, where the metabolic state of a cell is described by mass balance equations written for all the metabolite concentrations (*c*) that are mathematically expressed as follows in the matrix form:

$$\frac{dC}{dt} = S.v \qquad (1)$$

Here, the dilution effect caused by biomass growth is considered negligible. Metabolic interactions are expressed in the stoichiometric matrix ***S***, where each element S_{ij} represents the stoichiometric coefficient that indicates the participation of the i^{th} metabolite in the j^{th} reaction. Vector ***v*** corresponds to the reaction rates or fluxes through the metabolic reactions. Figure 1.4 shows an example.



$$\boxed{A: v_1 + v_4 - v_2}$$

$$\boxed{B: v_6 - v_4 - v_5}$$

$$\boxed{C: v_2 + v_5 - v_3}$$

$$\begin{bmatrix} \frac{dA}{dt} \\ \frac{dB}{dt} \\ \frac{dC}{dt} \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & -1 & 1 \\ 0 & 1 & -1 & 0 & 1 & 0 \end{bmatrix} . \begin{bmatrix} v1 \\ v2 \\ v3 \\ v4 \\ v5 \\ v6 \end{bmatrix}$$

Figure 1.4. Representation of the stoichiometry and mass balance equations in a metabolic network.

(a) Small reaction network consisting of three metabolites (A, B, and C) and six biochemical reactions (*v*). Material balances for each metabolite. (c) Metabolic model or material balance equations according to Eq. 1.

Under steady-state conditions, metabolite concentrations are constant and mass balances can be simplified to:

$$S.v = 0 \qquad (2)$$

When the number of fluxes is greater than the number of metabolites, then the system is mathematically considered underdetermined and the solution space is defined as the null space of *S*, where an infinite number of feasible flux distributions satisfy the mass balance equations (Edwards JS and Palsson BO, 1998). For that reason, constraints imposed by the thermodynamics (e.g. reaction reversibility/irreversibility) and enzyme kinetic properties (e.g. maximum reaction rates) can be included in the model to reduce the solution space (Covert MW *et al.*, 2003; Edwards JS *et al.*, 2002). These constraints can be introduced as linear inequalities:

$$\alpha_j \leq v_j \leq \beta_j, j = 1,...N \qquad (3)$$

Although constraints reduce the solution space of the system, they do not allow to define a single solution. FBA uses linear-optimization techniques to calculate an optimal flux distribution (a unique optimal solution) based on a pre-defined objective function (Edwards JS *et al.*, 2001; Edwards JS *et al.*, 2002; Pramanik J and Keasling JD, 1997; Price ND *et al.*, 2003; Reed JL and Palsson BO, 2003; Schilling CH *et al.*, 1999; Varma A and Palsson BO, 1993; Varner JD, 2000). The linear objective function (*Z*) can be maximized or minimized, according to the cellular function to be optimized (e.g., production of toxic by-products or cellular growth). The optimization problem can be formulated as:

$$\begin{gathered} Maximize \quad Z \qquad (4) \\ subject\ to \quad S.v = 0 \\ \alpha_j \leq v_j \leq \beta_j, j = 1,...N \end{gathered}$$

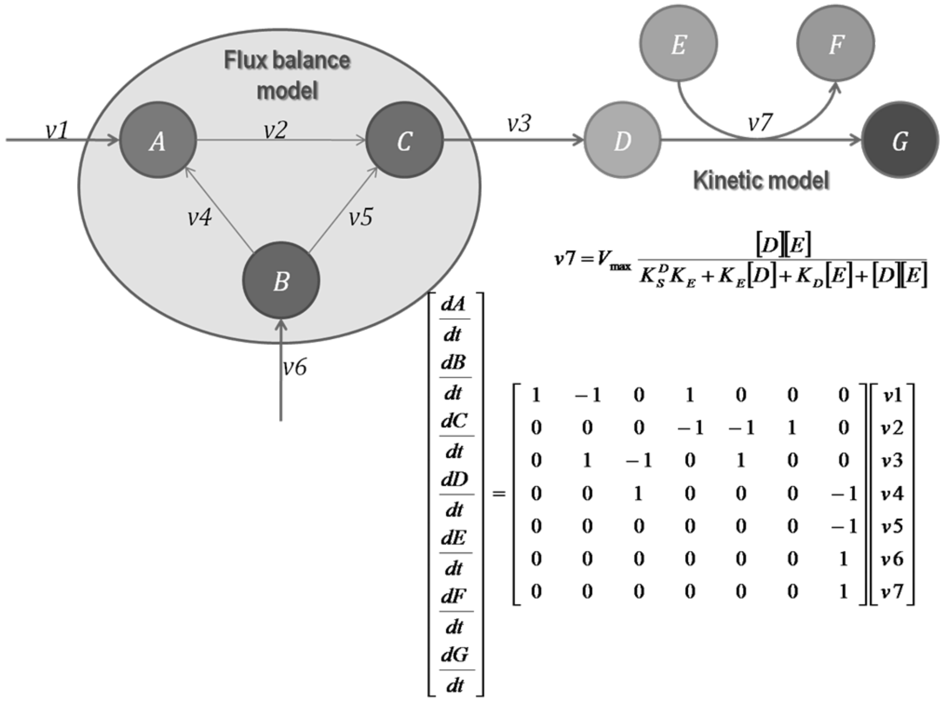
In nature and in specific circumstances, microbial cells have evolved towards maximization of biomass formation and, thus, this has been successfully used as an objective function in FBA simulations, to explore the capabilities and limitations of the metabolic network and predict cellular behaviour (Burgard AP and Maranas CD, 2003; Edwards JS *et al.*, 2001; Edwards JS and Palsson BO, 2000; Ibarra RU *et al.*, 2002; Schilling CH *et al.*, 2002).

The metabolic behaviour of recombinant systems has been investigated using flux balance approaches (Ow DS *et al.*, 2009; Weber J *et al.*, 2002). A genome-scale metabolic model for *E. coli* (Reed JL *et al.*, 2003) was complemented with mass balance equations for the expression of recombinant proteins and plasmid maintenance (Ozkan P *et al.*, 2005) and flux distributions provided access to information on metabolic pathway utilization and potential limitations. Flux distributions were quantitatively evaluated and revealed that metabolic perturbations imposed by the synthesis of recombinant proteins reduce cell growth rates and change the activity of catabolic pathways like the Embden-Meyerhof-Parnas (EMP) pathway and the tricarboxylic acid (TCA) cycle. Moreover, it was found that the maximizing growth rate as the cellular objective function does not provide the best description for the underlying cellular behaviour. Instead, maximization of maintenance energy (ATP_m) expenditure generates a metabolic flux distribution that explains better the physiological state of recombinant *E. coli* (Ow DS *et al.*, 2009).

Yet, predicted flux distributions by FBA simulations have shown certain inconsistencies between the model and experimental data. Some reasons have been pointed out, such as the lack of descriptions on regulatory phenomena or gaps in metabolic information (Breitling R *et al.*, 2008; Feist AM and Palsson BO, 2008; Orth JD and Palsson BO, 2010; Raman K and Chandra N, 2009). In most cases, the cell metabolism operates in a quasi- or pseudo-steady state, where metabolite concentrations do not change significantly over time, but some metabolic reactions can be heavily dependent on regulation. FBA models are therefore insufficient to predict metabolic activities; instead, they provide a snapshot of the metabolism at defined physiological states. The integration of transcriptional regulatory information into metabolic models using Boolean logic formulations has been already proposed (Covert MW *et al.*, 2001). Still, the dynamics of some metabolic processes cannot be neglected and therefore kinetic models may need to be used, despite the complexity of this modelling approach. Dynamic flux balance analysis (DFBA) was alternatively proposed (Mahadevan R *et al.*, 2002); however, the kinetic parameters of some important reactions are required, which is not always easy to get.

Clearly, there is no single modelling approach capable of representing all cellular phenomena, mainly because the level of detail of mathematical model descriptions depends largely on the information available about the system. Sometimes, the best modelling approach relies on the combination of different types of mathematical representations, according to the major purposes of the study. As an example, a three-level integrated approach, where the integration of a stoichiometric metabolic network with a Boolean transcriptional regulatory network (rFBA) and a set of ordinary differential equations (ODEs) to describe the dynamic variation of certain metabolite variables, was reported (Covert MW *et al.*, 2008). This modelling approach was able to model the *E. coli* central metabolism generating a metabolic flux distribution that was globally more accurate and informative.

As follows, in this thesis a two-step modelling approach is presented, combining a stoichiometric and a kinetic model to represent the metabolic perturbations imposed by recombinant protein synthesis and the ppGpp-induced response, respectively (see Chapter 3). Since a more detailed understanding on the ppGpp-induced response is expected with such approach, a kinetic model was implemented to represent the induction of the RelA activity upon depletion of the amino acids pool. Indeed, kinetic models are more suited to detail the dynamics of cellular processes that are highly dependent on regulation, such as protein translation processes (Zouridis H and Hatzimanikatis V, 2007) or folding and inclusion body formation in *E. coli* (Hoffmann F *et al.*, 2001). The combination of a flux balance model based on the stoichiometry of *E. coli* metabolic pathways (Reed JL *et al.*, 2003) and kinetic reactions for biomass formation, recombinant protein production, plasmid maintenance and ppGpp synthesis, are here proposed to get a first systematic representation of the effect of the accumulation of ppGpp in the cell physiology during recombinant protein production. It was aimed to develop a large scale modelling approach integrating different model representations that, ultimately, simplifies the computational efforts by restraining the number of variables that need to be evaluated by dynamic modelling. As exemplified in Figure 1.5, this approach reduces the need for kinetic parameters, while the dynamics of certain metabolite concentrations can be inferred using metabolic fluxes predicted by the flux balance model (e.g. metabolite *D*).



$$v7 = V_{\max} \frac{[D][E]}{K_S^D K_E + K_E[D] + K_D[E] + [D][E]}$$

$$\begin{bmatrix} \frac{dA}{dt} \\ \frac{dB}{dt} \\ \frac{dC}{dt} \\ \frac{dD}{dt} \\ \frac{dE}{dt} \\ \frac{dF}{dt} \\ \frac{dG}{dt} \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & -1 & 1 & 0 \\ 0 & 1 & -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} v1 \\ v2 \\ v3 \\ v4 \\ v5 \\ v6 \\ v7 \end{bmatrix}$$

Figure 1.5. Example of an integrated modelling approach.

The flux balance model describes the steady-state flux distributions (*v*1, *v*2, *v*3, *v*4, *v*5 and *v*6) through the metabolic network composed by metabolites *A*, *B* and *C*. Reaction rates *v*1 and *v*6 can be obtained from experimental measurements (e.g. substrate uptake rates). In the kinetic model, the expression for the reaction *v*7, is based on a Michaelis-Menten equation that describes the dynamic conversion of metabolites *D* and *E* into *F* and *G*.

1.4 METABOLOMICS APPROACHES FOR INVESTIGATING THE BEHAVIOUR OF RECOMBINANT *E. COLI* CELLS

The design of novel strategies to elucidate recombinant systems has been the primary goal of systems biotechnology (Bulter T *et al.*, 2003; Glick BR, 1995). The optimal design and development of upstream to downstream recombinant bioprocesses have recently been expanded mainly due to the advent of diverse omics technologies. The increasing availability of omics data has provided scientists with an unforeseen level of information on many different recombinant systems. For example, DNA microarrays have been used to investigate transcriptome profiles of recombinant *E. coli*, where thousands of genes were identified to change their expression levels along the fermentation process. Among these genes, those associated with the stress responses, like heat shock and stringent responses, were found to be up-regulated and genes related to translation processes and energy synthesis were down-regulated (Bonomo J and Gill RT, 2005; Haddadin FT and Harcum SW, 2005; Oh MK and Liao JC, 2000). Transcriptome analyses demonstrated that recombinant cultures are exposed to stressful conditions, which is denoted by the global changes in gene expression. To adapt to such conditions, cells engender a multitude of responses that often lead to important productivity losses, as well as cellular growth arrest.

The impact of recombinant protein production in *E. coli* cells has also been investigated by proteomic studies (Aldor IS *et al.*, 2005; Duerrschmid K *et al.*, 2008; Kabir MM and Shimizu K, 2001; Lee DH *et al.*, 2007; Wang YH *et al.*, 2005). Proteome changes reflected the physiological responses to heterologous protein production in recombinant *E. coli*, in particular the down-regulation of glycolytic enzymes, TCA cycle associated enzymes and ATP synthase, and up-regulation of cell protection proteins and some sugar transport proteins (Lee DH *et al.*, 2007). Apparently, the production of recombinant proteins can interfere with the performance of many metabolic pathways in the host metabolism, entailing a series of metabolic changes (Bentley WE *et al.*, 1990; Gill RT *et al.*, 2000; Harcum SW and Bentley WE, 1999). The association of these metabolic adjustments with stress-related processes, like the stringent response, has also been reported (Andersson L *et al.*, 1996; Chao YP *et al.*, 2002; Haddadin FT *et al.*, 2009; Hoffmann F and Rinas U, 2004; Schweder T *et al.*, 1995). The stringent response has been characterized by the down-regulation of nucleic acid and protein synthesis, and the simultaneous up-regulation of protein degradation (Ferullo DJ and Lovett ST, 2008; Jain V *et al.*, 2006; Magnusson LU *et al.*, 2005), which in turn would result in considerable losses during recombinant bioprocesses. The view of ppGpp as a global regulator of gene transcription has been expanded and it is now clear that cellular processes controlled by ppGpp are widespread and involve many cellular mechanisms important during cellular growth (Magnusson LU *et al.*, 2005). It has been reported (Schweder T *et al.*, 1995) that the accumulation of intracellular ppGpp, especially in slow-growing fed-batch conditions, seems to have a great impact in the recombinant protein synthesis, mainly due to the uncoupling between cell growth and protein production. Although important aspects regarding the physiology of recombinant systems have been found with transcriptome and proteome analysis, the metabolic alterations mediated by recombinant bioprocesses are still poorly uncovered. Moreover, it has been argued that the generation of hypotheses through these omics data alone, is incomplete and may lead to incorrect interpretations (Zhang W *et al.*, 2010).

As the metabolome level is the closest indicator of the cells' phenotype, metabolomics is now becoming the most relevant omics technology for understanding biological systems. Metabolomics aims at analysing and quantifying the complete set of metabolites (small organic molecules), providing substantial information on the organization of metabolism (see Table 1.1 for definitions used in the metabolomics field). In fact, metabolites play significant roles in the cell, as they are key participants in catabolic and anabolic pathways, regulate metabolic activities at the genetic (e.g. transcriptional effectors) or at the enzymatic (e.g. inhibitors/activators or cofactors) level and can be external (or internal) indicators of the environmental conditions. It is now widely recognized that metabolites are key biomolecules that govern the whole cell functioning, as elucidated by the complexity of interactions with many other functional units (see Figure 1.3). Metabolomics is a systematic analysis focused on this holistic view and encloses major advantages compared to other omics technologies, such as:

- Tractability: cells contain, in general, less metabolites than genes, transcripts, or proteins;

- Good discriminatory skills: changes in concentrations of metabolites are faster compared with the concentrations of proteins or transcripts and sometimes changes in the metabolite levels are not always detected in transcriptome or proteome profiles;

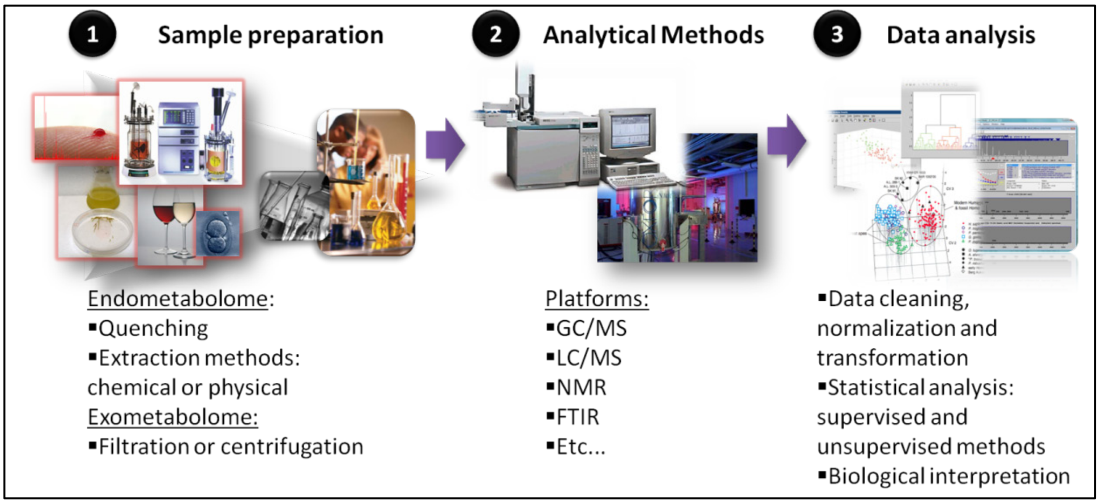
- Cost-effectiveness: costs *per* analyses are lower compared to proteomics and transcriptomics.

Table 1.1. Summary of some definitions used in the metabolomics field based on (Dunn WB and Ellis DI, 2005; Goodacre R *et al.*, 2004).

Term	Definition
Metabolome	The complete set of small metabolites present in a biological sample.
Endometabolome	Small metabolites found in the intracellular medium
Exometabolome	Small metabolites that are secreted by cells
Metabolomics	Non-biased identification and quantification of the whole metabolome in a sample under a given set of conditions.
Metabonomics	Analysis of the dynamic metabolic responses to biochemical perturbations caused by diseases, drugs and toxins, often employed to evaluate tissues and biological fluids.
Metabolic profiling	Identification and quantification focused on a group of metabolites, for example, a class of compounds such as carbohydrates, amino acids or those associated with a specific pathway.
Metabolic footprinting	Analysis of the set of metabolites that are secreted (or consumed) by cells into (or from) the extracellular medium.
Metabolic fingerprinting	High-throughput, rapid, global analysis providing sample classification based on the qualitative metabolic profiles exhibited by a biological sample.
Metabolite target analysis	Qualitative and quantitative analysis of a specific set of metabolites, for example, that are involved in a particular enzyme system that would be affected by genetic/environmental perturbations.

In the past few years, several studies have employed metabolomics approaches to evaluate the metabolic profiles of diverse microbial systems (Catchpole G *et al.*, 2009; Dobson G *et al.*, 2010; Dunn WB, 2008; Dunn WB and Ellis DI, 2005; Kell DB, 2004; Kol S *et al.*, 2010; Mashego MR *et al.*, 2007; Oldiges M *et al.*, 2007; Pasikanti KK *et al.*, 2008; van der Werf MJ, 2003; Yuliana ND *et al.*, 2010). Although the measurement of metabolites *per se* exists since the early days of biochemistry and techniques are basically the same used for years in analytical chemistry, the analysis of several metabolites in parallel in a unique sample is the novelty of this approach. Metabolite analysis is challenging due to the number of analytes present in biological samples, their concentration ranges and, most significantly, their chemical diversity. Progresses have been accomplished by the development of new experimental procedures and technological advances.

In a basic workflow procedure, there are three essential steps that define the capacity and success of the designed metabolome analysis (Figure 1.6). From the sample preparation that should take into consideration the chemical diversity and dynamic concentrations range of the metabolites under analysis, to the selection of the proper analytical platform, which can be comprised by a single analytical technique or the combination of various techniques, and the definition of data analysis methods, all these steps are crucial to obtain a more comprehensive analysis of the metabolome. These major steps will be further detailed below.



1 Sample preparation

Endometabolome:
▪Quenching
▪Extraction methods: chemical or physical
Exometabolome:
▪Filtration or centrifugation

2 Analytical Methods

Platforms:
▪GC/MS
▪LC/MS
▪NMR
▪FTIR
▪Etc...

3 Data analysis

▪Data cleaning, normalization and transformation
▪Statistical analysis: supervised and unsupervised methods
▪Biological interpretation

Figure 1.6. General workflow procedure of the metabolome analysis.

(1) The range of biological samples to be analysed is vast, which influences largely the applied methodologies to prepare theses sample. (2) With the latest technological advances, several analytical platforms became available increasing the sensitivity, quantitativeness and robustness of metabolomic analyses. (3) However, the growing application of these approaches in the identification and quantification of metabolites with a large chemical diversity, analytical software tools are facing some challenges. The need for specialized mathematical, statistical and bioinformatic tools is increasing.

1.4.1 Sample preparation

During this step, metabolites contained inside the cell (i.e. endometabolome) or the extracellular metabolite pool secreted into the culture broth (i.e. exometabolome) are collected. The endometabolome is more technically demanding than the exometabolome. Besides the difficulty in designing protocols capable of extracting metabolites from the intracellular milieu, a previous quenching step has to be devised to assure a rapid metabolic activity arrest. The quenching step is essential to assure that all metabolic activities inside the cell are arrested, and thus no further enzymatic conversions of metabolites are taking place. Usually, quenching protocols employ a rapid increase or decrease in temperature to prevent enzyme activity (Dunn WB, 2008), which is followed by the release of metabolites from the interior of the cell using heat or cold and occasionally acid or base solutions to assist cell lysis.

The first procedure combining fast quenching and subsequent separation of cells and supernatants, was developed by de Koning and van Dam (de Koning W and van Dam K, 1992). The quenching solution (60% methanol at -40°C) was used to halt yeast metabolism and cells were collected after centrifugation at -20°C. However, this method and others using cold methanol as a quenching agent have been contested due to cellular leakage (Taymaz-Nikerel H *et al.*, 2009; Villas-Boas SG and Bruheim P, 2007; Wittmann C *et al.*, 2004). During quenching procedures it is crucial that metabolites remain inside the cells, otherwise the levels of metabolites after cellular extraction would be underestimated. Therefore, several quenching protocols have been proposed, and successfully applied, to different microbial cells (see Table 1.2).

Table 1.2. List of some quenching solutions that have been applied in microbial metabolome analysis.

Quenching agent	Details	References
Cold methanol/water	60% (v/v) methanol at -40°C	(de Koning W and van Dam K, 1992)
Cold perchloric acid	35% (w/w) in water at -40°C	(Weuster-Botz D, 1997)
Liquid nitrogen	liquid N_2 at -150°C	(Chassagnole C *et al.*, 2002)
Cold methanol/water with ammonium carbonate or HEPES	60% (v/v) methanol buffered with 0.85% (w/v) ammonium carbonate or with 70 mM HEPES at -40°C	(Faijes M *et al.*, 2007)
Cold glycerol/saline	3:2 (v/v) glycerol:saline solution (13.5%NaCl) at -20°C	(Villas-Boas SG and Bruheim P, 2007)
Cold methanol/glycerol	3:2 (v/v) methanol:glycerol at -50°C	(Link H *et al.*, 2008)
Cold ethanol/sodium chloride	40% ethanol and 0.8% (w/v) sodium chloride at -20°C	(Spura J *et al.*, 2009)

Following the quenching procedure, cells are separated from the quenching solution by centrifugation or filtration, and subsequently subjected to extraction procedures through permeabilization of cell walls, usually with chemical or physical agents. These agents should allow for maximum extraction (i.e. as many metabolites as possible) with minimal alteration of the extracted metabolites. Logically, the choice for the most appropriate extraction method depends on the used microbial system, since cell wall and membrane structures, as well as composition, differ from organism to organism. Currently, the range of available chemical agents (Table 1.3) offers the possibility to optimize the extent and diversity of extracted metabolites in order to obtain meaningful metabolome data.

Table 1.3. List of chemical extraction methodologies applied to microbial cells to remove intracellular metabolites.

Extraction agent	Details	Microbial cells	References
Perchloric acid	Cold perchloric acid 50% (v/v) is added to samples and after rapid vortexing, the mixture is kept on ice for 10 min. Then, it is centrifuged for 10 min and extracts are neutralized with 15N KOH.	Bacteria	(Maharjan RP and Ferenci T, 2003)
Potassium hydroxide	Hot solution of 0.2N KOH at 80°C is added to samples and incubated for 10 min. After cooling on ice for 5 min, cell debris is removed by centrifugation and the extracts are neutralized with 0.1 mL perchloric acid (0.5 N).	Bacteria Filamentous fungus	(Hajjaj H *et al.*, 1998; Maharjan RP and Ferenci T, 2003)
Hot ethanol	Samples are boiled in 75% (v/v) ethanol at 80°C. After evaporation of the ethanol/water mixture, the pellet is resuspended in water.	Yeast	(Gonzalez B *et al.*, 1997)
Methanol/ chloroform	1:2 (v/v) methanol/ chloroform at -20°C is added to the samples and the mixture is vortexed for 30 s. Then samples are transfered to dry ice, where they are kept for 45 min and vortexed for 30 s every 15 min. The suspensions are then centrifuged for 10 min at 0°C.	Yeast	(de Koning W and van Dam K, 1992)
Hot methanol	2:1(v/v) methanol/water is added to the sample and incubated for 30 min at 70°C. The suspensions are then centrifuged for 10 min.	Bacteria	(Maharjan RP and Ferenci T, 2003)
Tris-H_2SO_4/EDTA	Samples are added to half the volume of Tris-H_2SO_4/EDTA (20 mM/2 mM), pH 7.75, at −25°C. After 1 min, samples are thawed and pipetted into another one-half volume of Tris-H_2SO_4/EDTA, pH 7.75, at 90°C. The mixture is then vortexed and after 10 min centrifuged.	Bacteria Yeast	(Buziol S *et al.*, 2002)
Cold methanol	Freeze-thawing cycles at low temperatures (-20°C). To enhance recovery, cells are washed with cold methanol once or twice.	Bacteria Yeast Filamentous fungus	(Maharjan RP and Ferenci T, 2003; Smart KF *et al.*, 2010; Villas-Boas SG *et al.*, 2005)

1.4.2 ANALYTICAL PLATFORMS

Once the biological samples are prepared there are several analytical platforms available that can be applied in metabolomics: gas chromatography (GC), capillary electrophoresis (CE), or liquid chromatography (LC) coupled with mass spectrometric detection (MS), direct infusion mass spectrometry (DIMS), Fourier transform-infrared (FT-IR) and nuclear magnetic resonance spectroscopy (NMR). Since the metabolome comprehends a vast range of chemical species with diverse physical and chemical properties and those analytes occur in a wide concentration range, analytical procedures are quite challenging. With the recent progresses in analytical technologies and the possibility to combine more than one in the analytical platforms, metabolomics experiments gained a wider coverage in terms of the type and number of metabolites analysed. Currently, the most popular method for metabolomics analysis is GC/MS (Coucheney E *et al.*, 2008; Jonsson P *et al.*, 2005; Kaspar H *et al.*, 2008; Koek MM *et al.*, 2006a; Kopka J, 2006; Pasikanti KK *et al.*, 2008; Sajewicz M *et al.*, 2009). In the GC equipment, a liquid sample is injected and rapidly vaporized and mixed with a carrier gas. Then, metabolites in the vaporized sample are separated on the GC column and follow to the MS device. Mass spectrometers operate in a three-step process: analytes are ionized in an ion source, either operating at atmospheric or vacuum pressures; ions are then separated according to the mass-to-charge (m/z) ratio in a mass analyser; and finally detected, either physically at a detector as an ion current or by the detection of orbital frequencies as image currents. Many advantages have been described for its application in metabolomic investigations, including its sensitivity (detection of μM concentrations) and the ability to identify metabolites through the fragmentation mass spectra. However, disadvantages like the requirement of chemical derivatization procedures, to convert some metabolites into less-polar, volatile and thermally stable derivatives before GC separation, are considerable. Chemical derivatization involves the substitution of active hydrogens in functional groups, such as –COOH, –OH, –NH, and –SH by other chemical groups via alkylation, acylation or silylation reactions. Silylation is the most traditional and results in stable derivatives with good reproducibility and wide range of metabolites that can be derivatized. However, it has been described as difficult to execute and time-consuming. Alkylation presents technical advantages compared to silylation, but the application range is limited. Nevertheless, important intermediates in the cellular metabolism, like carboxylic acids, amines, amino alcohols, and amino acids, are likely to be converted to volatile derivatives with this chemical derivatization. These chemical processes allow the detection of many more metabolite classes in a single GC/MS run. Although, some degree of variability can be introduced in the metabolome data, mainly because sample derivatization is a multi-step process, generally manual, GC/MS applications in metabolomic investigations are valuable in the identification of metabolites of microbial, plant and mammalian systems.

1.4.3 DATA ANALYSIS

Since the start of metabolomics, software approaches to automated data interpretation have been scarce, and most efforts have been made in individual laboratories to support their own scientific investigations. Mass spectral libraries are the main example, since mass spectra vary with the performed analytical procedures. For example, different derivatization agents produce different metabolite derivatives, which consequently generate altered mass spectra. Collaboration projects should be encouraged to promote the construction of larger and more complete libraries. The computerized matching of an unknown spectrum with a database/library would provide a very rapid and useful tool in metabolic profiling procedures. The Automated Mass Spectral Deconvolution and Identification System (AMDIS) (Stein SE, 1999) from the National Institute of Standards and Technology (NIST) has been largely applied to match unknown spectra with libraries in order to identify most detected components.

After identifying and quantifying the metabolome of a biological system, the main goal is to find relevant metabolites that discriminate between different phenotypic characteristics. By using statistical methods (e.g. multivariate data analysis or pattern recognition), differences between metabolomes can be identified and further contextualized to generate useful new knowledge. Most of the data analysis methods that have been applied in metabolomics, and also in many other omics data analyses, are based on unsupervised techniques, such as hierarchical cluster analysis (HCA) and principal component analysis (PCA) (Arbona V *et al.*, 2009; Boroczky K *et al.*, 2006; Griffin JL, 2004; Llorach-Asuncion R *et al.*, 2010; Pohjanen E *et al.*, 2006). Supervised methods are usually more powerful as the classification is based on prior knowledge (e.g. Fisher discriminant analysis); however, they have the disadvantage of requiring a training data set, which sometimes it is not easy to select and can overfit the model. Unsupervised methods, on the other hand, separate samples into classes without any training data set, but it is not always easy to interpret the basis for the generated clusters (Mendes P, 2002). Obviously, the most suitable data analysis method should be selected according to the aim of the study. Unsupervised methods are prioritized when prior information about the sample identity is unknown (e.g., in identification of a silent mutation phenotype) and the aim is sample classification. On the other hand, if sample identification is known and it is aimed to find characteristic metabolite profiles (e.g., search for groups of metabolites that define the phenotype of a microbial strain), supervised methods are ideal. Common statistical methods such as *t*-test or ANOVA, can also be used to find significant changes in the metabolite profiles. PCA analysis often uses only variables with significant changes tested by *t*-test or ANOVA. PCA analysis is mostly used for data visualization, where multivariate data is transformed into principal components and are projected on a 2D or 3D plots, facilitating data interpretation. An important aspect about this method is that PCA will emphasize variable components with high intensity levels, while lower metabolite levels are neglected, even if those are significant. Thus, data normalization is often performed before statistical analysis. Most books about chemometrics and statistics for analytical chemistry explore most of these subjects in detail, including data cleaning, normalization and transformation.

But, besides all the challenges in technical, computational or statistical analysis, the interpretation of data is perhaps the paramount dilemma in metabolomics. As stated by Professor Henry Nix, co-chair of the Australian Wild Country Science Council: "*Data does not equal information; information does not equal knowledge; and, most importantly of all, knowledge does not equal wisdom. We have oceans of data, rivers of information, small puddles of knowledge, and the odd drop of wisdom.*" Thus, to keep pace with this flood of data, procedures for storage, analysis and interpretation of omics data must be developed. Meanwhile, few attempts have been made to build tools capable to expedite these processes (Table 1.4).

Table 1.4. List of bioinformatics tools for storing, reporting, analysing and interpreting metabolomics data.

Tools	Description	References
Databases		
Madison Metabolomics Consortium Database (MMCD)[a]	Supports high-throughput NMR and MS data sets for the identification and quantification of metabolites present in biological samples.	(Cui Q *et al.*, 2008)
Human Metabolome Database[b]	The database covers more than 7900 metabolite entries with information about small molecule metabolites found in the human body. It contains three kinds of data: chemical data, clinical data and molecular biology/biochemistry data.	(Wishart DS *et al.*, 2009)
MassBank[c]	Public repository of mass spectra of small chemical compounds for life sciences (<3000 Da). The database contains data for 2337 authentic metabolites and for other synthetic compounds.	(Horai H *et al.*, 2010)
METLIN[d]	Metabolite database for metabolomics containing over 15,000 structures. Also, assists metabolite identification by providing public access to its repository of current and comprehensive mass spectral metabolite data.	(Smith CA *et al.*, 2005)
NIST[e]	The 2008 version of the NIST/EPA/NIH Mass Spectral Library (NIST 08) contains four libraries and a database of retention index values. Together they contain 220,460 spectra of 192,108 different chemical compounds.	(Stein SE, 1995)
Data analysis		
MetaboAnalyst[f]	Web-based tool for metabolomic data processing, data normalization, multivariate statistical analysis, graphing, metabolite identification and pathway mapping.	(Xia J *et al.*, 2009)
MetaFIND[g]	An application for 'post-feature selection' that aid metabolite signature elucidation, feature discovery and inference of metabolic correlations.	(Bryan K *et al.*, 2008)
MSEA[h]	A web-based tool to identify biologically meaningful patterns in quantitative metabolomic data.	(Xia J and Wishart DS, 2010b)
MEV[i]	MeV is an application for the analysis, visualization and data-mining of large-scale data. It was developed for the analysis of microarrays, but is a versatile analytical tool.	(Saeed AI *et al.*, 2006)
Data visualization		
MetPA[j]	A free and easy-to-use web application designed to perform pathway analysis and visualization of quantitative metabolomic data.	(Xia J and Wishart DS, 2010a)
MetExplore[k]	A web server that offers the possibility to link metabolites identified in untargeted metabolomics experiments within the context of genome-scale reconstructed metabolic networks.	(Cottret L *et al.*, 2010)
Metscape[l]	A plug-in for Cytoscape, used to visualize and interpret metabolomic data in the context of human metabolic networks.	(Gao J *et al.*, 2010)

[a]http://mmcd.nmrfam.wisc.edu/; [b]http://www.hmdb.ca/; [c]http://www.massbank.jp/; [d]http://metlin.scripps.edu/; [e]http://chemdata.nist.gov/; [f]http://www.metaboanalyst.ca/, [g]http://mlg.ucd.ie/metafind; [h]http://www.msea.ca; ihttp://www.tm4.org/mev/; [j]http://metpa.metabolomics.ca/; [k]http://metexplore.toulouse.inra.fr/; [l]http://metscape.ncibi.org/

Databases for storing detailed metabolite profiles, including raw data and detailed metadata, are being created. However, the implementation of standard formats for recording and reporting metabolomics data is not yet commonly used (Sansone SA *et al.*, 2007). Although some data analysis tools for metabolomics applications have became available in the last years, most researchers have found general software environments, like R or MATLAB, more versatile for statistical computing. Other software tools developed to analyse specific omics data, like MultiExperiment Viewer (MeV) for microarrays data (Saeed AI *et al.*, 2006), can also be useful when analysing metabolomics data. As presented in Chapters 4 and 5, clustering algorithms from MeV were quite valuable when classifying samples with different phenotypic characteristics. Most recently, computational tools for elucidating and visualizing metabolome data have emerged, namely MetPA (Xia J and Wishart DS, 2010a), MetExplore (Cottret L *et al.*, 2010) and Metscape (Gao J *et al.*, 2010). The depiction of metabolome information into diagrams that provide an overview of the metabolite neighbourhood and relationships is one of the potentialities of these tools. The shift from roadmaps of metabolic pathways into the construction of comprehensive metabolic networks, where all biochemical entities and their relationships are represented (Bersini H *et al.*, 2005; Borenstein E and Feldman MW, 2009; Jeong H *et al.*, 2000; Jinq Z *et al.*, 2006; Montanez R *et al.*, 2010), was perhaps the driving force behind these initiatives.

1.4.4 METABOLOMICS-DRIVEN ANALYSIS OF *E. COLI*

The *E. coli* metabolism has been extensively characterised, but only in the last decade metabolomics approaches were applied to evaluate (simultaneously) the global metabolite profiles, both qualitatively and quantitatively. One of the first approaches was developed by Tweeddale and co-workers (Tweeddale H *et al.*, 1998) and applied a two-dimensional thin-layer chromatography to analyse the slow growth metabolism of *E. coli* growing on glucose in minimal medium. Later on, metabolome analyses were performed using more sophisticated techniques, such as GC/MS or NMR, which have expanded our knowledge on *E. coli* metabolism (Koek MM *et al.*, 2006b). For instance, the *E. coli* metabolome has been investigated to understand key metabolic processes, such as the central nitrogen metabolism (Yuan J *et al.*, 2009) and glycolytic activities under different growth rates (Schaub J and Reuss M, 2008), or cellular responses to carbon and nitrogen starvation (Brauer MJ *et al.*, 2006). Other studies identified novel enzymatic activities*,* such as the hydroxybutyrate dehydrogenase that is involved in the metabolism of succinic semialdehyde, and other potentially toxic intermediates that may accumulate under stress conditions in *E. coli* (Saito N *et al.*, 2009a). Thus, metabolomic approaches can reveal new cellular metabolic processes, or bring important insights into the physiology of *E. coli.* The evaluation of metabolite levels and further association to enzymatic reactions has been fundamental in the elucidation of many biochemical functions, as well as to characterize enzymatic properties.

Often these biochemical reactions are organized into metabolic pathways as functional modules of organisms metabolic networks graph theory principles have been applied to study the structure and topological properties of these metabolic networks to dissect functional and behavioural features of many organisms (Almaas E, 2007; Barabasi AL and Oltvai ZN, 2004; Mazurie A *et al.*, 2010). As such, metabolic interpretations can be biased if the metabolic networks are incomplete or inaccurate. As stated (Montanez R *et al.*, 2010) "*a network is an abstraction of reality and its construction can determine the conclusions derived from it*". Therefore, computational approaches that evaluate the correlation between metabolite levels, in order to infer metabolic networks or even to discover novel pathways, have been explored in metabolomics-driven analysis (Steuer R *et al.*, 2003). It is remarkable how metabolite levels can be repeatedly correlated and more significantly, how frequently these correlations are shown between metabolites that are not neighbours in a metabolic pathway, but, most likely, are involved in regulatory mechanisms. To exemplify the potential of this approach, the metabolic correlations between metabolites analysed by GC/MS in the extracellular medium of an *E. coli* fermentation (metabolome data from experiments analysed in Chapter 6) were determined (Figure 1.7).

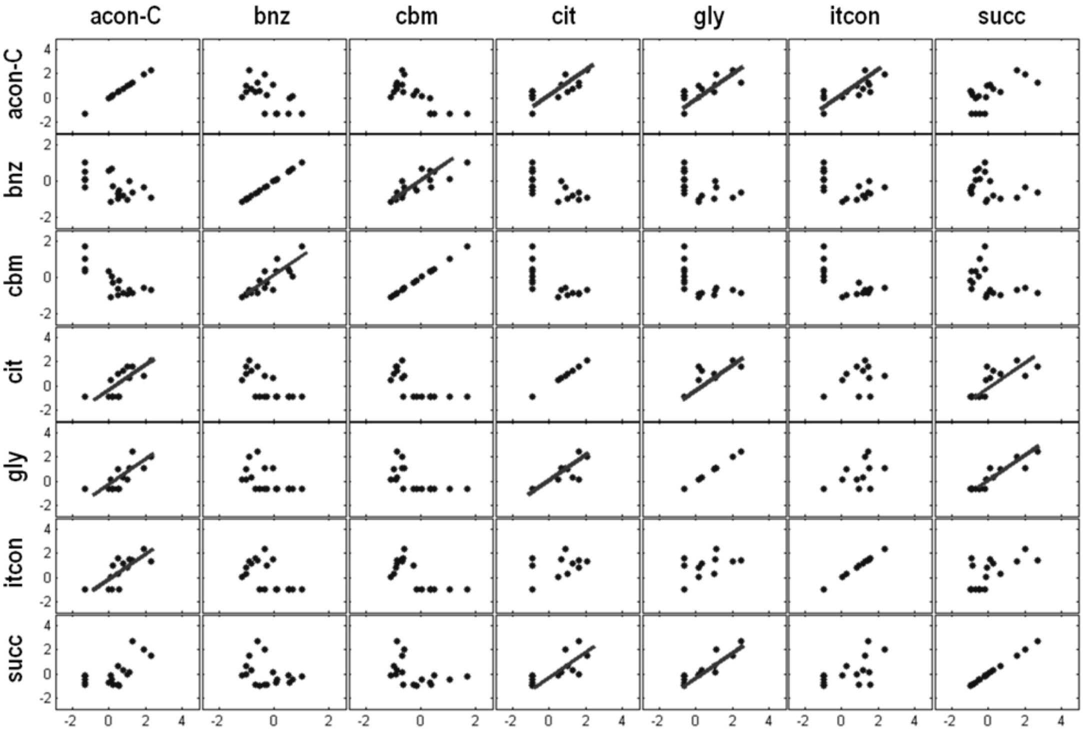


Figure 1.7. Metabolic correlations between extracellular metabolites detected by GC/MS during a recombinant *E. coli* fed-batch fermentation.

The relative concentrations of extracellular metabolites produced during a recombinant *E. coli* W3110 fed-batch fermentation were analysed by GC/MS (see Chapter 6 for details). Abbreviattions: acon-C – *cis*-aconitate; bnz – benzoate; cbm – carbamate; cit – citrate; gly – glycine; itcon – itaconate; succ – succinate.

As demonstrated, there are seven extracellular metabolites that show, at some extent, correlated patterns between their relative levels measured during the fermentation process. It indicates that a degree of association is likely to exist between these metabolites. Several reasons can be deduced, but the most reasonable is the participation in the same metabolic process or closely related processes. It should be emphasized that these interpretations are based on the assumption that extracellular metabolite levels are closely related to their intracellular concentrations, i.e. extracellular metabolic changes might be envisioned as changes in the intracellular metabolism, thus reflecting phenotypic alterations. To further interpret these metabolic correlations, metabolites were depicted into the *E. coli* metabolic network from EcoCyc (Keseler IM *et al.*, 2009) (Figure 1.8).

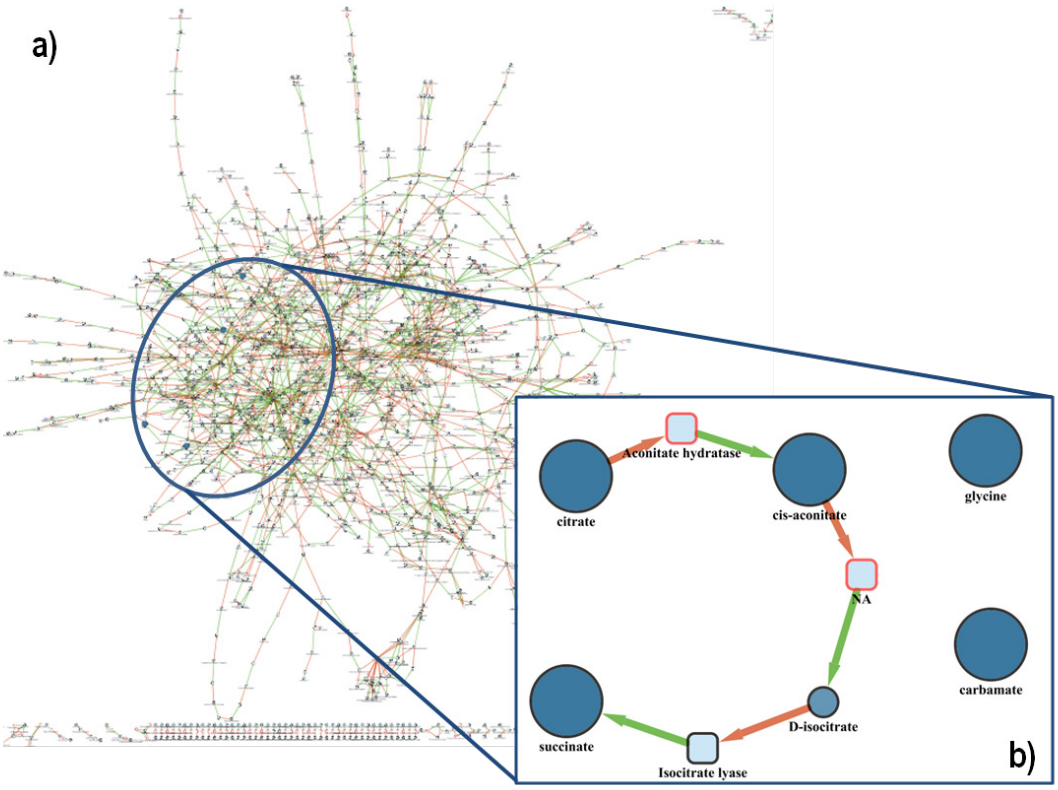


Figure 1.8. Diagram showing the neighbourhood of metabolites detected in the metabolite footprints.

a) The *E. coli* metabolic network from EcoCyc (Keseler IM *et al.*, 2009) was used to locate the metabolites found to have highly correlated profiles. b) Diagram showing the seven correlated metabolites; circles represent metabolites, in which the larger ones indicate those that were found to be correlated, while squares illustrate catalytic enzymes. Isolated metabolites have no links in the network associating these metabolites. Arrows indicate production (green) and consumption (orange) reactions. This exercise was performed using the MetExplore tool with a Cytoscape plug-in (Cottret L *et al.*, 2010).

As shown in Figure 1.8, at least three metabolites (citrate, *cis*-aconitate and succinate) are neighbours in the metabolic network, participating in the TCA cycle. Although *cis*-aconitate has a higher correlation coefficient with citrate than with succinate (i.e. its closest neighbour), it was shown that both are participants in the same metabolic pathway only two reactions distant. These reactions are connected through D-isocitrate that is an important node in the metabolic network, which has brought new considerations when examining the complete metabolite footprints studied in Chapter 6.

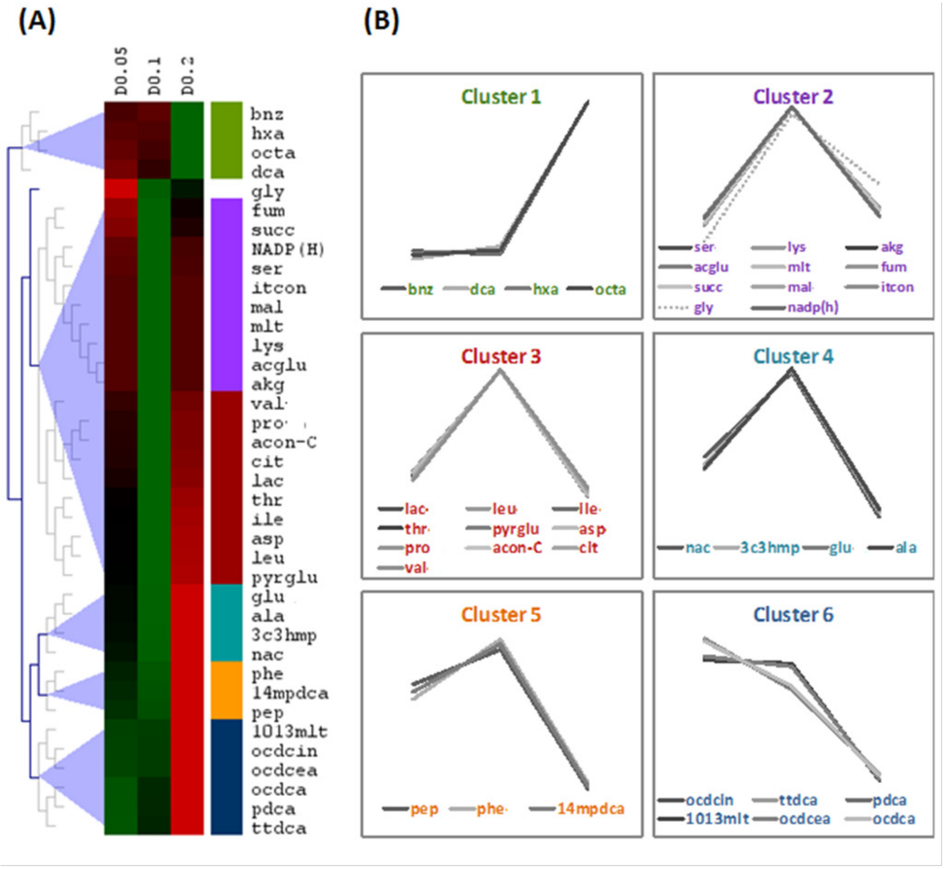


Figure 1.9. Clustering analyses of the *E. coli* W3110 endometabolome: (A) hierarchical and (B) *K*-means clustering.

Intracellular metabolites extracted from *E. coli* cells grown in chemostat cultures at three dilution rates (0.05, 0.1 and 0.2 h^{-1}) were analysed by GC/MS. Metabolite clusters were determined using MeV algorithms (Saeed AI *et al.*, 2006) for hierarchical clustering (HCL) and *K*-means clustering based on the Pearson's correlation metrics. Correlation coefficients (*r*) vary from 1 to -1 and coefficients higher than 0.8 indicate strong correlations between metabolite profiles. The dashed line delimits the correlation coefficient threshold below which metabolite profiles were considered uncorrelated.

Other methods that can used to find and visualize metabolite correlations are the hierarchical and *k*-means clustering analysis. As demonstrated in Figure 1.9, metabolite profiles from the endometabolome of *E. coli* W3110 cells grown in chemostat culture, were classified into six clusters. These clusters represent the set of metabolites that showed similar metabolic patterns when the culturing conditions were changed. For example, it was clear that at lower dilutions rates (0.05 and 0.1 h^{-1}) short chain fatty acids (hxa: hexanoate [n-C6:0], octa: octanoate [n-C8:0] and dca:decanoate [n-C10:0]) and benzoate were present in low levels, but increasing the dilution rate a significant increase in their levels was verified. The opposite was seen for long chain fatty acids. However, one pattern was similar among four clusters. It seems that, in general, at a dilution rate of 0.1h^{-1} metabolite levels increase. This is analysed in detail in Chapter 4.

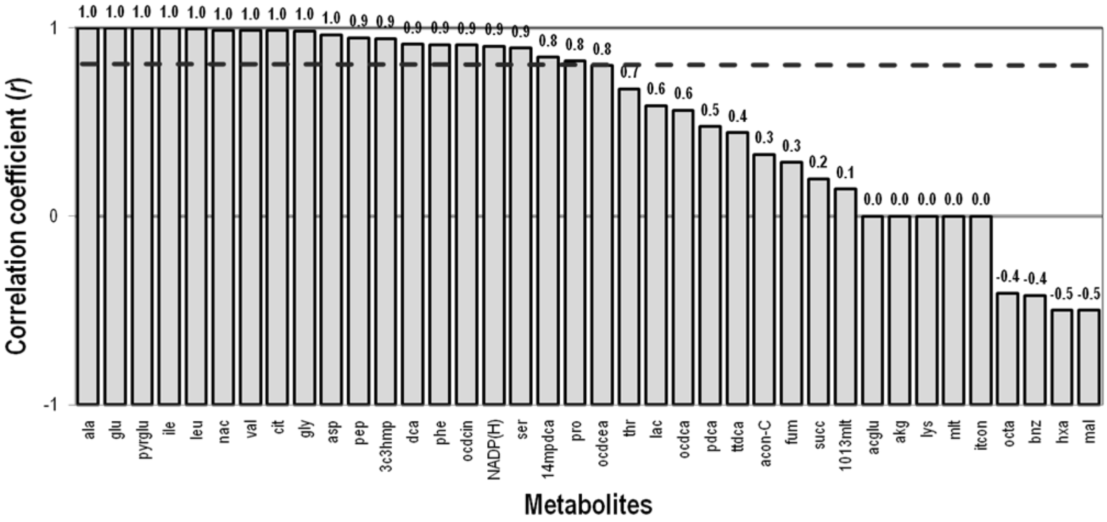


Figure 1.10. Pair-wise metabolite-metabolite correlations constructed from GC/MS measurements of intracellular metabolites from W3110 and Δ*rel*A *E. coli* cultures.

Metabolic correlations can also be explored for comparison of metabolite profiles produced by different *E coli* strains when growing in the same conditions (see Chapters 4 and 6). As exemplified in Figure 1.10, metabolic profiles of endometabolome samples from *E. coli* W3110 and Δ*rel*A cells were evaluated by pair-wise correlation coefficients (*r*), showing that almost half of the metabolites detected by GC/MS analysis showed strong correlated patterns (i.e. *r* above 0.8). Some metabolites revealed negative correlations, which mean that the intracellular accumulation of these metabolites followed an opposite pattern in one of the *E. coli* cultures, e.g. malate (mal) with *r* equal to -0.3.

As described, there are many ways to explore metabolome data and capture relevant metabolic patterns that explain some physiological characteristics observed during fermentation processes.

The fact that physiological responses to *stimulus* in microorganisms are often mediated by metabolic reactions entangled in intricate relationships in the metabolic network is pertinent. Moreover, alterations in quantitative levels of metabolites provide valuable information for understanding the dynamics of metabolic activities and the organizational principles underlying the metabolic networks. The access to metabolome data that can be further represented into large scale metabolic networks supports the holistic perspective of the metabolism that has been sustained by the systems biology view (Kitano H, 2002). However, it must be noted that interpretations are heavily shaped by the metabolomics approach performed to generated metabolome data and, consequently, the obtained metabolite coverage. For example, a metabolic footprinting approach is not capable to cover the same range of metabolites as an approach designed to measures intracellular metabolites. It is not expected that the entire set of metabolites produced under certain metabolic conditions are equally present inside and outside the cells. However, this approach offers several advantages, like technical simplicity, higher reproducibility and the possibility to find metabolic markers that could be used to monitor physiological alterations during bioprocesses that would not be otherwise detectable.

Metabolomic measurements have also been used to elucidate the function of the unknown and novel genes (Forster J *et al.*, 2002; Griffin JL, 2004; Raamsdonk LM *et al.*, 2001). Although genome sequencing projects have been successfully applied in the functional genomics field, the potential of metabolomics to determine gene functions has also been considered (Bino RJ *et al.*, 2004). Still today, of 4460 *E. coli* genes, only 2650 (59.3%) have experimentally defined functions (Keseler IM *et al.*, 2009), which denotes the necessity for new tools capable to uncover cellular functions of enzymes and other unassigned proteins. Although bioinformatics tools have been widely used to predict functional assignments, many metabolic activities are currently orphan. For example, metabolomics analysis presented in this thesis exposed that some metabolites detected by GC/MS in *E. coli* samples could not be linked to any known metabolic pathway. Metabolites, such as itaconate, malonate, 2-phenylglycine or benzoate, have no assigned metabolic reaction in the *E. coli* metabolic network, according to public databases, such as KEGG (Kanehisa M and Goto S, 2000) and EcoCyc (Keseler IM *et al.*, 2009), or genome-scale metabolic models (e.g. *i*AF1260 model of *E. coli* K12 (Feist AM *et al.*, 2007)). Yet, metabolome data confirmed that these metabolites are produced by *E. coli* cells and have characteristic metabolic profiles that were correlated with other metabolites in the metabolic network (see Figure 1.8).

Different approaches can be devised to confirm the role of these metabolites in the cellular metabolism. As detailed in Figure 1.11, metabolite profiling can be explored as a screening methodology to investigate and discover *in vitro* and *in vivo* activities of enzymes, by monitoring the changes in metabolite levels. Indeed, the 4-hydroxybutyrate dehydrogenase activity in *E. coli* was recently discovered using a metabolite profiling-based method (Saito N *et al.*, 2009b).

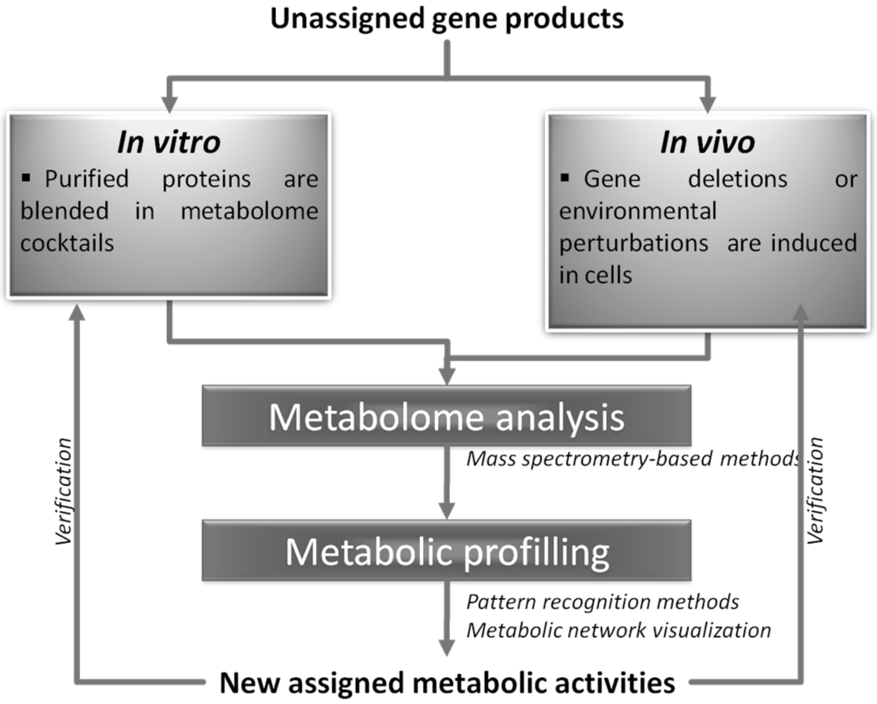

Unassigned gene products
In vitro
▪ Purified proteins are blended in metabolome cocktails
In vivo
▪ Gene deletions or environmental perturbations are induced in cells
Metabolome analysis
Mass spectrometry-based methods
Metabolic profilling
Pattern recognition methods
Metabolic network visualization
New assigned metabolic activities
Verification
Verification

Figure 1.11. Schematic representation of metabolomics-driven approaches to identify new enzymatic functions (Saito N *et al.*, 2010).

There are two fundamental approaches that have been implemented: *in vitro*, where gene products with unassigned biochemical functions, often overproduced using recombinant systems with fusion tags to facilitate the purification, are added to a metabolome cocktail; and *in vivo*, where cells are genetically modified (target deletion of specific genes) or are exposed to environmental perturbations to observe metabolic changes. Mass spectrometry-based methods are used for metabolite identification and metabolic profiling allows for the verification of metabolite patterns that might explain phenotypic features. It is always advised further experimental validations for the new assigned functions.

In this thesis *in vivo* metabolite profiling-based methods were used to determine the metabolic status of *E. coli* cells under various environmental and genetic conditions:

- Two *E. coli* strains were used. wild-type *E. coli* W3110 and the isogenic mutant Δ*rel*A;
- Both *E. coli* strains were cloned with a pTRC-His-AcGFP1 vector and were tested in producing and non-producing conditions;
- Cells were grown in different fermentation modes: fed-batch and chemostat cultivations;
- Several growth conditions were evaluated: from extremely low to excessive nutrient availability.

Although a significant number of metabolites, known to participate in central metabolic pathways of *E. coli,* were not measured by the implemented GC/MS approach, the obtained metabolome data allowed for important discoveries that will be detailed in Chapters 4, 5 and 6. As previously mentioned several attempts to improve the performance of *E. coli* cells during recombinant bioprocesses have been presented in the last years. In particular, omics technologies have been applied with some success to derive new hypothesis to engineer host cells improving recombinant protein productivity. However, metabolomics is still the less used technology, certainly because the implementation of these methodologies is more difficult, and thus, any interpretation based on transcriptome or proteome data lacks the link to phenotypic characteristics. To enable predictive metabolic engineering, analytical approaches might not only incorporate experimental information from gene expression profiles or protein abundances, but primarily metabolic data that could be derived from metabolomics or/and fluxomics.

1.5 References

1. Aldor IS *et al* (2005) Proteomic profiling of recombinant *Escherichia coli* in high-cell-density fermentations for improved production of an antibody fragment biopharmaceutical. *Applied and Environmental Microbiology* 71 (4):1717-1728.

2. Allen SP *et al* (1992) Two novel heat shock genes encoding proteins produced in response to heterologous protein expression in *Escherichia coli*. *Journal of Bacteriology* 174 (21):6938-6947.

3. Almaas E (2007) Biological impacts and context of network theory. *Journal of Experimental Biology* 210 (9):1548-1558.

4. Andersson L *et al* (1996) Impact of plasmid presence and induction on cellular responses in fed batch cultures of *Escherichia coli*. *Journal of Biotechnology* 46 (3):255-263.

5. Arbona V *et al* (2009) Plant Phenotype Demarcation Using Nontargeted LC-MS and GC-MS Metabolite Profiling. *Journal of Agricultural and Food Chemistry* 57 (16):7338-7347.

6. Baneyx F (1999) Recombinant protein expression in *Escherichia coli*. *Current Opinion in Biotechnology* 10 (5):411-421.

7. Barabasi AL and Oltvai ZN (2004) Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics* 5 (2):101-U15.

8. Baracchini E and Bremer H (1988) Stringent and Growth-Control of Ribosomal-Rna Synthesis in Escherichia-Coli Are Both Mediated by Ppgpp. *Journal of Biological Chemistry* 263 (6):2597-2602.

9. Barker MM, Gaal T, and Gourse RL (2001a) Mechanism of regulation of transcription initiation by ppGpp. II. Models for positive control based on properties of RNAP mutants and competition for RNAP. *Journal of Molecular Biology* 305 (4):689-702.

10. Barker MM *et al* (2001b) Mechanism of regulation of transcription initiation by ppGpp. I. Effects of ppGpp on transcription initiation in vivo and in vitro. *Journal of Molecular Biology* 305 (4):673-688.

11. Bentley WE *et al* (1990) Plasmid-encoded protein - The principal factor in the metabolic burden associated with recombinant bacteria. *Biotechnology and Bioengineering* 35 (7):668-681.

12. Bersini H, Lenaerts T, and Santos FC (2005) Growing biochemical networks: Identifying the intrinsic properties. *Advances in Artifical Life, Proceedings* 3630:864-873.

13. Bichara M *et al* (2007) RecA-mediated excision repair: a novel mechanism for repairing DNA lesions at sites of arrested DNA synthesis. *Molecular Microbiology* 65 (1):218-229.

14. Bino RJ *et al* (2004) Potential of metabolomics as a functional genomics tool. *Trends in Plant Science* 9 (9):418-425.

15. Bonomo J and Gill RT (2005) Amino acid content of recombinant proteins influences the metabolic burden response. *Biotechnology and Bioengineering* 90 (1):116-126.

16. Borenstein E and Feldman MW (2009) Topological signatures of species interactions in metabolic networks. *Journal of Computational Biology* 16 (2):191-200.

17. Boroczky K *et al* (2006) Cluster analysis as selection and dereplication tool for the identification of new natural compounds from large sample sets. *Chemistry & Biodiversity* 3 (6):622-634.

18. Borodina I, Krabben P, and Nielsen J (2005) Genome-scale analysis of *Streptomyces coelicolor* A3(2) metabolism. *Genome Research* 15 (6):820-829.

19. Brauer MJ *et al* (2006) Conservation of the metabolomic response to starvation across two divergent microbes. *Proceedings of the National Academy of Sciences of the United States of America* 103 (51):19302-19307.

20. Breitling R, Vitkup D, and Barrett MP (2008) New surveyor tools for charting microbial metabolic maps. *Nature Reviews Microbiology* 6 (2):156-161.

21. Brown L *et al* (2002) DksA affects ppGpp induction of RpoS at a translational level. *Journal of Bacteriology* 184 (16):4455-4465.

22. Bryan K, Brennan L, and Cunningham P (2008) MetaFIND: a feature analysis tool for metabolomics data. *BMC Bioinformatics* 9:470.

23. Bulter T, Bernstein JR, and Liao JC (2003) A perspective of metabolic engineering strategies: Moving up the systems hierarchy. *Biotechnology and Bioengineering* 84 (7):815-821.

24. Burgard AP and Maranas CD (2003) Optimization-based framework for inferring and testing hypothesized metabolic objective functions. *Biotechnology and Bioengineering* 82 (6):670-677.

25. Buziol S *et al* (2002) New bioreactor-coupled rapid stopped-flow sampling technique for measurements of metabolite dynamics on a subsecond time scale. *Biotechnology and Bioengineering* 80 (6):632-636.

26. Catchpole G *et al* (2009) Metabolic profiling reveals key metabolic features of renal cell carcinoma. *Journal of Cellular and Molecular Medicine*.

27. Chao YP, Chiang CJ, and Hung WB (2002) Stringent regulation and high-level expression of heterologous genes in *Escherichia coli* using T7 system controllable by the *ara*BAD promoter. *Biotechnology Progress* 18 (2):394-400.

28. Chassagnole C *et al* (2002) Dynamic modeling of the central carbon metabolism of *Escherichia coli*. *Biotechnology and Bioengineering* 79 (1):53-73.

29. Chatterji D and Ojha AK (2001) Revisiting the stringent response, ppGpp and starvation signaling. *Current Opinion in Microbiology* 4 (2):160-165.

30. Chou CP (2007) Engineering cell physiology to enhance recombinant protein production in *Escherichia coli*. *Applied Microbiology and Biotechnology* 76 (3):521-532.

31. Chung BK *et al* (2010) Genome-scale metabolic reconstruction and *in silico* analysis of methylotrophic yeast *Pichia pastoris* for strain improvement. *Microbial Cell Factories* 9 (1):50.

32. Cochran JW and Byrne RW (1974) Isolation and properties of a ribosome-bound factor required for ppGpp and pppGpp synthesis in *Escherichia coli*. *Journal of Biological Chemistry* 249 (2):353-360.

33. Cottret L *et al* (2010) MetExplore: a web server to link metabolomic experiments and genome-scale metabolic networks. *Nucleic Acids Research* 38 Suppl:W132-W137.

34. Coucheney E *et al* (2008) Gas chromatographic metabolic profiling: A sensitive tool for functional microbial ecology. *Journal of Microbiological Methods* 75 (3):491-500.

35. Covert MW, Famili I, and Palsson BO (2003) Identifying constraints that govern cell behavior: A key to converting conceptual to computational models in biology? *Biotechnology and Bioengineering* 84 (7):763-772.

36. Covert MW, Schilling CH, and Palsson B (2001) Regulation of gene expression in flux balance models of metabolism. *Journal of Theoretical Biology* 213 (1):73-88.

37. Covert MW *et al* (2008) Integrating metabolic, transcriptional regulatory and signal transduction models in *Escherichia coli*. *Bioinformatics* 24 (18):2044-2050.

38. Cui Q *et al* (2008) Metabolite identification via the Madison Metabolomics Consortium Database. *Nature Biotechnology* 26 (2):162-164.

39. de Koning W and van Dam K (1992) A method for the determination of changes of glycolytic metabolites in yeast on a subsecond time scale using extraction at neutral pH. *Analytical Chemistry* 204 (1):118-123.

40. Dedhia N *et al* (1997) Improvement in recombinant protein production in ppGpp-deficient *Escherichia coli*. *Biotechnology and Bioengineering* 53 (4):380-386.

41. Demain AL and Vaishnav P (2009) Production of recombinant proteins by microbes and higher organisms. *Biotechnology Advances* 27 (3):297-306.

42. Dobson G *et al* (2010) A metabolomics study of cultivated potato (*Solanum tuberosum*) groups Andigena, Phureja, Stenotomum, and Tuberosum using gas chromatography-mass spectrometry. *Journal of Agricultural and Food Chemistry* 58 (2):1214-1223.

43. Duarte NC, Herrgard MJ, and Palsson BO (2004) Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. *Genome Research* 14:1298-1309.

44. Duerrschmid K *et al* (2008) Monitoring of transcriptome and proteome profiles to investigate the cellular response of *E. coli* towards recombinant protein expression under defined chemostat conditions. *Journal of Biotechnology* 135 (1):34-44.

45. Dunn WB (2008) Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes. *Physical Biology* 5 (1):011001.

46. Dunn WB and Ellis DI (2005) Metabolomics: Current analytical platforms and methodologies. *Trac-Trends in Analytical Chemistry* 24 (4):285-294.

47. Durfee T *et al* (2008) Transcription profiling of the stringent response in *Escherichia coli*. *Journal of Bacteriology* 190 (3):1084-1096.

48. Edwards JS, Covert M, and Palsson B (2002) Metabolic modelling of microbes: the flux-balance approach. *Environmental Microbiology* 4 (3):133-140.

49. Edwards JS, Ibarra RU, and Palsson BO (2001) *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nature Biotechnology* 19 (2):125-130.

50. Edwards JS and Palsson BO (1998) How will bioinformatics influence metabolic engineering? *Biotechnology and Bioengineering* 58 (2-3):162-169.

51. Edwards JS and Palsson BO (2000) The *Escherichia coli* MG1655 *in silico* metabolic genotype: Its definition, characteristics, and capabilities. *Proceedings of the National Academy of Sciences of the United States of America* 97 (10):5528-5533.

52. Endo S *et al* (2006) Effects of E-coli chaperones on the solubility of human receptors in an in vitro expression system. *Molecular Biotechnology* 33 (3):199-209.

53. Faijes M, Mars AE, and Smid EJ (2007) Comparison of quenching and extraction methodologies for metabolome analysis of *Lactobacillus plantarum*. *Microbial Cell Factories* 6:27.

54. Feist AM *et al* (2007) A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Molecular Systems Biology* 3:121.

55. Feist AM and Palsson BO (2008) The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. *Nature Biotechnology* 26 (6):659-667.

56. Feist AM *et al* (2006) Modeling methanogenesis with a genome-scale metabolic reconstruction of *Methanosarcina barkeri*. *Molecular Systems Biology* doi:10.1038:msb4100046.

57. Ferrer-Miralles N *et al* (2009) Microbial factories for recombinant pharmaceuticals. *Microbial Cell Factories* 8:17.

58. Ferullo DJ and Lovett ST (2008) The stringent response and cell cycle arrest in *Escherichia coli*. *PLoS Genetics* 4 (12):e1000300.

59. Flores S *et al* (2004) Growth-rate recovery of Escherichia coli cultures carrying a multicopy plasmid, by engineering of the pentose-phosphate pathway. *Biotechnology and Bioengineering* 87 (4):485-494.

60. Forster J, Gombert AK, and Nielsen J (2002) A functional genomics approach using metabolomics and *in silico* pathway analysis. *Biotechnology and Bioengineering* 79 (7):703-712.

61. Gao J *et al* (2010) Metscape: a Cytoscape plug-in for visualizing and interpreting metabolomic data in the context of human metabolic networks. *Bioinformatics* 26 (7):971-973.

62. Gill RT, Valdes JJ, and Bentley WE (2000) A comparative study of global stress gene regulation in response to overexpression of recombinant proteins in *Escherichia coli*. *Metabolic Engineering* 2 (3):178-189.

63. Glick BR (1995) Metabolic Load and Heterologous Gene-Expression. *Biotechnology Advances* 13 (2):247-261.

64. Gnoth S *et al* (2008) Control of cultivation processes for recombinant protein production: a review. *Bioprocess and Biosystems Engineering* 31 (1):21-39.

65. Gonzalez B, Francois J, and Renaud M (1997) A rapid and reliable method for metabolite extraction in yeast using boiling buffered ethanol. *Yeast* 13 (14):1347-1355.

66. Goodacre R *et al* (2004) Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends in Biotechnology* 22 (5):245-252.

67. Gralla JD (2005) *Escherichia coli* ribosomal RNA transcription: regulatory roles for ppGpp, NTPs, architectural proteins and a polymerase-binding protein. *Molecular Microbiology* 55 (4):973-977.

68. Griffin JL (2004) Metabolic profiles to define the genome: can we hear the phenotypes? *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences* 359 (1446):857-871.

69. Gustafsson C, Govindarajan S, and Minshull J (2004) Codon bias and heterologous protein expression. *Trends Biotechnol* 22 (7):346-353.

70. Haddadin FT and Harcum SW (2005) Transcriptome profiles for high-cell-density recombinant and wild-type *Escherichia coli*. *Biotechnology and Bioengineering* 90 (2):127-153.

71. Haddadin FT, Kurtz H, and Harcum SW (2009) Serine hydroxamate and the transcriptome of high cell density recombinant *Escherichia coli* MG1655. *Applied Biochemistry and Biotechnology* 157 (2):124-139.

72. Hajjaj H *et al* (1998) Sampling techniques and comparative extraction procedures for quantitative determination of intra- and extracellular metabolites in filamentous fungi. *Fems Microbiology Letters* 164 (1):195-200.

73. Han MJ *et al* (2004) Roles and applications of small heat shock proteins in the production of recombinant proteins in *Escherichia coli*. *Biotechnology and Bioengineering* 88 (4):426-436.

74. Hannig G and Makrides S (1998) Strategies for optimizing heterologous protein expression in *Escherichia coli*. *Trends in Biotechnology* 16:54-60.

75. Harcum SW and Bentley WE (1999) Heat-shock and stringent responses have overlapping protease activity in *Escherichia coli*. Implications for heterologous protein yield. *Applied Biochemistry and Biotechnology* 80 (1):23-37.

76. Haseltin WA and Block R (1973) Synthesis of guanosine tetraphosphate and pentaphosphate requires presence of a codon-specific, uncharged transfer ribonucleic acid in acceptor site of ribosomes - (Stringent control ppGpp (Msi) and pppγpp (Msii) protein synthesis *Escherichia coli*). *Proceedings of the National Academy of Sciences of the United States of America* 70 (5):1564-1568.

77. Hoffmann F, Posten C, and Rinas U (2001) Kinetic model of in vivo folding and inclusion body formation in recombinant *Escherichia coli*. *Biotechnology and Bioengineering* 72 (3):315-322.

78. Hoffmann F and Rinas U (2004) Stress induced by recombinant protein production in *Escherichia coli*. *Advances in Biochemical Engineering / Biotechnology* 89:73-92.

79. Horai H *et al* (2010) MassBank: a public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry* 45 (7):703-714.

80. Ibarra RU, Edwards JS, and Palsson BO (2002) *Escherichia coli* K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth. *Nature* 420 (6912):186-189.

81. Jain V, Kumar M, and Chatterji D (2006) ppGpp: Stringent response and survival. *Journal of Microbiology* 44 (1):1-10.

82. Jensen KF and Pedersen S (1990) Metabolic growth rate control in *Escherichia coli* may be a consequence of subsaturation of the macromolecular biosynthetic apparatus with substrates and catalytic components. *Microbiological Reviews* 54 (2):89-100.

83. Jeong H *et al* (2000) The large-scale organization of metabolic networks. *Nature* 407 (6804):651-654.

84. Jinq Z *et al* (2006) Complex networks theory for analyzing metabolic networks. *Chinese Science Bulletin* 51 (13):1529-1537.

85. Jones KL, Kim SW, and Keasling JD (2000) Low copy plasmids can perform as well as or better than high copy plasmids for metabolic engineering of bacteria. *Metabolic Engineering* 2 (4):328-338.

86. Jonsson P *et al* (2005) High-throughput data analysis for detecting and identifying differences between samples in GC/MS-based metabolomic analyses. *Analytical Chemistry* 77 (17):5635-5642.

87. Jores L and Wagner R (2003) Essential steps in the ppGpp-dependent regulation of bacterial ribosomal RNA promoters can be explained by substrate competition. *Journal of Biological Chemistry* 278 (19):16834-16843.

88. Jurgen B *et al* (2000) Monitoring of genes that respond to overproduction of an insoluble recombinant protein in *Escherichia coli* glucose-limited fed-batch fermentations. *Biotechnology and Bioengineering* 70 (2):217-224.

89. Kabir MM and Shimizu K (2001) Proteome analysis of a temperature-inducible recombinant *Escherichia coli* for poly-beta-hydroxybutyrate production. *Journal of Bioscience and Bioengineering* 92 (3):277-284.

90. Kajitani M and Ishihama A (1984) Promoter selectivity of *Escherichia coli* Rna polymerase - Differential stringent control of the multiple promoters from ribosomal Rna and protein operons. *Journal of Biological Chemistry* 259 (3):1951-1957.

91. Kane JF (1995) Effects of rare codon clusters on high-level expression of heterologous proteins in *Escherichia coli*. *Current Opinion in Biotechnology* 6 (5):494-500.

92. Kanehisa M and Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* 28:27-30.

93. Kaspar H *et al* (2008) Automated GC-MS analysis of free amino acids in biological fluids. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences* 870 (2):222-232.

94. Kell DB (2004) Metabolomics and systems biology: making sense of the soup. *Current Opinion in Microbiology* 7 (3):296-307.

95. Keseler IM *et al* (2009) EcoCyc: a comprehensive view of *Escherichia coli* biology. *Nucleic Acids Research* 37 (Database issue):D464-D470.

96. Kim SY *et al* (2009) Improving the productivity of recombinant protein in *Escherichia coli* under thermal stress by coexpressing GroELS chaperone system. *Journal of Microbiology and Biotechnology* 19 (1):72-77.

97. Kitano H (2002) Systems biology: A brief overview. *Science* 295 (5560):1662-1664.

98. Koek MM *et al* (2006b) Microbial metabolomics with gas chromatography/mass spectrometry. *Analytical Chemistry* 78 (4):1272-1281.

99. Koek MM *et al* (2006a) Microbial metabolomics with gas chromatography/mass spectrometry. *Analytical Chemistry* 78 (4):1272-1281.

100. Kohda J *et al* (2002) Improvement of productivity of active form of glutamate racemase in *Escherichia coli* by coexpression of folding accessory proteins. *Biochemical Engineering Journal* 10 (1):39-45.

101. Kol S *et al* (2010) Metabolomic characterization of the salt stress response in *Streptomyces coelicolor*. *Applied and Environmental Microbiology* 76 (8):2574-2581.

102. Kopka J (2006) Current challenges and developments in GC-MS based metabolite profiling technology. *Journal of Biotechnology* 124 (1):312-322.

103. Kvint K *et al* (2003) The bacterial universal stress protein: function and regulation. *Current Opinion in Microbiology* 6 (2):140-145.

104. Kwon MJ *et al* (2002) Overproduction of *Bacillus macerans* cyclodextrin glucanotransferase in *E. coli* by coexpression of GroEL/ES chaperone. *Journal of Microbiology and Biotechnology* 12 (6):1002-1005.

105. Lee DH *et al* (2007) Proteome analysis of recombinant *Escherichia coli* producing human glucagon-like peptide-1. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences* 849 (1-2):323-330.

106. Lethanh H, Neubauer P, and Hoffmann F (2005) The small heat-shock proteins IbpA and IbpB reduce the stress load of recombinant *Escherichia coli* and delay degradation of inclusion bodies. *Microbial Cell Factories* 4 (1):6.

107. Lin HY *et al* (2004) Change of extracellular cAMP concentration is a sensitive reporter for bacterial fitness in high-cell-density cultures of *Escherichia coli*. *Biotechnology and Bioengineering* 87 (5):602-613.

108. Link H, Anselment B, and Weuster-Botz D (2008) Leakage of adenylates during cold methanol/glycerol quenching of *Escherichia coli*. *Metabolomics* 4 (3):240-247.

109. Llorach-Asuncion R *et al* (2010) Methodological aspects for metabolome visualization and characterization A metabolomic evaluation of the 24 h evolution of human urine after cocoa powder consumption. *Journal of Pharmaceutical and Biomedical Analysis* 51 (2):373-381.

110. Luli GW and Strohl WR (1990) Comparison of growth, acetate production, and acetate inhibition of *Escherichia coli* strains in batch and fed-batch fermentations. *Applied and Environmental Microbiology* 56:1004-1011.

111. Magnusson LU, Farewell A, and Nystrom T (2005) ppGpp: a global regulator in *Escherichia coli*. *Trends in Microbiology* 13 (5):236-242.

112. Mahadevan R, Edwards JS, and Doyle FJ, III (2002) Dynamic flux balance analysis of diauxic growth in *Escherichia coli*. *Biophysical Journal* 83 (3):1331-1340.

113. Maharjan RP and Ferenci T (2003) Global metabolite analysis: the influence of extraction methodology on metabolome profiles of *Escherichia coli*. *Analytical Biochemistry* 313 (1):145-154.

114. Majewski RA and Domach MM (1990) Simple constrained-optimization view of acetate overflow in *E. coli*. *Biotechnology and Bioengineering* 35:732-738.

115. Mashego MR *et al* (2007) Microbial metabolomics: past, present and future methodologies. *Biotechnology Letters* 29 (1):1-16.

116. Mazurie A *et al* (2010) Evolution of metabolic network organization. *BMC Systems Biology* 4:59.

117. Mendes P (2002) Emerging bioinformatics for the metabolome. *Briefings in Bioinformatics* 3 (2):134-145.

118. Mergulhao FJ, Summers DK, and Monteiro GA (2005) Recombinant protein secretion in *Escherichia coli*. *Biotechnology Advances* 23 (3):177-202.

119. Montanez R *et al* (2010) When metabolism meets topology: Reconciling metabolite and reaction networks. *Bioessays* 32 (3):246-256.

120. Murray KD and Bremer H (1996) Control of *spo*T-dependent ppGpp synthesis and degradation in *Escherichia coli*. *Journal of Molecular Biology* 259 (1):41-57.

121. Navid A and Almaas E (2009) Genome-scale reconstruction of the metabolic network in *Yersinia pestis*, strain 91001. *Molecular Biosystems* 5 (4):368-375.

122. Nogales J, Palsson BO, and Thiele I (2008) A genome-scale metabolic reconstruction of *Pseudomonas putida* KT2440: iJN746 as a cell factory. *BMC Systems Biology* 2:79.

123. Nystrom T (2004) Growth versus maintenance: a trade-off dictated by RNA polymerase availability and sigma factor competition? *Molecular Microbiology* 54 (4):855-862.

124. Oh MK and Liao JC (2000) DNA microarray detection of metabolic responses to protein overproduction in *Escherichia coli*. *Metabolic Engineering* 2 (3):201-209.

125. Oldiges M *et al* (2007) Metabolomics: current state and evolving methodologies and tools. *Applied Microbiology and Biotechnology* 76 (3):495-511.

126. Oliveira AP, Nielsen J, and Forster J (2005) Modeling *Lactococcus lactis* using a genome-scale flux model. *Bmc Microbiology* 5.

127. Orth JD and Palsson BO (2010) Systematizing the generation of missing metabolic knowledge. *Biotechnology and Bioengineering*.

128. Oshima T *et al* (2002) Transcriptome analysis of all two-component regulatory system mutants of *Escherichia coli* K-12. *Molecular Microbiology* 46 (1):281-291.

129. Ow DS *et al* (2009) Identification of cellular objective for elucidating the physiological state of plasmid-bearing *Escherichia coli* using genome-scale *in silico* analysis. *Biotechnology Progress* 25 (1):61-67.

130. Ozkan P *et al* (2005) Metabolic flux analysis of recombinant protein overproduction in *Escherichia coli*. *Biochemical Engineering Journal* 22 (2):167-195.

131. Pasikanti KK, Ho PC, and Chan ECY (2008) Gas chromatography/mass spectrometry in metabolic profiling of biological fluids. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences* 871 (2):202-211.

132. Paul BJ *et al* (2004a) DksA: A critical component of the transcription initiation machinery that potentiates the regulation of rRNA promoters by ppGpp and the initiating NTP. *Cell* 118 (3):311-322.

133. Paul BJ *et al* (2004b) rRNA transcription in *Escherichia coli*. *Annual Review of Genetics* 38:749-770.

134. Perederina A *et al* (2004) Regulation through the secondary channel-structural framework for ppGpp-DksA synergism during transcription. *Cell* 118 (3):297-309.

135. Peti W and Page R (2007) Strategies to maximize heterologous protein expression in *Escherichia coli* with minimal cost. *Protein Expression and Purification* 51 (1):1-10.

136. Pohjanen E *et al* (2006) Statistical multivariate metabolite profiling for aiding biomarker pattern detection and mechanistic interpretations in GC/MS based metabolomics. *Metabolomics* 2 (4):257-268.

137. Ponce E (1999) Effect of growth rate reduction and genetic modifications on acetate accumulation and biomass yields in *Escherichia coli*. *Journal of Bioscience and Bioengineering* 87 (6):775-780.

138. Pramanik J and Keasling JD (1997) Stoichiometric model of *Escherichia coli* metabolism: Incorporation of growth-rate dependent biomass composition and mechanistic energy requirements. *Biotechnology and Bioengineering* 56 (4):398-421.

139. Price ND *et al* (2003) Genome-scale microbial *in silico* models: the constraints-based approach. *Trends in Biotechnology* 21 (4):162-169.

140. Raamsdonk LM *et al* (2001) A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nature Biotechnology* 19 (1):45-50.

141. Raghavan A and Chatterji D (1998) Guanosine tetraphosphate-induced dissociation of open complexes at the *Escherichia coli* ribosomal protein promoters *rpl*J and *rps*A P1: Nanosecond depolarization spectroscopic studies. *Biophysical Chemistry* 75 (1):21-32.

142. Raman K and Chandra N (2009) Flux balance analysis of biological systems: applications and challenges. *Briefings in Bioinformatics* 10 (4):435-449.

143. Reed JL and Palsson BO (2003) Thirteen years of building constraint-based *in silico* models of *Escherichia coli*. *Journal of Bacteriology* 185 (9):2692-2699.

144. Reed JL *et al* (2003) An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biology* 4 (9):R54.

145. Roberts JW (2009) Promoter-specific control of *E. coli* RNA polymerase by ppGpp and a general transcription factor. *Genes & Development* 23 (2):143-146.

146. Saeed AI *et al* (2006) TM4 microarray software suite. *Methods in Enzymology* 411:134-193.

147. Sahdev S, Khattar SK, and Saini KS (2008) Production of active eukaryotic proteins through bacterial expression systems: a review of the existing biotechnology strategies. *Molecular and Cellular Biochemistry* 307 (1-2):249-264.

148. Saito N *et al* (2010) Unveiling cellular biochemical reactions via metabolomics-driven approaches. *Current Opinion in Microbiology* 13 (3):358-362.

149. Saito N *et al* (2009a) Metabolite profiling reveals YihU as a novel hydroxybutyrate dehydrogenase for alternative succinic semialdehyde metabolism in *Escherichia coli*. *Journal of Biological Chemistry* 284 (24):16442-16451.

150. Saito N *et al* (2009b) Metabolite profiling reveals YihU as a novel hydroxybutyrate dehydrogenase for alternative succinic semialdehyde metabolism in *Escherichia coli*. *Journal of Biological Chemistry* 284 (24):16442-16451.

151. Sajewicz M *et al* (2009) GC-MS study of the performance of different techniques for isolating the volatile fraction from Sage (*Salvia L.*) species, and comparison of seasonal differences in the composition of this fraction. *Acta Chromatographica* 21 (3):453-471.

152. San KY *et al* (2002) Metabolic engineering through cofactor manipulation and its effects on metabolic flux redistribution in *Escherichia coli*. *Metabolic Engineering* 4 (2):182-192.

153. Sanden AM *et al* (2003) Limiting factors in *Escherichia coli* fed-batch production of recombinant proteins. *Biotechnology and Bioengineering* 81 (2):158-166.

154. Saner U, Heinzle E, Bonvin D (1992) Computation of stoichiometric models for bioprocess.Oxford

155. Sansone SA *et al* (2007) The metabolomics standards initiative. *Nature Biotechnology* 25 (8):846-848.

156. Schaub J and Reuss M (2008) In vivo dynamics of glycolysis in *Escherichia coli* shows need for growth-rate dependent metabolome analysis. *Biotechnology Progress* 24 (6):1402-1407.

157. Schilling CH *et al* (2002) Genome-scale metabolic model of *Helicobacter pylori* 26695. *Journal of Bacteriology* 184 (16):4582-4593.

158. Schilling CH, Edwards JS, and Palsson BO (1999) Toward metabolic phenomics: Analysis of genomic data using flux balances. *Biotechnology Progress* 15 (3):288-295.

159. Schmidt FR (2004) Recombinant expression systems in the pharmaceutical industry. *Applied Microbiology and Biotechnology* 65 (4):363-372.

160. Schweder T, Hofmann K, and Hecker M (1995) *Escherichia coli* K12 *rel*A strains as safe hosts for expression of recombinant DNA. *Applied Microbiology and Biotechnology* 42 (5):718-723.

161. Schweder T *et al* (2002) Role of the general stress response during strong overexpression of a heterologous gene in *Escherichia coli*. *Applied Microbiology and Biotechnology* 58 (3):330-337.

162. Seo JH, Kang DG, and Cha HJ (2003) Comparison of cellular stress levels and green-fluorescent-protein expression in several *Escherichia coli* strains. *Biotechnology and Applied Biochemistry* 37 (Pt 2):103-107.

163. Sheikh K, Forster J, and Nielsen LK (2005) Modeling hybridoma cell metabolism using a generic genome-scale metabolic model of *Mus musculus*. *Biotechnology Progress* 21 (1):112-121.

164. Shiloach J *et al* (1996) Effect of glucose supply strategy on acetate accumulation, growth, and recombinant protein production by *Escherichia coli* BL21 (λDE3) and *Escherichia coli* JM109. *Biotechnology and Bioengineering* 49 (4):421-428.

165. Smart KF et al.Analytical platform for metabolome analysis of microbial cells using gas chromatography-mass spectrometry (GC-MS). Nature Protocols (in press)

166. Smith CA *et al* (2005) METLIN: a metabolite mass spectral database. *Therapeutic Drug Monitoring* 27 (6):747-751.

167. Sorensen HP and Mortensen KK (2005) Advanced genetic strategies for recombinant protein expression in *Escherichia coli*. *Journal of Biotechnology* 115 (2):113-128.

168. Spura J *et al* (2009) A method for enzyme quenching in microbial metabolome analysis successfully applied to gram-positive and gram-negative bacteria and yeast. *Analytical Biochemistry* 394 (2):192-201.

169. Srinivasan S, Barnard GC, and Gerngross TU (2003) Production of recombinant proteins using multiple-copy gene integration in high-cell-density fermentations of *Ralstonia eutropha*. *Biotechnology and Bioengineering* 84 (1):114-120.

170. Srivatsan A and Wang JD (2008) Control of bacterial transcription, translation and replication by (p)ppGpp. *Current Opinion in Microbiology* 11 (2):100-105.

171. Stein SE (1995) Chemical substructure identification by mass spectral library searching. *Journal of the American Society for Mass Spectrometry* 6 (8):644-655.

172. Stein SE (1999) An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *Journal of the American Society for Mass Spectrometry* 10 (8):770-781.

173. Stephanopoulos G, Aristidou A, Nielsen J (1998) Metabolic engineering. Academic Press: San Diego.

174. Stephens JC, Artz SW, and Ames BN (1975) Guanosine 5'-diphosphate 3'-diphosphate (ppGpp) - Positive effector for histidine operon transcription and general signal for amino acid deficiency. *Proceedings of the National Academy of Sciences of the United States of America* 72 (11):4389-4393.

175. Steuer R *et al* (2003) Observing and interpreting correlations in metabolomic networks. *Bioinformatics* 19 (8):1019-1026.

176. Suarez DC and Kilikian BV (2000) Acetic acid accumulation in aerobic growth of recombinant *Escherichia coli*. *Process Biochemistry* 35 (9):1051-1055.

177. Taymaz-Nikerel H *et al* (2009) Development and application of a differential method for reliable metabolome analysis in *Escherichia coli*. *Analytical Biochemistry* 386 (1):9-19.

178. Thomas JG and Baneyx F (1996) Protein misfolding and inclusion body formation in recombinant *Escherichia coli* cells overexpressing Heat-shock proteins. *Journal of Biological Chemistry* 271 (19):11141-11147.

179. Trautinger BW *et al* (2005) RNA polymerase modulators and DNA repair activities resolve conflicts between DNA replication and transcription. *Molecular Cell* 19 (2):247-258.

180. Traxler MF, Chang DE, and Conway T (2006) Guanosine 3 ',5 '-bispyrophosphate coordinates global gene expression during glucose-lactose diauxie in *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America* 103 (7):2374-2379.

181. Turner C, Gregory ME, and Turner MK (1994) A study of the effect of specific growth rate and acetate on recombinant protein production of *Escherichia coli* JM107. *Biotechnology Letters* 16 (9):891-896.

182. Tweeddale H, Notley-McRobb L, and Ferenci T (1998) Effect of slow growth on metabolism of *Escherichia coli*, as revealed by global metabolite pool ("metabolome") analysis. *J Bacteriol* 180 (19):5109-5116.

183. Urbanczik R and Wagner C (2005) Functional stoichiometric analysis of metabolic networks. *Bioinformatics* 21 (22):4176-4180.

184. van der Werf MJ (2003) Metabolomics: A revolutionary tool to optimize microbial processes. *Abstracts of Papers of the American Chemical Society* 226:U84.

185. Varma A and Palsson BO (1993) Metabolic capabilities of *Escherichia coli*: II. Optimal growth patterns. *Journal of Theoretical Biology* 165:503-522.

186. Varma A and Palsson BO (1994) Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. *Applied and Environmental Microbiology* 60:3724-3731.

187. Varner JD (2000) Large-scale prediction of phenotype: Concept. *Biotechnology and Bioengineering* 69 (6):664-678.

188. Villas-Boas SG and Bruheim P (2007) Cold glycerol-saline: the promising quenching solution for accurate intracellular metabolite analysis of microbial cells. *Analytical Biochemistry* 370 (1):87-97.

189. Villas-Boas SG *et al* (2005) High-throughput metabolic state analysis: the missing link in integrated functional genomics of yeasts. *Biochemical Journal* 388:669-677.

190. Vinella D *et al* (2005) Iron limitation induces SpoT-dependent accumulation of ppGpp in *Escherichia coli*. *Molecular Microbiology* 56 (4):958-970.

191. Wang YH *et al* (2005) Proteomic profiling of *Escherichia coli* proteins under high cell density fed-batch cultivation with overexpression of phosphogluconolactonase. *Biotechnology Progress* 21 (5):1401-1411.

192. Weber J, Hoffmann F, and Rinas U (2002) Metabolic adaptation of *Escherichia coli* during temperature-induced recombinant protein production: 2. Redirection of metabolic fluxes. *Biotechnology and Bioengineering* 80 (3):320-330.

193. Weuster-Botz D (1997) Sampling tube device for monitoring intracellular metabolite dynamics. *Analytical Biochemistry* 246 (2):225-233.

194. Wishart DS *et al* (2009) HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Research* 37 (Database issue):D603-D610.

195. Wittmann C *et al* (2004) Impact of the cold shock phenomenon on quantification of intracellular metabolites in bacteria. *Analytical Biochemistry* 327 (1):135-139.

196. Wong MS *et al* (2008) Reduction of acetate accumulation in *Escherichia coli* cultures for increased recombinant protein production. *Metabolic Engineering* 10 (2):97-108.

197. Xia J *et al* (2009) MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Research* 37 (Web Server issue):W652-W660.

198. Xia J and Wishart DS (2010b) MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Res* 38 Suppl:W71-W77.

199. Xia J and Wishart DS (2010a) MetPA: a web-based metabolomics tool for pathway analysis and visualization. *Bioinformatics*.

200. Xiao H *et al* (1991) Residual guanosine 3',5'-bispyrophosphate synthetic activity of *rel*A null mutants can be eliminated by *spo*T null mutations. *Journal of Biological Chemistry* 266 (9):5980-5990.

201. Yokoyama K, Kikuchi Y, and Yasueda H (1998) Overproduction of DnaJ in Escherichia coli improves in vivo solubility of the recombinant fish-derived transglutaminase. *Bioscience Biotechnology and Biochemistry* 62 (6):1205-1210.

202. Yuan J *et al* (2009) Metabolomics-driven quantitative analysis of ammonia assimilation in *E. coli*. *Molecular Systems Biology* 5:302.

203. Yuliana ND *et al* (2010) Metabolomics for bioactivity assessment of natural products. *Phytotherapy Research*.

204. Zhang W, Li F, and Nie L (2010) Integrating multiple 'omics' analysis for microbial biology: application and methodologies. *Microbiology* 156 (Pt 2):287-301.

205. Zhou YN and Jin DJ (1998) The *rpo*B mutants destabilizing initiation complexes at stringently controlled promoters behave like "stringent" RNA polymerases in *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America* 95 (6):2908-2913.

206. Zouridis H and Hatzimanikatis V (2007) A model for protein translation: polysome self-organization leads to maximum protein synthesis rates. *Biophysical Journal* 92 (3):717-730.

CHAPTER 2

STRINGENT RESPONSE OF *ESCHERICHIA COLI*:

REVISITING THE BIBLIOME USING LITERATURE MINING

> *"Nearly 40 years ago two spots appeared on autoradiograms, as if by magic..."*
>
> By K. Potrykus and M. Cashel in (p)ppGpp: Still Magical?,
> Annu. Rev. Microbiol. 2008. 62:35-51

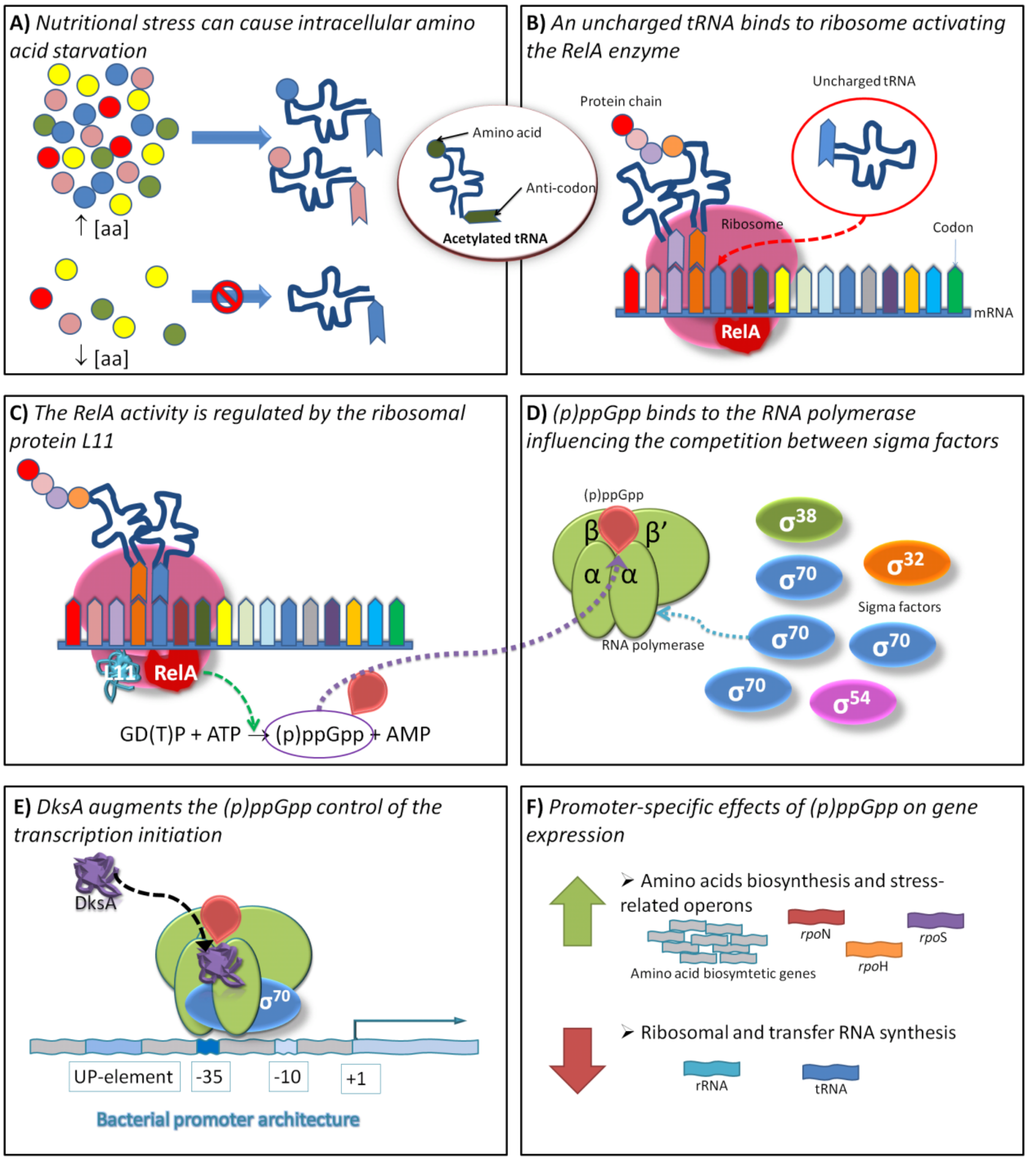
2.1 ABSTRACT

This work aims to present a novel approach to the large-scale compilation, processing, and analysis of literature on the stringent response in *Escherichia coli*. This cellular response consists of a range of (p)ppGpp-induced cellular activities triggered by amino acid depletion, and involving a multitude of molecular components, namely genes, tRNAs, mRNAs, gene products and small molecules, that ensure a coordinated and effective response. Specialised controlled vocabulary supported the automatic recognition of molecular components whereas statistical co-occurrence analysis suggested the most likely to be biologically engaged. As a result, the cellular processes affected by the activity of (p)ppGpp nucleotides were identified and further investigated, complementing existing reviews on this pleiotropic cellular response. The RelA and SpoT enzymes that control the basal levels of (p)ppGpp nucleotides and the RNA polymerase to which these nucleotides bind, were the most represented components. However, the identification of less annotated components revealed that some (p)ppGpp-induced functional activities are still unclear or largely overlooked. The suggested literature mining approach offers a more comprehensive analysis of the stringent response in *E. coli*, enhancing the process of compiling and processing relevant literature, as well as enabling the incremental extension of the knowledge base, following up research breakthroughs.

2.2 INTRODUCTION

The survival capacities of all living organisms are dependent on the ability to sense and respond to environmental changes. Organisms can face several stress conditions, such as nutritional and energetic starvation, excess of toxic substances and presence of inhibitory agents, all of which can affect normal growth (Nystrom T 1999; Mukherjee TK *et al.* 1998). The identification of the specific mechanisms involved in microbial survival under stress conditions is expected to provide insight into stress response systems across life forms. In particular, the stringent response of the *E. coli* serves as a paradigm for understanding global responses to sudden nutritional starvation (Chatterji D and Ojha AK 2001; Durfee T *et al.* 2008). Moreover, there is an industrial motivation for understanding the mechanisms of the stringent response in *E. coli*, since it is sometimes triggered during recombinant protein production, leading to decreased productivities (Harcum SW 2002; Haddadin FT *et al.* 2009).

In recent years, numerous publications have discussed the mechanisms involved in this response (Jain V *et al.* 2006; Battesti A and Bouveret E 2009; Srivatsan A and Wang JD 2008; Chang DE *et al.* 2002; Murray KD and Bremer H 1996; Xiao H *et al.* 1991). Studies indicate that the accumulation of unusual guanosine nucleotides, collectively called (p)ppGpp, is the hallmark of the stringent response of *E. coli* (Magnusson LU *et al.* 2005) (Figure 2.1). Such accumulation is known to be controlled by the activity of two enzymes, RelA and SpoT. Upon depletion of amino acids, the ribosome-bound RelA enzyme (ppGpp synthetase I) is induced to synthesise (p)ppGpp nucleotides when an uncharged tRNA binds to the acceptor site of the translating ribosome (Torok I and Kari C 1980). In turn, the bifunctional SpoT enzyme (ppGpp synthetase II), which also possesses weak synthetase activity, is responsible for maintaining the intracellular levels of (p)ppGpp nucleotides via enzymatic degradation (Johnson GS *et al.* 1979).



A) *Nutritional stress can cause intracellular amino acid starvation*

B) *An uncharged tRNA binds to ribosome activating the RelA enzyme*

C) *The RelA activity is regulated by the ribosomal protein L11*

D) *(p)ppGpp binds to the RNA polymerase influencing the competition between sigma factors*

E) *DksA augments the (p)ppGpp control of the transcription initiation*

F) *Promoter-specific effects of (p)ppGpp on gene expression*

➢ Amino acids biosynthesis and stress-related operons
rpoN
rpoS
rpoH
Amino acid biosymtetic genes
➢ Ribosomal and transfer RNA synthesis
rRNA
tRNA

Figure 2.1. The (p)ppGpp-mediated stringent response.

(A) Low amino-acid concentrations lead to decreased charging of the corresponding tRNAs. (B) The translational machinery depends on the translocation along the mRNA whereby a new acetylated-tRNA is positioned in the ribosome. Whenever an uncharged tRNA binds to the ribosome, the elongation of the polypeptide chain is stalled. (C) The stringent factor RelA is then activated in the presence of the ribosomal protein L11, catalyzing the synthesis of (p)ppGpp nucleotides. (D) These nucleotides bind directly to the RNA polymerase and affect the binding abilities of sigma factors to the core RNA polymerase. (E) The co-factor DksA also binds to the RNA polymerase and augments the (p)ppGpp regulation of the transcription initiation at certain σ70-dependent promoters, functioning both as negative and positive regulators. (F) These regulators change the gene expression: (i) decreasing the transcription activity of genes involved in the translational activity; (ii) and increasing the transcription of stress-related operons and genes encoding for enzymes needed for the synthesis and the transport of amino acids.

The (p)ppGpp-mediated response involves the control of the genetic expression by direct interaction of the (p)ppGpp nucleotides with the RNA polymerase (RNAP) (Chatterji D *et al.* 1998; Artsimovitch I *et al.* 2004). Typically, transcription depends on the recognition of the promoter elements by the complex RNAP-sigma (σ) factor and the DNA-binding activity of transcription factors. During stringent response (p)ppGpp nucleotides, in conjunction with the DksA protein, bind to RNAP activating the transcription of the genes coding for stress-associated sigma factors and amino acid biosynthesis and inhibiting the transcription of stable RNAs (rRNA and tRNA) (Paul BJ *et al.* 2004).

This (p)ppGpp-mediated scenario is quite complex and many fundamental details, such as the mechanisms underlying the activation of transcription by (p)ppGpp, remain unknown or uncertain (Potrykus K and Cashel M 2008; Wu J and Xie J 2009). Experimental results keep being produced, and therefore continuous and incremental literature review is required to gain a better understanding of the process. In this work, we propose the development of a literature mining approach that complements manual literature review. This semi-automatic information extraction approach aims at speeding up the review process, but, far more important, takes advantage of public database information and ontology assignments to provide for large-scale enrichment and contextualisation of textual evidences. Manual curation was on top of the automatic annotation process and addressed the validity and consistency of the extracted information.

A corpus, i.e., a set of documents, related to the stringent response of *E. coli* (published till 2009) was retrieved using NCBI PubMed tools. EcoCyc database (Keseler IM *et al.* 2009), a key resource for *E. coli* studies, provided for most of the controlled vocabulary used for the identification of relevant entities in the documents, namely genes, gene products and small molecules. Moreover, EcoCyc gene and gene product assignments to Gene Ontology (GO) (Ashburner M *et al.* 2000) and MultiFun ontology (Serres MH and Riley M 2000) enabled the annotation of biological processes and molecular functions. Additionally, the Proteomics Standards Initiative-Molecular Interactions (PSI-MI) ontology (Hermjakob H *et al.* 2004) supported the annotation of experimental techniques.

The analysis of corpus annotations aimed at (i) corroborating existing knowledge about key players and processes involved in the stringent response of *E. coli* and (ii) unveiling knowledge that has been overlooked in the existing reviews. Analysis was based on the assumption that entities, processes and functions of interest could be identified by finding those terms (i.e. the set of name variants that identify a given biological entity or experimental technique) that were significantly present among all terms found on the corpus. As such, the frequency and variance of term annotation supported the identification of key players, whereas GO and MultiFun assignments to gene products assisted on the inspection of the involved biological processes. Finally, a decade-by-decade retrospective analysis was performed to evaluate the influence of technology advances on these findings and to get a broader perspective on the current knowledge of the stringent response.

2.3 EXPERIMENTAL PROCEDURES

The semi-automatic information extraction approach designed to review existing literature on the stringent response of *E. coli*, which integrated automatic document retrieval and entity recognition processes, manual curation and corpus analysis, is shown in Figure 2.2.

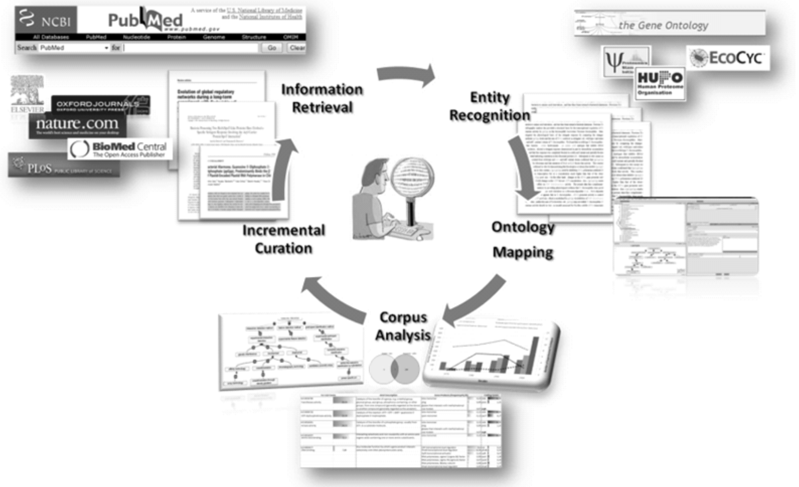


Figure 2.2. Semi-automatic information extraction approach.

The first step encompasses the retrieval of relevant documents that are then processed to recognize biological entities and map ontological concepts. The corpus analysis enables the identification of key players or significant information by an incremental curation that can further deliver information for retrieving new relevant documents.

2.3.1 Semi-automatic information extraction approach

A keyword-based search in PubMed, using the query ((“*Escherichia coli*” OR “*E. coli*”) AND “stringent response”) was used to compile the set of documents analysed in this work. @Note (Lourenco A *et al.* 2009), a workbench for Biomedical Text Mining enabled the use of common literature mining techniques, namely dictionary and rule-based techniques, in the recognition and annotation of genetic components, gene products and small molecules. Its regular expression module enabled the identification of genes and proteins that adhere to standard gene and protein naming conventions for *E. coli* (e.g. three lower case letters followed by a fourth letter in upper case or a term consisting of four digits preceded by character ‘b’ are candidates for gene names).

The annotated corpus was stored into XML files for enabling further computer processing. Manual curation was regarded as the second stage of annotation and aimed at validating automatic annotations and coping with automatic annotation flaws and limitations. Each XML file was manipulated by the curator using the @Note’s manual curation environment. The curator was able to add, remove and correct annotations, evaluating possible improvements in the entity recognition process, namely the refinement of the vocabulary and the adjustment of the set of regular expressions.

2.3.2 Controlled vocabulary

EcoCyc database (version 13.0, released in March 2009) provided for most of the controlled vocabulary. It supported the automatic identification of genetic components, gene products and small molecules as follows: common names and extensive name variants (synonyms) were used for recognising all entities in the corpus; entity name variants were normalised by associating the corresponding database record identifier to the annotation; and database assignments to GO and MultiFun categories enabled the mapping of annotated terms to the associated molecular functions and biological processes. Additional vocabulary was extracted from PSI-MI ontology for the annotation of experimental techniques.

2.3.3 Analysis methodology

The number of annotations of a term (or entity), the number of documents that contributed for those annotations and the number of documents composing the corpus constituted the baseline of the statistical metrics used in the analyses. Let *D* be the set of documents in the corpus and *T* be the set of annotated terms in *D*. For every $t_i \in T$, the frequency, the mean, the standard deviation of annotation, and the variance-to-the mean ratio (or coefficient of dispersion) were computed as described in Table 2.1.

Table 2.1. Annotation statistics used in the analysis.

Frequency:	$freq_{t_i} = \frac{D_{t_i}}{D} = \frac{\#docs_{t_i}}{\#docs}$	(Eq. 1)
Mean:	$mean_{t_i} = \mu_{t_i} = \frac{\#annots_{t_i}}{\#docs_{t_i}}$	(Eq. 2)
Standard deviation:	$std_{t_i} = \sigma_{t_i} = \sqrt{1/\#docs_{t_i} \sum_{j=1}^{\#docs_{t_i}} (\#annots_{t_i,doc_j} - \mu_{t_i})^2}$	(Eq. 3)
Variance-to-mean:	$VMR_{t_i} = \frac{\sigma_{t_i}^{\;2}}{\mu_{t_i}}$	(Eq. 4)
Co-annotation:	$freq_{t_i,t_j} = \frac{D_{t_i \cap t_j}}{D_{t_i}} = \frac{\#docs_{t_i \cap t_j}}{\#docs_{t_i}}$	(Eq. 5)

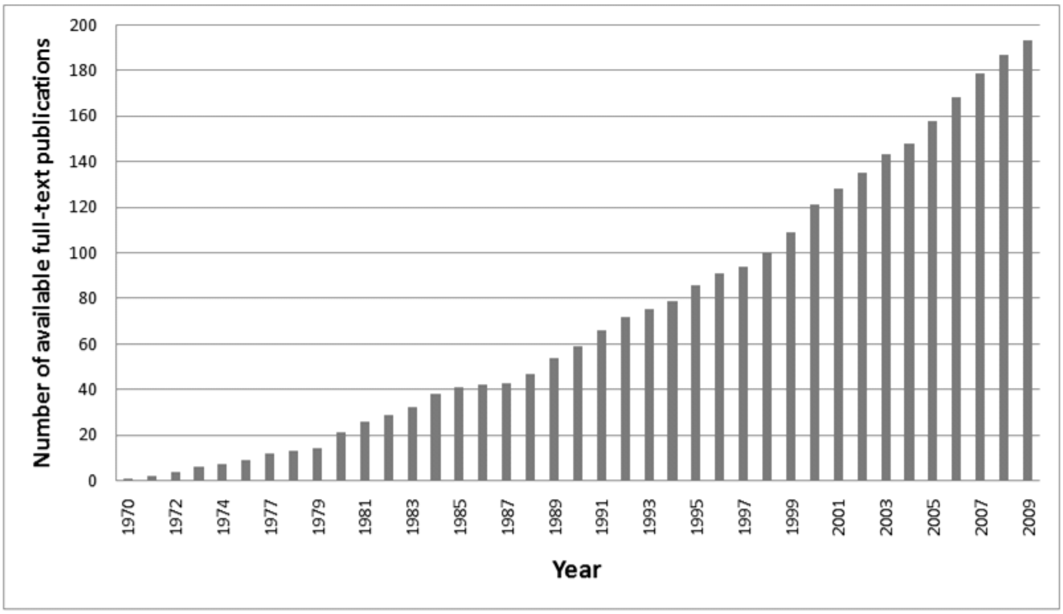
The frequency of annotation, *freq*$_{ti}$, is the fraction of documents in *D* that refers to the term t_i. In turn, the mean, μ_{ti}, and the standard deviation, σ_{ti}, weight the number of annotations of a term, *#annots*$_{ti}$, in the documents in *D* that include those annotations, *docs*$_{ti,}$ and measure the average or dispersion of the annotations, respectively. The mean indicates the representativeness of the term in the subset *docs*$_{ti}$ whereas the standard deviation indicates the variability of annotation in the subset. The variance-to-mean ratio (also called index of dispersion), *VMR*$_{ti}$, is a quantitative measure of the degree of clustering of term annotations. A ratio that is greater than 1 indicates a clustered distribution, i.e., term annotations are unevenly distributed in the subset $docs_{ti}$; less than 1 indicates an evenly dispersed distribution, i.e., term annotations are evenly distributed in the subset $docs_{ti}$; equal to 1, a random distribution; and, equal to 0, indicates a constant distribution, i.e., the number of term annotations is the same in all documents that refer to the term. Finally, the frequency of co-annotation relates two different entities, assuming that entities that are often co-annotated (in the same document) are biologically engaged. The frequency of co-annotation was estimated and these interactions were illustrated using the Cytoscape biomolecular interaction viewer and analyser (Shannon P *et al.* 2003).

The statistics of the ontology assignments of the annotated terms were studied for process and function analysis. The functions of the annotated gene products and the involved biological processes are reflected in their GO and MultiFun annotations and thus, processes and functions of interest were identified by finding statistically enriched terms (mainly by looking into the frequency of annotation). Similar assessments were taken over PSI-MI assignments towards the identification of the techniques that have contributed the most to the study of the stringent response. Apart from the systematic analysis of the set of annotations in corpus, a retrospective analysis of annotations per decade was also undertaken (i.e. frequency of annotation per decade). Such analysis aimed at looking into the evolving experimental techniques that contributed to the study of the stringent response over the decades and, in particular, evaluating the impact that the technological evolution had in the identification of molecular participants.

Finally, a set of recent documents that review the literature on the subject were manually retrieved from PubMed. Their contents were evaluated in terms of annotations of genetic components, gene products and small molecules and further compared to the annotations retrieved from the corpus.

2.4 RESULTS

Since the aim of this approach was to extract detailed information on the stringent response, the process of document retrieval was focused on full-texts. A keyword-based search on PubMed retrieved a total of 251 documents, from which 231 documents had full-texts available under different subscription policies. Open-access journals provided for 129 documents whereas 64 documents were retrieved using team's institutional journal subscriptions. At the end, an overall of 193 full-text documents, most of which published in recent years, were available for annotation (Figure 2.3).



	Decades			
	1970	1980	1990	2000
Number of documents	14	40	55	84

Figure 2.3. Number of retrieved full-text publications per year (accumulated).

The automatic identification of biological entities in the documents, so-called entity recognition, was based on EcoCyc controlled vocabulary for major biological classes and involved the recognition of genetic components (genes, RNA and DNA molecules), gene products (proteins, including transcription factors and enzymes) and small molecules. Additionally, a hand-crafted dictionary supported the recognition of experimental techniques and their association to PSI-MI ontology concepts. The manual curation process consisted on reviewing document annotations, i.e., the text markups (XML tags) for recognised entities, to ensure the corpus quality and consistency. Errors of the automated recognition process such as the annotation of false entities (e.g. the words 'release' or 'crease' were annotated as enzymes based on common enzyme suffix 'ase'), homonyms (e.g. the same term 'elongation factor Tu' to designate two different polypeptides 'TufA' and 'TufB') and PDF-to-text format conversion typos (e.g. '4azaleucine' and '9galactosidase' were corrected to '4-azaleucine' and 'β-galactosidase', respectively) were manually curated. After manual curation, the corpus consisted of 93893 annotations for 2474 entities.

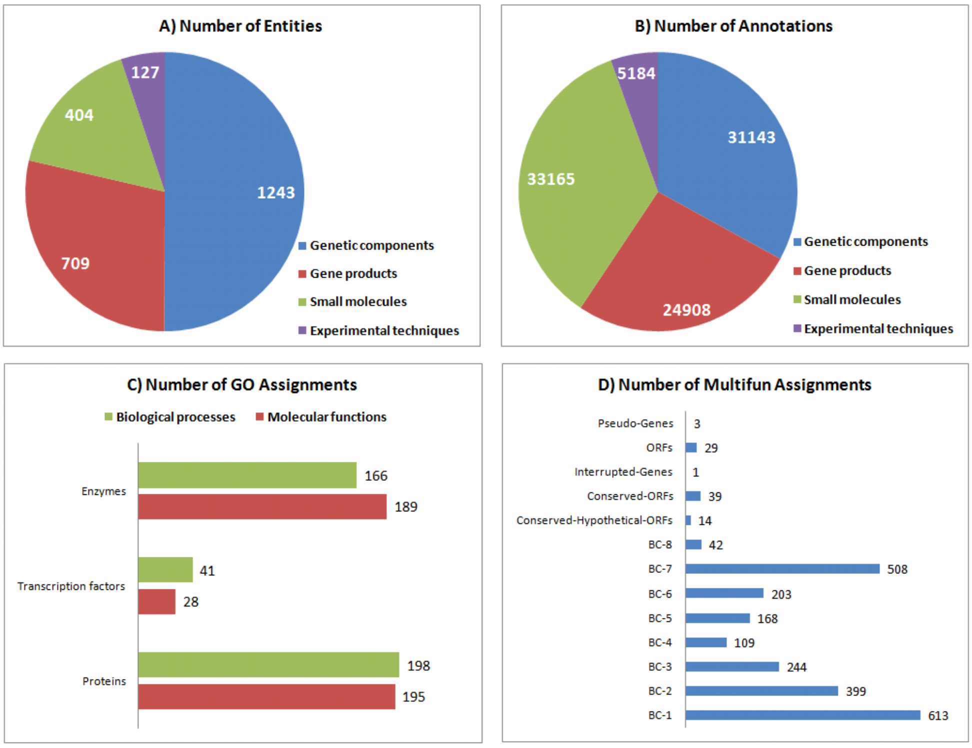


Figure 2.4. Corpus annotation contents.

Overview of the extent of entities (A) and entity annotations (B) corresponding to each class. GO assignments (C) for molecular functions and biological processes mapped for each set of gene products (i.e. enzymes, transcription factors and other proteins) and MultiFun gene assignments (D) for different functional roles (BC-1 to Metabolism, BC-2 to Information transfer, BC-3 to Regulation, BC-4 to Transport, BC-5 to Cell processes, BC-6 to Cell structure, BC-7 to Location of gene products and BC-8 to Extrachromosomal origin) were recognized in the corpus.

Gene and gene products ontology assignments to MultiFun and GO were straightforward since all entries in the EcoCyc dictionary keep the corresponding database assignments to these ontologies. Similarly, ontology assignments to PSI-MI concerning laboratory techniques were enabled by the hand-crafted dictionary. Annotations were distributed as follows (Figure 2.4): (i) most of the annotated entities represent genetic components (50% of the entities), whereas small molecules accounted for the largest number of annotations (35% of the overall number of annotations); (ii) enzymes and proteins contributed to most GO assignments; and (iii) the MultiFun cellular function categories 'Metabolism'(BC-1) and 'Location of gene products' (BC-7) related to most of the annotated genes.

The analysis of the corpus was based in the assumption that relevant entities could be identified by finding frequent terms. Thus, the analysis was primarily based on the number of documents where each entity was annotated and the number of entity annotations per document. First, the frequency of annotation ($freq_{ti}$) assessed the recurrent mention of an entity throughout the corpus and the mean ($mean_{ti}$) and standard deviation (std_{ti}) of annotation showed its relevance in the discussion. Then, the variance-to-mean ratio (VMR_{ti}) indicated the existence of document clusters, which put emphasis on particular entities that are much less discussed on the rest of the documents. Additionally, the frequency of co-annotation ($freq_{ti,tj}$) was calculated for highly representative entities (entities with high frequency and mean) to identify those entities that were most probably engaged to them.

To understand the evolving of the topic throughout the years, and in particular the impact of technology-driven advances, this analysis was extended to the study of annotations per decade. The analysis of gene and gene product assignments to MultiFun and GO ontologies pursued the comprehension of the molecular functions and biological processes involved in the stringent response of *E. coli*.

2.4.1 BIOLOGICAL ENTITIES

The systems-level investigation of the stringent response involved the search for three main biological classes: genetic components (genes, RNA and DNA molecules), gene products (proteins, transcription factors and enzymes) and small molecules.

Table 2.2. Annotations of the genetic components in the corpus.

Individual genetic components (i.e. genes, DNAs and RNAs) were evaluated considering the number of documents where these entities were annotated and their number of annotations in the corpus. Statistical measurements are detailed in the Methods and Materials section. *VMR*: variance-to-mean; *Std*: standard deviation.

Class	Term	Number of Annotations	Number of Documents	Frequency (%)$^{\Psi}$	*Mean*	*Std*	*VMR*
Gene	*rel*A	3163	138	71.50	22.92	27.23	33.14
	*spo*T	1315	88	45.60	14.94	27.42	52.07
	lac	354	63	32.64	5.620	19.42	72.20
	*lac*Z	534	50	25.91	10.68	17.16	28.90
	thi	91	47	24.35	1.940	0.050	4.000
	rel	523	47	24.35	11.13	20.68	36.36
	*rec*A	82	39	20.21	2.100	1.810	0.5000
	*rps*L	95	36	18.65	2.640	3.530	4.500
	thr	84	36	18.65	2.330	3.760	4.500
	*rps*G	103	34	17.62	3.030	7.250	16.33
	leu	98	34	17.62	2.880	6.800	18.00
	*rpo*S	205	33	17.10	6.210	10.83	16.67
	kan	308	33	17.10	9.330	16.61	28.44
	*gln*V	42	31	16.06	1.350	0.7400	0
	*rpo*B	389	30	15.54	12.97	17.60	24.08
	*pts*G	240	30	15.54	8.000	21.54	55.13
	trp	144	25	12.95	5.760	14.73	39.20
	*car*A	60	20	10.36	3.000	3.810	3.000
	*hsd*R	23	19	9.840	1.210	0.5600	0
DNA	DNA	1839	137	70.98	13.42	16.31	19.69
	plasmid DNA	193	36	18.65	5.360	12.31	28.80
	chromosomal DNA	63	24	12.44	2.630	2.440	2.000
	cDNA	125	23	11.92	5.430	5.820	5.000
RNA	RNA	4193	140	72.54	29.95	38.21	49.79
	uncharged tRNA	1168	117	60.62	9.980	19.64	40.11
	rRNA	1116	97	50.26	11.51	25.97	56.82
	a mRNA	999	91	47.15	10.98	19.52	36.10
	rrnA	911	87	45.08	10.47	22.51	48.40
	stable RNA	430	87	45.08	4.940	8.030	16.00
	a charged tRNA	140	43	22.28	3.260	4.200	5.330
	rrnB	301	26	13.47	11.58	19.30	32.82
	rrn	321	26	13.47	12.35	30.42	75.00
	16s-rRNAs	156	25	12.95	6.240	9.090	13.50

$^{\Psi}$A threshold of 10% of the frequency of annotation was set for each genetic component category.

The analysis of the frequency of annotation of genetic components (see Table 2.2) evidenced that entities like the *rel*A gene and the RNA and DNA molecules were annotated in more than 70% of the documents. Though the representativeness of such entities in these documents was considerably high (i.e. high mean of annotation), the annotations were over-dispersed (*VMR*>1). For example, the *rel*A gene has a mean of over 22 annotations per document and a *VMR* of over 33, meaning that a small part of the documents presents a very high number of annotations, which suggests that these are focused on the discussion of the role of this gene in the stringent response.

Similarly, the analysis of gene product annotations (see Table 2.3) exposed RelA, RNAP and ribosomes as highly annotated entities (present in over than 50% of the documents) with a considerable degree of over-dispersion (*VMR*>1). More interesting, the Fis transcriptional dual regulator, which modulates several cellular processes, such as the transcription of stable RNA (PMID: 2209559; PMID: 9973355)[1] (Ross W *et al.* 1990; Walker KA *et al.* 1999), presented a low frequency of annotation (less than 10% of the documents), but was highly annotated in the associated documents (with a mean of almost 50 annotations per document). An extremely high value of *VMR* (over 150) pointed out that some of these documents are really devoted to the discussion of this biological entity.

[1] The PubMed Unique Identifiers (PMIDs) indicate which documents from the corpus supported the evidences.

Table 2.3. Annotations of the gene products in the corpus.

Individual gene products (i.e. enzymes, transcription factors and other proteins) were evaluated considering the number of documents where these entities were annotated and their number of annotations in the corpus. Statistical measurements are detailed in the Methods and Materials section.

Class	Term	Number of Annotations	Number of Documents	Frequency (%)[Ψ]	Mean	*Std*	*VMR*
Proteins	Ribosome	1643	128	66.32	12.84	23.57	44.08
	Rel	1021	62	32.12	16.50	36.60	81.00
	LacZ	543	53	27.46	10.30	17.44	28.90
	Sigma 38 factor	392	42	21.76	9.330	15.40	25.00
	Sigma factor	112	35	18.13	3.200	5.870	8.330
	UvrD	56	35	18.13	1.600	1.300	1.000
	RpoB	252	35	18.13	7.200	11.50	17.29
	RecA	99	31	16.06	3.190	4.260	5.330
	EF-Tu	223	26	13.47	8.580	17.32	36.13
	Der	51	25	12.95	2.040	2.140	2.000
	Sigma 70 factor	134	21	10.88	6.380	11.19	20.17
Transcription factors	Fis	888	18	9.330	49.33	86.88	150.9
	Fur	56	13	6.740	4.310	9.260	20.25
	CRP	279	12	6.220	23.25	36.28	56.35
	DnaA	121	11	5.700	11.00	23.00	48.09
	H-NS	73	11	5.700	6.640	10.73	16.67
	LexA	101	10	5.180	10.10	18.32	32.40
	IHF	54	9	4.660	6.000	5.250	4.170
Enzymes	RelA	4138	152	78.76	27.22	31.16	35.59
	RNAP	1873	117	60.62	16.01	28.08	49.00
	SpoT	1024	60	31.09	17.07	42.19	103.8
	EcoRI	215	53	27.46	4.060	4.970	4.000
	β-galactosidase	294	47	24.35	6.260	6.550	6.000
	BamHI	149	43	22.28	3.470	5.870	8.330
	HindIII	114	41	21.24	2.780	2.160	2.000
	RNase	109	36	18.65	3.030	4.280	5.330
	YbcS	50	23	11.92	2.170	2.620	2.000
	Reverse transcriptase	34	21	10.88	1.620	1.050	1.000
	tRNA synthetase	54	20	10.36	2.700	2.630	2.000
	Endonuclease I	29	20	10.36	1.450	1.400	1.000

[Ψ] A threshold of 10% of the frequency of annotation was set for enzymes and other proteins, whereas a threshold of 5% was set for transcription factors. *VMR*: variance-to-mean; *Std*: standard deviation.

The analysis of the annotation of small molecules (see Table 2.4) revealed that, though almost 83% of the documents discussed the general role of amino acids and nucleotides, the mean of annotation of specific nucleotides and amino acids was quite low (less than 10 annotations per document in most cases). The two exceptions were the nucleotides ppGpp and (p)ppGpp (the collective reference for ppGpp and pppGpp). A high frequency of annotation (75% and 37%, respectively) and a high mean of annotation (29 and 43, respectively) confirm that these nucleotides are central in the stringent response in *E. coli*. Indeed, during amino acid starvation (p)ppGpp nucleotides coordinate several cellular activities by influencing gene expression. As a result, further analysis based on the frequency of co-annotation of these nucleotides (and also the related pppGpp nucleotide) with gene products was issued to support the identification of more key players (Figure 2.5).

Table 2.4. Annotations of the small molecules in the corpus.

Individual small molecules were evaluated considering the number of documents where these entities were annotated and the number of annotations in the corpus. Statistical measurements are detailed in the Methods and Materials section.

Term	Number of Annotations	Number of Documents	Frequency (%)[Ψ]	Mean	*Std*	*VMR*
Amino acids	1557	160	82.90	9.730	13.83	18.78
Nucleotides	1230	145	75.13	8.480	9.290	10.13
ppGpp	4159	145	75.13	28.68	31.00	34.32
β-D-glucose	792	123	63.73	6.440	10.63	16.67
Pi	662	113	58.55	5.860	12.60	28.80
Guanosine	407	112	58.03	3.630	3.540	3.000
ATP	587	100	51.81	5.870	7.410	9.800
GTP	748	91	47.15	8.220	13.85	21.13
AMP	598	90	46.63	6.640	10.09	16.67
PPi	447	87	45.08	5.140	5.180	5.000
H_2O	210	83	43.01	2.530	2.430	2.000
Tris	261	82	42.49	3.180	2.800	1.330
Carbon	288	80	41.45	3.600	4.850	5.330
Chloramphenicol	435	77	39.90	5.650	8.250	12.80
pppGpp	632	74	38.34	8.540	13.61	21.13
(p)ppGpp	3127	72	37.31	43.43	56.00	72.93
NaCl	189	67	34.72	2.820	2.790	2.000
L-lactate	413	65	33.68	6.350	20.84	66.67
Glycerol	145	65	33.68	2.230	1.850	0.5000
Ethanol	189	65	33.68	2.910	4.400	8.000
Na^+	145	63	32.64	2.300	2.100	2.000
Ampicillin	321	62	32.12	5.180	12.74	28.80
EDTA	142	60	31.09	2.370	1.680	0.5000
L-methionine	248	59	30.57	4.200	6.670	9.000
L-histidine	183	59	30.57	3.100	5.410	8.330
L-valine	396	57	29.53	6.950	11.90	20.17
Formate	136	57	29.53	2.390	2.360	2.000

[Ψ] A threshold of 30% of the frequency of annotation was set for compounds. *VMR*: variance-to-mean; *Std*: standard deviation.

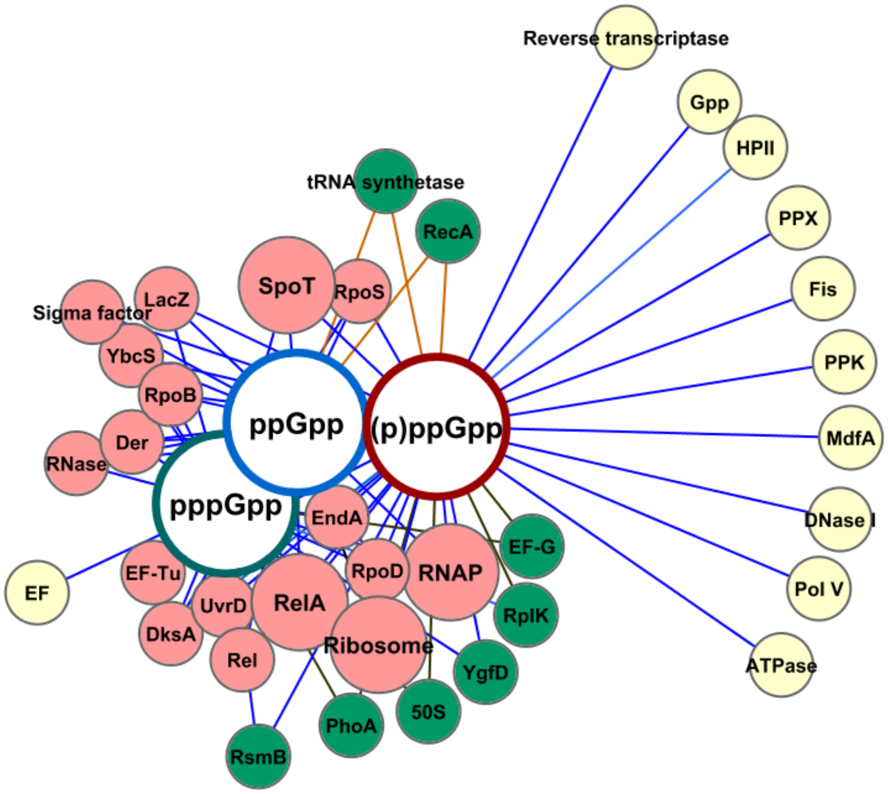


Figure 2.5. Proteins co-annotated with ppGpp, pppGpp and the collective (p)ppGpp entities.

Edges indicate the proteins co-annotated with these core nucleotides and nodes represent proteins with frequency of co-annotation higher than 10%. Highly co-annotated proteins are represented by nodes with a larger size (frequencies of co-annotation greater than 50%). Pink nodes represent the proteins that were co-annotated with the three entities, while green and yellow nodes indicate the proteins that are co-annotated with only two and one of the nucleotides, respectively.

The (p)ppGpp nucleotides were found to be considerably co-annotated with highly representative proteins, namely: the RelA and SpoT enzymes that control the basal levels of the nucleotides (around 93% and 67%, respectively); ribosomes that are affected by the nucleotides activity (around 79%); RNAP (around 64%); and the RpoS, the alternative sigma factor σ^{38} that acts as the master regulator of the general stress response (around 40%) (PMID: 9326588) (Shiba T *et al.* 1997). Some proteins were co-annotated with only one or two of the terms. For instance, the Gpp enzyme that converts pppGpp into ppGpp (PMID: 8531889; PMID: 6130093) (Condon C *et al.* 1995; Hara A and Sy J 1983) was essentially co-annotated with pppGpp. In turn, the RecA protein, which catalyses DNA strand exchange reactions (PMID: 17590232) (Manganelli R 2007), and the tRNA synthetase were co-annotated with (p)ppGpp and ppGpp with a frequency higher than 10%, whereas other proteins were mainly co-annotated with (p)ppGpp and pppGpp: the elongation factor (EF) G, known to facilitate the translocation of the ribosome along the mRNA molecules (PMID: 8531889) (Condon C *et al.* 1995); the RplK (or 50S ribosomal subunit protein L11) that was reported to be essential when the 30S ribosomal initiation complex joins to the 50S ribosomal subunit and in the EF-G-dependent GTPase activity (PMID: 17095013; PMID: 12419222) (Wendrich TM *et al.* 2002; Jenvert RM and Schiavone LH 2007); and the enzyme PhoA known to be involved in the acquisition and transport of phosphate (PMID: 9555903) (Rao NN *et al.* 1998). Additionally, results pointed out potentially interesting associations with less represented proteins, such as: the Fur transcriptional activator that controls the transcription of genes involved in iron homeostasis (PMID: 15853883) (Vinella D *et al.* 2005); the HN-S transcriptional dual regulator that is capable of condensing and supercoiling DNA (PMID: 10966109) (Johansson J *et al.* 2000); the DnaA protein implicated in the chromosomal replication initiation (PMID:1690706) (Chiaramello AE and Zyskind JW 1990); the DinJ-YafQ complex involved in the inhibition of protein synthesis and growth (PMID:12123445) and the MazE antitoxin of the MazF-MazE toxin-antitoxin system involved in translation inhibition processes (PMID:12123445) (Chang DE *et al.* 2002).

2.4.2 Ontology term enrichment

Recent developments in the functional annotation of genomes using biological ontologies provided the means to contextualise literature mining outputs, i.e. to disclose the biological meaning behind the annotated entities. The automatic mapping of annotated entities to ontology concepts was possible because EcoCyc supports the assignment of MultiFun and GO ontology concepts to genes and gene products in its curation procedures. MultiFun ontology classifies gene products according to their cellular function, namely: metabolism, information transfer, regulation, transport, cell processes, cell structure, location, extra-chromosomal origin, DNA site, and cryptic gene. In turn, GO embraces three separate ontologies: cellular components, i.e. the parts of a cell or its extracellular environment; molecular functions, i.e. the basic activities of a gene product at the molecular level, such as binding or catalysis; and biological processes, i.e. the set of molecular events related to the integrated functioning of cells, tissues, organs or organisms.

Here, the analysis of ontology assignments was focused on MultiFun cellular functions and GO biological processes (Table 2.5 and Table 2.6). The aim was to identify the processes and functions to which annotated entities contributed the most. As such, the frequency of ontology assignment represented the fraction of documents in the corpus that included entities with assignments to the concept. The frequency of assignment estimated the particular contribution of an entity to the overall assignment of the ontology concept and the frequency of annotation of the entity indicated the number of documents in which that entity was annotated. Since one entity can be associated to several ontology concepts and the representation of an ontology concept depends on the number of annotations of the embraced entities, the frequency of assignment evaluated the most contributing entities to ontology concepts whereas the frequency of annotation provided indication of whether those entities are considerably discussed in the corpus or not.

The analysis of MultiFun cellular function assignments (Table 2.5) evidenced gene functions related to central metabolism processes, post-transcriptional processes and transcription related functions (covered by over 50% of the documents). The most assigned MultiFun cellular functions, namely metabolic functions related to nucleotide and nucleoside conversions (BC-1.7.33) and proteolytic cleavage of compounds (BC-3.1.3.4), derived from the highly annotated *rel*A and *spo*T genes. The *lac*Z gene, another highly annotated gene (26% of the documents), that encodes the β-galactosidase enzyme responsible for the hydrolysis of β-galactosides into monosaccharides, contributed significantly (almost 50% of assignments) to the annotation of cellular functions implicated in the metabolism of carbon compounds (BC-1.1.1). The gene *fis* that encodes the Fis transcriptional dual regulator and the gene *rpo*B coding for the β subunit of the RNAP, contributed the most to the annotation of transcriptional related functions (BC-2.2.2). Similarly, genes like *dks*A that encodes the DksA protein, *rpl*K that encodes the 50S ribosomal subunit L11, and *rps*G and *rps*L coding for the 30S ribosomal subunitsS7 and S12, respectively, contributed the most to the annotation of transcription related processes (BC-

2.3.2). By looking into the frequency of annotation of the corresponding encoded products, it was verified that there is a discrepancy of annotation between the genes contributing to enriched ontology terms and the corresponding gene products. Therefore, the use of the MultiFun ontology not only pointed out relevant gene function assignments, but also disclosed the participation of several gene products that, even though presenting extremely low frequency of annotation, were highlighted by functional association. Some examples are 30S ribosomal subunit protein S12 and 30S ribosomal subunit protein S7 encoded by *rps*L and *rps*G, respectively.

On the other hand, the analysis of GO biological process assignments (Table 2.6) highlighted metabolic and genetic information transfer processes as the most frequent in the corpus (over 50% of the documents). Although the general concept of metabolic process had the highest frequency (89% of the documents), two particular metabolic processes: the nucleobase, nucleoside and nucleotide interconversion (GO:0015949) and the guanosine tetraphosphate metabolic (GO:0015969) processes, had high frequencies (around 80% of the documents) as well. The gene product that contributed the most to the annotation of metabolic-related processes was the RelA enzyme, with over 80% of the assignments. Regarding genetic information transfer, transcription (GO:0006350), DNA-dependent transcription regulation (GO:0006355) and translation (GO:0006412) were the most represented processes (56%, 52% and 40% of the documents, respectively). The σ^{38} factor, the CRP transcriptional dual regulator, known to participate in the transcriptional regulation of genes involved in the degradation of non-glucose carbon sources (PMID: 10966109) (Johansson J *et al.* 2000) and the Mfd protein, found to be responsible for ATP-dependent removal of stalled RNAPs from DNA (PMID: 7968917) (Selby CP and Sancar A 1994), contributed similarly to the annotation of transcription and DNA-dependent transcription regulation processes, ranging between 10% and 20% of the assignments. Translation process assignments were derived from the RplK (or 50S ribosomal subunit protein L11) and the DksA proteins, with 28% of the assignments each, and the Elongation Factor Tu (EF-Tu), which mediates the entry of the aminoacyl tRNA into the ribosome, with 13% of the assignments.

Table 2.5. MultiFun cellular function assignments.

Gene annotations were used to identify MultiFun concepts. A threshold of 30% of documents was considered for ontology assignment and a threshold of 10% was used to point out the genes that most contributed to such assignment.

MultiFun Concepts	Frequency of Ontology Annotation	Brief Description	Genes			Gene Products	
			Name	Frequency of Assignment	Frequency of Annotation	Name	Frequency of Annotation
BC-1.7.33 Nucleotide and nucleoside conversions	76%	The chemical reactions involved in the central carbon metabolism by which a nucleobase, nucleoside or nucleotide is converted from another nucleobase, nucleoside or nucleotide.	*rel*A	68%	72%	RelA	79%
			*spo*T	28%	46%	SpoT	31%
BC-3.1.3.4 Proteases, cleavage of compounds	55%	Proteins that hydrolysates a peptide bond or bonds within a protein during posttranscriptional regulatory processes.	*spo*T	91%	46%	SpoT	31%
BC-2.2.2 Transcription related functions	51%	The information tranfer related functions involved in the synthesis of RNA on a template of DNA.	*fis*	22%	6%	Fis	9%
			*rpo*B	17%	16%	RpoB	18%
BC-2.3.2 Translation	48%	The cellular metabolic process by which a protein is formed, using the sequence of a mature mRNA molecule to specify the sequence of amino acids in a polypeptide chain.	*dks*A	23%	3%	DksA	8%
			*rpl*K	17%	6%	RplK	7%
			*rps*G	12%	18%	RpsG	NA
			*rps*L	11%	19%	RpsL	3%
BC-5.5.3 Starvation	47%	A state or activity of a cell or an organism as a result to the adaptation to starvation.	*spo*T	85%	46%	SpoT	31%
			*dks*A	12%	3%	DksA	8%
BC-1.1.1 Carbon compounds	46%	The metabolic reactions by which living organisms utilises carbon compounds.	*lac*Z	49%	26%	LacZ	28%
			*pts*G	22%	16%	PtsG	1%
BC-2.3.8 Ribosomal proteins	44%	Proteins that associate to form a ribosome involved in genetic information transfer in cells.	*rpl*K	24%	6%	RplK	7%
			*rps*G	18%	18%	RpsG	NA
			*rps*L	17%	19%	RpsL	3%
BC-3.1.2.3 Repressor	40%	Any transcription regulator that prevents or downregulates transcription.	*fis*	41%	6%	Fis	9%
BC-3.1.2.2 Activator	32%	Any transcription regulator that induces or upregulates transcription.	*fis*	45%	6%	Fis	9%

NA: corresponds to non-annotated gene products in the corpus.

Table 2.6. GO biological processes assignments.

Gene product annotations were used to identify the most assigned GO terms. A threshold of 35% of documents was considered for ontology assignment and a threshold of 10% was used to point out the gene products that most contributed to such assignment.

Gene Ontology concepts	Frequency of Ontology Annotation	Brief Description	Gene Products			Coding Genes	
			Name	Frequency of Assignment	Frequency of Annotation	Name	Frequency of Annotation
GO:0008152 Metabolic process	89%	The chemical reactions and pathways, including anabolism and catabolism, by which living organisms transform chemical substances.	RelA	80%	79%	*rel*A	72%
			LacZ	10%	24%	*lac*Z	26%
GO:0015949 Nucleobase, nucleoside and nucleotide interconversion	80%	The chemical reactions and pathways by which a nucleobase, nucleoside or nucleotide is synthesized from another nucleobase, nucleoside or nucleotide.	RelA	80%	79%	*rel*A	72%
			SpoT	20%	31%	*spo*T	46%
GO:0015969 Guanosine tetraphosphate metabolic process	80%	The chemical reactions and pathways involving guanine tetraphosphate (5'-ppGpp-3'), a derivative of guanine riboside with four phosphates.	RelA	80%	79%	*rel*A	72%
			SpoT	20%	31%	*spo*T	46%
GO:0006350 Transcription	56%	The synthesis of either RNA on a template of DNA or DNA on a template of RNA.	RpoS	16%	22%	*rpo*S	17%
			CRP	12%	6%	*crp*	4%
			RpoB	10%	18%	*rpo*B	16%
			Mfd	10%	2%	*mfd*	2%
GO:0006355 Regulation of transcription, DNA-dependent	52%	Any process that modulates the frequency, rate or extent of DNA-dependent transcription.	RpoS	20%	22%	*rpo*S	17%
			CRP	14%	6%	*crp*	4%
			Mfd	12%	2%	*mfd*	2%
GO:0006412 Translation	40%	The cellular metabolic process by which a protein is formed, using the sequence of a mature mRNA molecule to specify the sequence of amino acids in a polypeptide chain.	RplK	28%	7%	*rpl*K	6%
			DksA	28%	8%	*dks*A	3%
			EF-Tu	13%	14%	*tuf*B	3%
GO:0006950 Response to stress	39%	A change in state or activity of a cell or an organism as a result of a disturbance in organismal or cellular homeostasis, usually, but not necessarily, exogenous .	RecA	20%	16%	*rec*A	20%
			RelB	16%	4%	*rel*B	3%
			NusA	10%	5%	*nus*A	4%
GO:0042594 Response to starvation	39%	A change in state or activity of a cell or an organism as a result of a starvation stimulus, deprivation of nourishment.	SpoT	67%	31%	*spo*T	46%
			DksA	30%	8%	*dks*A	3%

Table 2.6. (continued).

Gene Ontology concepts	Frequency of Ontology Annotation	Brief Description	Gene Products			Coding Genes	
			Name	Frequency of Assignment	Frequency of Annotation	Name	Frequency of Annotation
GO:0006970 Response to osmotic stress	38%	A change in state or activity of a cell or an organism as a result of a stimulus indicating an increase or decrease in the concentration of solutes outside the organism or cell.	RpoS EF-Tu	59% 34%	22% 14%	*rpo*S *tuf*B	17% 3%
GO:0005975 Carbohydrate metabolic process	36%	The chemical reactions and pathways involving carbohydrates, any of a group of organic compounds based of the general formula Cx(H2O)y.	LacZ	94%	24%	*lac*Z	26%
GO:0006974 Response to DNA damage stimulus	36%	A change in state or activity of a cell or an organism as a result of a stimulus indicating damage to its DNA from environmental insults or errors during metabolism.	Mfd RecA RecG	28% 11% 11%	2% 16% 3%	*mfd* *rec*A *rec*G	2% 20% NA
GO:0006281 DNA repair	36%	The process of restoring DNA after damage that include direct reversal, base excision repair, nucleotide excision repair, photoreactivation, bypass, double-strand break repair pathway, and mismatch repair pathway.	Mfd RecA RecG	28% 11% 11%	2% 16% 3%	*mfd* *rec*A *rec*G	2% 20% NA

NA: corresponds to non-annotated genes in the corpus.

In terms of stress-specific ontology annotations, responses to stress (GO: 0006950), starvation (GO: 0042594), osmotic stress (GO: 0006970) and DNA damage stimulus (GO:0006974) were the most frequently assigned cellular response processes (present in almost 40% of the documents). The response to stress was mostly assigned by the RecA regulatory protein, the RelB transcriptional repressor and the transcription antitermination protein NusA (frequencies of assignment of 20%, 16% and 10%, respectively). In the response to starvation, SpoT enzyme detached from other contributing gene products (almost 70% of the assignments). The σ^{38} factor and the EF-Tu protein were the main contributors to the assignment of osmotic stress responses (59% and 34% of the assignments, respectively), while the response to DNA damage stimulus was derived from the annotation of proteins Mfd, RecA and RecG (28%, 11% and 11% of the assignments, respectively). The process of restoring DNA after damage, called DNA repair (GO:0006281), was also assigned by the aforementioned annotated entities.

Since results evidenced significant assignment of stress-related processes, it was considered interesting to explore in detail the functional annotations of gene products related to *E. coli* stress responses (Table 2.7). A decade-by-decade analysis was performed to evaluate the extent of documents that study entities associated with these functional annotations. As shown, the response to starvation (GO:0042594) was mostly evidenced in the last decade, being assigned in almost 70% of the documents of this decade. The response to DNA damage stimulus (GO:0006974) and osmotic stress (GO:0006970) were also considerably assigned in the last decade (50% of the documents). On the contrary, the defense response to bacterium (GO:0042742) was less assigned in the documents from the last two decades (less than 10% of the documents) and the stringent response (GO:0015968) was poorly assigned in the last decade, probably because GO only associates this biological process to the 50S ribosomal subunit protein L11, which only recently has been studied in the context of this stress.

Table 2.7. Assignment of GO concepts related to stress responses.

The frequency of annotation of stress response-related concepts was estimated for documents published in the four decades analysed (from 1970 to 2009).

GO Concept	GO Description	Frequency of Ontology Annotation			
		1970	1980	1990	2000
Response to starvation	A change in state or activity of a cell or an organism as a result of a starvation stimulus, deprivation of nourishment.	-	15%	28%	68%
Response to DNA damage stimulus	A change in state or activity of a cell or an organism as a result of a stimulus indicating damage to its DNA.	7%	26%	31%	50%
Response to osmotic stress	A change in state or activity of a cell as a result of an increase or decrease in the concentration of solutes outside the cell.	21%	28%	35%	50%
Response to stress	A change in state or activity of a cell or an organism as a result of a disturbance in organismal or cellular homeostasis.	-	31%	46%	46%
Response to oxidative stress	A change in state or activity of a cell or an organism as a result of oxidative stress.	-	3%	20%	45%
SOS response	An error-prone process for repairing damaged microbial DNA.	-	23%	30%	45%
Response to antibiotic	A change in state or activity of a cell or an organism as a result of an antibiotic stimulus.	-	23%	30%	15%
Response to drug	A change in state or activity of a cell or an organism as a result of a drug stimulus.	-	15%	35%	11%
Response to temperature stimulus	A change in state or activity of a cell or an organism as a result of a temperature stimulus.	-	-	9%	11%
Defense response to bacterium	Reactions triggered in response to the presence of a bacterium that act to protect the cell or organism.	14%	26%	9%	8%
Stringent response	A specific global change in the metabolism of a bacterial cell as a result of starvation.	-	13%	7%	6%
Response to heat	A change in state or activity of a cell or an organism as a result of a heat stimulus.	-	3%	13%	5%
Response to cold	A change in state or activity of a cell or an organism as a result of a cold stimulus.	-	-	-	3%
Response to toxin	A change in state or activity of a cell or an organism as a result of a toxin stimulus.	-	-	-	3%
Cellular response to starvation	A change in state or activity of a cell as a result of deprivation of nourishment.	-	-	-	1%
Response to copper ion	A change in state or activity of a cell or an organism as a result of a copper ion stimulus.	-	-	-	1%
Response to desiccation	A change in state or activity of a cell or an organism as a result of a desiccation stimulus.	-	-	4%	-
Response to methotrexate	A change in state or activity of a cell or an organism as a result of a methotrexate stimulus.	-	-	2%	-

2.4.3 EXAMINING LESS-REPORTED ENTITIES

In the present corpus, most of the biological entities identified as major participants in the *E. coli* stringent response, were also extensively cited in recent reviews (Wu J and Xie J 2009; Srivatsan A and Wang JD 2008; Potrykus K and Cashel M 2008; Jain V *et al.* 2006; Magnusson LU *et al.* 2005). As illustrated in Figure 2.6, biological entities considered to be key components in the important reviews, namely the enzymes RelA and SpoT and the RNAP, were also evidenced by the semi-automatic information extraction approach.

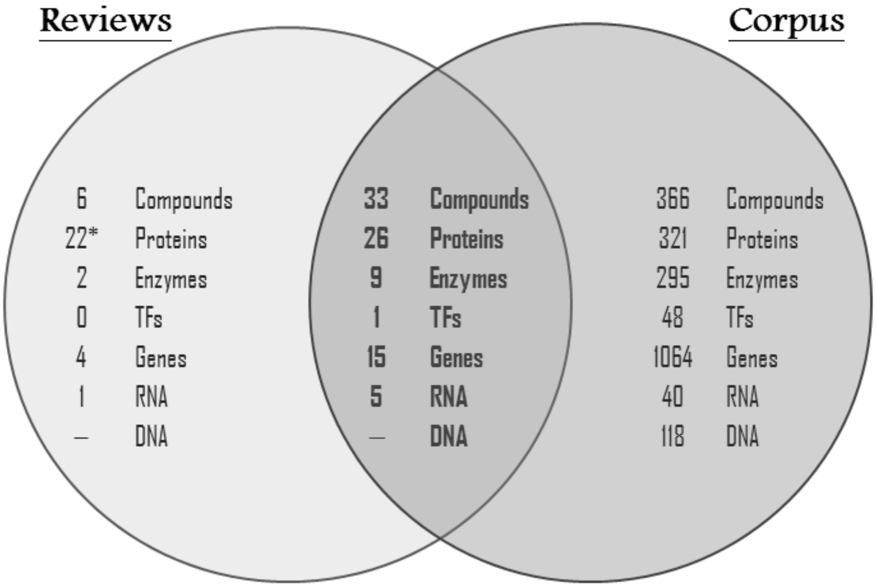


Figure 2.6. Venn diagram comparing annotations from corpus and selected reviews.

This diagram indicates the number of entities per class that were retrieved from the corpus and from the latest reviews considered to be relevant in this subject. The intersecting zone gives the number of entities that were simultaneously reported in the two set of documents.

However, when examining the extent of annotations from the selected reviews and the corpus, it was evident that many biological entities have been disregarded or less reported in the reviews. Biological entities, such as transcriptional factors and other gene products like stress-related proteins, were not described in the selected reviews. As exemplified in Table 2.8, the role of some biological entities that are directly (or indirectly) associated with the stringent response, is often underestimated by most literature revisions.

Table 2.8. Some examples of less-reported entities (namely in recent reviews), which are relevant in the *E. coli* stringent response.

Biological entities	Freq (%)	Details	References
DnaJ - chaperone with DnaK	3.11	Chaperone protein that assists the DnaJ/DnaK/GrpE system of *E. coli*. The overproduction of ppGpp has shown to induce the accumulation of these chaperones.	(Jones PG *et al.* 1992)
ClpB chaperone	1.55	ClpB, together with the DnaJ/DnaK/GrpE chaperone system, is able to resolubilize aggregated proteins.	(Mogk A *et al.* 2003)
GroEL-GroES chaperonin complex	0.52	GroEL and GroES are both induced by heat and when ppGpp is overproduced in *E. coli*.	(Jones PG *et al.* 1992)
RuvB - repair helicase	1.55	Component of the RuvABC enzymatic complex that promotes the rescue of stalled (often formed by ppGpp) or broken DNA replication forks in *E. coli*.	(Shinagawa H *et al.* 1988)
CsrA - carbon storage regulator	1.04	Regulator of carbohydrate metabolism, which activates UvrY, responsible for the transcription of *csr*B that, in turn, inhibits the CsrA activity.	(Sabnis NA *et al.* 1995)
*uvr*Y	0.52	Encodes the UvrY protein that has been shown to be the cognate response regulator for the BarA sensor protein. This regulator participates in controlling several genes involved in the DNA repair system (e.g. CsrA) and carbon metabolism.	(Pernestig AK *et al.* 2001)
*cst*A	0.52	Gene encoding the CstA peptide transporter, which expression is induced by carbon starvation and requires the CRP-cAMP transcriptional regulator. The CstA translation is regulated by the CsrA that occludes ribosome binding to the *cst*A mRNA.	(Dubey AK *et al.* 2003)
CspD - DNA replication inhibitor	0.52	CspD is a toxin that appears to inhibit the DNA replication. ppGpp is one of the positive factors for the expression of *csp*D.	(Yamanaka K and Inouye M 1997)
FabH - β-ketoacyl-ACP synthase III	0.52	A key enzyme in the initiation of fatty acids biosynthesis that is stringently regulated by ppGpp.	(Podkovyrov SM and Larson TJ 1996)
FadR transcriptional dual regulator	1.55	Regulates the fatty acid biosynthesis and fatty acid degradation at the level of transcription. ppGpp has been shown to be also involved in the regulation of these pathways	(Podkovyrov SM and Larson TJ 1996)
NtrC-Phosphorylated transcriptional dual regulator	1.04	Regulatory protein involved in the assimilation of nitrogen and in slow growth caused by N-limited condition. It was reported that ppGpp levels increase upon nitrogen starvation.	(Peterson CN *et al.* 2005)
dps	2.59	Gene encoding the Dps protein that is highly abundant in the stationary-phase and is required for the starvation responses. It was found to be regulated by ppGpp and RpoS.	(Gong L *et al.* 2002)
*psi*F	2.07	Gene induced during phosphate starvation that has been associated with the accumulation of ppGpp.	(Rao NN *et al.* 1998)
*chp*R	2.07	Encodes the MazE antitoxin, a component of the MazE-MazF system that causes a "programmed cell death" in response to stresses, including starvation. Genes *maz*E and *maz*F are located in the *E. coli rel* operon and are regulated by ppGpp.	(Aizenman E *et al.* 1996)
*maz*G	0.52	Encodes the MazG nucleoside pyrophosphohydrolase that limits the detrimental effects of the MazF toxin under nutritional stress conditions. Overexpression of *maz*G inhibits cell growth and negatively affects accumulation of ppGpp.	(Gross M *et al.* 2006)

The recognition of these proteins in the corpus was invaluable, allowing to uncover various stress-responsive proteins that are normally associated with other stress responses, such as chaperones (e.g. DnaJ, ClpB or the GroEL-GroES chaperonin complex) and toxin-antitoxin systems (e.g. protein encoded by *chp*R). The description of such entities as participants in the stringent response discloses a more insightful overview of the complexity of these entangled cellular processes. For example, the identification of entities related to certain metabolic pathways, like the fatty acids biosynthesis (e.g. FabH and FabR), or DNA processes, like DNA replication (e.g. CspD) and DNA repair (e.g. *uvr*Y), expanded the characterization of stringently regulated activities that were not evident in most reviews analysed.

2.4.4 EVOLUTION OF TECHNOLOGY

The evolution of experimental techniques was expected to intersect points of turnover on the study of the stringent response. This assumption was confirmed by comparing the number of annotations of biological entities (genetic components, gene products and small molecules) and experimental techniques (grouped into major PSI-MI classes) per decade (Figure 2.7).

The analysis evidenced that the repertoire of experimental techniques has been growing significantly and the study is ever more dedicated to genetic components. In particular, results showed the use of an ever-growing number of experimental interaction detection methods (MI:0045) and a considerable number of experimental participant identification (MI:0661) and experimental feature detection (MI:0659) methods.

The analysis of the frequency of annotation (Table 2.9) evidenced that the chromatography technology (MI:0091), experimental feature detection (MI:0657), genetic interference (MI:0254) and primer specific polymerase chain reaction (PCR) (MI:0088) techniques were annotated in more than 40% of the documents. Most techniques were referred roughly two times per document, but primer-specific PCR (MI:0088) and array technology (MI:0008) presented a considerable mean of annotation (with over 10 and 8 annotations per document, respectively) and high *VMR* values (22.5 and 6.13, respectively), which indicated that these techniques were essentially discussed in a given set of documents.

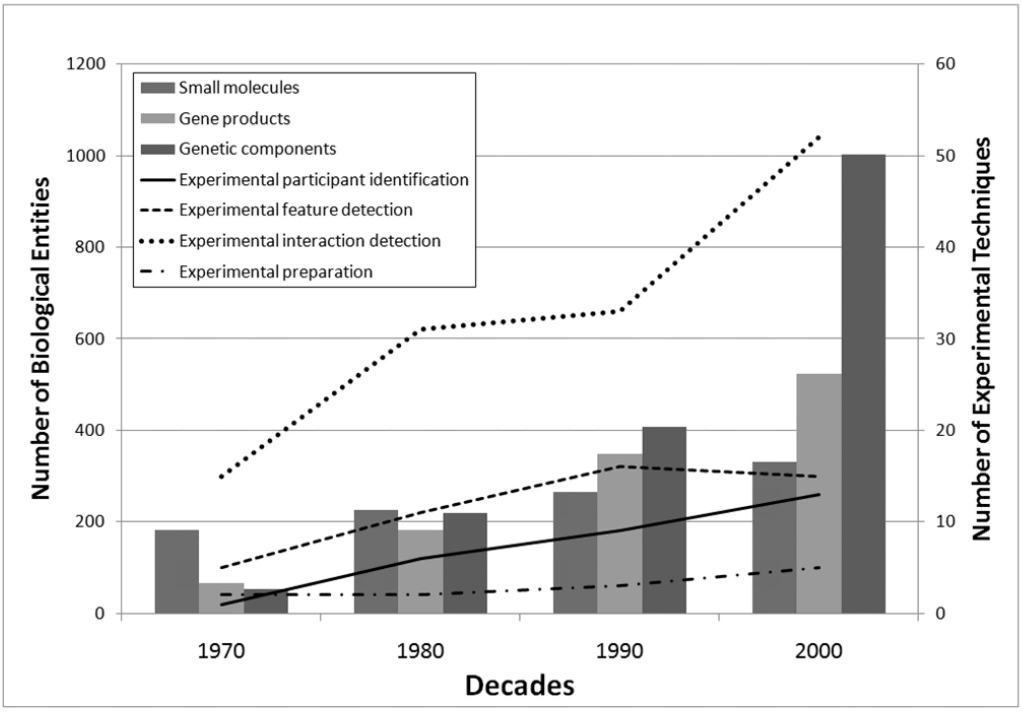


Figure 2.7. Comparison of the expansion of knowledge to the applied experimental techniques.

Bars represent the number of biological entities (left Y axis) found for the three major biological classes, i.e., genetic components (genes, RNAs and DNAs), gene products (proteins, transcription factors and enzymes) and small molecules. Lines plot the number of experimental techniques (right Y axis) associated to the annotated PSI-MI classes.

Also, a detailed look into the frequency of annotation per decade points out that some of the techniques used in early studies have a reduced application today and highlights the increasing influence of high-throughput technologies in recent studies. For instance, experimental interaction detection methods (MI:0045) such as the scintillation proximity assay (MI:0099), the molecular sieving (MI:0071), the filter trap assays (MI:0928) and the cosedimentation through density gradient (MI:0029) showed a higher frequency of annotation in the first decade (1970-1980) whereas the comigration in gel electrophoresis (MI:0807) and enzymatic studies (MI:0415) experienced an increase in the frequency of annotation throughout the decades.

Table 2.9. PSI-MI assignments to annotated experimental techniques.

Techniques			**Statistics over the corpus**				**Frequency per Decade**			
PSI-MI Class	**PSI-MI**	**PSI-MI name**	**Freq**	**Mean**	**Std**	**VMR**	**1970**	**1980**	**1990**	**2000**
MI:0659 experimental feature detection	MI:0659	experimental feature detection*	55%	2.44	2.76	2	64%	65%	67%	67%
	MI:0833	autoradiography	25%	1.65	1.4	1	29%	35%	22%	22%
	MI:0113	western blot	21%	4.95	3.38	2.25	-	13%	18%	44%
	MI:0074	mutation analysis	20%	3.34	3.24	3	14%	20%	27%	32%
	MI:0114	x-ray crystallography	4%	1.63	1.46	1	-	-	2%	8%
	MI:0811	insertion analysis	4%	1.14	0.38	0	-	-	7%	4%
MI:0045 experimental interaction detection	MI:0091	chromatography technology	50%	4.23	4.14	4	100%	85%	55%	64%
	MI:0254	genetic interference	42%	2.68	2.35	2	-	28%	69%	64%
	MI:0807	comigration in gel electrophoresis	37%	2.51	2.13	2	21%	48%	51%	49%
	MI:0045	experimental interaction detection*	36%	3.1	3.12	3	14%	45%	47%	41%
	MI:0808	comigration in sds page	27%	2.02	1.61	0.5	7%	23%	33%	29%
	MI:0099	scintillation proximity assay	24%	1.94	1.65	1	64%	35%	20%	15%
	MI:0051	fluorescence technology	16%	1.84	1.8	1	7%	20%	5%	22%
	MI:0071	molecular sieving	15%	2.25	3.19	4.5	29%	13%	15%	13%
	MI:0217	phosphorylation reaction	13%	2.72	3.74	4.5	7%	8%	16%	14%
	MI:0415	enzymatic study	12%	1.96	2.11	4	7%	8%	11%	19%
	MI:0008	array technology	10%	8.47	7.47	6.13	-	-	-	22%
	MI:0928	filter trap assay	9%	2.24	2.47	2	36%	13%	4%	6%
	MI:0004	affinity chromatography technology	8%	2.33	2.44	2	-	5%	5%	13%
	MI:0428	imaging technique	7%	1.57	1.41	1	-	8%	4%	11%
	MI:0047	far western blotting	6%	1.5	0.82	0	-	3%	7%	8%
	MI:0435	protease assay	6%	3.92	4.01	5.33	-	5%	5%	8%
	MI:0017	classical fluorescence spectroscopy	6%	1.08	0.29	0	-	-	5%	11%
	MI:0089	protein array	6%	1.64	1.62	1	-	3%	2%	11%
	MI:0029	cosedimentation through density gradient	5%	5.22	3.94	1.8	43%	10%	-	2%
	MI:0040	electron microscopy	4%	2.57	1.31	0.5	-	10%	-	4%
	MI:0676	tandem affinity purification	3%	3.8	5.18	8.33	-	-	-	7%
	MI:0054	fluorescence-activated cell sorting	3%	5.8	4.6	3.2	-	-	4%	4%
	MI:0413	electrophoretic mobility shift assay	3%	1.6	0.77	0	-	-	5%	2%
	MI:0012	bioluminescence resonance energy transfer	2%	6.25	5.68	4.17	-	-	4%	2%
	MI:0018	two hybrid	2%	2	1.41	0.5	-	-	-	7%
	MI:0053	fluorescence polarization spectroscopy	2%	5	2.94	0.8	-	3%	-	2%
	MI:0397	two hybrid array	2%	2	1.41	0.5	-	-	2%	2%
	MI:0227	reverse phase chromatography	2%	3.25	2.87	1.33	-	-	5%	1%
	MI:0226	ion exchange chromatography	1%	1	0	0	-	-	-	1%
	MI:0031	protein cross-linking with a bifunctional reagent	1%	7	4	2.29	-	3%	-	1%
	MI:0052	fluorescence correlation spectroscopy	1%	1	0	0	-	-	-	1%
	MI:0416	fluorescence microscopy	1%	2.5	0.71	0	-	-	-	2%
	MI:0016	circular dichroism	1%	1.5	0.71	0	-	-	-	2%
	MI:0225	chromatin immunoprecipitation array	1%	1	0	0	-	-	-	1%
	MI:0872	atomic force microscopy	1%	1	0	0	-	-	-	1%
	MI:0049	filter binding	1%	1	0	0	-	3%	2%	-
	MI:0426	light microscopy	1%	1	0	0	-	-	-	2%
MI:0661 experimental participant identification	MI:0088	primer specific pcr	40%	10.38	15.87	22.5	-	8%	29%	95%
	MI:0080	partial dna sequence identification by hybridization	27%	3.75	3.47	3	14%	30%	29%	26%
	MI:0078	nucleotide sequence identification	20%	1.77	1.26	1	-	15%	25%	22%
	MI:0103	southern blot	14%	3.04	2.05	1.33	-	8%	15%	19%
	MI:0929	nothern blot	8%	5.56	4.8	3.2	-	3%	11%	11%
	MI:0421	identification by antibody	6%	1.82	1.51	1	-	8%	5%	6%
	MI:0427	identification by mass spectrometry	5%	1.67	0.94	0	-	-	4%	8%
	MI:0082	peptide massfingerprinting	2%	1.5	0.71	0	-	-	-	5%
	MI:0093	protein sequence identification	1%	1	0	0	-	-	2%	
	MI:0411	enzyme linked immunosorbent assay	1%	4	2	1	-	-	2%	1%
MI:0346 experimental preparation	MI:0714	nucleic acid transduction	26%	2.22	2.82	2	14%	15%	31%	29%
	MI:0715	nucleic acid conjugation	6%	1.73	1.41	1	7%	3%	5%	7%
	MI:0308	electroporation	5%	1.89	1.56	1	-	-	2%	9%
	MI:0343	cdna library	3%	1	0	0	-	-	-	6%
MI:0190 interaction	MI:0194	cleavage reaction	1%	1	0	0	-	3%	-	-
MI:0116 feature type	MI:0373	dye label	5%	1.2	0.45	0	7%	8%	5%	4%

*These PSI-MI general classes were used to identify techniques that did not map into any particular technique within the class

2.5 DISCUSSION

The aim of this work was to use literature mining to complement manual curation in the revision, systematisation and interpretation of current knowledge on the stringent response of *E. coli*. Literature mining was expected to speed up the retrieval of relevant literature and help on the identification of important biological players and their molecular functions. The controlled vocabulary extracted from EcoCyc repository (GO and MultiFun assignments) and PSI-MI ontology was expected to support large-scale information processing and biological contextualisation.

At present, manual curation can extract more detailed information from literature than it is possible by mining approaches, and more accurately define the participants and their roles. However, to achieve a broad coverage, both approaches can efficiently complement each other. Results suggested that: (i) automatic literature retrieval is able to provide documents of interest whereas controlled vocabulary from publicly available databases can support the identification of relevant entities; (ii) database ontology assignments enable entity contextualisation into cellular functions and biological processes, delivering a more comprehensive and biologically meaningful scenario; and (iii) statistical analysis identifies biological entities of interest and facilitates document indexing for additional manual curation. Ultimately, the literature mining approach presented clues on entities and associations of interest and suggested which documents in the corpus should be further inspected for details on given entities or processes.

The analysis evidenced the (p)ppGpp nucleotides as some of the most annotated biological entities: the ppGpp nucleotide was annotated in 75% of the documents, and the term (p)ppGpp exhibited the highest average of annotations per document (Table 2.4). The extensive number of documents supporting these annotations evidenced that the role of (p)ppGpp nucleotides in the stringent response has been extensively studied. Corpus analysis disclosed that the synthesis of ppGpp was first associated in 1970 to the activity of *rel*A gene product (PMID: 4315151) (Cashel M and Kalbache B 1970), which, during amino acid deprivation, promotes the accumulation of this nucleotide above the basal levels. Later, in 1980, the ppGpp level was found to be controlled by the SpoT enzyme via GTP hydrolysis activity (PMID: 6159345)

(Lagosky PA and Chang FN 1980). From then on, studies have been detailing the role of these entities on the (p)ppGpp-mediated response.

Most studies have been focused on the activity of the *rel*A gene and its product RelA (over 70% of the documents), while a fewer set of documents (roughly 30% of the documents) has reported the activity of the *spo*T gene and the SpoT enzyme (see Tables 2.2 and 2.3). In part, because RelA was the first enzyme discovered to be involved in the stringent response, but mostly because it is the first biological entity to respond to the amino acid starvation. Accordingly, nucleobase, nucleoside and nucleotide metabolic processes emerged as the most assigned functional class associated with the metabolic response to amino acid starvation. The *lac*Z gene was also highly annotated in the corpus, but it was surmised to be associated with the involvement of this gene in most genetic manipulation procedures described in many studies.

Transcriptional and translational processes were also highlighted by the analysis. The acknowledgment that (p)ppGpp nucleotides manipulate gene expression, so that gene products with important roles in the starvation survival are favoured at the expense of those required for growth and proliferation, has been widely reported (PMID:12123445; PMID:10809680) (Chang DE *et al.* 2002; Liang ST *et al.* 2000). *In vitro* studies demonstrated that (p)ppGpp bind directly to the RNAP, affecting the transcription of many genes (PMID:4553835) (Irr JD 1972). Also, studies hypothesised that the configuration of the RNAP is altered, decreasing the affinity of the housekeeping sigma factor (i.e. σ^{70}) to RNAP and thus, allowing other sigma factors to compete and influence promoter selectivity (PMID:12023304) (Jishage M *et al.* 2002). As covered by the corpus analysis, besides RNAP (annotated in over 60% of the documents), four of the existing sigma factors in *E. coli* were also annotated: the σ^{38} that acts as the master regulator of the general stress response (annotated in 22% of the documents); the σ^{70} that is the primary sigma factor during exponential growth (annotated in 11% of the documents); the σ^{54} that controls the expression of nitrogen-related genes (annotated in 4% of the documents); and the σ^{32} that controls the heat shock response during log-phase growth (annotated in 3% of the documents). Although the regulation of transcription initiation is not yet fully understood, current knowledge suggests that these four sigma factors may interact with the RNAP during stringent control.

Further inspection of the functional annotations of transcription-related processes helped in the characterisation of other annotated entities to understand their roles in this scenario, namely: the β subunit of the RNAP (RpoB) to which (p)ppGpp nucleotides bind (PMID:9501189) (Zhou YN and Jin DJ 1998); the CRP transcriptional dual regulator that is activated in response to starvation conditions (PMID:10966109) (Johansson J *et al.* 2000); the Fis transcriptional dual regulator, whose gene promoter is inhibited during the transcription initiation by the (p)ppGpp-bound RNAP polymerase (PMID:2209559; PMID:9973355) (Ross W *et al.* 1990; Walker KA *et al.* 1999); and the Mfd protein that releases the arrested RNAP-DNA complexes after (p)ppGpp nucleotides induce the transcription elongation pausing, protecting genome integrity during transient stress conditions (PMID:7968917) (Selby CP and Sancar A 1994). In resume, (p)ppGpp nucleotides not only modulate the RNAP activity, either by reducing the expression of genes like *fis* (which in turn modulates the expression of the *crp* gene) or increasing the expression of the σ^{38} gene, but also mediate the inhibition of the RNAP replication-elongation, which afterwards requires the Mfd protein to remove the stalled RNAPs (PMID:16039593; PMID:7968917) (Selby CP and Sancar A 1994; Trautinger BW *et al.* 2005). Although most studies have focused on the influence of the (p)ppGpp nucleotides on the mechanisms that regulate transcription initiation activities, their regulatory effects on the elongation of DNA transcription are also important. The combined control of the DNA transcription initiation and elongation are central to a prompter cellular response to nutritional starvation (Srivatsan A and Wang JD 2008).

Similarly, (p)ppGpp also influence certain translation-related processes. Studies showed that (p)ppGpp inhibits translation by repressing the expression of ribosomal proteins and also potentially inhibiting the activity of the particular proteins (PMID:7021151; PMID:11673421; PMID:6358217) (Yang X and Ishiguro EE 2001; Pingoud A and Block W 1981; Pingoud A *et al.* 1983). Corpus analysis evidenced the annotation of ribosomal proteins, such as the 50S ribosomal subunit protein L11 and the 30S ribosomal subunit proteins S7 and S12, as well as the EF-Tu and the non-ribosomal DksA protein. The 50S ribosomal subunit protein L11 has been indirectly implicated in the feedback inhibition of (p)ppGpp, because ribosomes lacking this protein are unable to stimulate the synthesis of these nucleotides (PMID:11673421; PMID:17095013) (Jenvert RM and Schiavone LH 2007; Yang X and Ishiguro EE 2001). The involvement of the DksA protein in translation processes was inferred through the inspection of functional assignments. As reported (PMID:16824105) (Nakanishi N *et al.* 2006), DksA regulates the posttranscriptional stability of σ^{38} factor, which increases dramatically when (p)ppGpp levels are high. Although these are the main (p)ppGpp interactions at the translational level, the impact of these nucleotides in the translation apparatus was further analysed based on the frequency of co-annotation of gene products with (p)ppGpp nucleotides that unveiled additional participants at this level. As a result, it was possible to perceive the relevance of specific translation GTPases known to be inhibited by (p)ppGpp nucleotides, namely: the Der protein that stabilises the 50S ribosomal subunit and the EF-G that facilitates the translocation of the ribosome along the mRNA molecules (PMID:8531889) (Condon C *et al.* 1995).

Apart from identifying and contextualising numerous biological participants in the stringent response, the proposed analysis (in particular, the analysis of GO functional assignments) suggested that some of the biological entities involved in the stringent response may also participate in other stress responses. Responses to starvation, DNA damage and osmotic, oxidative and SOS stresses, which are gaining increasing attention in the last decades, are some examples of stress responses that were also evidenced in the analysis of the corpus (over 30% of the documents). Yet, it was striking to notice that the stringent response concept was barely assigned, probably because few biological entities are currently associated with this GO concept. In fact, the 50S ribosomal subunit protein L11 was the only entity in this corpus associated with the concept. Nevertheless, several biological entities that interplay between different responses to stress were identified. For example: the link between the stringent response and the response to osmotic and oxidative stresses was demonstrated via the involvement of the σ^{38} factor and EF-Tu protein; the response to DNA damage stimulus was assigned by the RecA, RecG and Mfd proteins that intervene in the early dissociation of the elongation complex stalled by ppGpp (PMID: 16039593) (Trautinger BW *et al.* 2005); and finally, the RecA regulator and the UvrABC nucleotide excision repair complex have been implicated in the DNA repair process and SOS response (Bichara M *et al.* 2007).

With the extensive list of biological players retrieved from the corpus it was possible to recognize and investigate most of the (p)ppGpp induced cellular processes. The basic participants in the stringent response were highlighted by their frequency of annotation and their representativeness in the corpus. The (p)ppGpp nucleotides and the RelA and SpoT enzymes that control (p)ppGpp basal levels, along with the RNAP, were pointed as the most significant entities in the corpus. Nonetheless, corpus analysis also revealed the involvement of entities that have been disregarded or less reported in most recent revisions (Wu J and Xie J 2009; Srivatsan A and Wang JD 2008; Potrykus K and Cashel M 2008; Jain V *et al.* 2006; Magnusson LU *et al.* 2005). In most cases, this is due to the fact that the reviews are not focused on the detailed description of the molecular mechanisms involved in the stringent response. They reflect the current state of knowledge, including the different levels of cellular processes that are triggered during this stress response, but do not specify which biological entities are involved in these processes. However, researchers often need to compile this information, not only for experimental purposes, but also for computational modelling or to better understand the complexity of the response. Hence, in this study, the stringent response was broadly described considering the large variety of biological entities that were directly or indirectly affected by the (p)ppGpp within specific metabolic, transcriptional and translational processes.

Following a retrospective analysis, the remarkable advances accomplished in the investigation of the (p)ppGpp-induced starvation response in *E. coli* became evident. The (p)ppGpp metabolism has been investigated since the 70s, as well as the transcriptional, translational and DNA replication control by these nucleotides. Technological developments have promoted the discovery of many new entities and have clarified their roles in the stringent response. At the early stage of the study of the stringent response, some traditional experimental techniques were considered decisive in the identification of the main metabolic participants (see Figure 2.7 and Table 2.9). Yet, in the last decades, research efforts have been focused on the newest molecular biology techniques, namely high-throughput detection methods. In particular, techniques based on array technology have addressed the rapid screening of biological entities as well as molecular interactions (PMID: 18039766; PMID: 17233676) (Durfee T *et al.* 2008; Chatterji D *et al.* 2007). DNA microarrays have been used to inspect the genome-wide transcriptional profiles of *E. coli* (PMID: 18039766) (Durfee T *et al.* 2008). This technology has also provided information on transcriptional regulation, determining negatively controlled promoters (typically involved in cell growth and DNA replication) and positively controlled promoters (the amino acid biosynthesis, the transcription factors, and/or alternative sigma factor genes). Although the reconstruction of the transcriptional regulatory structure of the stringent response is far from complete, these recent advances have brought a closer view of the pleiotropic nature of the response.

Finally, results showed that it is possible to scale-up conventional manual curation coping with the ever-increasing publication rate and, at the same time, provide automatic means of identifying and contextualising participants of interest. Beyond the accomplishments of the approach on this particular study, its extension to the analysis of other stress responses and/or organisms is fairly easy and interesting. Adaptation to other scenarios implicates the compilation of sets of related documents and eventually other controlled vocabularies (when considering other organisms). As future work, the incorporation of an automatic relation extraction process is foreseen as a most valuable support to the inspection and understanding of involved interactions.

2.6 REFERENCES

1. Aizenman E, Engelberg-Kulka H, and Glaser G (1996) An *Escherichia coli* chromosomal "addiction module" regulated by guanosine 3',5'-bispyrophosphate: a model for programmed bacterial cell death. *Proceedings of the National Academy of Sciences of the United States of America* 93 (12):6059-6063.

2. Artsimovitch I *et al* (2004) Structural basis for transcription regulation by alarmone ppGpp. *Cell* 117 (3):299-310.

3. Ashburner M *et al* (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics* 25 (1):25-29.

4. Battesti A and Bouveret E (2009) Bacteria possessing two RelA/SpoT-like proteins have evolved a specific stringent response involving the Acyl Carrier Protein-SpoT interaction. *Journal of Bacteriology* 191 (2):616-624.

5. Bichara M *et al* (2007) RecA-mediated excision repair: a novel mechanism for repairing DNA lesions at sites of arrested DNA synthesis. *Molecular Microbiology* 65 (1):218-229.

6. Cashel M and Kalbache B (1970) Control of ribonucleic acid synthesis in *Escherichia coli* 5. Characterization of a nucleotide associated with stringent response. *Journal of Biological Chemistry* 245 (9):2309-&.

7. Chang DE, Smalley DJ, and Conway T (2002) Gene expression profiling of *Escherichia coli* growth transitions: an expanded stringent response model. *Molecular Microbiology* 45 (2):289-306.

8. Chatterji D, Fujita N, and Ishihama A (1998) The mediator for stringent control, ppGpp, binds to the beta-subunit of *Escherichia coli* RNA polymerase. *Genes to Cells* 3 (5):279-287.

9. Chatterji D *et al* (2007) The role of the omega subunit of RNA polymerase in expression of the *rel*A gene in *Escherichia coli*. *Fems Microbiology Letters* 267 (1):51-55.

10. Chatterji D and Ojha AK (2001) Revisiting the stringent response, ppGpp and starvation signaling. *Current Opinion in Microbiology* 4 (2):160-165.

11. Chiaramello AE and Zyskind JW (1990) Coupling of DNA replication to growth rate in *Escherichia coli*: a possible role for guanosine tetraphosphate. *Journal of Bacteriology* 172 (4):2013-2019.

12. Condon C, Squires C, and Squires CL (1995) Control of ribosomal Rna transcription in *Escherichia coli*. *Microbiological Reviews* 59 (4):623-&.

13. Dubey AK *et al* (2003) CsrA regulates translation of the *Escherichia coli* carbon starvation gene, *cst*A, by blocking ribosome access to the *cst*A transcript. *Journal of Bacteriology* 185 (15):4450-4460.

14. Durfee T *et al* (2008) Transcription profiling of the stringent response in *Escherichia coli*. *Journal of Bacteriology* 190 (3):1084-1096.

15. Gong L, Takayama K, and Kjelleberg S (2002) Role of *spo*T-dependent ppGpp accumulation in the survival of light-exposed starved bacteria. *Microbiology* 148 (Pt 2):559-570.

16. Gross M, Marianovsky I, and Glaser G (2006) MazG - a regulator of programmed cell death in *Escherichia coli*. *Molecular Microbiology* 59 (2):590-601.

17. Haddadin FT, Kurtz H, and Harcum SW (2009) Serine hydroxamate and the transcriptome of high cell density recombinant *Escherichia coli* MG1655. *Applied Biochemistry and Biotechnology* 157 (2):124-139.

18. Hara A and Sy J (1983) Guanosine 5'-triphosphate, 3'-diphosphate 5'-phosphohydrolase. Purification and substrate specificity. *Journal of Biological Chemistry* 258 (3):1678-1683.

19. Harcum SW (2002) Structured model to predict intracellular amino acid shortages during recombinant protein overexpression in *E. coli*. *Journal of Biotechnology* 93 (3):189-202.

20. Hermjakob H *et al* (2004) The HUPOPSI's Molecular Interaction format - a community standard for the representation of protein interaction data. *Nature Biotechnology* 22 (2):177-183.

21. Irr JD (1972) Control of nucleotide metabolism and ribosomal ribonucleic acid synthesis during nitrogen starvation of Escherichia coli. *Journal of Bacteriology* 110 (2):554-561.

22. Jain V, Kumar M, and Chatterji D (2006) ppGpp: Stringent response and survival. *Journal of Microbiology* 44 (1):1-10.

23. Jenvert RM and Schiavone LH (2007) The flexible N-terminal domain of ribosomal protein L11 from *Escherichia coli* is necessary for the activation of stringent factor. *Journal of Molecular Biology* 365 (3):764-772.

24. Jishage M *et al* (2002) Regulation of or factor competition by the alarmone ppGpp. *Genes & Development* 16 (10):1260-1270.

25. Johansson J *et al* (2000) Nucleoid proteins stimulate stringently controlled bacterial promoters: A link between the cAMP-CRP and the (p)ppGpp regulons in *Escherichia coli*. *Cell* 102 (4):475-485.

26. Johnson GS *et al* (1979) Role of the *spo*T gene product and manganese ion in the metabolism of guanosine 5'-diphosphate 3'-diphosphate in *Escherichia coli*. *Journal of Biological Chemistry* 254 (12):5483-5487.

27. Jones PG *et al* (1992) Function of a relaxed-like state following temperature downshifts in *Escherichia coli*. *Journal of Bacteriology* 174 (12):3903-3914.

28. Keseler IM *et al* (2009) EcoCyc: A comprehensive view of *Escherichia coli* biology. *Nucl Acids Res* 37:D464-D470.

29. Lagosky PA and Chang FN (1980) Influence of amino acid starvation on guanosine 5'-diphosphate 3'-diphosphate basal-level synthesis in *Escherichia coli*. *Journal of Bacteriology* 144 (2):499-508.

30. Liang ST *et al* (2000) mRNA composition and control of bacterial gene expression. *Journal of Bacteriology* 182 (11):3037-3044.

31. Lourenco A *et al* (2009) @Note: A workbench for Biomedical Text Mining. *Journal of Biomedical Informatics* 42 (4):710-720.

32. Magnusson LU, Farewell A, and Nystrom T (2005) ppGpp: a global regulator in *Escherichia coli*. *Trends in Microbiology* 13 (5):236-242.

33. Manganelli R (2007) Polyphosphate and stress response in mycobacteria. *Molecular Microbiology* 65 (2):258-260.

34. Mogk A *et al* (2003) Small heat shock proteins, ClpB and the DnaK system form a functional triade in reversing protein aggregation. *Molecular Microbiology* 50 (2):585-595.

35. Mukherjee TK, Raghavan A, and Chatterji D (1998) Shortage of nutrients in bacteria: The stringent response. *Current Science* 75 (7):684-689.

36. Murray KD and Bremer H (1996) Control of spoT-dependent ppGpp synthesis and degradation in Escherichia coli. *Journal of Molecular Biology* 259 (1):41-57.

37. Nakanishi N *et al* (2006) ppGpp with DksA controls gene expression in the locus of enterocyte effacement (LEE) pathogenicity island of enterohaemorrhagic *Escherichia coli* through activation of two virulence regulatory genes. *Molecular Microbiology* 61 (1):194-205.

38. Nystrom T (1999) Starvation, cessation of growth and bacterial aging. *Current Opinion in Microbiology* 2 (2):214-219.

39. Paul BJ *et al* (2004) DksA: A critical component of the transcription initiation machinery that potentiates the regulation of rRNA promoters by ppGpp and the initiating NTP. *Cell* 118 (3):311-322.

40. Pernestig AK, Melefors O, and Georgellis D (2001) Identification of UvrY as the cognate response regulator for the BarA sensor kinase in *Escherichia coli*. *Journal of Biological Chemistry* 276 (1):225-231.

41. Peterson CN, Mandel MJ, and Silhavy TJ (2005) *Escherichia coli* starvation diets: essential nutrients weigh in distinctly. *Journal of Bacteriology* 187 (22):7549-7553.

42. Pingoud A and Block W (1981) The elongation factor Tu . guanosine tetraphosphate complex. *European Journal of Biochemistry* 116 (3):631-634.

43. Pingoud A *et al* (1983) The elongation factor Tu from *Escherichia coli*, aminoacyl-tRNA, and guanosine tetraphosphate form a ternary complex which is bound by programmed ribosomes. *Journal of Biological Chemistry* 258 (23):14200-14205.

44. Podkovyrov SM and Larson TJ (1996) Identification of promoter and stringent regulation of transcription of the *fab*H, *fab*D and *fab*G genes encoding fatty acid biosynthetic enzymes of *Escherichia coli*. *Nucleic Acids Research* 24 (9):1747-1752.

45. Potrykus K and Cashel M (2008) (p)ppGpp: Still Magical? *Annual Review of Microbiology* 62:35-51.

46. Rao NN, Liu SJ, and Kornberg A (1998) Inorganic polyphosphate in *Escherichia coli*: the phosphate regulon and the stringent response. *Journal of Bacteriology* 180 (8):2186-2193.

47. Ross W *et al* (1990) *E.coli* Fis protein activates ribosomal RNA transcription *in vitro* and *in vivo*. *EMBO J* 9 (11):3733-3742.

48. Sabnis NA, Yang H, and Romeo T (1995) Pleiotropic regulation of central carbohydrate metabolism in *Escherichia coli* via the gene *csr*A. *Journal of Biological Chemistry* 270 (49):29096-29104.

49. Selby CP and Sancar A (1994) Mechanisms of transcription-repair coupling and mutation frequency decline. *Microbiological Reviews* 58 (3):317-329.

50. Serres MH and Riley M (2000) MultiFun, a multifunctional classification scheme for *Escherichia coli* K-12 gene products. *Microbial & Comparative Genomics* 5 (4):205-222.

51. Shannon P *et al* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* 13 (11):2498-2504.

52. Shiba T *et al* (1997) Inorganic polyphosphate and the induction of *rpo*S expression. *Proceedings of the National Academy of Sciences of the United States of America* 94 (21):11210-11215.

53. Shinagawa H *et al* (1988) Structure and regulation of the *Escherichia coli ruv* operon involved in DNA repair and recombination. *Journal of Bacteriology* 170 (9):4322-4329.

54. Srivatsan A and Wang JD (2008) Control of bacterial transcription, translation and replication by (p)ppGpp. *Current Opinion in Microbiology* 11 (2):100-105.

55. Torok I and Kari C (1980) Accumulation of ppGpp in a *rel*A mutant of *Escherichia coli* during amino acid starvation. *Journal of Biological Chemistry* 255 (9):3838-3840.

56. Trautinger BW *et al* (2005) RNA polymerase modulators and DNA repair activities resolve conflicts between DNA replication and transcription. *Molecular Cell* 19 (2):247-258.

57. Vinella D *et al* (2005) Iron limitation induces SpoT-dependent accumulation of ppGpp in *Escherichia coli*. *Molecular Microbiology* 56 (4):958-970.

58. Walker KA, Atkins CL, and Osuna R (1999) Functional determinants of the *Escherichia coli fis* promoter: roles of -35, -10, and transcription initiation regions in the response to stringent control and growth phase-dependent regulation. *Journal of Bacteriology* 181 (4):1269-1280.

59. Wendrich TM *et al* (2002) Dissection of the mechanism for the stringent factor RelA. *Molecular Cell* 10 (4):779-788.

60. Wu J and Xie J (2009) Magic spot: (p) ppGpp. *Journal of Cellular Physiology* 220 (2):297-302.

61. Xiao H *et al* (1991) Residual Guanosine 3',5'-Bispyrophosphate Synthetic Activity of RelA Null Mutants Can be Eliminated by SpoT Null Mutations. *Journal of Biological Chemistry* 266 (9):5980-5990.

62. Yamanaka K and Inouye M (1997) Growth-phase-dependent expression of *csp*D, encoding a member of the CspA family in *Escherichia coli*. *Journal of Bacteriology* 179 (16):5126-5130.

63. Yang X and Ishiguro EE (2001) Involvement of the N terminus of ribosomal protein L11 in regulation of the RelA protein of *Escherichia coli*. *Journal of Bacteriology* 183 (22):6532-6537.

64. Zhou YN and Jin DJ (1998) The rpoB mutants destabilizing initiation complexes at stringently controlled promoters behave like "stringent" RNA polymerases in *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America* 95 (6):2908-2913.

CHAPTER 3

A SYSTEMATIC MODELLING APPROACH TO ELUCIDATE THE TRIGGERING OF THE STRINGENT RESPONSE IN RECOMBINANT *E. COLI* SYSTEMS

"The main strengths of metabolic modelling are in testing where knowledge is inadequate..."

By P. Cronjé and E. A. Boomker in Ruminant physiology:
digestion, metabolism, growth, and reproduction,
CABI, 2000

3.1 Abstract

A hybrid modelling approach, combining a stoichiometric model of the *E. coli* metabolic network and kinetic-based descriptions for the production of the recombinant protein, cell growth and ppGpp synthesis, was applied to describe metabolic bottlenecks associated with recombinant processes. The model represents the triggering of the stringent response upon the deprivation of amino acids caused by the additional drainage of biosynthetic precursors for the production of recombinant proteins. The equation for ppGpp synthesis allows to estimate the accumulation of this molecule above its basal levels once amino acid shortages occur. The capability to predict these stress-responsive events might be crucial in the design of optimal cultivation strategies.

3.2 Background

The production of recombinant proteins can challenge cells with different levels of stress and metabolic burden (Hoffmann F and Rinas U, 2004; Schweder T *et al.*, 2002). The cellular processes for plasmid DNA replication and expression require the drainage of biosynthetic precursors, energy and other cellular resources that are shared with the host cell's metabolic processes. Consequently, the competition for a limited pool of cellular resources will provoke serious perturbations in metabolism (Bentley WE *et al.*, 1990; Glick BR, 1995). Typically, it promotes the reduction of cellular growth and protein productivity in recombinant cells (Glick BR, 1995). In particular, the metabolic load imposed by the overexpression of recombinant proteins, with an amino acid composition that is often different from the average composition of biomass proteins, leads to the imbalance of the cellular metabolism resulting in the over-accumulation of some metabolites and shortage of others, like some amino acids. In the past few years, the association of this metabolic burden with other cellular events, like the stringent response, has been demonstrated (Andersson L *et al.*, 1996; Chao YP *et al.*, 2002; Haddadin FT *et al.*, 2009; Hoffmann F and Rinas U, 2004; Schweder T *et al.*, 1995) and the understanding of the whole scenario underlying these cellular events in recombinant bioprocesses is fundamental for productivity purposes. For example, the stringent response has been characterized by the down-regulation of stable RNAs (i.e., tRNA and rRNA) and protein synthesis, and the simultaneous up-regulation of protein degradation (Ferullo DJ and Lovett ST, 2008; Jain V *et al.*, 2006; Magnusson LU *et al.*, 2005), which results in considerable losses during recombinant bioprocesses.

The unusual accumulation of a guanosine nucleotide, termed GDP 3'-diphosphate or GTP 3'-diphosphate, collectively called (p)ppGpp, was identified as one of the key factors that provide bacteria the ability to survive under hostile conditions (Chatterji D and Ojha AK, 2001; Jain V *et al.*, 2006; Magnusson LU *et al.*, 2005; Mukherjee TK *et al.*, 1998; Torok I and Kari C, 1980; Wu J and Xie J, 2009), such as the lack of amino acids. The discovery that these unusual guanosine nucleotides are accumulated in response to starvation was followed by extensive studies on the downstream pathways of the stringent response, which suggested that accumulating ppGpp is an important link between nutritional stress and bacterial adaptation (Cashel M and Kalbache B, 1970; Chatterji D and Ojha AK, 2001; Foster PL, 2005; Mukherjee TK *et al.*, 1998; Nystrom T, 2004b; Nystrom T, 2002).

It was initially proposed (Haseltin WA and Block R, 1973) that the ratio of aminoacylated transfer RNAs (tRNAs) to free tRNAs is one of the critical parameters that regulate the synthesis of ppGpp. When free tRNA is encountered at the A-site of the 50S ribosome, protein synthesis is delayed, resulting in an idling reaction in which ribosome-bound RelA is activated to synthesize ppGpp (pppGpp is produced first and then it is converted to ppGpp). Thus, an increase in the population of free tRNA during starvation leads to an accumulation of ppGpp (Chatterji D and Ojha AK, 2001). The ppGpp levels are controlled by two enzymes, RelA and SpoT. RelA is the major ppGpp synthase (also known as ppGpp synthetase I) whose activity occurs in the presence of Mg^{2+}, whereas the SpoT protein mainly carries out a pyrophosphate hydrolase activity in the presence of Mn^{2+}, although it has also mild synthetic activity (ppGpp synthetase II).

The pleiotropic effects of the intracellular accumulation of ppGpp have been explored and major alterations in the transcription of many stringently controlled genes were found (Magnusson LU *et al.*, 2005; Nystrom T, 2004a; Srivatsan A and Wang JD, 2008). In general, genes involved in cell proliferation and growth are negatively regulated by ppGpp, whereas genes implicated in maintenance and stress defence are positively regulated by this regulator.

3.3 Modelling amino acid shortages towards the synthesis of ppGpp

The view of ppGpp primarily as a regulator of gene transcription has been expanded and it is now clear that the response controlled by ppGpp is crucial for cell survival during the adaptation to stressful conditions (Magnusson LU *et al.*, 2005). To understand the impact of recombinant processes in metabolism it is crucial to capture the molecular basis of the ppGpp synthesis in response to metabolic *stimuli*, like the deprivation of amino acids. However, this stress-responsive process is still poorly understood, though it is acknowledged that during recombinant processes the accumulation of this regulator above a threshold level promotes the stringent response.

Here we propose a model to examine two fundamental events: the amino acid deprivation caused by the unbalanced drainage of biosynthetic resources imposed by the recombinant protein production; and the consequent induction of the ppGpp synthesis. A deterministic mathematical modelling approach based on a hybrid system, i.e. a system in which discrete events are combined with kinetic expressions, was implemented to capture the behaviour of these cellular phenomena.

3.3.1 Model description

Our aim is to analyse the initial steps of the *E. coli* stringent response by creating a modelling structure that acts as a basis to analyse the impact of the synthesis of heterologous proteins on the cellular behaviour, specifically in the intracellular accumulation of ppGpp. A combined modelling approach was developed to represent the key metabolic bottlenecks in the *E. coli* metabolism during recombinant processes and, subsequently, the induction of the ppGpp biosynthetic pathway (Figure 3.1).

First, the genome-scale metabolic model of *E. coli i*JR904 (Reed JL *et al.*, 2003) was used for simulations using Flux Balance Analysis (FBA) (Price ND *et al.*, 2003; Schilling CH *et al.*, 1999; Varma A and Palsson BO, 1994) to determine the metabolic fluxes leading to the biosynthesis of amino acids. This approach simulates the *E. coli* metabolic network under steady-state conditions, based on the mass balance of metabolites constrained by stoichiometry and thermodynamics, and estimates the optimal metabolic flux distribution subjected to an objective function. Details about the mathematical formulations were presented in Chapter 1. In this work, the FBA simulation was performed using the OptFlux tool (Rocha I *et al.*, 2010) defining the maximization of biomass formation (μ) as the objective function. The predicted amino acid biosynthetic fluxes (r_{aa}) were then included in the dynamic model that was used for simulations using the Systems Biology Toolbox 2 (Schmidt H and Jirstrand M, 2006) implemented in MATLAB (version 2009b, The Mathworks, Inc). The ODE model is presented in Appendix A.

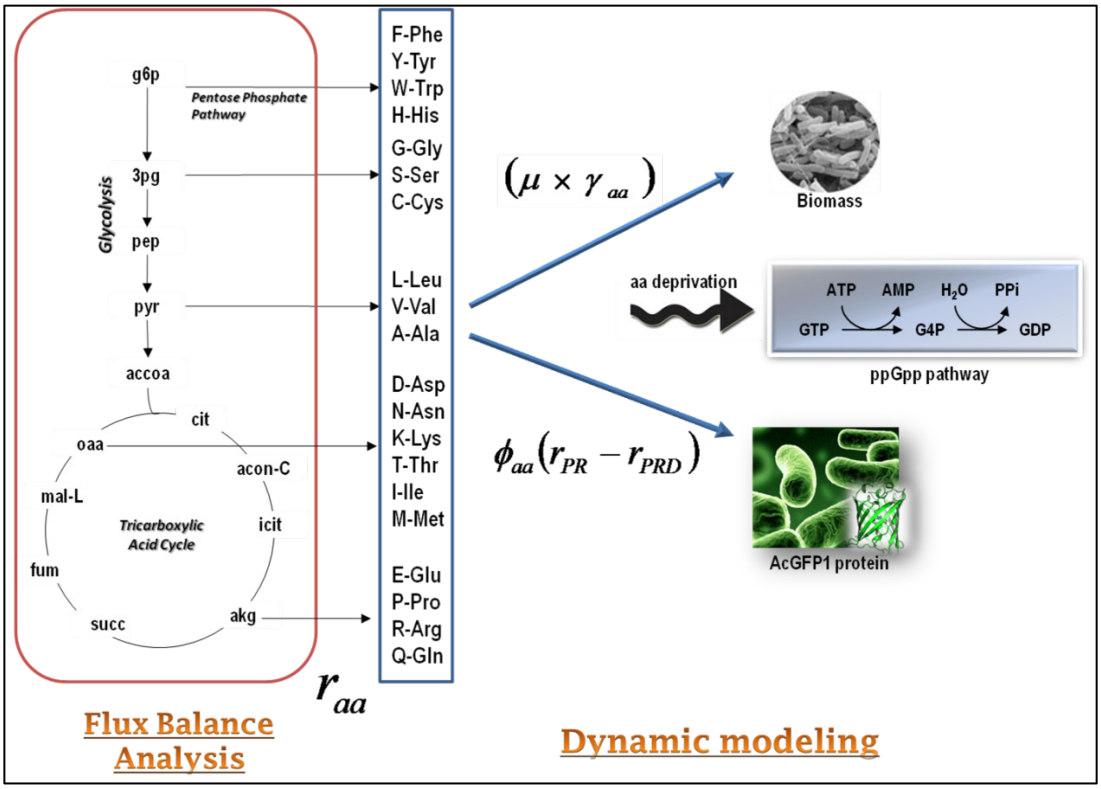


Figure 3.1. Proposed modelling approach that combines Flux Balance Analysis (FBA) with the genome-scale metabolic model of *E. coli i*JR904 (Reed JL *et al.*, 2003) and dynamic rate equations for the intracellular pools: amino acids (*aa*), biomass, recombinant protein (*PR*) and ppGpp (*G4P*).

The amino acid reactions (r_{aa}) are depicted as boundary reactions linking the stoichiometric model and the dynamic model.

Variables considered in this model are presented in Table 3.1.

Table 3.1. Variables included in the model.

Model variables	Units
Biomass (X)	$gX.L^{-1}$
Carbon source (S)	$gS.L^{-1}$
Recombinant protein (PR)	$gPR.gX^{-1}$
Amino acids (aa)	$gaa.gX^{-1}$
ppGpp (G4P)	$pmolG4P.gX^{-1}$
Amino acid reactions (r_{aa})	$gaa.gX^{-1}.h^{-1}$
Biomass-associated stoichiometric coefficients (γ_{aa})	$gaa.gX^{-1}$
Recombinant protein-associated stoichiometric coefficients (ϕ_{aa})	$gaa.gPR^{-1}$

Differential mass balance equations for amino acid biosynthesis reaction rates were based on the previously developed model by Bentley and Kompala (Bentley WE and Kompala DS, 1989):

$$\frac{d[aa]}{dt} = r_{aa} - (\mu \times \gamma_{aa}) - \phi_{aa}(r_{PR} - r_{PRD}) \qquad (1)$$

Changes in the intracellular amino acids pool were calculated by subtracting the term associated with the consumption of a specific amino acid (*aa*) for biomass formation ($\mu \times \gamma_{\mathrm{aa}}$) and the term associated with the consumption of that amino acid for recombinant protein polymerization ($\phi_{\mathrm{aa}} \times r_{\mathrm{PR}}$) and its degradation ($\phi_{\mathrm{aa}} \times r_{\mathrm{PRD}}$), to the amino acid synthesis flux (r_{aa}). The stoichiometric coefficients (γ_{aa} and ϕ_{aa}) describe the mass requirements of that amino acid for the production of the biomass-associated proteins and recombinant AcGFP1 protein (given in Table 3.2). The r_{aa} term represents the reaction fluxes for each amino acid pool predicted by the FBA approach, which were also designated as boundary reactions. Amino acid biosynthesis fluxes were calculated by subtracting fluxes that account for the consumption of a particular amino acid (e.g. glutamate is often used as a precursor in other reactions) from those that lead to the production of that amino acid (in Table 3.3 the values of these fluxes are given for a wild-type strain under the conditions detailed in (Reed JL *et al.*, 2003)). All rates are specific for biomass concentration, i.e., they are expressed per grams of biomass.

Table 3.2. Stoichiometric coefficients of amino acids in the reactions leading to biomass (X) and recombinant protein (PR) formation, calculated on the basis of the amino acids composition based on (Neidhardt FC *et al.*, 1990) and on AcGFP data, respectively.

Amino acid (aa)	γ_{aa} (g.g$_X^{-1}$)	ϕ_{aa}(g.g$_{PR}^{-1}$)
A-Ala	0.0434	0.0366
R-Arg	0.0489	0.0417
N-Asn	0.0302	0.0678
D-Asp	0.0305	0.0774
C-Cys	0.0105	0.0083
Q-Gln	0.0365	0.0350
E-Glu	0.0368	0.0755
G-Gly	0.0437	0.0642
H-His	0.0140	0.0849
I-Ile	0.0362	0.0673
L-Leu	0.0561	0.0808
K-Lys	0.0476	0.0900
M-Met	0.0218	0.0408
F-Phe	0.0290	0.0735
P-Pro	0.0242	0.0433
S-Ser	0.0215	0.0611
T-Thr	0.0287	0.0693
W-Trp	0.0110	0.0070
Y-Tyr	0.0237	0.0744
V-Val	0.0470	0.0601

Table 3.3. Amino acid biosynthesis reactions based on the metabolic model of *E. coli* *i*JR904 (Reed JL *et al.*, 2003) and their flux, calculated for a wild-type strain.

Amino acid	Producing reactions	Consuming reactions	Overall synthesis flux $(g.g_x^{-1}.h^{-1})$
A-Ala	r_{ALATA_L}	r_{ALAR} r_{VPAMT} r_{UAMAS}	0.0400
R-Arg	r_{ARGSL}		0.0451
N-Asn	r_{ASNS2}		0.0279
D-Asp	r_{ASPTA}	r_{ASNS2} r_{ASPO3} r_{ADSS} r_{ASPCT} r_{ARGSS} r_{ASP1DC} r_{PRASCS} r_{ASPK}	0.0281
C-Cys	r_{CYSS}	r_{PPNCL2} r_{SHSL1}	0.0097
Q-Gln	r_{GLNS}	r_{GF6PTA} r_{IG3PS} r_{GLUPRT} r_{CTPS2} r_{ADCS} r_{GMPS2} r_{PRFGS} r_{ANS}	0.0337
E-Glu	r_{GF6PTA} r_{IG3PS} r_{PRFGS} r_{ADCS} r_{GMPS2} r_{CTPS2} r_{GLUDy} r_{GLUPRT} r_{ANS}	r_{ALATA_L} r_{ACOTA} r_{GLNS} r_{TYRTA} r_{ASPTA} r_{GLU5K} r_{PSERT} r_{PHETA1} r_{UNK3} r_{GLUR} r_{HSTPT} r_{ILETA} r_{ACGS} r_{DHFS} r_{SDPTA} r_{LEUTAi}	0.0339
G-Gly	r_{GHMT2}	r_{PRAGSr}	0.0760
H-His	r_{HISTD}		0.0129
I-Ile	r_{ILETA}		0.0333
L-Leu	r_{LEUTAi}		0.0517
K-Lys	r_{DAPDC}		0.0439
M-Met	r_{UNK3} r_{METS}	r_{METAT}	0.0201
F-Phe	r_{PHETA1}		0.0268
P-Pro	r_{P5CR}		0.0223
S-Ser	r_{PSP_L}	r_{SERAT} r_{GHMT2} r_{PSSA_EC}	0.0198
T-Thr	r_{THRAr}	r_{THRD_L}	0.0264
W-Trp	r_{TRPAS2}		0.0102
Y-Tyr	r_{TYRTA}		0.0219
V-Val	r_{VPAMT}		0.0434

This model describes cellular growth and recombinant protein production in a batch fermentation mode with constant volume, where the only available carbon source is glucose and maintenance coefficients were not considered. The specific growth rate (μ) for the recombinant bacteria was estimated as a function of the substrate concentration (*S*) based on the Monod equation:

$$\mu = \mu_0 \times \frac{S}{K_s + S} \tag{2}$$

where μ_0 is referred to the maximum specific growth rate of the wild-type cells predicted by the FBA simulation and K_s is the Monod affinity constant corresponding to an approximated value from recombinant *E. coli* growth measurements. The mass balance equations for biomass (*X*) and substrate (*S*) concentrations were defined as:

$$\frac{dX}{dt} = \mu \times X \tag{3}$$

$$\frac{dS}{dt} = -\mu \times \frac{X}{Y_{X/S}} \tag{4}$$

where $Y_{X/S}$ is the biomass yield on substrate for recombinant *E. coli* cells.

The recombinant protein formation was induced at time 0 and the rates of synthesis and degradation are described using the following expressions based on (Palaiomylitou MA *et al.*, 2002):

$$r_{PR} = K_e \times \frac{PR}{K_t + PR} \tag{5}$$

$$r_{PRD} = K_{PRD} * PR \tag{6}$$

where *PR* is the recombinant protein concentration, K_e is the maximal rate of protein synthesis, incorporating the rate constants for transcription and translation and K_t is the saturation constant, which depends on the host-plasmid system used. For protein degradation, the protein denaturation rate constant (K_{PRD}) is given independently of the growth rate. The mass balance equation for the recombinant protein concentration was given by the protein synthesis rate and an empirical function of the intracellular recombinant protein concentration:

$$\frac{dPR}{dt} = r_{PR} - r_{PRD} \tag{7}$$

At last, the model describes the accumulation of ppGpp or *G4P* in response to the shortage of any amino acid pool (eqs. 8 to 11). The intracellular concentration of ppGpp is given by the rates of degradation (r_{G4PD}) and synthesis (r_{G4P}). The ppGpp synthesis rate was defined as a function of the intracellular levels of amino acids given by *f(aa)* and the parameter K_{G4P}.

$$\frac{dG4P}{dt} = r_{G4P} - r_{G4PD} \tag{8}$$

$$r_{G4PD} = K_{G4PD} \times G4P \tag{9}$$

$$r_{G4P} = K_{G4P} \times f(aa) \tag{10}$$

$$f(aa) = 0.2 \times e^{-1E4 \times aa} \tag{11}$$

The terms K_{G4P} and K_{G4PD} refer to parameters for the synthesis and degradation of ppGpp. The accumulation of ppGpp was empirically described as an exponential function that defines a relation *f* between the stimulus (i.e., levels of *aa*) and the expected value of the response (i.e., levels of ppGpp or *G4P*) (Torok I and Kari C, 1980).

All model parameters are given in Table 3.4.

Table 3.4. Model parameters.

Parameter	Value	References
$\mu_0\ (h^{-1})$	0.36	Inferred from experimental data
$K_s\ (g\ L^{-1})$	0.05	(Harcum SW, 2002)
$Y_{X/S}\ (g_X\ g_S^{-1})$	0.40	Inferred from experimental data
$K_e\ (g_{PR}\ g_X^{-1}\ h^{-1})$	4.09	Adapted from (Bentley WE and Kompala DS, 1989; Palaiomylitou MA *et al.*, 2002)
$K_t\ (g_{PR}\ g_X^{-1})$	5.39	Adapted from (Bentley WE and Kompala DS, 1989; Palaiomylitou MA *et al.*, 2002)
$K_{PRD}\ (h^{-1})$	0.04	Adapted from (Bentley WE and Kompala DS, 1989; Palaiomylitou MA *et al.*, 2002)
$K_{G4P}\ (h^{-1})$	0.2	Adapted from (Dedhia N *et al.*, 1997; Torok I and Kari C, 1980)
$K_{G4PD}\ (h^{-1})$	0.002	Adapted from (Dedhia N *et al.*, 1997; Torok I and Kari C, 1980)

3.4 Simulation results

This model can be used to estimate amino acid shortages caused by recombinant processes and the consequent activation of the ppGpp synthesis. The systematic evaluation of the deprivation of amino acids during the translation process gives a preliminary recognition of potential metabolic bottlenecks that ultimately lead to the activation of the stringent response. The stoichiometric coefficients associated with biomass and recombinant protein formation requirements were fundamental to estimate the mass balance dynamics for each amino acid pool. Information on these parameters and kinetic data on the synthesis of biomass (μ) and recombinant protein (r_{PR}) formation, allowed estimating the evolution over time of the intracellular amino acid levels. To simulate the stringent response, a function determining the relation of ppGpp accumulation and decreasing levels of amino acids, as well as an expression (or event) defining the delay on the formation of biomass and recombinant protein were implemented in the computation.

3.4.1 Amino acids deprivation

In this model, the abundance of amino acids was estimated over time (eq. 1) to detect at what extent the withdrawn of amino acids for biomass and recombinant protein formation exceeds the biosynthetic capacities of *E. coli* cells. To illustrate the impact of recombinant protein production, the concentration of amino acids was allowed to become negative in this exercise (although, clearly, this has no biological meaning). As shown in Figure 3.2, it is clear that the deprivation of most amino acids would be extensive, if cells were not capable to counteract these events.

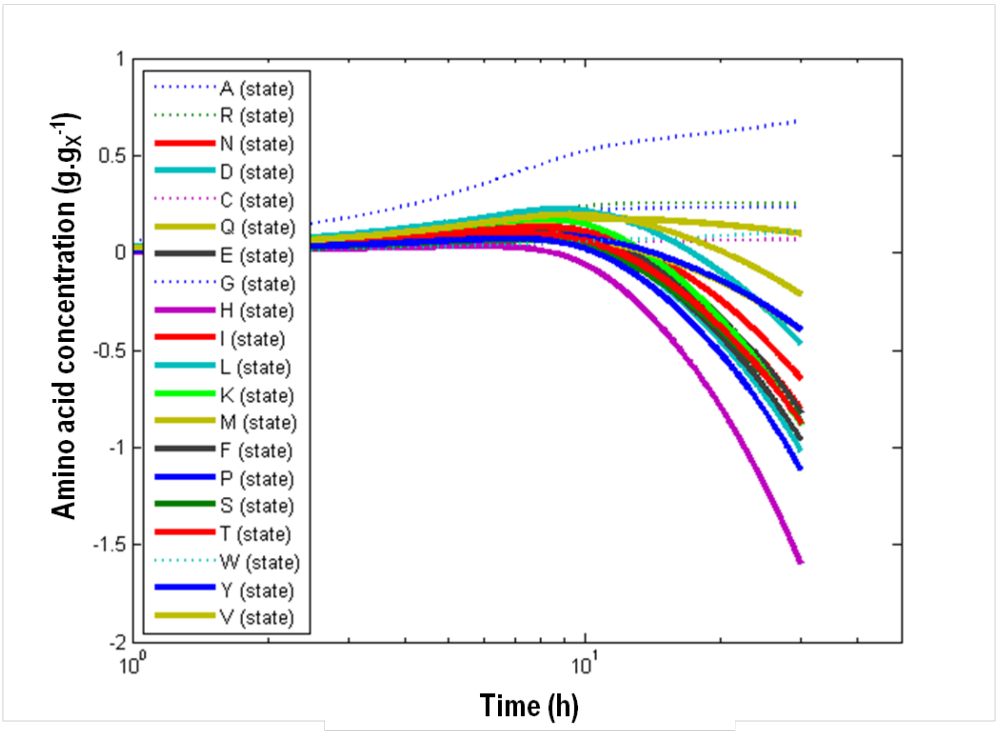


Figure 3.2. Dynamics of the intracellular concentrations of amino acids.

Dashed lines indicate those amino acid pools that seem not to be deprived over time. Concentrations below zero indicate that amino acids consumption exceeds the synthesis rate.

3.4.2 ppGpp biosynthesis

To simulate the triggering of the stringent response caused by the deprivation of amino acids shown above, the dynamics of ppGpp was included in the simulation (eqs 8 to 11). In Figure 3.3, the cellular response to perturbations in the amino acids pools is demonstrated by the increasing levels of ppGpp (*G4P*) when histidine (*H*) reached concentration levels close to zero. The synthesis of this signalling molecule is described as a function of the first amino acid to be depleted in the intracellular pool (eq. 10 and 11). In these particular conditions, histidine (*H*) was the first amino acid to reach levels close to zero, determining the exponential increase in ppGpp levels. At high levels, the concentration of the ppGpp regulator is controlled by the degradation rate (eq. 9) of the SpoT enzyme.

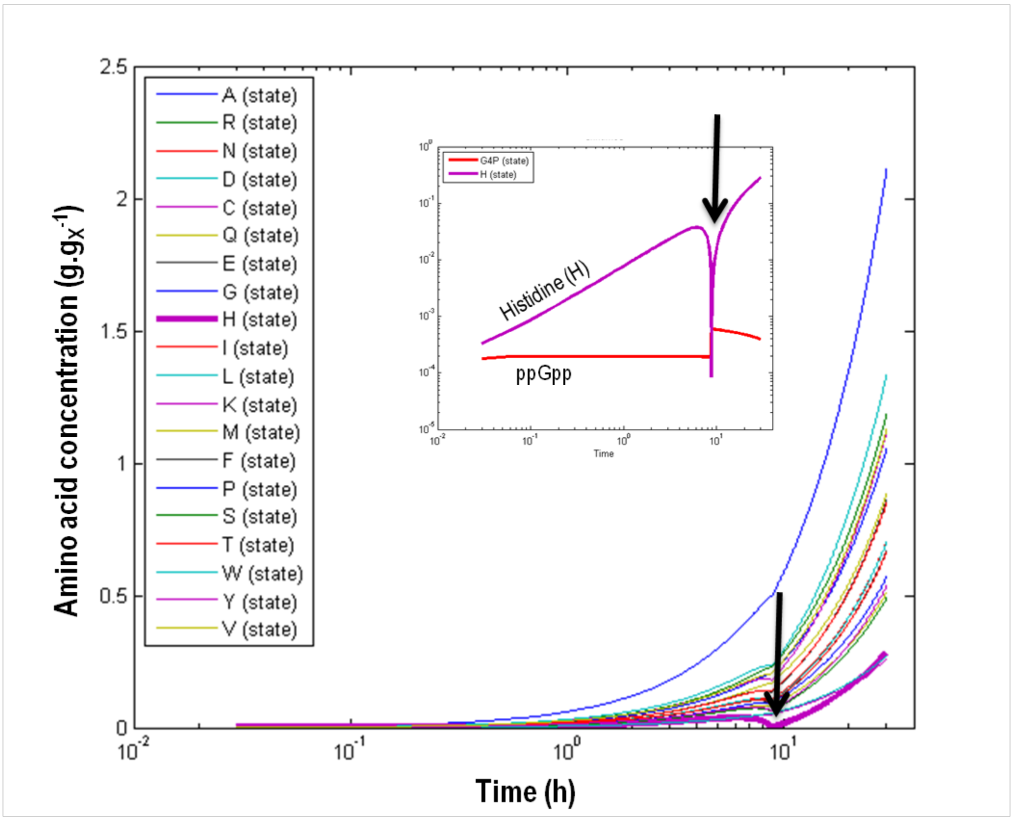


Figure 3.3. Intracellular pool concentrations of amino acids and ppGpp (*G4P*).

The arrow indicates when histidine (*H*) concentration level falls to zero.

To illustrate the impact of the ppGpp accumulation in the biomass and recombinant protein formation, an event was defined, i.e. an expression to simulate discrete state changes when a given condition is fulfilled. A simple syntax expression was used to determine that when the histidine (*H*) pool falls below zero, the parameters for the degradation of ppGpp (K_{G4PD}), maximum specific growth rate (μ_0) and the transcription and translation rate saturation constant (K_t) were set to 0.02, 0.2 and 100, respectively (Eq. 12).

$$event = lt(H,0.0),K_{G4PD},0.02,\ \mu_0,0.2,\ K_t,100 \qquad (12)$$

The results of the addition of this event are shown in Figure 3.4.

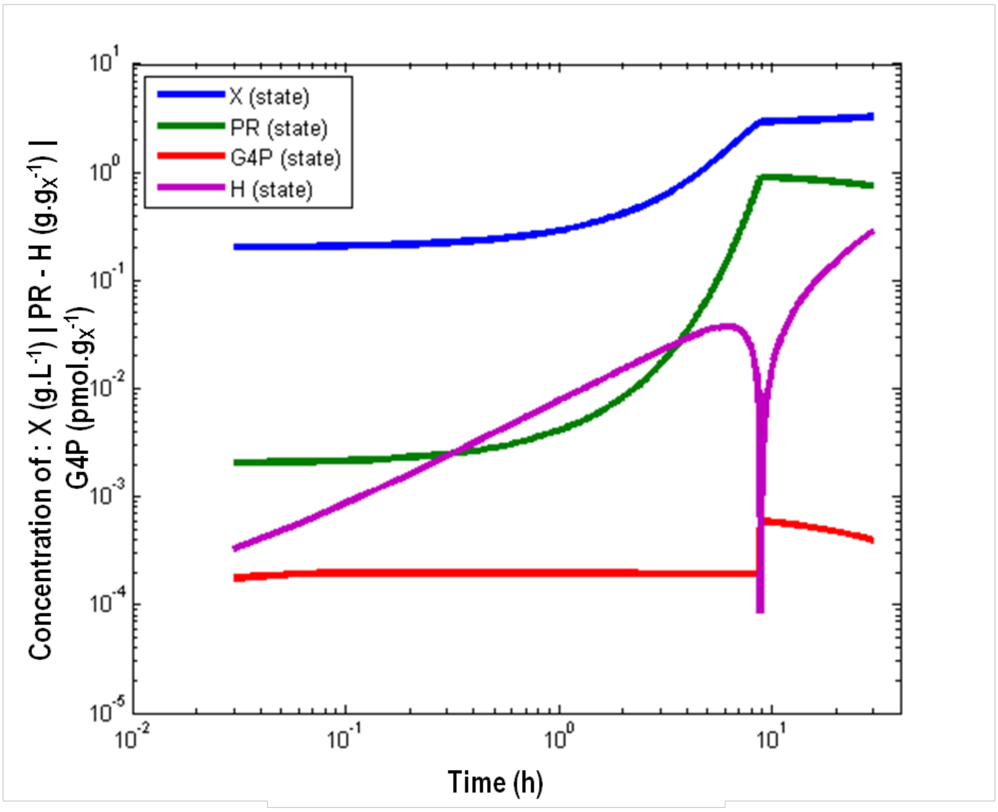


Figure 3.4. Intracellular concentrations of histidine (*H*), ppGpp (*G4P*), recombinant protein (*PR*), and biomass (*X*) accumulation during the recombinant bioprocess simulation.

Basal levels for ppGpp are maintained until the amino acid histidine (*H*) drops to zero. At that point, the biomass formation and the recombinant protein production are stalled (see eq. 12).

3.5 Discussion

Our current understanding on the behaviour of recombinant systems is based on empirical descriptions that disregard the involvement of cellular events, like stress-responsive mechanisms. As observed in most cellular systems, in particular in recombinant *E. coli* cells (Sanden AM *et al.*, 2003; Tedin K and Bremer H, 1992), stimulus-responses (e.g. amino acid shortages) are fundamental to sense and react to metabolic perturbations.

The proposed model aims at providing a systematic approach capable to predict amino acid shortages based on the biosynthetic capabilities of the *E. coli* metabolism when induced to produce recombinant proteins. A combined modelling approach based on the FBA simulation of the *E. coli* metabolic network and a kinetics-based dynamic method to simulate the behaviour of the intracellular amino acids pools during the recombinant bioprocess were implemented. The definition of amino acid biosynthetic reactions as boundary reactions was central to follow the dynamics of amino acids concentrations, which may constitute metabolic bottlenecks during recombinant bioprocess. The existence of one or more bottlenecks is expected to affect most cellular processes, including the synthesis of recombinant products, if cells are deprived of mechanisms capable to circumvent and respond to such events.

The accumulation of ppGpp during recombinant bioprocesses has been pointed as a consequence of amino acid shortages due to the overexpression of recombinant proteins. As the amino acids composition of these proteins is frequently different from the average composition of biomass proteins, the unbalanced drainage of these biosynthetic resources is often indicated as the major cause for the commonly observed amino acid shortages. In the proposed model, stoichiometric coefficients determining the amount of amino acids that are drained from the intracellular pools toward biomass and recombinant protein formation, establish the basis for estimating possible metabolic bottlenecks in recombinant process. The dynamic equations for the amino acid intracellular pools describe the time evolution of their concentrations and, once one amino acid reaches concentration levels close to zero, the model estimates the accumulation of ppGpp above its basal levels. This stimulus-response model integrates an expression that outlines the production of ppGpp as a negative exponential function of the intracellular concentration of amino acids. When cells sense the depletion of amino acids, the activity of the RelA enzyme is sharply increased. Normally, it is the ratio of aminoacylated transfer RNAs (tRNAs) to free tRNAs that triggers the synthesis of ppGpp. Yet, to simplify the model representation, we did not include this ratio as the triggering factor for the stringent response, given that the limiting substrate in any tRNA charging reaction is its specific amino acid. The depletion of the amino acids results, logically, in the increase of free tRNAs that will bind to the ribosome and induce the accumulation of ppGpp. The pleiotropic effects of this global regulator have been described (Chang DE *et al.*, 2002; Durfee T *et al.*, 2008; Traxler MF *et al.*, 2006), but the most significant in recombinant bioprocesses are growth and protein synthesis decline. Thus, the kinetic parameters used to describe these reaction rates were changed to reproduce such effects.

The possibility to predict such dynamic phenomena provides an important advantage when designing recombinant fermentation processes. The optimization of cultivation processes to produce recombinant proteins is still a difficult task. Despite the advances in novel expression systems, it is acknowledged that culture control procedures have a significant impact in the productivity of the bioprocess. And thus, it is fundamental to achieve a finer balance between the expression levels of the desired protein, along with the tight control of the culture performance. In order to predict the process dynamics that hinders the productivity of the recombinant process, the development of systematic modelling approaches capable to describe these systems is central. Most of the modelling strategies used for optimization and control of bioprocesses are based on empirical models that do not sufficiently reflect these dynamic processes. The design of optimal recombinant cultivation processes should, however, consider the complexity behind these cellular processes to enhance protein productivity.

3.6 Future work

Studies in the *E. coli* stringent response have been disclosing the complexity of the mechanisms underlying this stress response. The role of ppGpp in microbial cells has been expanded from the coordination of cellular activities to cope with amino acid starvation, to many other cellular processes, like the coordination of processes associated with pathogenesis, virulence, sporulation, etc (Chatterji D and Ojha AK, 2001; Wu J and Xie J, 2009).

The representation of the dynamics of the overall processes leading to the ppGpp activity and subsequent coordinated activities is still complex, but the development of modelling approaches capable to describe at least some parts of the system, is hoped to bring major advances. The proposed modelling scheme, combining a large scale stoichiometric network simulated by FBA and kinetic-based equations, shows some advantages when describing the dynamic behaviour of several metabolic variables with limited information on kinetic parameters. However, it is still limited to the early stages of the stringent response and transcriptional, translational and posttranscriptional events were not included. The inclusion of regulatory information that, finally, will reflect alterations in the metabolic flux distributions, is foreseen. Some approaches have already been described, like the implementation of Boolean logic rules (Covert MW *et al.*, 2001), structure-oriented analyses based on network topology (Stelling J *et al.*, 2002) or the representation of dynamic descriptions for metabolic reactions and gene regulation (Varner JD, 2000). Whilst the dynamic mathematical modelling offers detailed descriptions of the system's behaviour, the lack of data and unknown kinetic parameter values, limits its application. However, the combination of mathematical approaches that integrate stoichiometry, kinetics and gene regulation seems a promising strategy to analyse large scale models.

3.7 References

1. Andersson L *et al* (1996) Impact of plasmid presence and induction on cellular responses in fed batch cultures of *Escherichia coli*. *Journal of Biotechnology* 46 (3):255-263.

2. Bentley WE and Kompala DS (1989) A novel structured kinetic modeling approach for the analysis of plasmid instability in recombinant bacterial cultures. *Biotechnology and Bioengineering* 33:49-61.

3. Bentley WE *et al* (1990) Plasmid-encoded protein - The principal factor in the metabolic burden associated with recombinant bacteria. *Biotechnology and Bioengineering* 35 (7):668-681.

4. Cashel M and Kalbache B (1970) Control of ribonucleic acid synthesis in *Escherichia coli* 5. Characterization of a nucleotide associated with stringent response. *Journal of Biological Chemistry* 245 (9):2309-&.

5. Chang DE, Smalley DJ, and Conway T (2002) Gene expression profiling of *Escherichia coli* growth transitions: an expanded stringent response model. *Molecular Microbiology* 45 (2):289-306.

6. Chao YP, Chiang CJ, and Hung WB (2002) Stringent regulation and high-level expression of heterologous genes in *Escherichia coli* using T7 system controllable by the *ara*BAD promoter. *Biotechnology Progress* 18 (2):394-400.

7. Chatterji D and Ojha AK (2001) Revisiting the stringent response, ppGpp and starvation signaling. *Current Opinion in Microbiology* 4 (2):160-165.

8. Covert MW, Schilling CH, and Palsson B (2001) Regulation of gene expression in flux balance models of metabolism. *Journal of Theoretical Biology* 213 (1):73-88.

9. Dedhia N *et al* (1997) Improvement in recombinant protein production in ppGpp-deficient *Escherichia coli*. *Biotechnology and Bioengineering* 53 (4):380-386.

10. Durfee T *et al* (2008) Transcription profiling of the stringent response in *Escherichia coli*. *Journal of Bacteriology* 190 (3):1084-1096.

11. Ferullo DJ and Lovett ST (2008) The stringent response and cell cycle arrest in *Escherichia coli*. *PLoS Genetics* 4 (12):e1000300.

12. Foster PL (2005) Stress responses and genetic variation in bacteria. *Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis* 569 (1-2):3-11.

13. Glick BR (1995) Metabolic Load and Heterologous Gene-Expression. *Biotechnology Advances* 13 (2):247-261.

14. Haddadin FT, Kurtz H, and Harcum SW (2009) Serine hydroxamate and the transcriptome of high cell density recombinant *Escherichia coli* MG1655. *Applied Biochemistry and Biotechnology* 157 (2):124-139.

15. Harcum SW (2002) Structured model to predict intracellular amino acid shortages during recombinant protein overexpression in *E. coli*. *Journal of Biotechnology* 93 (3):189-202.

16. Haseltin WA and Block R (1973) Synthesis of guanosine tetraphosphate and pentaphosphate requires presence of a codon-specific, uncharged transfer ribonucleic acid in acceptor site of ribosomes - (Stringent control ppGpp (Msi) and pppGpp (Msii) protein synthesis *Escherichia coli*). *Proceedings of the National Academy of Sciences of the United States of America* 70 (5):1564-1568.

17. Hoffmann F and Rinas U (2004) Stress induced by recombinant protein production in *Escherichia coli*. *Advances in Biochemical Engineering / Biotechnology* 89:73-92.

18. Jain V, Kumar M, and Chatterji D (2006) ppGpp: Stringent response and survival. *Journal of Microbiology* 44 (1):1-10.

19. Magnusson LU, Farewell A, and Nystrom T (2005) ppGpp: a global regulator in *Escherichia coli*. *Trends in Microbiology* 13 (5):236-242.

20. Mukherjee TK, Raghavan A, and Chatterji D (1998) Shortage of nutrients in bacteria: The stringent response. *Current Science* 75 (7):684-689.

21. Neidhardt FC, Ingraham JL, Schaechter M (1990) Physiology of the bacterial cell - a molecular approach. 1st ed. Sinauer Associates: Sunderland, USA.

22. Nystrom T (2004a) Growth versus maintenance: a trade-off dictated by RNA polymerase availability and sigma factor competition? *Molecular Microbiology* 54 (4):855-862.

23. Nystrom T (2004b) Stationary-phase physiology. *Annual Review of Microbiology* 58:161-181.

24. Nystrom T (2002) Aging in bacteria. *Current Opinion in Microbiology* 5 (6):596-601.

25. Palaiomylitou MA *et al* (2002) A kinetic model describing cell growth and production of highly active, recombinant ice nucleation protein in *Escherichia coli*. *Biotechnology and Bioengineering* 78 (3):321-332.

26. Price ND *et al* (2003) Genome-scale microbial *in silico* models: the constraints-based approach. *Trends in Biotechnology* 21 (4):162-169.

27. Reed JL *et al* (2003) An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biology* 4 (9):R54.

28. Rocha I *et al* (2010) OptFlux: an open-source software platform for in silico metabolic engineering. *BMC Systems Biology* 4:45.

29. Sanden AM *et al* (2003) Limiting factors in *Escherichia coli* fed-batch production of recombinant proteins. *Biotechnology and Bioengineering* 81 (2):158-166.

30. Schilling CH, Edwards JS, and Palsson BO (1999) Toward metabolic phenomics: Analysis of genomic data using flux balances. *Biotechnology Progress* 15 (3):288-295.

31. Schmidt H and Jirstrand M (2006) Systems Biology Toolbox for MATLAB: a computational platform for research in systems biology. *Bioinformatics* 22 (4):514-515.

32. Schweder T, Hofmann K, and Hecker M (1995) *Escherichia coli* K12 *rel*A strains as safe hosts for expression of recombinant DNA. *Applied Microbiology and Biotechnology* 42 (5):718-723.

33. Schweder T *et al* (2002) Role of the general stress response during strong overexpression of a heterologous gene in *Escherichia coli*. *Applied Microbiology and Biotechnology* 58 (3):330-337.

34. Srivatsan A and Wang JD (2008) Control of bacterial transcription, translation and replication by (p)ppGpp. *Current Opinion in Microbiology* 11 (2):100-105.

35. Stelling J *et al* (2002) Metabolic network structure determines key aspects of functionality and regulation. *Nature* 420 (6912):190-193.

36. Tedin K and Bremer H (1992) Toxic effects of high-levels of ppGpp in *Escherichia coli* are relieved by *rpo*B mutations. *Journal of Biological Chemistry* 267 (4):2337-2344.

37. Torok I and Kari C (1980) Accumulation of ppGpp in a *rel*A mutant of *Escherichia coli* during amino acid starvation. *Journal of Biological Chemistry* 255 (9):3838-3840.

38. Traxler MF, Chang DE, and Conway T (2006) Guanosine 3 ',5 '-bispyrophosphate coordinates global gene expression during glucose-lactose diauxie in *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America* 103 (7):2374-2379.

39. Varma A and Palsson BO (1994) Metabolic Flux Balancing - Basic Concepts, Scientific and Practical Use. *Bio-Technology* 12 (10):994-998.

40. Varner JD (2000) Large-scale prediction of phenotype: Concept. *Biotechnology and Bioengineering* 69 (6):664-678.

41. Wu J and Xie J (2009) Magic spot: (p) ppGpp. *Journal of Cellular Physiology* 220 (2):297-302.

3.8 Appendix A

```
********** MODEL NAME
Stringent response in recombinant E. coli systems
********** MODEL NOTES

********** MODEL STATES
d/dt(Sg) = -miu*X/Yxs
d/dt(X) = miu*X
d/dt(PR) = rPR-rPRD
d/dt(G4P) = rG4P-rG4PD
d/dt(A) = rA-(miu*gamaA)-fiA*(rPR-rPRD)
d/dt(R) = rR-(miu*gamaR)-fiR*(rPR-rPRD)
d/dt(N) = rN-(miu*gamaN)-fiN*(rPR-rPRD)
d/dt(D) = rD-(miu*gamaD)-fiD*(rPR-rPRD)
d/dt(C) = rC-(miu*gamaC)-fiC*(rPR-rPRD)
d/dt(Q) = rQ-(miu*gamaQ)-fiQ*(rPR-rPRD)
d/dt(E) = rE-(miu*gamaE)-fiE*(rPR-rPRD)
d/dt(G) = rG-(miu*gamaG)-fiG*(rPR-rPRD)
d/dt(H) = rH-(miu*gamaH)-fiH*(rPR-rPRD)
d/dt(I) = rI-(miu*gamaI)-fiI*(rPR-rPRD)
d/dt(L) = rL-(miu*gamaL)-fiL*(rPR-rPRD)
d/dt(K) = rK-(miu*gamaK)-fiK*(rPR-rPRD)
d/dt(M) = rM-(miu*gamaM)-fiM*(rPR-rPRD)
d/dt(F) = rF-(miu*gamaF)-fiF*(rPR-rPRD)
d/dt(P) = rP-(miu*gamaP)-fiP*(rPR-rPRD)
d/dt(S) = rS-(miu*gamaS)-fiS*(rPR-rPRD)
d/dt(T) = rT-(miu*gamaT)-fiT*(rPR-rPRD)
d/dt(W) = rW-(miu*gamaW)-fiW*(rPR-rPRD)
d/dt(Y) = rY-(miu*gamaY)-fiY*(rPR-rPRD)
d/dt(V) = rV-(miu*gamaV)-fiV*(rPR-rPRD)

Sg(0) = 10
X(0) = 0.20000000000000001
PR(0) = 0.002
G4P(0) = 1.0000000000000001e-005
A(0) = 0.00023729999999999999
R(0) = 0.000232
N(0) = 0.00022000000000000001
D(0) = 0.00059409999999999997
C(0) = 8.0699999999999996e-005
Q(0) = 0.0002433
E(0) = 0.00024499999999999999
G(0) = 0.00020000000000000001
H(0) = 0.0001033
I(0) = 0.00021829999999999999
L(0) = 0.0003057
K(0) = 0.00022389999999999999
M(0) = 4.9700000000000002e-005
F(0) = 0.00022000000000000001
P(0) = 0.0001917
S(0) = 0.00024499999999999999
T(0) = 0.0013844
W(0) = 0.000136
Y(0) = 0.00042230000000000002
V(0) = 0.000234

********** MODEL PARAMETERS
miuwt = 0.36000000000000004
ks = 0.05
Yxs = 0.4
ke = 4.0999999999999988
kt = 5.3999999999999995
kPRD = 0.050000000000000002
kg4p = 0.20000000000000001
kg4pd = 0.002

********** MODEL VARIABLES
rA = 0.04004199
rR = 0.04507818
rN = 0.02786916
```

```
rD = 0.02807896
rC = 0.00970541
rQ = 0.03365008
rE = 0.03388203
rG = 0.0759915
rH = 0.0128619
rI = 0.03333426
rL = 0.05169129
rK = 0.0438803
rM = 0.0200554
rF = 0.0267729
rP = 0.02226515
rS = 0.019845
rT = 0.02644061
rW = 0.01015716
rY = 0.02185937
rV = 0.04336254
fiA = 0.036580604000000003
fiR = 0.041718328999999998
fiN = 0.067817974000000003
fiD = 0.077442646000000004
fiC = 0.0082888630000000005
fiQ = 0.035005034999999997
fiE = 0.075524562000000003
fiG = 0.064221565999999994
fiH = 0.084943724999999998
fiI = 0.067304200999999994
fiL = 0.080765041999999995
fiK = 0.090012946999999996
fiM = 0.040827790000000003
fiF = 0.073469471999999994
fiP = 0.043328150000000003
fiS = 0.061138931000000001
fiT = 0.069290788000000006
fiW = 0.0069873060000000004
fiY = 0.074394262000000003
fiV = 0.060111386000000003
gamaA = 0.0434
gamaR = 0.0489
gamaN = 0.0302
gamaD = 0.0305
gamaC = 0.0105
gamaQ = 0.0365
gamaE = 0.0368
gamaG = 0.0437
gamaH = 0.0140
gamaI = 0.0362
gamaL = 0.0561
gamaK = 0.0476
gamaM = 0.0218
gamaF = 0.0290
gamaP = 0.0242
gamaS = 0.0215
gamaT = 0.0287
gamaW = 0.0110
gamaY = 0.0237
gamaV = 0.0470

********** MODEL REACTIONS
miu = miuwt*Sg/(ks+Sg)
rPR = ke*PR/(kt+PR)
rPRD = kPRD*PR
rG4P = kg4p*f(H)
rG4PD = kg4pd*G4P

********** MODEL FUNCTIONS
f(x) = 0.2*exp(-10000*x)

********** MODEL EVENTS
event = lt(H,0.0),kg4pd,0.02,miuWT,0.2,kt,100
```

CHAPTER 4

Metabolic profiling of the *E. coli rel*A mutant under nutrient-limited growth conditions

> *"Being exposed to stress has become an everyday experience..."*
>
> By G. Storz and R. Hengge-Aronis in Bacterial stress responses,
> ASM Press, 2000

4.1 Abstract

The bacterial stringent response was found to be triggered by the ribosome-associated RelA enzyme in response to nutrient deprivation conditions. Here a metabolic profiling analysis of *E. coli* W3110 and the isogenic Δ*rel*A mutant cells is proposed to characterize the activity of this enzyme under different growth conditions. Metabolic profiles evaluated by gas chromatography–mass spectrometry (GC-MS) revealed that there were significant differences in metabolic activities between *E. coli* strains and predominantly among different growth conditions. Major differences were detected in the relative concentration levels of most metabolites when cells were grown at a dilution rate of 0.1 h^{-1}, for which higher levels were obtained, though the accumulation of fatty acids was more significant at a dilution rate of 0.05 h^{-1}. These metabolic differences were less pronounced in the Δ*rel*A mutant cells, denoting that RelA must be involved in such metabolic activities. This suggests that under nutrient limiting conditions the RelA-dependent stringent response promotes key changes in the *E. coli* metabolism.

4.2 Introduction

Under low nutrient conditions cells engage a multitude of cellular processes allowing for survival until growth can restart. Typically, the coordination of these cellular responses involves the global regulator guanosine-3',5'-bis-pyrophosphate (ppGpp), a core molecule that triggers the stringent response (Chatterji D and Ojha AK, 2001; Haseltin WA and Block R, 1973; Jain V *et al.*, 2006; Magnusson LU *et al.*, 2005). In *E. coli*, the stringent control mediated by ppGpp is a key regulatory process governing gene transcription, but also protein translation, enzyme activation and growth arrest (Artsimovitch I *et al.*, 2004; Dennis PP *et al.*, 2004; Paul BJ *et al.*, 2004). This phenomenon is mainly triggered by the activation of the ribosome-associated enzyme encoded by the *rel*A gene, which catalyzes the conversion of cellular GTP to ppGpp (Torok I and Kari C, 1980). Although the synthesis of ppGpp has been mainly associated with the activity of the RelA enzyme in response to amino acid starvation, studies indicate that ppGpp also accumulates during carbon starvation (Traxler MF *et al.*, 2006). A second ppGpp synthetase, SpoT, has been described to be involved in the ppGpp accumulation during carbon starvation, but its activity was shown to be much weaker than the RelA enzyme (Xiao H *et al.*, 1991). Hypotheses indicate that these two phenomena might be strictly related and the exhaustion of carbon often results in the rapid limitation of amino acids availability (Traxler MF *et al.*, 2006). Thus, it is expected that RelA, directly or indirectly, interferes in the definition of cellular responses to carbon starvation conditions. Otherwise, in the absence of the *rel*A gene, bacterial cells would respond just as the wild-type strain to such environmental conditions. To dissect the role of the RelA enzyme in the *E. coli* responses to nutrient-limited growth conditions, a metabolomics approach was applied in this study. The intracellular metabolite profiles measured by gas chromatography–mass spectrometry (GC-MS) were exploited to characterize the main metabolic differences when cells are grown at different dilution rates.

So far, studies have been focused on the gene expression profiling to investigate the RelA activity in response to environmental growth perturbations. For instance, changes in the transcriptome during glucose-lactose diauxic conditions were investigated in the wild-type *E. coli* and Δ*rel*A mutant cells (Traxler MF *et al.*, 2006), revealing that cells lacking the *rel*A gene show a delayed diauxie and a deficient induction of RpoS and Crp regulons. Other studies (Durfee T *et al.*, 2008; Haddadin FT *et al.*, 2009) evaluated gene expression profiles of *E. coli* cells after treatment with "artificial" inducers that mimic amino starvation (e.g. serine hydroxymate) leading to the induction of the stringent response. Results of this study showed that Δ*rel*A mutant transcriptional responses are consistent with the "relaxed" phenotype of double null mutant (Δ*rel*A/Δ*spo*T) and fail to adapt their physiology to the new conditions.

In general, gene expression patterns revealed that the accumulation of ppGpp can occur either during carbon or amino acid starvation conditions in wild-type strains and that the RelA is involved in cellular adaptation responses to both conditions. From transcriptome data, some common features were observed, like the downregulation of the translational apparatus, including rRNAs, tRNAs and ribosomal genes and induction of the RpoS-dependent general stress response. As a result, stringent response promotes the rapid decay of cellular growth, since the translational apparatus is significantly reduced, and invokes a manifold of cellular survival mechanisms that limits most processes for cell division and reproduction.

It has been proposed (Dedhia N *et al.*, 1997) that *E. coli* strains with relaxed phenotypes (e.g., Δ*rel*A mutants) can be quite useful as host strains in the synthesis of recombinant proteins. In fact, some successful applications of ppGpp-defective strains in recombinant bioprocesses have been reported (Dedhia N *et al.*, 1997; Sanden AM *et al.*, 2003), showing that these recombinant systems can limit ppGpp accumulation, thus reducing growth arrest and productivity losses. However, our understanding on the ppGpp effects in the organization of the metabolic network under nutrient-limiting conditions is still incomplete. Some studies have used microarrays to profile the ppGpp effects on gene expression and then infer metabolic alterations (Chang DE *et al.*, 2002; Durfee T *et al.*, 2008), but the unbiased determination of metabolite levels can bring better insights to unveil the mechanism of stringent response in the metabolism. For that reason our goals in this study are twofold: to investigate the metabolic behaviour of *E. coli* W3110 strains at different growth conditions; and to elucidate how the *rel*A gene mutation affects these metabolic behaviours.

4.3 Material and Methods

4.3.1 Bacterial strains and growth conditions

E. coli K12 W3110 (F-, *LAM-*, *IN*[*rrnD-rrnE*]*1*, *rph-1*) and the isogenic mutant Δ*rel*A (obtained from M. Cashel (Xiao H *et al.*, 1991)) were grown under controlled conditions in a chemostat culture at 37°C, pH 7 and dissolved oxygen above 30%. The minimal medium consisted of 5 g.kg^{-1} of glucose, 6 g.kg^{-1} of Na_2HPO_4, 3 g.kg^{-1} of KH_2PO_4, 0.5 g.kg^{-1} of NaCl, 1 g.kg^{-1} of NH_4Cl, 0.015 g.kg^{-1} of $CaCl_2$, 0.12 g.kg^{-1} of $MgSO_4.7H_2O$, 0.34 g.kg^{-1} of thiamine, 2 mL.kg^{-1} of trace-element solution (described elsewhere (Rocha I and Ferreira EC, 2002)) and 2 mL.kg^{-1} of vitamins solution (described elsewhere (Rocha I and Ferreira EC, 2002)). The minimal medium supplemented with 20 mg.kg^{-1} of L-isoleucine was used to grow the W3110 strain, while the same medium with further addition of 20 mg L^{-1} L-valine and 25 mg L^{-1} kanamycin was used to grow the Δ*rel*A mutant strain.

Chemostat cultivations were operated in a 3 L fermentor (BioFlo 3000, New Brunswick Scientific, USA) with a working volume of 1.5 L. The described minimal medium was continuously fed, at least for five residence times, at a given dilution rate (0.05, 0.1 and 0.2 h^{-1}), and the working volume was kept constant by withdrawing the culture broth through level control. Steady-state conditions were verified by constant optical density and glucose measurements. The pH of the culture was maintained at 7.0 by adding 2.0 M NaOH and 2.0 M HCl. Dissolved oxygen was maintained above 30% saturation through a cascade mode controlling the agitation speed and airflow.

4.3.2 Analytical techniques

Biomass was determined by measuring culture turbidity (OD_{600nm}) and cell dry weight (CDW). In order to determine CDW, 10 mL of broth were filtered with 0.2 μm filters and the filtrate was dried in the microwave to a constant weight. For glucose and acetate analysis, culture broth was centrifuged at 8000 rpm for 15 min to remove the cell debris and the supernatant was collected. The glucose concentration in the culture broth was determined by the dinitrosalicylic acid (DNS) colorimetric method (Miller GL, 1959). The concentrations of acetic acid in the culture broth were determined with an enzymatic test kit (R-Biopharm AG, Germany).

4.3.2.1 Quenching and metabolite extraction

For metabolomics analysis, samples (50 mL) were taken quickly from the fermentor and immediately quenched, to halt cellular metabolism, in 200 mL of glycerol/saline solution (60%, v/v) at -23 °C, followed by quick homogenization. Samples were centrifuged at 10,000 rpm for 20 min at -20 °C using a refrigerated centrifuge. The supernatants were discarded and the cell pellets were resuspended in 2 mL of cold glycerol/saline solution (50%, v/v) at -23 °C, followed by a second centrifugation at 8,000 rpm for 30 min at -20 °C. The supernatants were again discarded and the cell pellets were dissolved in 2.5 mL of cold methanol/water solution (50%, v/v) at -30 °C and stored at -80°C for subsequent intracellular metabolite extraction. For that, samples were subjected to three freeze–thaw cycles with 1 min of vigorous mixing using a vortex between each cycle. After the third cycle, samples were centrifuged at 8,000 rpm for 30 min at -20 °C and the supernatants were collected and stored at -80 °C. The extracted pellet was then resuspended in another 2.5 ml of cold methanol/water solution (50%, v/v) and centrifuged at 8,000 rpm for 30 min at -20 °C. The supernatant was collected and pooled with the first one and kept at -80 °C and, afterwards lyophilized.

4.3.2.2 Derivatization and GC-MS analysis

For GC-MS analysis samples were further treated as follows. The dried intracellular metabolite extracts were resuspended in 200 µL of sodium hydroxide (1 M) and derivatized using the methyl chloroformate (MCF) method (Smart KF *et al.*, 2010). The derivatized samples were then analyzed with a GC-MS systems - GC7890 coupled to a MSD5975 (Agilent Technologies, Inc., Santa Clara, CA, USA) equipped with a ZB-1701 GC capillary column, 30m x 250mm id x 0.15 mm (film thickness) with a 5 m guard column (Phenomenex, Inc., Torrance, CA, USA) at a constant flow rate of 1.0 mL/min of helium. The oven temperature was initially held at 45°C for 2 min. Thereafter the temperature is raised with a gradient of 9°C/min until 180°C. This temperature (180°C) is held for 5 min. Then the temperature is raised with a gradient of 40°C/min until 220°C. The temperature is again held for 5 min. Then the temperature is raised with a gradient of 40°C/min until 240°C and this temperature is held for 11.5 min. Finally the temperature is raised with a gradient of 40°C/min until 280°C, which is held for 2 min. The temperature of the inlet is 290°C, the interface temperature 250°C, and the quadrupole temperature 200°C. Sample (1 µL) was injected onto the column under pulsed splitless mode (1.8 bars until 1 min, 20 mL/min split flow after 1.01 min) and the detector was set with a scan interval of 1.47 seconds and m/z range of 38-650.

4.3.3 Statistical analysis

Data from GC-MS analyses were deconvoluted using the AMDIS spectral deconvolution software (Stein SE, 1999) to identify the compounds through matching with a library constructed by using analytical chemical standards. Peak intensity values corresponding to each identified compound were corrected for the recovery of the internal standard (D-4-alanine) and normalized with respect to biomass concentration. The corrected and normalized peak intensity values were thereafter transformed into Z-scores, by subtracting the average peak intensity corresponding to a metabolite *k* among all the *n* samples (including replicates) in the set of experiments, from the peak intensity value (I_{MK}) of that metabolite in one sample, and dividing that result by the standard deviation ($SD_{M1..Mn}$) of all measured *n* peak intensities, according to:

$$Z - score, k = \frac{\left(I_{Mk} - mean\ I_{Mk_1 \ldots Mk_n}\right)}{SD_{Mk_1 \ldots Mk_n}}$$

Further data processing and statistical analysis were performed with MATLAB (version 2009b, The Mathworks, Inc) and MultiExperiment Viewer (MeV) (Saeed AI *et al.*, 2003). The nonparametric two-way method, Mack-Skillings test, was used for testing the null hypothesis (H_o) of no differences among experiments and to look for the main differences between metabolic profiles that are related to either of the two factors: strain (Factor A) and dilution rate (Factor B). As exemplified by Table 4.1, this two factor design allows the combination of one or more observations for each factor with an uneven number of replicates per observation. *p*-values were calculated to determine if the null hypothesis, which establishes that metabolic profiles are not affected neither by factors A and/or B, is rejected. Thus, metabolites with *p*-value lower than 0.01 show statistically significant differences (with 99% of confidence) in their relative concentration between different experimental conditions when considering factors A or B.

Table 4.1. Mack-Skillings test for two factor design.

The design matrix captures the disproportionate number of replicates (either 3 or 4) in the group of experiments performed in this study.

		Factor B		
		0.05	0.1	0.2
Factor A	W3110	4	3	4
	Δ*rel*A	3	3	4

Hierarchical clustering (HCL) was used to cluster samples and metabolites that showed significant changes according to the Mack-Skillings test. The construction of hierarchical trees was based on the Pearson correlation metrics. Pearson's correlation coefficients (*r*) were also used to evaluate the degree of association between the metabolite profiles produced by the W3110 and Δ*rel*A *E. coli* cultures. *p*-values associated with each Pearson correlation coefficient was calculated using a Student's t distribution to test the null hypothesis (H_0) of no significant correlation between the metabolite profiles from the two cultures, against the alternative hypothesis (H_1) that establishes a significant correlation between the profiles.

4.4 Results

This study aimed at evaluating the impact of different dilution rates (0.05, 0.1 and 0.2 h^{-1}) in the metabolome in a wild-type and a relaxed strain of *E. coli*. Assuming that ppGpp accumulates under very slow growth conditions and the RelA enzyme is involved in the ppGpp-induced response, metabolic patterns of the *E. coli* W3110 and the isogenic Δ*rel*A mutant were analysed to explore the metabolic activities influenced by the ppGpp-induced response.

4.4.1 Growth parameters of *E. coli* W310 wild-type and Δ*rel*A cells under different dilution rates

Chemostat cultures of *E. coli* W3110 and the isogenic Δ*rel*A mutant were run at different dilution rates (0.05, 0.1, and 0.2 h^{-1}) and growth parameters were determined (Table 4.2). It was observed that only at higher dilution rates residual concentrations of glucose and acetate were detected in the extracellular medium. Biomass yields increased with the dilution rate and the mutant strain exhibited a biomass yield slightly higher than that of the W3110 strain in the same conditions.

Table 4.2. Growth parameters of the W3110 and Δ*rel*A mutant *E. coli* strains in aerobic glucose-limited continuous culture.

	W3110			**Δ*rel*A mutant**		
Dilution rate (h^{-1})	**0.05**	**0.10**	**0.20**	**0.05**	**0.10**	**0.20**
Biomass yield ($g\ g^{-1}$)	0.36±0.056	0.44±0.15	0.55±0.10	0.46±0.063	0.46±0.064	0.67±0.3
Biomass ($g\ L^{-1}$)	1.8±0.28	2.2±0.34	2.7±0.43	2.3±0.31	2.3±0.32	3.3±0.45
Glucose ($g\ L^{-1}$)	(1)	0.029±0.0086	0.040±0.0033	(1)	(1)	0.023±0.010
$q_{Glucose}$ ($g\ g^{-1}\ h^{-1}$)	0.14±0.021	0.23±0.076	0.36±0.063	0.11±0.015	0.22±0.030	0.30±0.13
Acetate ($g\ L^{-1}$)	(1)	(1)	0.34	(1)	(1)	0.02
$q_{Acetate}$ (10^{-3})($g\ g^{-1}\ h^{-1}$)	-	-	25±3.8	-	-	1.1±0.15

(1) Undeterminable traces.

4.4.2 Metabolic profiling of *E. coli*

In this work, the *E. coli* metabolic profiles were determined through the analysis of intracellular chemical molecules detected by GC-MS. The chemical derivatization procedure was chosen in order to quantify the main amino acids and their precursors generated in the central carbon metabolism and fatty acid biosynthesis. The overall list of detected metabolites is presented in Table 4.3.

Table 4.3. List of the intracellular metabolites detected by GC-MS.

The metabolites abbreviations are discriminated inside brackets.

TCA cycle	Fatty acids and derivatives biosynthesis	Amino acid biosynthesis	Others
2-ketoglutarate (akg) *cis*-Aconitate (acon-C) Citrate (cit) Fumarate (fum) Malate (mal) Succinate (succ)	Hexanoate (hxa) Octanoate (octa) Decanoate (dca) Tetradecanoate (ttdca) 10,13-Dimethyltetradecanoate (1013mlt) Pentadecanoate (pdca) 14-Methylpentadecanoate (14mpdca) Octadecanoate (ocdca) Octadecenoate (ocdcea) 9-*cis*,12-*cis*-Octadecadienoate (ocdcin)	Aspartate (asp) Isoleucine (ile) Lysine (lys) Threonine (thr) Alanine (ala) Leucine (leu) Valine (val) Glycine (gly) Serine (ser) Glutamate (glu) Proline (pro) Phenylalanine (phe) (*2S*)-2-isopropylmalate (3c3hmp) N-Acetyl-L-glutamate (acglu)	Benzoate$^{\Psi}$ (bnz) NADP(H) Nicotinate (nac) Phosphoenolpyruvate (pep) 5-oxo-D-proline$^{\Psi}$ (pyrglu) Malonate$^{\Psi}$ (mlt) Itaconate$^{\Psi}$ (itcon) Lactate (lac)

$^{\Psi}$ Metabolites unknown to be synthesized by *E. coli*

Metabolomic analysis showed that the metabolite profiles were largely influenced by the growth rate conditions and the genetic characteristics of the bacteria (i.e., *rel*A gene deletion). Nevertheless, Mack-Skillings test demonstrated that the amount of metabolites whose profiles were significantly changed with the dilution rate is greater than the ones that are different among the different strains (i.e. *E. coli* W3110 and Δ*rel*A mutant). As illustrated in Figure 4.1, only 15 metabolites (20% of the total metabolites detected by GC-MS) changed significantly with the alteration of the strain, while 20 metabolites were found to be considerably changing their profile with the alteration of the dilution rate.

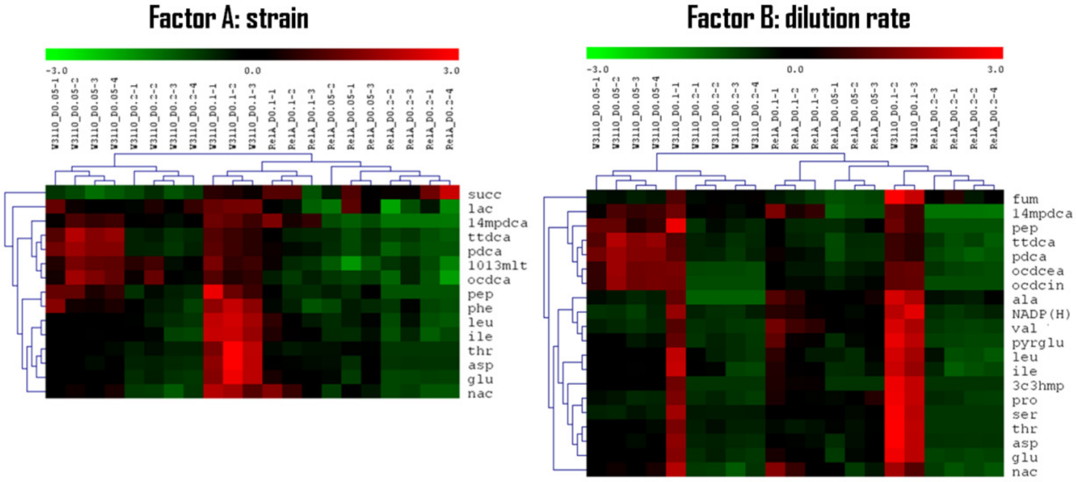


Figure 4.1. Hierarchical tree representing metabolites that were significantly changed (*p*-value <0.01) with both experimental factors: strain or dilution rate.

4.4.2.1 The impact of the dilution rate in the *E. coli* metabolome

Before analysing the metabolic differences between *E. coli* W3110 and Δ*rel*A mutant cultures, the metabolic states of the non-relaxed cells were examined in detail. This analysis was performed in order to verify if the RelA-modulated stringent response is triggered at any of the growth conditions evaluated and how this stress response affects *E. coli* metabolic activities. In this analysis, only those metabolites that presented *p*-values less than 0.01 in the Mack-Skillings test were considered.

Hierarchical clustering (HCL) analysis was used to group metabolites into different clusters that congregate those metabolites that behave in a similar manner under different growth conditions (Figure 4.2), and that are most likely engaged in similar metabolic processes. The overall trend of metabolite profiles was characterized by a maximum accumulation at the dilution rate of 0.1 h^{-1} and decreased levels below and above this dilution rate, except for fatty acids (clusters 1 and 2). Medium- and long-chain fatty acids (included in clusters 1 and 2, respectively) presented higher relative concentrations at lower dilution rates when compared with high dilution rates. This seems to indicate that those compounds are associated with key metabolic activities when cells are growing at nutrient-limiting conditions and are less required when increasing the availability of nutrients.

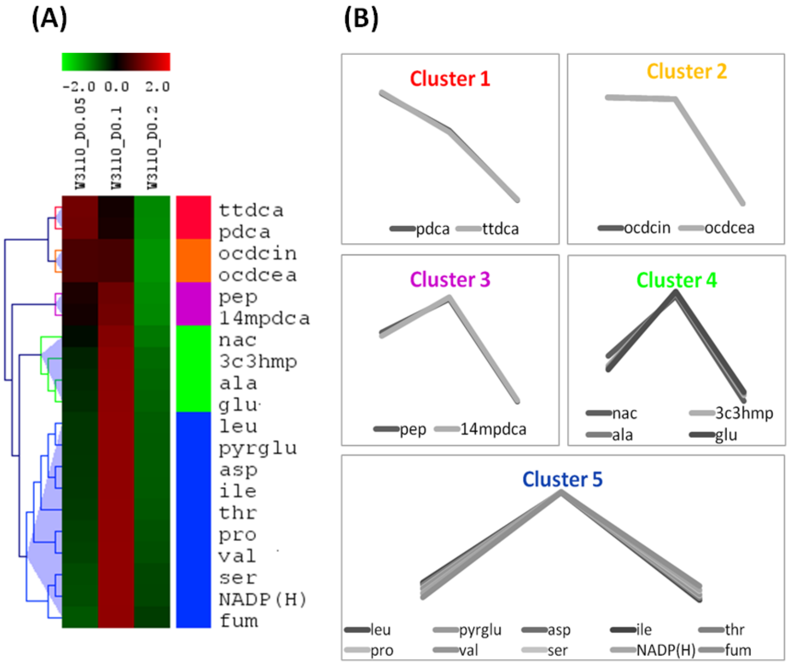


Figure 4.2. Metabolite profiles of the *E. coli* W3110 strain exhibited under three dilutions rates (0.05, 0.1 and 0.2 h^{-1}) were clustered by HCL analysis.

For abbreviations see Table 4.3.

On the other hand, clusters 4 and 5 grouped most of the TCA metabolites, amino acids, including amino acid derivatives like, 5-oxo-D-proline (*pyrglu*) and (*2S*)-2-isopropylmalate (*3c3hmp*) and the NAD salvage pathway intermediaries, nicotinate (*nac*) and NADP(H), presenting higher levels at the dilution rate of 0.1 h^{-1}. The third cluster presented slightly higher relative metabolite concentrations at the dilution rate 0.05 h^{-1}, compared to the fourth and fifth clusters, but the overall metabolic patterns were similar. Most likely, these clusters are at some extent linked and biochemical reactions involving these metabolites may be synchronized.

To understand how these alterations in the relative metabolite concentrations are related with changes in biochemical activities, metabolites detected in the *E. coli* metabolome were represented into a metabolic map (Figure 4.3).

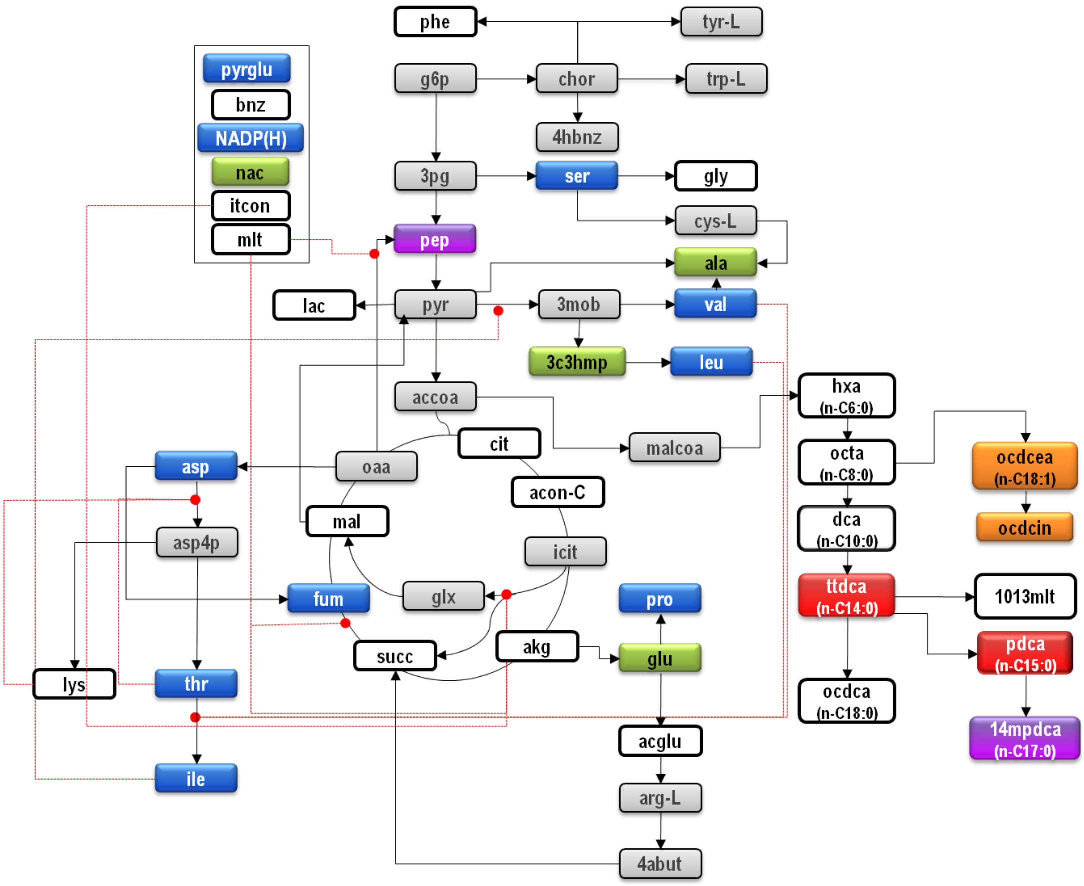


Figure 4.3. Metabolic map representing the metabolome of the *E. coli* W3110 according to the clusters obtained.

Grey boxes represent those metabolites that were not detected by the GC-MS analysis and the coloured boxes display the sets of metabolites that were clustered according to the HCL analysis (each colour corresponds to a cluster, according to Figure 4.2) and that were statistically significant for the dilution rate factor. Non-coloured boxes represent those metabolites that were detected by the GC-MS analysis, but were not significantly changed with the dilution rate. Red dashed lines indicate enzymatic inhibition activities performed by various metabolites.

As illustrated in Figure 4.3, the strong correlation between medium-chain fatty acids, pentadecanoate (p*dca*) and tetradecanoate (*ttdca*), seems derived from their proximity in the metabolic network (i.e. pentadecanoate is produced via saturated fatty acid elongation cycle that uses tetradecanoate as an intermediary). Likewise, amino acids, leucine (*leu*), isoleucine (*ile*) and valine (*val*) were also clustered together and are participants in the two pathways that are highly correlated. As depicted, the pathway of isoleucine biosynthesis is subject to regulation by valine and leucine, whereas the first step in the pathway of valine biosynthesis is also regulated by isoleucine. Although it is not illustrated, several enzymes are simultaneously involved in the synthesis of the three amino acids. In addition, amino acids, fumarate (*fum*) and the NADP(H) molecules were also clustered with these metabolites, but not all presented clear metabolic associations. For instance, serine (*ser*) and proline (*pro*) are not neighbours in the metabolic network, but glutamate is a common participant in both biosynthetic pathways (i.e. glutamate is the amino donor in the biosynthesis of serine and the biosynthesis of proline involves the reduction of L-glutamate).

4.4.2.2 The impact of the Δ*rel*A mutation in the *E. coli* metabolome

Having evaluated the metabolic states exhibited by *E. coli* W3110 cells, the metabolic profiles exhibited by the two strains at different dilution rates were compared. Here the aim was to determine the key metabolic activities that were influenced by the *rel*A gene mutation under the tested growth conditions. Although the relative concentrations of most metabolites were significantly changed (see Appendix A), when comparing the metabolite profiles between the two *E. coli* strains, it was found that their trend were similar in some cases. Therefore, pairwise correlation coefficients were calculated for each metabolite that was found to be significant by the Mack-Skillings test (Figure 4.4). As illustrated, almost half of the metabolites showed poorly correlated patterns (i.e. *r* below 0.8). Moreover, some metabolites revealed to have negatively correlated profiles, which means that the intracellular accumulation of these metabolites follows an opposite pattern in one of the cultures, e.g. succinate (*succ*) with *r* equal to -0.3.

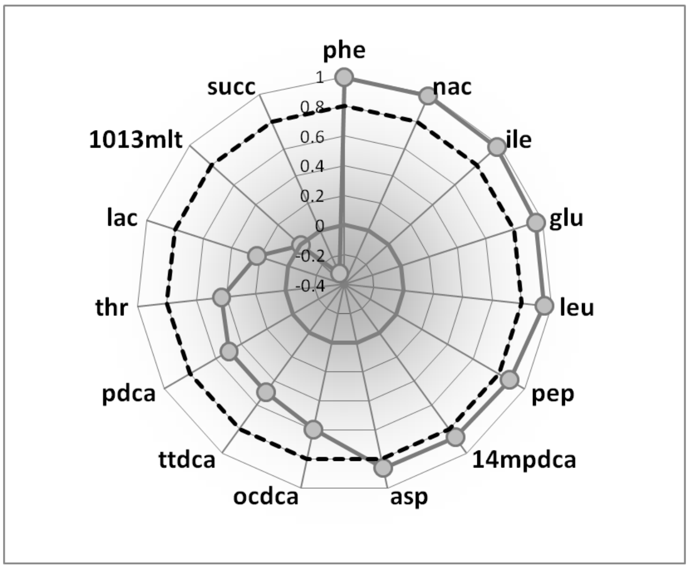


Figure 4.4. Representation of the pairwise correlation coefficients (*r*) determined between metabolite profiles generated in the W3110 and Δ*rel*A *E. coli* cultures.

Only metabolites that presented significant changes according to the Mack-Skillings test for the strain factor were considered. Dashed line delimits the correlation coefficient threshold below which metabolite profiles produced by the two strains were considered uncorrelated.

Profiles of those uncorrelated metabolites are then shown in Figure 4.5. In general, relative concentrations were significantly different at lower dilutions rates and have a tendency to become similar as the dilution rate increases. For example, octadecanoate (*ocdca*), tetradecanoate (*ttdca*), pentadecanoate (*pdca*) and 10,13-dimethyltetradecanoate (*1013mlt*) presented quite divergent trends when the dilution rate was lower. In contrast, metabolites like succinate (*succ*), threonine (*thr*) and lactate (*lac*) showed particularly asymmetrical metabolic patterns, which indicate that metabolite accumulation in *E. coli* mutant cells was quite different compared to the W3110 strain.

Other interesting differences between strains are represented by five metabolites that, although using the Mack-Skillings test were not considered significant, were uniquely detected in the W3110 culture at a dilution rate of 0.1 h^{-1} (see Appendix B): N-acetyl-L-glutamate (*acglu*), lysine (*lys*), malate (*mal*), 2-ketoglutarate (*akg*) and the two inhibitors of the isocitrate lyase, itaconate (*itcon*) and malonate (*mlt*). It is worth to notice that each one of these metabolites is somehow related to the TCA cycle, in particular the 2-ketoglutarate (*akg*), itaconate (*itcon*) and malonate (*mlt*), which are directly linked to the isocitrate node, which is a key metabolic valve that controls the carbon flux through the TCA cycle or the glyoxylate shunt depending on the metabolic requirements.

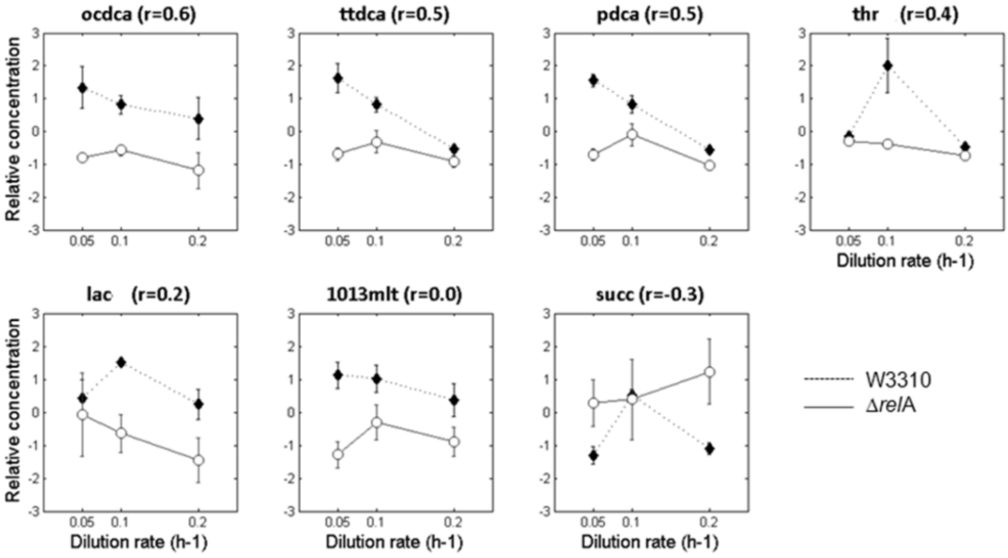


Figure 4.5. Metabolic profiles of the metabolites that were uncorrelated between *E. coli* strains.

4.5 Discussion

The growth rate-dependent regulation of the metabolism is fundamental to fine-tune the fuelling and biosynthetic reactions, in such a way that cells can rapidly adapt to the existing environmental conditions. The effect of nutrient-limiting conditions on the balance of the catabolic and anabolic fluxes and energetic efficiency of the *E. coli* cells has been the focus of previous investigations (Farmer IS and Jones CW, 1976; Ko Y-F *et al.*, 1994; Neijssel OM *et al.*, 1990; Russell JB and Cook GM, 1995) and it was confirmed that several key metabolic activities are systematically activated under those conditions. In this study, it was aimed to analyse this growth rate-dependent behaviour and observe how it is affected by the deletion of the *rel*A gene, which encodes a ppGpp synthetase described as a key element for the cellular responses to nutrient limiting conditions.

First, the physiological parameters of *E. coli* W3110 and the isogenic Δ*rel*A mutant were assessed. Apparently, biomass yields are lower as the dilution rate decreases and the biomass yields of the mutant strain were consistently higher compared to the *E. coli* strain W3110, in particular at lower dilutions rate (D=0.05 h^{-1}). Indeed, it was expected that at lower dilution rates the metabolic and physiological behaviour of these two strains would be different, since it is assumed that part of the maintenance energy is spent in cellular processes to counteract nutritional limitations, typically through RelA-dependent processes. To understand how these cellular processes impact the *E. coli* metabolism, a metabolic profiling approach was applied to study the intracellular metabolome of *E. coli* W3110 and Δ*rel*A mutant cells growing at three different dilution rates.

It was verified that metabolite pools were strongly affected by the dilution rate and the accumulation of most metabolites was higher at a dilution rate of 0.1 h^{-1}, which is in good agreement with other studies performed in glucose-limited chemostat experiments (Nanchen A *et al.*, 2006; Prasad MR *et al.*, 2005; Sauer U *et al.*, 1999). According to Nanchen *et al* (Nanchen A *et al.*, 2006), at a dilution rate of 0.1 h^{-1}, large flux variations are verified in the metabolic network, in particular at the oxaloacetate node where two anaplerotic reactions converge. The carbon flux through the glyoxylate cycle (i.e. an anaplerotic pathway that converts isocitrate to succinate or to malate via glyoxylate) is maximum at this dilution rate and decreases at higher dilution rates (Fischer E and Sauer U, 2003; Nanchen A *et al.*, 2006; Prasad MR *et al.*, 2005). It was proposed (Fischer E and Sauer U, 2003; Nanchen A *et al.*, 2008; Sauer U and Eikmanns BJ, 2005) that at nutrient starvation conditions the cAMP-mediated catabolite repression of enzymes in the glyoxylate cycle is limited and the activity of the competing enzyme, i.e. the isocitrate dehydrogenase, is decreased. As such, it is believed that anaplerotic reactions are stimulated in hungry *E. coli* cells and, at higher dilution rates, are restrained as a consequence of the increasing glucose concentrations and the catabolite repression.

This phenomenon has been associated with the stringent response induced by the ppGpp accumulation that should require the activity of the RelA enzyme (Ozkan P *et al.*, 2005; Pao CC and Dyess BT, 1981; Traxler MF *et al.*, 2006). The levels of this global regulator, ppGpp, increase rapidly when cells are grown in nutrient-deprived conditions, which potentiates the expression of several stress response genes, namely the transcriptional regulator cAMP receptor protein (Crp) that governs the catabolite repression (Traxler MF *et al.*, 2006). Thus, it was expected that Δ*rel*A mutants would be less effective to increase metabolic activities at a dilution rate of 0.1h^{-1}. Although it seemed that the catabolite derepression was invoked in both *E. coli* strains, it was clear that it was induced to a much lesser extent in Δ*rel*A mutant cells, as the relative concentrations of most metabolites were much lower (see Appendix A). Besides the overall relative metabolite concentrations at these conditions, the absence of the isocitrate lyase inhibitors is indicative of a less effective ppGpp-induced catabolite derepression of anaplerotic activities. It is believed that anaplerotic functions in "less stringent" phenotypes are limited, and for that reason enzymatic inhibitors, such as isocitrate lyase inhibitors, are not needed to control such metabolic activities. In contrast, these inhibitors were found to be overproduced in the W3110 *E. coli* strain under this particular dilution rate (0.1 h^{-1}), which denotes that they can be physiologically determinant for the regulation of enzymes at the metabolic branch-point at isocitrate under nutrient-limiting conditions. Although it is only an hypothesis, the failure to induce the catabolite derepression of isocitrate lyase (i.e. anaplerotic enzyme) and the concomitant absence of isocitrate lyase inhibitors, suggest that "relaxed" phenotypes are less effective to stimulate anaplerotic functions under nutrient limiting conditions.

At dilutions rates lower than 0.1 h^{-1}, the strong dependence of metabolic activities on carbon source availability was evidenced by generally lower relative concentrations of most metabolites, even when anaplerotic reactions were derepressed. However, a group of metabolites were highly accumulated at these conditions (D=0.05 h^{-1}). Fatty acids, like the tetradecanoate (*ttdca*), pentadecanoate (*pdca*), octadecenoate (*ocdcea*) and 9-cis,12-cis-octadecadienoate (*ocdcin*), were highly accumulated in the intracellular milieu of the *E. coli* W3110 cells suggesting the involvement of key metabolic changes in the structure of cell membranes. Fatty acids biosynthesis was definitely the most significant change occurred at these growth conditions (0.05 h^{-1}) and represented the utmost opposite metabolic behaviour between the *E. coli* W3110 and Δ*rel*A cells. While fatty acids were highly accumulated in starved *E. coli* W3110 cells, the opposite was observed in the Δ*rel*A mutant cells. As such, it is clear that the RelA activity is fundamental to define the cellular response during extreme starvation conditions. It is known that fatty acid biosynthetic genes (e.g. *fab*HDG and *fab*A) are stringently controlled by the ppGpp and the alternative sigma factor, RpoS (Dong T and Schellhorn HE, 2009). Since the activity of RpoS is augmented by high levels of ppGpp, it is expected that in the absence of the RelA, the RpoS-transcriptional control of these genes is reduced. Thus, the accumulation of fatty acids at lower dilution rates might be associated with the nutrient starvation responses and apparently is lessen in Δ*rel*A mutant cells.

One of the most reported effects provoked by the stringent response (Chatterji D and Ojha AK, 2001; Durfee T *et al.*, 2008; Jain V *et al.*, 2006; Jishage M *et al.*, 2002; Magnusson LU *et al.*, 2005; Mukherjee TK *et al.*, 1998; Traxler MF *et al.*, 2006) is the overexpression of genes coding for enzymes that participate in amino acid biosynthetic pathways. Although the curtailment of most amino acids was not observed in the intracellular milieu, Δ*rel*A mutant cells showed some changes in metabolic profiles related with the biosynthesis of amino acids. For example, it was identified that lysine (*lys*) and *N*-acetyl-L-glutamate (*acglu*) were not detected in *E. coli* Δ*rel*A mutant at a dilution rate of 0.1 h^{-1}, but relative concentration levels of these amino acids were determined in the *E. coli* W3110 strain. The only common feature between these metabolites is the connectivity to glutamate, one of the most required amino donors in the metabolism, whose relative concentration levels were found to be much lower in the *E. coli* Δ*rel*A mutant strain. Furthermore, 2-ketoglutarate (*akg*), the metabolic precursor of glutamate, was also undetected in this mutant strain at the same conditions, which indicates the relative concentrations of these metabolites were severely reduced in mutant cells. This may suggest that *E. coli* Δ*rel*A mutant cells present a “relaxed” phenotype that may have led to important shortages in certain metabolites.

4.6 Conclusions

Even with detailed knowledge about the overall metabolic reactions and their regulation, the interpretation of metabolic patterns is still not a trivial task. Analytical limitations in the detection of the whole set of metabolites within the cellular milieu, are still a problem to fully characterize the metabolic system. For example, key metabolic nodes like the isocitrate (*icit*), oxaloacetate (*oaa*) or glyoxylate (*glx*), would be helpful in the verification if anaplerotic functions were indeed activated under catabolite derepression. However, it was possible to unveil crucial metabolic alterations in response to different nutrient-limiting conditions, and more importantly to confirm that the RelA activity is fundamental in the coordination of several cellular responses, like anaplerosis and fatty acid biosynthesis. These two metabolic activities were associated with the most remarkable differences between the two *E. coli* strains and exposed some limitations that "less stringent" phenotypes might exhibit. Nevertheless, there are no evidences suggesting that the *rel*A mutation leads to impaired metabolic performances and are entirely devoid of survival mechanisms. In fact, it was observed that biomass yields were higher in Δ*rel*A mutant cells. It can be hypothesized that ppGpp synthesized by SpoT guarantees the most basic responses to maintain cell growth and survival and the activity of RelA would be involved in the reorganization of the global network that operates in response to several environmental perturbations entailing a multitude of cellular processes, namely cellular metabolism.

4.7 References

1. Artsimovitch I *et al* (2004) Structural basis for transcription regulation by alarmone ppGpp. *Cell* 117 (3):299-310.

2. Chang DE, Smalley DJ, and Conway T (2002) Gene expression profiling of *Escherichia coli* growth transitions: an expanded stringent response model. *Molecular Microbiology* 45 (2):289-306.

3. Chatterji D and Ojha AK (2001) Revisiting the stringent response, ppGpp and starvation signaling. *Current Opinion in Microbiology* 4 (2):160-165.

4. Dedhia N *et al* (1997) Improvement in recombinant protein production in ppGpp-deficient *Escherichia coli*. *Biotechnology and Bioengineering* 53 (4):380-386.

5. Dennis PP, Ehrenberg M, and Bremer H (2004) Control of rRNA synthesis in *Escherichia coli*: a systems biology approach. *Microbiology and Molecular Biology Reviews* 68 (4):639-+.

6. Dong T and Schellhorn HE (2009) Control of RpoS in global gene expression of *Escherichia coli* in minimal media. *Molecular Genetics and Genomics* 281 (1):19-33.

7. Durfee T *et al* (2008) Transcription profiling of the stringent response in *Escherichia coli*. *Journal of Bacteriology* 190 (3):1084-1096.

8. Farmer IS and Jones CW (1976) Energetics of *Escherichia coli* During Aerobic Growth in Continuous Culture. *European Journal of Biochemistry* 67 (1):115-122.

9. Fischer E and Sauer U (2003) A novel metabolic cycle catalyzes glucose oxidation and anaplerosis in hungry *Escherichia coli*. *Journal of Biological Chemistry* 278 (47):46446-46451.

10. Haddadin FT, Kurtz H, and Harcum SW (2009) Serine hydroxamate and the transcriptome of high cell density recombinant *Escherichia coli* MG1655. *Applied Biochemistry and Biotechnology* 157 (2):124-139.

11. Haseltin WA and Block R (1973) Synthesis of guanosine tetraphosphate and pentaphosphate requires presence of a codon-specific, uncharged transfer ribonucleic acid in acceptor site of ribosomes - (Stringent control ppGpp (Msi) and pppGpp (Msii) protein synthesis *Escherichia coli*). *Proceedings of the National Academy of Sciences of the United States of America* 70 (5):1564-1568.

12. Jain V, Kumar M, and Chatterji D (2006) ppGpp: Stringent response and survival. *Journal of Microbiology* 44 (1):1-10.

13. Jishage M *et al* (2002) Regulation of σ factor competition by the alarmone ppGpp. *Genes & Development* 16 (10):1260-1270.

14. Ko Y-F, Bentley WE, and Weigand WA (1994) A metabolic model of cellular energetics and carbon flux during aerobic *Escherichia coli* fermentation. *Biotechnology and Bioengineering* 43:847-855.

15. Magnusson LU, Farewell A, and Nystrom T (2005) ppGpp: a global regulator in *Escherichia coli*. *Trends in Microbiology* 13 (5):236-242.

16. Miller GL (1959) Use of dinitrosalicylic acid reagent for the determination of reducing sugar. *Analytical Chemistry* 31:426-428.

17. Mukherjee TK, Raghavan A, and Chatterji D (1998) Shortage of nutrients in bacteria: The stringent response. *Current Science* 75 (7):684-689.

18. Nanchen A *et al* (2008) Cyclic AMP-dependent catabolite repression is the dominant control mechanism of metabolic fluxes under glucose limitation in *Escherichia coli*. *Journal of Bacteriology* 190 (7):2323-2330.

19. Nanchen A, Schicker A, and Sauer U (2006) Nonlinear dependency of intracellular fluxes on growth rate in miniaturized continuous cultures of *Escherichia coli*. *Applied and Environmental Microbiology* 72 (2):1164-1172.

20. Neijssel OM, Buurman ET, and Demattos MJT (1990) The role of futile cycles in the energetics of bacterial growth. *Biochimica et Biophysica Acta* 1018 (2-3):252-255.

21. Ozkan P *et al* (2005) Metabolic flux analysis of recombinant protein overproduction in *Escherichia coli*. *Biochemical Engineering Journal* 22 (2):167-195.

22. Pao CC and Dyess BT (1981) Effect of unusual guanosine nucleotides on the activities of some *Escherichia coli* cellular enzymes. *Biochim Biophys Acta* 677 (3-4):358-362.

23. Paul BJ *et al* (2004) rRNA transcription in *Escherichia coli*. *Annual Review of Genetics* 38:749-770.

24. Prasad MR *et al* (2005) The role of isocitrate lyase and the glyoxylate cycle in *Escherichia coli* growing under glucose limitation. *Research in Microbiology* 156 (2):178-183.

25. Rocha I and Ferreira EC (2002) On-line simultaneous monitoring of glucose and acetate with FIA during high cell density fermentation of recombinant *E. coli*. *Analytica Chimica Acta* 462 (2):293-304.

26. Russell JB and Cook GM (1995) Energetics of bacterial growth: balance of anabolic and catabolic reactions. *Microbiological Reviews* 59 (1):48-62.

27. Saeed AI *et al* (2003) TM4: A free, open-source system for microarray data management and analysis. *Biotechniques* 34 (2):374-378.

28. Sanden AM *et al* (2003) Limiting factors in *Escherichia coli* fed-batch production of recombinant proteins. *Biotechnology and Bioengineering* 81 (2):158-166.

29. Sauer U and Eikmanns BJ (2005) The PEP-pyruvate-oxaloacetate node as the switch point for carbon flux distribution in bacteria. *Fems Microbiology Reviews* 29 (4):765-794.

30. Sauer U *et al* (1999) Metabolic flux ratio analysis of genetic and environmental modulations of *Escherichia coli* central carbon metabolism. *Journal of Bacteriology* 181 (21):6679-6688.

31. Smart KF et al.Analytical platform for metabolome analysis of microbial cells using gas chromatography-mass spectrometry (GC-MS). Nature Protocols (in press)

32. Stein SE (1999) An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *Journal of the American Society for Mass Spectrometry* 10 (8):770-781.

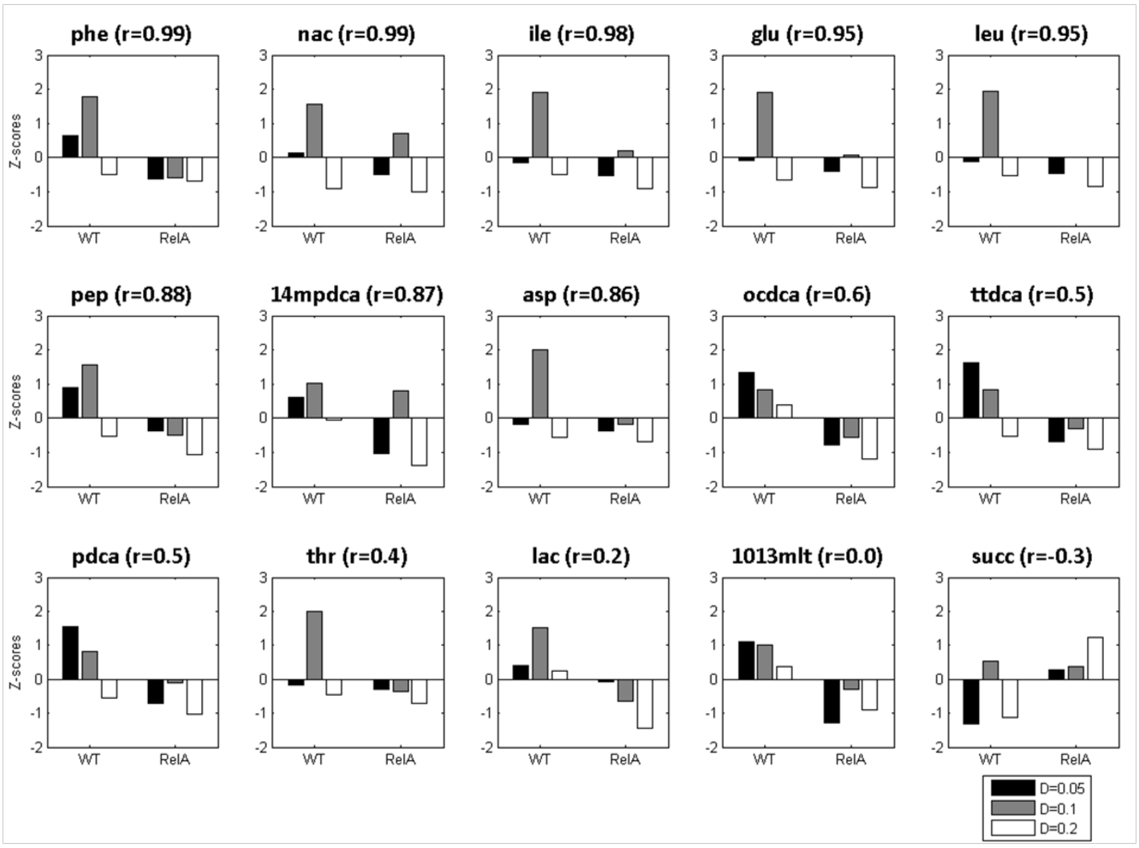
33. Torok I and Kari C (1980) Accumulation of ppGpp in a *rel*A mutant of *Escherichia coli* during amino acid starvation. *Journal of Biological Chemistry* 255 (9):3838-3840.

34. Traxler MF, Chang DE, and Conway T (2006) Guanosine 3 ',5 '-bispyrophosphate coordinates global gene expression during glucose-lactose diauxie in *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America* 103 (7):2374-2379.

35. Xiao H *et al* (1991) Residual guanosine 3',5'-bispyrophosphate synthetic activity of *rel*A null mutants can be eliminated by *spo*T null mutations. *Journal of Biological Chemistry* 266 (9):5980-5990.

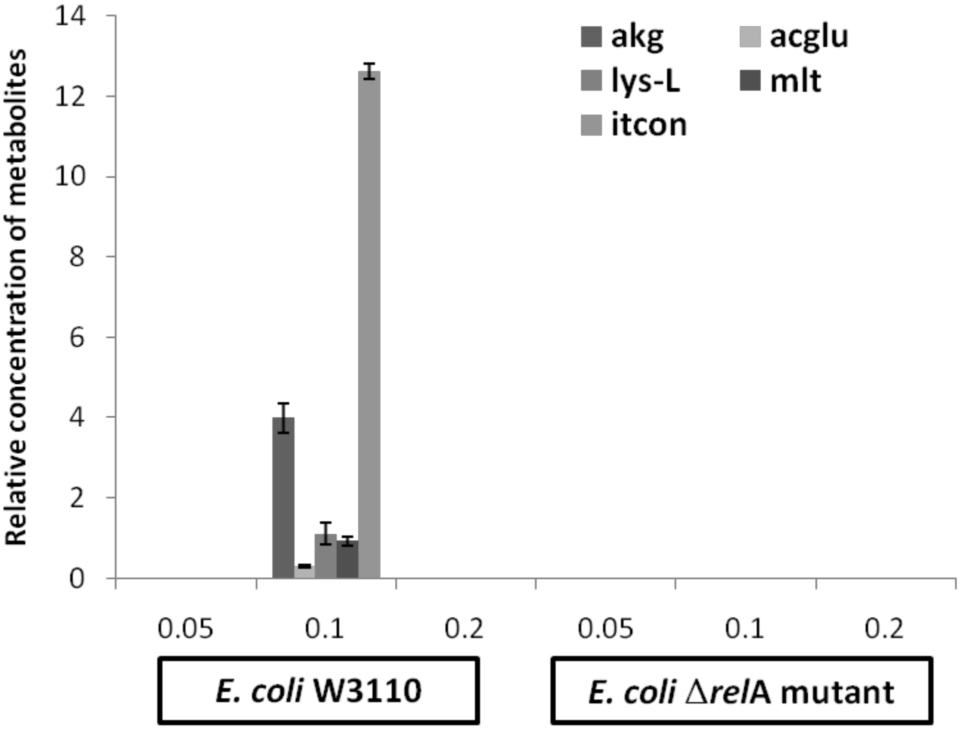
4.8 Appendix A

Z-scores of metabolites that presented significant changes according to the strain factor. R-values indicate the correlation between the relative concentrations of these metabolites in the two cultures: E. coli W310 (WT) and E. coli ΔrelA (RelA).



4.9 Appendix B

Relative concentrations of metabolites that were undetected in chemostat cultures, except in the *E. coli* W3110 grown at a dilution rate of 0.1 h-1.



4.10 Appendix C

Relative concentrations of metabolites detected by GC/MS analysis.

	W3110_D0.05-1	W3110_D0.05-2	W3110_D0.05-3	W3110_D0.05-4	W3110_D0.1-1	W3110_D0.1-2	W3110_D0.1-3	W3110_D0.2-1	W3110_D0.2-2	W3110_D0.2-3	W3110_D0.2-4	RelA_D0.05-1	RelA_D0.05-2	RelA_D0.05-3	RelA_D0.1-1	RelA_D0.1-2	RelA_D0.1-3	RelA_D0.2-1	RelA_D0.2-2	RelA_D0.2-3	RelA_D0.2-4
NADP(H)	0.082	0.085	0.064	0.099	0.363	0.406	0.458	0.1	0.084	0.081	0.106	0.147	0.13	0.152	0.229	0.215	0.184	0.072	0.087	0.091	0.101
bnz	0.449	0.352	0.271	0.371	0.339	0.284	0.406	0.42	0.62	0.411	0.481	0.906	0.377	0.897	0.405	0.331	0.357	0.586	0.335	0.603	0.22
ala	4.219	3.639	2.869	3.262	10.9	16.55	14.62	0	0	0	0	4.799	5.974	7.229	10.21	8.469	5.185	3.625	3.357	4.056	4.592
gly	1.159	1.423	1.254	1.336	2.283	2.511	2.092	1.801	1.826	1.522	1.37	0.691	1.08	1.098	1.499	1.084	1.143	1.838	0.002	1.91	1.49
dca	0.2	0.273	0.325	0.257	0.319	0.233	0.31	0.627	0.753	0.617	0.711	0.587	0.322	0.548	0.448	0.524	0.414	0.522	0.473	0.812	0.385
3c3hmp	0.178	0.173	0.181	0.2	0.462	0.768	0.667	0	0	0	0	0	0	0	0.268	0.219	0.216	0	0	0	0
val	3.77	3.125	2.406	3.17	12.34	20.22	17.56	2.573	2.775	2.105	1.931	6.038	5.351	6.184	14.21	11.1	10.01	1.745	1.563	2.287	1.9
akg	0	0	0	0	4.013	4.355	3.618	0	0	0	0	0	0	0	0	0	0	0	0	0	0
leu	2.546	2.286	2.336	2.409	4.953	5.167	4.615	2.264	1.93	1.771	1.658	1.971	1.776	2.243	3.215	2.552	1.976	1.35	1.412	2.093	1.225
ile	1.431	1.243	1.23	1.249	2.695	2.722	2.401	1.395	1.207	0.817	0.798	0.958	0.971	1.145	1.635	1.445	1.41	0.645	0.728	1.004	0.789
pep	0.647	0.642	0.427	0.505	1.066	0.664	0.517	0.104	0.222	0.16	0.126	0.113	0.164	0.318	0.247	0.129	0.117	0	0	0	0
pro	2.617	2.399	1.911	2.017	6.511	9.34	8.241	1.819	2.061	1.711	1.478	2.676	3.424	4.475	4.227	3.32	2.789	1.806	1.479	1.935	1.42
acon-C	0.282	0.47	0.304	0.357	2.081	2.861	2.051	0.139	0.315	0.193	0.124	0.09	0.108	0.144	0.15	0.121	0.1	0.196	0.143	0.18	0.197
thr	0.481	0.45	0.396	0.311	1.446	2.556	1.931	0.173	0.277	0.154	0.151	0.347	0.263	0.355	0.311	0.205	0.255	0	0	0	0
ser	0.054	0.046	0.086	0.061	0.627	0.852	0.689	0.077	0.066	0.067	0.075	0.137	0.117	0.14	0.21	0.172	0.17	0	0	0	0
pyrglu	5.492	5.684	3.737	4.476	13.55	22.39	19.35	1.929	3.061	2.475	1.647	2.332	4.713	6.82	13.45	7.732	7.546	1.077	0.744	1.203	0.793
cit	5.712	6.975	3.812	4.559	46.65	56.75	44.29	1.016	1.816	1.305	1.075	1.065	1.262	2.12	2.051	1.388	1.222	1.721	1.191	1.69	1.683
asp	24.25	24.17	15.79	19.49	76.57	115.8	92.02	8.035	9.259	6.551	5.691	5.018	15.24	22.08	27.17	18.25	16.03	2.887	2.635	3.623	2.951
ttdca	15.02	21	18.18	19.1	15.04	13.62	12.56	6.298	6.82	5.33	5.958	3.776	5.77	5.761	9.464	6.297	5.908	3.454	4.267	5.013	2.653
glu	22.45	22.21	17.01	18.69	51.86	74.65	63.19	7.581	10.91	6.954	5.493	4.813	15.59	18.95	28.64	19.15	22.49	3.76	2.314	2.917	3.115
pdca	8.86	10.42	9.546	9.542	8.074	7.048	6.271	2.657	2.592	1.915	2.618	1.32	2.297	2.225	5.309	3.494	3.186	0.93	0.401	1.409	0.824
acglu	0	0	0	0	0.307	0.306	0.252	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1013mlt	38.87	50.12	47.21	46.58	49.98	40.81	42.42	39.68	43.15	28.91	34.71	11.19	19.09	19.46	35.82	25.6	24.09	21.03	23.8	26.32	13.81
phe	1.023	0.684	0.67	0.714	1.052	1.201	1.033	0.542	0.481	0.447	0.359	0.466	0.347	0.431	0.524	0.453	0.3	0.35	0.386	0.534	0.325
14mpdca	13.4	18.95	17.03	17.8	21.72	20.55	17.81	10.93	13.33	9.216	12.17	2.626	4.151	4.301	23.05	14.82	17.03	0.951	0.878	0.965	0.686
ocdcea	36.1	58.15	48.74	49.75	50.68	45.52	46.23	5.992	5.801	5.467	4.333	10.29	14.13	15.75	24.28	15.27	15.16	6.703	5.557	9.307	4.329
ocdcin	14.25	23.69	19.9	20.61	20.31	18.63	18.84	0	0	0	0	4.086	5.74	6.54	9.686	6.084	6.085	0	0	0	0
ocdca	2.803	4.205	3.862	3.566	3.428	2.943	3.075	2.714	3.622	2.542	2.243	1.682	1.784	1.77	2.029	1.747	2.019	1.479	1.323	1.999	0.759
lys	0	0	0	0	0.846	1.396	1.078	0	0	0	0	0	0	0	0	0	0	0	0	0	0
mlt	0	0	0	0	0.795	1.01	0.979	0	0	0	0	0	0	0	0	0	0	0	0	0	0
hxa	2.311	2.184	2.149	2.61	2.9	1.518	2.298	2.697	3.869	2.907	3.027	9.296	5.364	8.235	3.34	2.926	2.574	3.375	2.702	5.059	2.386
fum	0.838	0.897	0.739	0.826	1.831	3.254	2.766	0.894	0.973	0.96	1.262	0.363	0.416	0.481	0.809	0.688	0.535	1.537	0.802	1.111	1.096
succ	1.971	1.693	1.228	1.486	3.619	3.46	3.196	1.685	2.05	1.56	1.872	4.095	2.379	3.066	4.162	4.126	1.563	4.48	3.161	3.158	5.691
lac	12.63	10.22	9.531	9.327	12.66	13.04	13.1	8.811	10.72	9.275	11.35	12.34	5.69	9.907	9.138	8.563	6.313	8.384	4.376	6.727	5.15
mal	0	0	0	0	1.244	2.258	1.758	0	0	0	0	0.349	0.344	0.442	0	0	0	0	0	0	0
octa	3.245	2.877	2.787	3.266	3.817	2.212	2.895	4.924	5.268	4.022	5.277	6.978	3.782	6.377	3.971	3.733	2.528	4.533	3.22	4.973	3.351
itcon	0	0	0	0	12.46	12.6	12.82	0	0	0	0	0	0	0	0	0	0	0	0	0	0
nac	1.594	1.78	1.629	1.865	3.123	2.98	2.353	0.955	0.945	0.687	1.004	1.429	0.951	1.24	2.747	2.114	1.592	0.734	0.861	0.933	0.72

CHAPTER 5

MAPPING METABOLIC-INDUCED CHANGES BY PLASMID BURDEN AND RECOMBINANT PROTEIN PRODUCTION IN *E. COLI*

"Can we create Escherichia coli cells with the ability to secrete silk thread?"

By Y. Murooka and T. Imanaka in Recombinant microbes for
industrial and agricultural applications, CRC Press, 1994

5.1. ABSTRACT

The expression of foreign DNA in bacterial cells has been explored by using recombinant plasmids. The impact of the expression and maintenance of these plasmids in cells is often associated with the metabolic burden phenomenon, which consists in the additional drainage of cellular resources that often leads to metabolic imbalances.

We observed that *Escherichia coli* growth yields and metabolic changes were significant in cells harbouring the plasmid pTRC-AcGFP1. The GC-MS metabolome analysis showed that the intracellular amino acid pool was significantly lower, except for unsaturated long-chain fatty acids. Less pronounced changes were observed when cells were induced to express the recombinant protein, but the expression levels were not considerable. However, significant changes in the relative concentrations of fatty acids were verified. Some amino acids and TCA metabolites levels were also altered.

Here we demonstrated that metabolic profiling represents a useful method to observe metabolic imbalances caused by plasmids maintenance and expression. Alterations in the metabolic state of host cells harbouring plasmids are significant and might impair important cellular functions.

5.2. INTRODUCTION

Recombinant DNA technology offers today the ability to produce foreign proteins in several microorganisms, such as the bacterium *Escherichia coli.* The introduction of a plasmid into bacterial cells to express genes from other sources can perturb the host cellular functions at many levels. The adjustment of most metabolic activities, to cope with the additional biosynthetic requirements, and the consequent adaptation of the transcriptional regulation, to fine-tune the levels of catabolic enzymes, interfere with the physiology of the host cell. Most notably, drastic changes in the specific growth rate can occur when the recombinant protein synthesis is induced. This physiological perturbation is often associated with the metabolic burden that has been defined as "the portion of host cell's resources – either in the form of energy such as ATP or GTP, or raw materials such as nucleotides and amino acids – that is required to maintain and express foreign DNA in the cell" (Hoffmann F and Rinas U, 2004). Cellular growth and expression of foreign gene products in recombinant cells compete for the use of intracellular resources, such as amino acids, nucleic acids and metabolic energy. To overproduce these recombinant proteins, cellular resources are extensively consumed and a shortage of even one of these resources can cause alterations in the cellular metabolism, a condition known as starvation. Since protein synthesis, or more specifically the polymerization of amino acids, is the biggest energy-consuming process in the cell, with more than 50% of the ATP consumption for biosynthetic purposes (Stouthamer AH, 1973), energy generation may become critical in recombinant-protein overproducing cells.

Besides metabolic burden, several other consequences of the production of heterologous proteins have been referred, like the triggering of various stress responses. The transcription of various stress-encoded proteins might be stimulated when cells are exposed to sudden disturbing conditions, like the abnormal protein synthesis (Bukau B, 1993; Harcum SW and Bentley WE, 1999) or the accumulation of inclusion bodies (Ho JGS and Middelberg APJ, 2004; Jurgen B *et al.*, 2000; Panda AK *et al.*, 1999), as well as the shortage of biosynthetic precursors, such as amino acids (Andersson L *et al.*, 1996; Harcum SW and Bentley WE, 1999). In the last case, if the composition of the recombinant protein considerably differs from the average composition of *E. coli* proteins, then, at some point, the intracellular pool of certain amino acids might get depleted and the ratio of charged to uncharged tRNAs decreases. This increases the possibility of an unchaged tRNA binding to a translating ribosome, which induces the activity of the RelA enzyme (Haseltin WA and Block R, 1973). RelA synthesizes an unusual nucleotide, ppGpp, which in a process called stringent response, reprograms gene expression towards the transcription of genes involved in amino acids biosynthesis and stress proteins and inhibits the majority of stable RNAs (rRNA and tRNA) (Chatterji D and Ojha AK, 2001; Irr J and Gallant J, 1969; Jain V *et al.*, 2006; Magnusson LU *et al.*, 2005; Srivatsan A and Wang JD, 2008). It also interferes with the control of plasmid DNA replication (Potrykus K *et al.*, 2000), although the molecular mechanisms of this ppGpp-mediated inhibition of replication remains unclear (Herman A *et al.*, 1994; Wegrzyn G, 1999). Thus, the overall alteration in metabolism is the result of a series of cellular processes triggered to respond to the metabolic load imposed by the production of heterologous proteins, which ultimately leads to the arrest of cellular growth and to the reduction of the recombinant protein productivity (Bentley WE *et al.*, 1990).

In this study, the isogenic mutant Δ*rel*A of *E. coli* W3110 was chosen as the host cell to study the impact of the metabolic burden caused by plasmid DNA maintenance and expression without the interference of other subsequently elicited cellular processes. Mutants in the *rel*A gene do not contain functional RelA (ppGpp synthetase I), which will limit the ppGpp-induced responses, namely the transcriptional activation of stress-regulons associated with cellular mechanisms to survive starvation conditions. As detailed above, regulatory responses induced by the ppGpp-stringent control might be a common problem during recombinant protein overproduction (Dedhia N *et al.*, 1997; Sanden AM *et al.*, 2003) and the use of relaxed mutants (Δ*rel*A) may overcome some of these issues and isolate the metabolic burden phenomenon.

To better understand and control recombinant cultures, it is important to have a more profound knowledge of how the cell couples energy generation and carbon supply in the catabolic pathways with the anabolic requirements at different growth rates under producing and non-producing conditions. Continuous cultivation has been demonstrated (Kim E *et al.*, 1992; Maresova H *et al.*, 2001; Vaiphei ST *et al.*, 2009; Zhang WH *et al.*, 2004) to be a useful bioprocess tool to study the kinetics of recombinant cultures and to identify the primary metabolic bottlenecks associated with the recombinant protein expression. Thus, continuous cultivation of the host strain (*E. coli* W3110 Δ*rel*A) and plasmid-bearing cells (*E. coli* W3110 Δ*rel*A/pTRC-AcGFP1) were performed in a glucose-limited chemostat at a dilution rate of 0.2 h^{-1}.

We aimed to investigate the metabolic profiles at steady-state conditions to evaluate the effects of plasmid DNA maintenance and recombinant protein expression in the absence of stress responses. A GC/MS-based metabolomics approach to measure the intracellular levels of metabolites associated with central metabolic pathways, such as amino and fatty acids, was applied.

5.3. EXPERIMENTAL PROCEDURES

5.3.1. BACTERIAL STRAINS AND GROWTH CONDITIONS

The Δ*rel*A251 mutant *E. coli* strain (obtained from M. Cashel (Xiao H *et al.*, 1991)) was transformed with the cloned pTRC-HisA vector (Invitrogen Corporation and Applied Biosystems Inc, USA) carrying a *gfp* gene amplified from the pAcGFP1 plasmid for the green fluorescent protein AcGFP1, a derivative of GFP from *Aequorea coerulescens* (Clontech, Takara Bio Company, USA). The pTRC-AcGFP1 contains a *trc* promoter for high-level expression of the fusion protein and an ampicillin resistance gene for propagation and selection in *E. coli*.

Three different experiments were performed in a minimal medium consisting of 5 g.kg^{-1} of glucose, 6 g.kg^{-1} of Na_2HPO_4, 3 g.kg^{-1} of KH_2PO_4, 0.5 g.kg^{-1} of NaCl, 1 g.kg^{-1} of NH_4Cl, 0.015 g.kg^{-1} of $CaCl_2$, 0.12 g.kg^{-1} of $MgSO_4.7H_2O$, 0.34 g.kg^{-1} of thiamine, 2 mL.kg^{-1} of trace-element solution (described elsewhere (Rocha I and Ferreira EC, 2002)) and 2 mL.kg^{-1} of vitamins solution (described elsewhere (Rocha I and Ferreira EC, 2002)). The minimal medium was further supplemented with 20 mg.kg^{-1} of L-isoleucine, 20 mg.kg^{-1} of L-valine, 100 mg kg^{-1} of ampicillin and 25 mg kg^{-1} of kanamycin. The first experiment was conducted with the strain *E. coli* W3110 (Δ*rel*A), while *E. coli* W3110 (Δ*rel*A)/pTRC-AcGFP1) was used in the second and the third one. In this last one, induction of AcGFP1 production was performed with 1.5 mM IPTG (isopropyl b-D-thiogalactoside) when the culture optical density (OD_{600nm}) reached a constant value.

Chemostat cultivations were operated at 37°C in a 3 L fermenter (BioFlo 3000, New Brunswick Scientific, USA) with a working volume of 1.5 L. The described minimal medium was continuously fed into the vessel, at least for five residence times, at a dilution rate of 0.2 h^{-1}, and the working volume was kept constant by withdrawing the culture broth through a level control system. The pH of the culture was maintained at 7.0 by adding 2.0M NaOH and 2.0M HCl. Dissolved oxygen was maintained above 30% saturation through a cascade mode, controlling the agitation speed and airflow.

5.3.2. ANALYTICAL TECHNIQUES

Biomass was determined by measuring culture turbidity (OD_{600nm}) and cell dry weight (CDW). In order to determine CDW, 10 mL of broth were filtered by 0.2 µm filters and the filtrate was dried in the microwave to a constant weight. For glucose and acetate analysis, culture broth was centrifuged at 8000 rpm for 15 min to remove the cell debris and the supernatant was collected. The glucose concentration in the culture broth was determined by the dinitrosalicylic acid (DNS) colorimetric method (Miller GL, 1959). The concentrations of acetic acid in the culture broth were determined with an enzymatic test kit (R-Biopharm AG, Germany).

The expression level of AcGFP1 was determined by fluorescence measurements at a 2104 EnVision® Multilabel Reader (PerkinElmer, USA) with excitation and emission wavelengths of 475 and 505 nm, respectively, and a bandwidth of 8 nm. His-Tag purification of the AcGFP1 was performed with HiTrap columns (GE Healthcare Bio-Sciences AB, Sweden) and the concentration was determined by the Bradford method using BSA as standard. Plasmids from cell samples were isolated using the Illustra™ plasmid Prep Mini Spin Kit (GE Healthcare UK Limited, UK) and quantified by a Quant-it assay (Invitrogen Corporation and Applied Biosystems Inc, USA).

5.3.2.1. QUENCHING AND METABOLITE EXTRACTION

For metabolomics analysis, samples (50 mL) were taken quickly from the fermentor and immediately quenched, to halt cellular metabolism, in 200 mL of glycerol/saline solution (60%, v/v) at -23 °C, followed by quick homogenization. Samples were centrifuged at 10,000 rpm for 20 min at -20 °C using a refrigerated centrifuge. The supernatants were discarded and the cell pellets were resuspended in 2 mL of cold glycerol/saline solution (50%, v/v) at -23 °C, followed by a second centrifugation at 8,000 rpm for 30 min at -20 °C. The supernatants were again discarded and the cell pellets were dissolved in 2.5 mL of cold methanol/water solution (50%, v/v) at -30 °C and stored at -80°C for subsequent intracellular metabolite extraction. For that, samples were subjected to three freeze–thaw cycles with 1 min of vigorous mixing using a vortex between each cycle. After the third cycle, samples were centrifuged at 8,000 rpm for 30 min at -20 °C and the supernatants were collected and stored at -80 °C. The extracted pellet was then resuspended in another 2.5 ml of cold methanol/water solution (50%, v/v) and centrifuged at 8,000 rpm for 30 min at -20 °C. The supernatant was collected and pooled with the first one and kept at -80 °C and, afterwards lyophilized.

5.3.2.2. DERIVATIZATION AND GC-MS ANALYSIS

For GC-MS analysis samples were further treated as follows. The dried intracellular metabolite extracts were resuspended in 200 µL of sodium hydroxide (1 M) and derivatized using the methyl chloroformate (MCF) method (Villas-Boas SG *et al.*, 2003). The derivatized samples were then analyzed with a GC-MS systems - GC7890 coupled to MSD5975 (Agilent Technologies, Inc., Santa Clara, CA, USA) equipped with a ZB-1701 GC capillary column, 30m x 250mm id x 0.15 mm (film thickness) with a 5 m guard column (Phenomenex, Inc., Torrance, CA, USA) at a constant flow rate of 1.0 mL/min of helium. The oven temperature was initially held at 45°C for 2 min. Thereafter the temperature is raised with a gradient of 9°C/min until 180°C. This temperature (180°C) is held for 5 min. Then the temperature is raised with a gradient of 40°C/min until 220°C. The temperature is again held for 5 min. Then the temperature is raised with a gradient of 40°C/min until 240°C and this temperature is held for 11.5 min. Finally the temperature is raised with a gradient of 40°C/min until 280°C, which is held for 2 min. The temperature of the inlet is 290°C, the interface temperature 250°C, and the quadrupole temperature 200°C. Sample (1 µL) was injected onto the column under pulsed splitless mode (1.8 bars until 1 min, 20mL/min split flow after 1.01min) and the detector was set with a scan interval of 1.47 seconds and m/z range of 38-650. The mass fragmentation spectrum was analysed in The Automated Mass Spectral Deconvolution and Identification System (AMDIS)

(Stein SE, 1999) to identify the compounds through matching with a library constructed by using analytical chemical standards.

5.3.3. STATISTICAL ANALYSIS

The mass fragmentation spectrum was analysed using the Automated Mass Spectral Deconvolution and Identification System (AMDIS) (Stein SE, 1999) to identify the compounds through matching with a library constructed using analytical chemical standards. The peak intensity values from the AMDIS analysis were refined and corrected and normalized for the recovery of the internal standard (D-4-alanine) and the corresponding biomass concentration. Values were then transformed into Z-scores, by subtracting the average peak intensity corresponding to a metabolite *k* among all the *n* samples (including replicates) in the set of experiments, from the peak intensity value (I_{MK}) of that metabolite in one sample, and dividing that result by the standard deviation ($SD_{M1..Mn}$) of all measured *n* peak intensities, according to:

$$Z - score, k = \frac{\left(I_{Mk} - mean\ I_{Mk_1 \ldots Mk_n}\right)}{SD_{Mk_1 \ldots Mk_n}}$$

Data analysis was carried out in MATLAB 7.1 (Mathworks, USA). Normality of the data and homogeneity of variance were studied using Kolmogorov-Smirnov (KS) and Levene's tests, respectively. One-way ANOVA was applied to compare the means of the relative metabolite concentrations, where each sample was representing a particular experimental condition:

(A) Chemostat cultivation of *E. coli* W3110 (Δ*rel*A) grown at a dilution rate of 0.2 h^{-1};

(B) Chemostat cultivation of *E. coli* W3110 (Δ*rel*A)/pTRC-AcGFP1 grown at a dilution rate of 0.2 h^{-1};

(C) Chemostat cultivation of *E. coli* W3110 (Δ*rel*A)/pTRC-AcGFP1 grown at a dilution rate of 0.2 h^{-1} and IPTG-induced to express AcGFP1.

5.4. Results

5.4.1. Physiological parameters

Three aerobic glucose-limited continuous steady-state cultures were performed with the *E. coli* W3110 (Δ*rel*A) strain and the plasmid-bearing *E. coli* W3110 (Δ*rel*A)/pTRC-AcGFP1 strain. Initially, cells were grown in a batch mode and then the culture was fed at the rate of 300 mL.h^{-1} to achieve the desired dilution rate of 0.2 h^{-1}. When the cultivation system reached a steady state (i.e., when OD_{600} becomes constant) samples were collected to determine the cell dry weight (CDW), AcGFP1 protein, pTRC-AcGFP1 plasmid, glucose and acetate concentrations. Steady-state parameters determined for each culture are summarized in Table 5.1.

Table 5.1. Comparison of the physiological parameters of *E. coli* W3110 (Δ*rel*A) strains during glucose-limited chemostats using a dilution rate of 0.2h^{-1}.

The three *E. coli* cultures are identified as: (A) W3110 (Δ*rel*A) host cells; (B) W3110 (Δ*rel*A) harbouring the pTRC-AcGFP1 plasmid; and W3110 (Δ*rel*A)/pTRC-AcGFP1 induced with IPTG 1.5 mM.

	Strains		
		W3110 (ΔrelA)/pTCR-AcGFP1	
	(A) **W3110 (Δ*rel*A)**	(B) **Non IPTG-induced**	(C) **IPTG-induced**
Dilution rate (h^{-1})	0.2	0.2	0.2
CDW (g.L^{-1})	3.33±0.45	2.69±0.46	2.29±0.39
$Y_{X/G}$ (g.g^{-1})	0.67±0.30	0.54±0.19	0.46±0.12
Glucose (g.L^{-1})	0.023±0.01	0.031±0.0093	0.059±0.011
q_G (g.g^{-1}.h^{-1})	0.3±0.13	0.37±0.13	0.43±0.11
AcGFP1 (g.L^{-1})	-	-	0.18±0.02
q_{AcGFP1} (×10^{-6}) (g.g^{-1}.h^{-1})	-	-	15.78±3.21
Plasmid (mg.L^{-1})	-	21.7±0.28	38.4±1.67
Acetate (g.L^{-1})	0.020	0.18±0.0077	0.52±0.024
q_A (×10^{-3}) (g.g^{-1}.h^{-1})	1.08±0.15	13.15±2.31	45.86±8.14

CDW, cell dry weight; q_G, specific glucose consumption rate (the maintenance coefficient was not considered); q_A, specific acetate consumption rate; q_{AcGFP1}, specific AcGFP1 production rate; $Y_{X/G}$, mass yield coefficient of biomass/glucose.

It was observed that:

- biomass concentration of *E. coli* host cells is higher than of the plasmid-bearing strain (W3110 (Δ*rel*A)/pTCR-AcGFP1);
- the specific glucose consumption rate (q_G) increased for plasmid-bearing and producing cells;
- acetate accumulation increased approximately fifty times in the AcGFP1 producing culture;
- intracellular concentration of the pAcGFP1 is higher when cells are induced to express the recombinant protein.

5.4.2. Effects of plasmid and AcGFP1 biosynthesis on the *E. coli* metabolism

To analyse effects of plasmid DNA maintenance and recombinant protein production on the metabolism of *E. coli*, intracellular metabolites were extracted and analysed by GC-MS. In the present study, fatty acids and the main amino acids and their precursors generated in the central carbon metabolism were measured in order to evaluate the main metabolic changes imposed by the recombinant bioprocess in the host cell. Metabolite profiles corresponding to each chemostat culture are depicted in Figure 5.1.

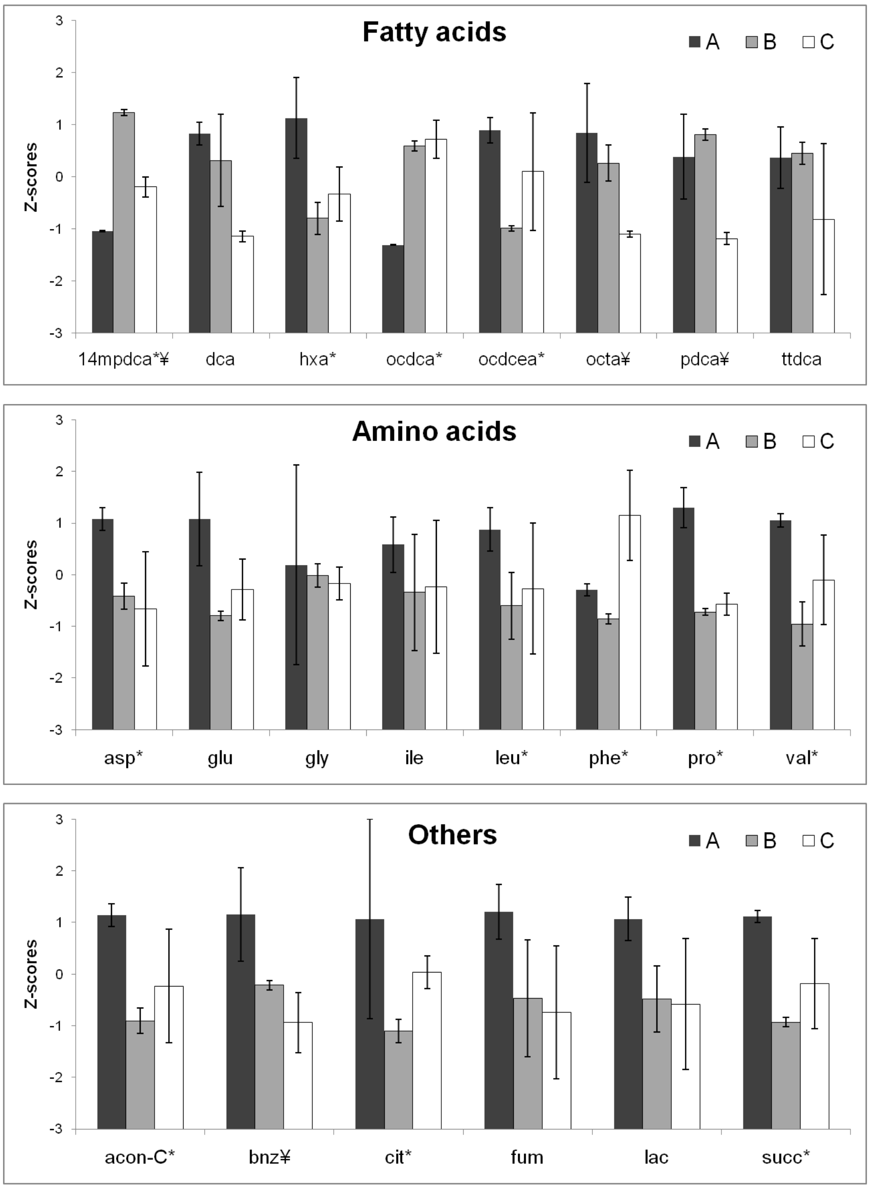


Figure 5.1. Comparison of metabolite profiles during chemostat cultures with a dilution rate of $0.2h^{-1}$ with the following *E. coli* strains: (A) W3110 (Δ*rel*A) host cells; (B) W3110 (Δ*rel*A) harbouring the pTRC-AcGFP1 plasmid; and (C) W3110 (Δ*rel*A)/pTRC-AcGFP1 induced with IPTG 1.5 mM.

Metabolite relative concentrations were normalized into *Z*-scores and are shown as the average of three measurements with the standard deviations. Metabolites that presented significant changes (*p*-values≤0.05) between experiments A/B and B/C are indicated with (*) and (¥), respectively. **Abbreviations:** *Fatty acids* – 14mpdca, 14-methylpentadecanoate; dca, decanoate; hxa, hexanoate; ocdca, octadecanoate; ocdcea, octadecenoate; octa, octanoate; pdca, pentadecanoate; ttdca, tetradecanoate. *Amino acids* – asp, aspartate; glu, glutamate; gly, glycine; ile, isoleucine; leu, leucine; phe, phenylalanine; pro, proline; val, valine. *Others* – acon-C, *cis*-aconitate; bnz, benzoate; cit, citrate; fum, fumarate; lac, lactate; succ, succinate.

To investigate the differences between the three experiments (A, B and C), a one-way ANOVA analysis was performed (Figure 5.2). Normality (KS test) and homogeneity of variances (Levene's test) were tested before using ANOVA and both tests returned that the null hypotheses (i.e., data come from normal distributions with the same variance) cannot be rejected at the 5% significance level (*p*-values were 0.37 and 0.17, respectively).

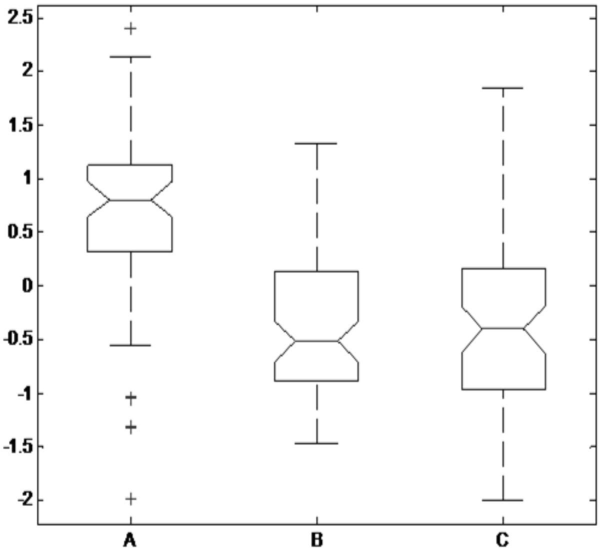


Figure 5.2. One-way ANOVA analysis (*p*-value of 5.2E-13).

It was found that metabolic profiles from cells harbouring the recombinant plasmid (B and C) and host cells (A) show a significant difference (*p*-value less than 0.05). However, differences between the IPTG induced culture (C) and non-induced were not significant. It can be inferred that plasmid burden overlaps the extent of the metabolic load imposed by the expression of the recombinant protein. To further investigate these metabolic changes, the intracellular metabolite pools produced in culture A were compared to culture B to evaluate the impact of the plasmid maintenance in the host cell, and metabolite pools from culture B were compared to culture C to determine the solely impact of the expression of the recombinant protein in the cellular metabolism.

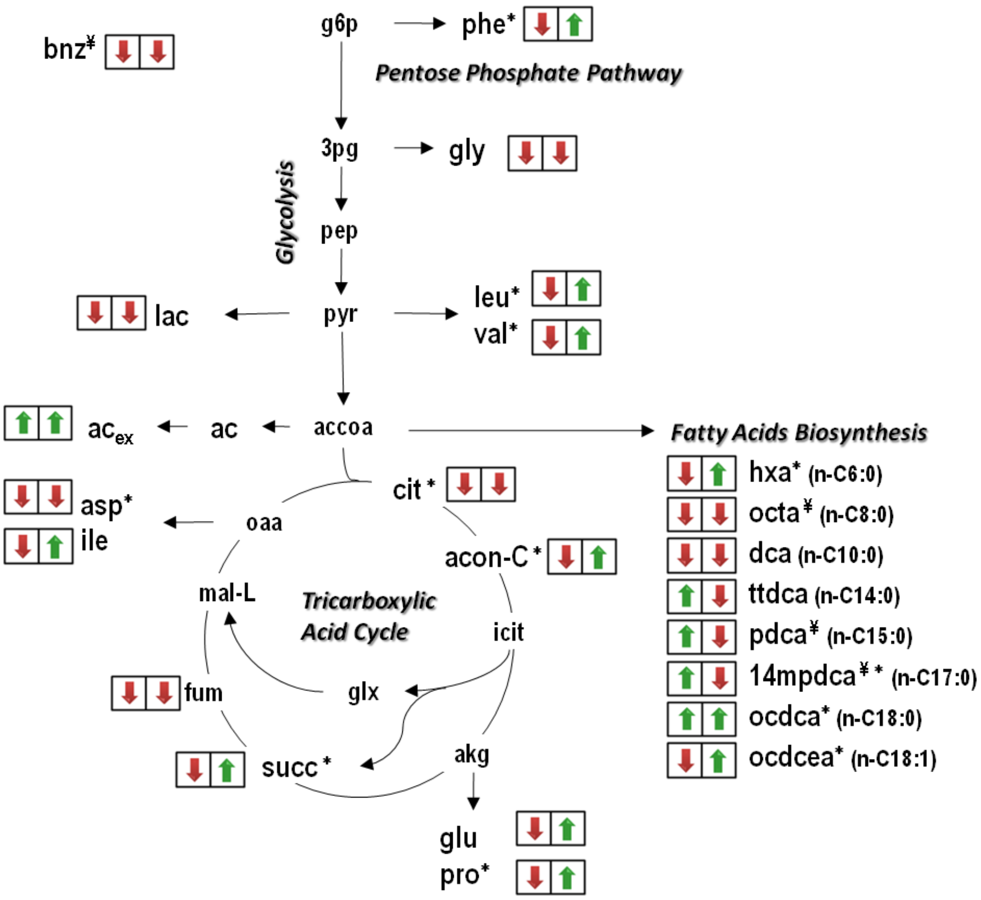


Figure 5.3. Metabolic map illustrating the metabolic changes between the metabolite profiles analysed by GC/MS.

Alterations in the intracellular metabolite concentrations are depicted by a green arrow pointed up, if the ratio of metabolite levels of culture B to culture A (first arrow) and culture C to culture B (second arrow) are superior to 1, and a red arrow pointed down, if ratios are less than 1. Metabolites that presented significant changes (*p*-values≤0.05) between experiments A/B and B/C are indicated with (*) and (¥), respectively. Acetate was also represented as an extracellular metabolite (ac_{ex}). **Abbreviations:** g6p, glucose-6-phosphate; 3pg, 3-phospho-D-glycerate; pep, phosphoenolpyruvate; pyr, pyruvate; accoa, acetyl-CoA; ac, acetate; icit, isocitrate; akg, α-ketoglutarate; glx, glyoxylate; mal-L, L-malate; oaa, oxaloacetate. For other abbreviations see Figure 5.1.

As illustrated in Figure 5.3, most of the metabolite pools were significantly decreased for the cells harbouring the pTRC-AcGFP1 plasmid, except for unsaturated long-chain fatty acids. However, when comparing metabolite pools from cultures B and C, it appears that some metabolites presented higher levels when cells are induced to express the recombinant protein. Though these changes were not significant (p-values higher than 0.05), which poses some uncertainties regarding these variations, differences in the amino acid pools provide important information about the metabolic load associated with the protein biosynthetic process. As shown, only two of the detected amino acids (aspartate and glycine) presented lower levels when cells were producing the recombinant protein, which may indicate that amino acid pools are increased to balance the additional biosynthetic requirements. Fatty acids levels were also significantly changed when cells were induced with IPTG. For example, octanoate and pentadecanoate presented significant lower levels in the protein producing culture. A detailed inspection of Figure 5.1 reveals that, although not significant, most fatty acids presented lower levels during recombinant protein production.

5.5. Discussion

Continuous culture in a chemostat is a practical way to study the state of host cells during the production of recombinant proteins under defined growth conditions. The physiological effects of the plasmid burden on the metabolism and the maximum biosynthetic capacity of the host cells can be evaluated during a continuous growth for many generations, which present some advantages compared to other cultivation modes (Maresova H *et al.*, 2001; Zhang WH *et al.*, 2004).

In the present investigation, the estimated physiological parameters, such as q_{G}, q_{A} and q_{AcGFP1}, revealed that culture performance by cells harbouring the pAcGFP1 plasmid is quite different from the chemostat culture with host cells. Besides the higher q_{G}, the accumulation of the metabolic by-product acetate in the plasmid-bearing cultures was significant. It has been described (Holms H, 1996; Majewski RA and Domach MM, 1990) that when increasing the consumption of glucose, carbon flux through the glycolysis pathway exceeds the TCA capabilities and acetyl-CoA is over-accumulated, resulting in the excretion of acetate. Indeed, it is expected that to cope with the additional biosynthetic requirements, cells enhance their ability to uptake the carbon source (e.g. glucose). As a consequence, glucose overflow leads to the accumulation of certain metabolic by-products, like acetate, that are afterwards excreted to the extracellular medium. This not only represents a diversion of carbon that might otherwise have been used for biomass or recombinant protein synthesis, but, at certain levels a toxicity agent for cells.

Reduction of the growth yields and protein expression has been indicated as a result of this metabolic state. In fact, biomass concentration decreased from 3.33±0.45 g.L^{-1} in the host *E. coli* culture (A) to 2.69±0.46 g.L^{-1} and 2.29±0.39 g.L^{-1} in non-induced (B) and IPTG-induced (C) plasmid-bearing chemostat cultures, respectively. Moreover, acetate levels in the IPTG-induced culture reached toxic levels (0.52±0.02 g.L^{-1}), which may be deleterious for cells. The specific acetate production rate (q_A) increased almost 45 times compared to the culture with host cells.

This metabolic imbalance was also observed in the metabolite profiles of plasmid-bearing chemostat cultures. Metabolic profiling demonstrated that most metabolites from the host cell culture (A) presented significantly different levels from those analysed in the plasmid-bearing cultures (B and C). These differences were stronger in the amino acids pool, where aspartate, isoleucine, leucine, phenylalanine, proline and valine presented lower levels for the plasmid-bearing culture (between 20-70% less). It seems that either amino acid biosynthesis decreased, or the drainage of metabolic resources for plasmid maintenance was superior to their synthesis. Plasmids are non-essential DNA molecules, which require additional intracellular resources and energy to be maintained in host cells, including for DNA replication and expression of resistance gene products (e.g. proteins encoded by an antibiotic resistance marker gene). Several studies (Flores S *et al.*, 2004; Ozkan P *et al.*, 2005; Wang ZJ *et al.*, 2006) indicate that the presence of plasmids significantly influences various metabolic pathways in the host cells, e.g. glycolysis and the pentose phosphate (PP) pathway. On another study, however, TCA enzymatic activities at a dilution rate of 0.2 h^{-1} did not change significantly between plasmid-bearing and non-bearing cultures (Wang ZJ *et al.*, 2006). Nevertheless, in the present experiments, the level of TCA intermediaries in cultures harbouring the pAcGFP1 plasmid decreased, and more significantly for citrate, *cis*-aconitate and succinate. This might be associated with the drainage of amino acid precursors to fulfil the extra requirement for the expression of resistance gene products, which are not completely replenished by anaplerotic reactions.

By combining metabolic flux analyses and DNA microarrays, another study found that the phosphotransferase system for glucose uptake and glycolytic and PP pathway genes were up-regulated in a plasmid-bearing culture when compared with host cells (Wang ZJ *et al.*, 2006). These pathways are involved in the generation of NADPH for biosynthetic pathways, recruiting essential metabolites for nucleic acids, amino acids and vitamins, and the generation of ingredients of the cell lipopolysaccharide layer (Wang ZJ *et al.*, 2006). Therefore, to supply cells for the additional biosynthetic processes, metabolic fluxes through these pathways are probably up-regulated. The increased carbon uptake is channelled to the glycolysis and PP pathway and, although no experimental data was obtained in our study to confirm this, metabolome data indicate that plasmid-induced changes up-regulate certain metabolic enzymes in recombinant cells, which could in turn lead to the atypical increase of flux in some pathways. As the glycolytic activity increases, the TCA capacity to consume the excess of acetyl-CoA is insufficient and acetate is accumulated as a by-product. This is in a good agreement with the experimental results that confirmed the increase in q_G and q_A for plasmid-bearing cultures and the simultaneous failure to increase the level of TCA metabolites, most likely because TCA intermediates are being consumed for the synthesis of amino acids.

It is evident that plasmid burden causes the shortage of certain biosynthetic products, namely amino acids, generating a metabolic imbalance that might also be responsible for the accumulation of some by-products. Yet, it was expected that the metabolism of the bacteria would be even more largely affected by the expression of the AcGFP1 under the control of a high-level expression promoter. The plasmid concentration almost doubled compared with the non-induced culture, which would represent an increase in the amino acids required for the production of the Lac repressor and β-lactamase enzyme, and nucleotides and energy for the plasmid DNA replication. In fact, besides cellular resources needed for the synthesis of AcGFP1, plasmid maintenance would increase the metabolic load in the IPTG-induced culture. However, the observed metabolite profiles did not reflect this phenomenon and only three metabolites, benzoate, octanoate and pentadecanoate presented significant different levels between the ITPG-induced and non-induced cultures. Most fatty acids were characterized by a small decrease when cells were induced to produce the recombinant protein, but there are no indications that fatty acids biosynthesis is affected during recombinant processes. The increase in some amino acids and TCA metabolites levels were not significant, and thus, it is suggested that the increase in q_G was a consequence of metabolic requisites to replenish precursors for plasmid replication and expression, and small differences in metabolite levels resulted from the metabolic unbalance caused by these recombinant processes. This study demonstrates that plasmid burden is of major importance when expressing foreign proteins in bacterial cells and, thus strategies to enhance recombinant cultivations must consider this as critical.

5.6. REFERENCES

1. Andersson L *et al* (1996) Impact of plasmid presence and induction on cellular responses in fed batch cultures of *Escherichia coli*. *Journal of Biotechnology* 46 (3):255-263.

2. Bentley WE *et al* (1990) Plasmid-encoded protein - The principal factor in the metabolic burden associated with recombinant bacteria. *Biotechnology and Bioengineering* 35 (7):668-681.

3. Bukau B (1993) Regulation of the *Escherichia coli* Heat-Shock Response. *Molecular Microbiology* 9 (4):671-680.

4. Chatterji D and Ojha AK (2001) Revisiting the stringent response, ppGpp and starvation signaling. *Current Opinion in Microbiology* 4 (2):160-165.

5. Dedhia N *et al* (1997) Improvement in recombinant protein production in ppGpp-deficient *Escherichia coli*. *Biotechnology and Bioengineering* 53 (4):380-386.

6. Flores S *et al* (2004) Growth rate recovery of *Escherichia coli* cultures carrying a multicopy plasmid, by engineering of the pentose-phosphate pathway. *Biotechnology and Bioengineering* 87 (4):485-494.

7. Harcum SW and Bentley WE (1999) Heat-shock and stringent responses have overlapping protease activity in *Escherichia coli*. Implications for heterologous protein yield. *Applied Biochemistry and Biotechnology* 80 (1):23-37.

8. Haseltin WA and Block R (1973) Synthesis of guanosine tetraphosphate and pentaphosphate requires presence of a codon-specific, uncharged transfer ribonucleic acid in acceptor site of ribosomes - (Stringent control ppGpp (Msi) and pppGpp (Msii) protein synthesis *Escherichia coli*). *Proceedings of the National Academy of Sciences of the United States of America* 70 (5):1564-1568.

9. Herman A, Wegrzyn A, and Wegrzyn G (1994) Differential replication of plasmids during stringent and relaxed response of *Escherichia coli*. *Plasmid* 32 (1):89-94.

10. Ho JGS and Middelberg APJ (2004) Estimating the potential refolding yield of recombinant proteins expressed as inclusion bodies. *Biotechnology and Bioengineering* 87 (5):584-592.

11. Hoffmann F and Rinas U (2004) Stress induced by recombinant protein production in *Escherichia coli*. *Advances in Biochemical Engineering / Biotechnology* 89:73-92.

12. Holms H (1996) Flux analysis and control of the central metabolic pathways in *Escherichia coli*. *Fems Microbiology Reviews* 19 (2):85-116.

13. Irr J and Gallant J (1969) Control of ribonucleic acid synthesis in *Escherichia coli* .2. Stringent control of energy metabolism. *Journal of Biological Chemistry* 244 (8):2233-&.

14. Jain V, Kumar M, and Chatterji D (2006) ppGpp: Stringent response and survival. *Journal of Microbiology* 44 (1):1-10.

15. Jurgen B *et al* (2000) Monitoring of genes that respond to overproduction of an insoluble recombinant protein in *Escherichia coli* glucose-limited fed-batch fermentations. *Biotechnology and Bioengineering* 70 (2):217-224.

16. Kim E *et al* (1992) Expression of human epidermal growth factor by *Escherichia coli* in continuous culture. *Biotechnology Letters* 14 (5):339-344.

17. Magnusson LU, Farewell A, and Nystrom T (2005) ppGpp: a global regulator in *Escherichia coli*. *Trends in Microbiology* 13 (5):236-242.

18. Majewski RA and Domach MM (1990) Simple constrained-optimization view of acetate overflow in *E. coli*. *Biotechnology and Bioengineering* 35:732-738.

19. Maresova H, Stepanek V, and Kyslik P (2001) A chemostat culture as a tool for the improvement of a recombinant E. coli strain over-producing penicillin G acylase. *Biotechnology and Bioengineering* 75 (1):46-52.

20. Miller GL (1959) Use of dinitrosalicylic acid reagent for the determination of reducing sugar. *Analytical Chemistry* 31:426-428.

21. Ozkan P *et al* (2005) Metabolic flux analysis of recombinant protein overproduction in *Escherichia coli*. *Biochemical Engineering Journal* 22 (2):167-195.

22. Panda AK *et al* (1999) Kinetics of inclusion body production in batch and high cell density fed-batch culture of *Escherichia coli* expressing ovine growth hormone. *Journal of Biotechnology* 75 (2-3):161-172.

23. Potrykus K *et al* (2000) Replication of oriJ-based plasmid DNA during the stringent and relaxed responses of *Escherichia coli*. *Plasmid* 44 (2):111-126.

24. Rocha I and Ferreira EC (2002) On-line simultaneous monitoring of glucose and acetate with FIA during high cell density fermentation of recombinant *E. coli*. *Analytica Chimica Acta* 462 (2):293-304.

25. Sanden AM *et al* (2003) Limiting factors in *Escherichia coli* fed-batch production of recombinant proteins. *Biotechnology and Bioengineering* 81 (2):158-166.

26. Srivatsan A and Wang JD (2008) Control of bacterial transcription, translation and replication by (p)ppGpp. *Current Opinion in Microbiology* 11 (2):100-105.

27. Stein SE (1999) An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *Journal of the American Society for Mass Spectrometry* 10 (8):770-781.

28. Stouthamer AH (1973) A theoretical study on the amount of ATP required for synthesis of microbial cell material. *Antonie Van Leeuwenhoek* 39 (3):545-565.

29. Vaiphei ST, Pandey G, and Mukherjee KJ (2009) Kinetic studies of recombinant human interferon-gamma expression in continuous cultures of *E. coli*. *Journal of Industrial Microbiology & Biotechnology* 36 (12):1453-1458.

30. Villas-Boas SG *et al* (2003) Simultaneous analysis of amino and nonamino organic acids as methyl chloroformate derivatives using gas chromatography-mass spectrometry. *Analytical Biochemistry* 322 (1):134-138.

31. Wang ZJ *et al* (2006) Effects of the presence of ColEl plasmid DNA in *Escherichia coli* on the host cell metabolism. *Microbial Cell Factories* 5.

32. Wegrzyn G (1999) Replication of plasmids during bacterial response to amino acid starvation. *Plasmid* 41 (1):1-16.

33. Xiao H *et al* (1991) Residual guanosine 3',5'-bispyrophosphate synthetic activity of *rel*A null mutants can be eliminated by *spo*T null mutations. *Journal of Biological Chemistry* 266 (9):5980-5990.

34. Zhang WH *et al* (2004) Optimization of cell density and dilution rate in *Pichia pastoris* continuous fermentations for production of recombinant proteins. *Journal of Industrial Microbiology & Biotechnology* 31 (7):330-334.

5.7. Appendix A

Relative concentrations of metabolites measured by GC/MS in triplicate samples from chemostat cultures: (A) W3110 (Δ*rel*A) host cells; (B) W3110 (Δ*rel*A) harbouring the pTRC-AcGFP1 plasmid; and (C) W3110 (Δ*rel*A)/pTRC-AcGFP1 induced with IPTG 1.5 mM.

	A1	**A2**	**A3**	**B1**	**B2**	**B3**	**C1**	**C2**	**C3**
14mpdca	0.186429	0.185507	0.136943	5.283406	5.47573	5.255288	2.041313	1.687848	2.564628
acon-C	0.038408	0.030225	0.025485	0.010684	0.011733	0.00967	0.01763	0.011852	0.023088
asp	0.565733	0.557056	0.514258	0.368158	0.377643	0.319164	0.389346	0.161777	0.418649
bnz	0.114883	0.070905	0.085612	0.055245	0.047175	0.044189	0.026036	0.028837	0.025808
cit	0.337218	0.251821	0.23983	0	0	0	0.146447	0.078008	0.211557
dca	0.102219	0.100025	0.115247	0.124137	0.073598	0.061794	0.027894	0.034146	0.034248
fum	0.301198	0.169623	0.157737	0.060372	0.032003	0.03859	0.021242	0.013923	0.014997
glu	0.736768	0.489133	0.414076	0.188132	0.215817	0.187001	0.334468	0.167956	0.372613
gly	0.360081	0.000339	0.271089	0.200151	0.205155	0.165167	0.190983	0.140309	0.194851
hxa	0.661241	0.571264	0.718048	0.472048	0.436159	0.493936	0.504508	0.564231	0.465243
ile	0.126371	0.153941	0.142446	0.117106	0.087624	0.146101	0.143367	0.081522	0.134098
lac	1.64279	0.925121	0.954811	0.693951	0.578771	0.61849	0.610835	0.676584	0.495655
leu	0.264534	0.298567	0.297116	0.216668	0.191265	0.250048	0.282915	0.17038	0.25006
ocdca	0.289862	0.279734	0.283674	2.376452	2.457453	2.250954	2.364764	2.189778	2.956353
ocdcea	1.313383	1.174697	1.321075	0.625673	0.640545	0.658477	0.90643	0.688921	1.421259
octa	0.888126	0.680638	0.705869	0.73251	0.684583	0.649707	0.533402	0.520965	0.522719
pdca	0.182145	0.084711	0.200047	0.178676	0.193723	0.191308	0.035338	0.027714	0.045461
phe	0.068627	0.081695	0.075779	0.038283	0.042032	0.048652	0.174938	0.102329	0.199001
pro	0.353901	0.312707	0.274698	0.111703	0.099403	0.108202	0.141432	0.098132	0.124664
succ	0.877927	0.668273	0.448312	0.140408	0.089389	0.089135	0.251111	0.247176	0.436227
ttdca	0.676711	0.902067	0.711598	0.780771	0.822851	0.735005	0.430089	0.278018	0.852923
val	0.341881	0.3304	0.32463	0.226384	0.167964	0.18704	0.290087	0.18373	0.284151

CHAPTER 6

METABOLIC FOOTPRINT ANALYSIS OF RECOMBINANT *E. COLI* STRAINS DURING FED-BATCH FERMENTATIONS

"...in metabolic footprinting the medium is the message."

By D. B. Kell. *et al* in Metabolic footprinting and systems biology:
the medium is the message,
Nat Rev Microbiol. 2005 Jul;3(7):557-65

6.1 ABSTRACT

Metabolic footprinting has become a valuable analytical approach for the characterization of phenotypes and the distinction of specific metabolic states resulting from environmental and/or genetic alterations. The metabolic impact of heterologous protein production in *Escherichia coli* cells is of particular interest, since there are numerous cellular stresses triggered during this process that limit the overall productivity. Because the knowledge on the metabolic responses in recombinant bioprocesses is still scarce, metabolic footprinting can provide relevant information on the intrinsic metabolic adjustments. Thus, the metabolic footprints generated by *Escherichia coli* W3110 and the Δ*rel*A mutant strain during recombinant fed-batch fermentations at different experimental conditions, were measured and interpreted. The IPTG-induction of the heterologous protein expression resulted in the rapid accumulation of inhibitors of the glyoxylate shunt in the culture broth, suggesting the clearance of this anaplerotic route to replenish the TCA intermediaries withdrawn for the additional formation of heterologous protein. Nutritional shifts were also critical in the recombinant cellular metabolism, indicating that cells employ diverse strategies to counteract imbalances in the cellular metabolism, including the secretion of certain metabolites that are, most likely, used as a metabolic relief to survival processes.

6.2 INTRODUCTION

The optimization of bioprocesses using recombinant microorganisms is still restrained by the lack of information available on the metabolic responses induced by various stress conditions (Gnoth S *et al.*, 2008). Significant knowledge could be gained from a comprehensive analysis of the metabolic footprint (i.e. extracellular metabolite profiling) by inspecting key metabolic changes and understanding their relation with environmental conditions (Allen J *et al.*, 2003; Allen J *et al.*, 2004; Kell DB *et al.*, 2005; Kell DB *et al.*, 2003; Pope GA *et al.*, 2007; Villas-Boas SG *et al.*, 2006; Villas-Boas SG *et al.*, 2008).

Some work has been published addressing various metabolic responses during the production of heterologous proteins in *E. coli* (Chou CP, 2007). Experimental studies showed that the host cell metabolism undergoes a severe metabolic burden, resulting in rapid exhaustion of essential precursors and cellular energy (Aldor IS *et al.*, 2005). Typically, strong expression systems are employed to assure the production of large amounts of heterologous proteins by the host, which uses a large quantity of metabolic and energy resources in order to maintain and express the foreign DNA (Seo JH *et al.*, 2003). Heterologous protein production is also believed to diminish flow in the TCA cycle through the withdrawal of intermediates that serve as precursors for amino acid biosynthesis (Jurgen B *et al.*, 2000). Moreover, the difference usually observed in amino acid composition of foreign proteins and the average composition of the host proteins contributes to this metabolic imbalance (Bentley WE *et al.*, 1990; Bonomo J and Gill RT, 2005; Glick BR, 1995; Harcum SW, 2002).

Stringent response has been associated with the stress phenomenon caused by the depletion of certain metabolic resources, namely amino acids (Harcum SW and Bentley WE, 1999). The increased level of free tRNA molecules, due to the lack of amino acids, triggers this stress response that is characterized by the arrest of the ribosomal translation process and the rapid RelA-mediated accumulation of ppGpp (Jain V *et al.*, 2006). This nucleotide has been found (Chatterji D *et al.*, 1998; Toulokhonov II *et al.*, 2001; Wu J and Xie J, 2009) to bind directly to the RNA polymerase, adjusting the transcriptional activity from the expression of genes required for rapid growth, to stress-related genes and amino acid biosynthetic operons (Durfee T *et al.*, 2008; Magnusson LU *et al.*, 2005; Roberts JW, 2009; Srivatsan A and Wang JD, 2008). The regulatory mechanisms of this ppGpp-induced stress response are known in some detail (Artsimovitch I *et al.*, 2004; Chatterji D and Ojha AK, 2001; Jishage M *et al.*, 2002), but the impact of this response on the cellular metabolism has been less studied. Knowledge on how these responses take place and how to dodge them is of great importance, since *E. coli* has become one of the most used microbial systems to produce heterologous proteins.

Here, we aimed at investigate the physiological and metabolic changes in *E. coli* cultures during the production of heterologous proteins by performing a metabolic footprinting analysis. Furthermore, the focus of the study was not only to evaluate the changes of the extracellular metabolite pools during heterologous protein production, but also to assess the effect of removing a gene closely related to the initiation of the stringent response (*rel*A) on the cellular behaviour. Thus, the W3110 and the isogenic mutant (Δ*rel*A) *E. coli* strains were grown and induced to produce a heterologous protein (AcGFP1) at different nutritional conditions in a controlled fed-batch mode.

6.3 EXPERIMENTAL PROCEDURES

6.3.1 MICROBIAL STRAINS

Escherichia coli strains W3110 (F-, *LAM-*, *IN*[*rrnD-rrnE*]*1*, *rph-1*) and the isogenic mutant containing the Δ*rel*A251*::kan* allele (obtained from M. Cashel, National Institute of Health, USA) were transformed with the cloned pTRC-HisA-AcGFP1 plasmid encoding the expression of the recombinant AcGFP1 protein. The *gfp* gene was amplified from the pAcGFP1 plasmid (Clontech, Takara Bio Company, USA) that encodes for the green fluorescent protein AcGFP1, a derivative of AcGFP from *Aequorea coerulescens*. The PCR product was then cloned into the pTRC-HisA vector (Invitrogen Corporation and Applied Biosystems Inc, USA) that contains a *trc* promoter for high-level expression of the fusion protein and an ampicillin resistance gene for propagation and selection in *E. coli*.

6.3.2 Growth conditions

Precultures were prepared in 500-mL shaking flasks filled with 300 mL of minimal medium consisting of 5 g.kg^{-1} of glucose, 6 g.kg^{-1} of Na_2HPO_4, 3 g.kg^{-1} of KH_2PO_4, 0.5 g.kg^{-1} of NaCl, 1 g.kg^{-1} of NH_4Cl, 0.015 g.kg^{-1} of $CaCl_2$, 0.12 g.kg^{-1} of $MgSO_4.7H_2O$, 0.34 g.kg^{-1} of thiamine, 2 mL.kg^{-1} of trace-element solution (described elsewhere(Rocha I and Ferreira EC, 2002)) and 2 mL.kg^{-1} of vitamins solution (described elsewhere(Rocha I and Ferreira EC, 2002)). The minimal medium containing additional 20 mg.kg^{-1} of L-isoleucine and 100 mg kg^{-1} of ampicillin was used to grow the recombinant wild-type strain, while this same medium with further addition of 20 mg.kg^{-1} L-valine and 25 mg.kg^{-1} kanamycin was used to grow the Δ*rel*A mutant strain. Cells were thereafter washed and transferred to a 5-L fermenter (Biostat MD, Sartorius) with a working volume of 2 L containing the same minimal medium, except glucose. The fed-batch operation was started immediately after inoculation at 37?C, pH 7 and dissolved oxygen (DO) above 30%. The feed media used contained 50 g.kg^{-1} of glucose, 10 g.kg^{-1} of NH_4Cl, 4 g.kg^{-1} of $MgSO_4.7H_2O$ and the additional requirements for amino acids and antibiotics as described before. The induction of AcGFP1 production was performed with 1.5 mM IPTG (isopropyl b-D-thiogalactoside) when the microbial culture reached an OD_{600nm} of 2.3. Fermentation conditions were monitored and controlled via a computer control system. A closed-loop feeding control algorithm was employed to maintain a constant specific growth rate (μ) in the fed-batch culture (Rocha I *et al.*, 2008). The algorithm is based on a Monod kinetic model using glucose as the only growth-limiting substrate. The model combines terms for cell growth ($X.\mu$), glucose consumption ($Y_{X/S}.S_f$) and the online measurement of culture medium weight (W_R) to control the feeding profile, represented by:

$$F = \frac{X.\mu.W_R}{Y_{X/S}.S_f}$$

The biomass concentration (X) was initially measured by optical density and estimated at each acquisition time (every 3 minutes). The predicted growth yields on glucose ($Y_{X/S}$) were set to 0.35 and 0.2, when the specific growth rates were set to 0.1 and 0.2 h^{-1}, respectively. The fed-batch experiments were at first conducted at μ = 0.1 h^{-1}, corresponding to the pre-induction (A) and the post-induction (B) phases. Afterwards μ was changed to 0.2 h^{-1}, which was kept during almost 4 hours, corresponding to a nutritional upshift phase (phase C). When the feeding was ceased (glucose limitation phase or phase D), growth was followed until the OD_{600nm} dropped.

6.3.3 SAMPLING AND ANALYTICAL PROCEDURES

Cell growth was monitored by measuring optical density (OD_{600nm}) and cell dry weight. In order to determine cell dry weight, 10 mL of broth were centrifuged at 10000 g for 20 min at 4°C, washed twice with deionised water and dried at 105 °C to constant weight. The expression level of AcGFP1 was determined by fluorescence measurements at a Jasco FP-6200 spectrofluorometer with excitation and emission wavelengths of 475 and 505 nm, respectively, a bandwidth of 10 nm and a high sensitivity response in 0.1 seconds. His-Tag purification of the AcGFP1 was performed with HiTrap columns (GE Healthcare Bio-Sciences AB, Sweden) and the concentration was determined by the Bradford method using BSA as standard. For further analysis, culture samples were centrifuged (15 min, 3000 rpm, 4°C) and the resulting supernatants were immediately filtered and collected. Afterwards the samples were stored at -20°C for subsequent analysis and lyophilisation. Glucose and acetate were analysed by HPLC with a refractive index detector (Jasco, Canada) and a Chrompack organic acids column (Varian, USA) at 35°C. The mobile phase consisted in a 0.01N solution of H_2SO_4 at a flow rate of 0.6 mL/min.

6.3.3.1 Derivatization and GC-MS analysis

One millilitre of the extracellular samples was lyophilized in triplicate. The lyophilized samples were then derivatized using the methyl chloroformate (MCF) method (Villas-Boas SG *et al.*, 2003) and analyzed with a GC-MS system - GC7890 coupled to an MSD5975 - (Agilent Technologies, Inc., Santa Clara, CA, USA) equipped with a ZB-1701 GC capillary column, 30 m x 250 mm id x 0.15 mm (film thickness) with 5 m guard column (Phenomenex, Inc., Torrance, CA, USA), at a constant flow rate of 1.0 mL/min of helium. Samples (1 µL) were injected onto the column under a pulsed splitless mode (1.8 bars until 1 min, 2 0mL/min split flow after 1.01 min) and the detector was set with a scan interval of 1.47 seconds and m/z range of 38-650. The oven temperature was initially held at 45°C for 2 minutes. Thereafter, the temperature was raised with a gradient of 9°C/min until 180°C and held at this value for 5 minutes. The temperature was raised again at a gradient of 40°C/min in three steps: until 220°C (held for 5 min), 240°C (held for 11.5 min) and finally 280°C (held for 2 min). The temperature of the inlet was 290°C, the interface temperature 250°C, and the quadrupole temperature 200°C.

6.3.4 Data processing and statistical analysis

The mass fragmentation spectrum was analysed using the Automated Mass Spectral Deconvolution and Identification System (AMDIS) (Stein SE, 1999) to identify the compounds through matching with a library constructed by using analytical chemical standards. The peak intensity values from the AMDIS analysis were refined and corrected for the recovery of the internal standard (D-4-alanine) and normalized with the corresponding biomass concentration. The corrected and normalized peak intensity values were thereafter transformed into Z-scores for each dataset, which corresponds to a single fermentation. Z-score values were calculated by subtracting the average peak intensity among all the *n* samples (including replicates) of a fermentation from the peak intensity values (I_{MK}) corresponding to each metabolite *k,* and dividing that result by the standard deviation ($SD_{M1..Mn}$) of all measured *n* peak intensities, according to:

$$Z - score, k = \frac{\left(I_{Mk} - mean\ I_{Mk_1 \ldots Mk_n}\right)}{SD_{Mk_1 \ldots Mk_n}}$$

Further data processing and statistical analysis were performed with MATLAB (version 2009b, The Mathworks, Inc) and MultiExperiment Viewer (MeV) (Saeed AI *et al.*, 2003). Hierarchical clustering (HCL) was used to cluster the samples and metabolites based on the Pearson correlation metrics and principal component analysis (PCA) was used to visualize whether samples could be discriminated based on their metabolic footprints. Component coefficients were computed to expose the metabolites that contributed the most to discriminate between sample clusters determined by the HCL analysis, i.e to characterize the metabolic shifts, sample clusters defined by HCL were used to determine which metabolites suffered the highest alterations. Pearson's correlation coefficients (*r*) were computed to evaluate the degree of association between the metabolite profiles produced by the W3110 and Δ*rel*A *E. coli* cultures. *p*-values associated with each Pearson correlation coefficient was calculated using a Student's t distribution to test the null hypothesis (H_0) of no significant correlation between the metabolite profiles from the two cultures, against the alternative hypothesis (H_1) that establishes a significant correlation between the profiles. The p-value for Pearson's correlation coefficient is based on the test statistic, *s*, with *n*-2 degrees of freedom:

$$s = \frac{r \times \sqrt{n-2}}{\sqrt{1-r^2}}$$

6.4 RESULTS

6.4.1 FED-BATCH FERMENTATIONS OF W3110 AND Δ*REL*A *E. COLI* CELLS

To elucidate the physiological responses of the W3110 and Δ*rel*A mutant *E. coli* strains during recombinant fed-batch fermentation, cells were grown aerobically with a closed-loop feeding control to maintain a quasi-steady state growth. Fed-batch cultures were started with a low specific growth rate setpoint (between 0.09 and 0.16 h^{-1}) and growth characteristics were determined prior and after IPTG induction (phases A and B, respectively). Then, the glucose feeding rate was increased to maintain the specific growth rate around 0.2 h^{-1} (phase C) to evaluate the impact of nutritional upshift on the *E. coli* cultures during heterologous protein production. Nutrient limitation by ceasing the glucose feeding (phase D) was finally examined in these *E. coli* cultures until growth arrested.

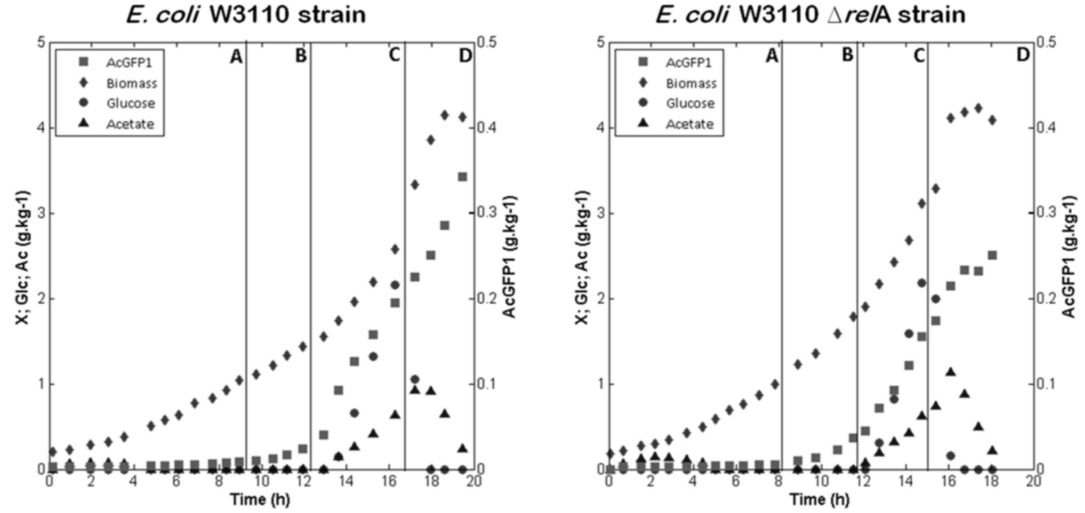


Figure 6.1. Growth curves for the W3110 [pTRC-His-AcGFP1] ***E. coli*** **and the isogenic derivative Δ*rel*A251*::kan* strains cultured in fed-batch fermentations with closed-loop feeding control.**

Phases: A- prior to induction; B- after IPTG induction; C- after growth upshift; and D- after nutrient downshift.

In Figure 6.1, biomass, glucose, acetate and AcGFP1 concentrations are depicted for each fed-batch culture. Table 6.1 shows the growth parameters determined at each experimental condition. As demonstrated, prior to IPTG induction the specific growth rates were similar for both strains (around 0.16 h^{-1}), but after IPTG induction the growth of the wild-type strain was significantly reduced (0.092 h^{-1}) while the Δ*rel*A mutant had a specific growth rate of 0.13 h^{-1}. In contrast, the AcGFP1 production increased after IPTG induction to a maximum of around 7×10^{-3} g.g^{-1}DW.h^{-1} for both cultures. Upon the nutritional upshift (phase C), both strains increased their growth rates and the AcGFP1 production rates. The maximum AcGFP1 production was determined to be close to 20×10^{-3} g.g^{-1}DW.h^{-1} in both cultures. At these conditions acetate accumulation was detected in both cultures at rates below 90×10^{-3} g.g^{-1}DW.h^{-1}. When the glucose feeding was stopped (phase D) acetate was consumed and growth was arrested. However, AcGFP1 production continued until the carbon sources (i.e. glucose and acetate) were completely depleted from the medium.

Table 6.1. Growth parameters of the fed-batch cultures of the W3110 and Δ*rel*A *E. coli* strains.

	Phases	Specific growth rate set point (h^{-1})	Expected biomass yield ($g_{DW}.g^{-1}$substrate)	Specific rates ($g.g^{-1}DW.h^{-1}$)				Biomass yield ($g_{DW}.g^{-1}$substrate)
		μ	$Y_{X/S}$	μ	q_{AcGFP1} ($\times 10^{-3}$)	q_{Gluc}	q_{Ac}	$Y_{X/S}$
W3110 strain	A	0.1	0.35	0.16±0.020	0.60±0.12	0.30±0.067	-	0.59±0.038
	B	0.1	0.35	0.092±0.023	6.8±4.5	0.29±0.052	-	0.36±0.065
	C	0.2	0.2	0.17±0.023	19±6.0	0.64±0.11	0.089±0.0074	0.26±0.068
	D	-	-	-	5.3±5.8	0.24±0.043	-0.49±0.27	0.31±0.069[a]
Δ*rel*A mutant strain	A	0.1	0.35	0.16±0.019	0.60±0.45	0.21±0.016	-	0.62±0.059
	B	0.1	0.35	0.13±0.026	6.9±3.8	0.25±0.036	-	0.48±0.022
	C	0.2	0.2	0.20±0.030	17±1.8	0.56±0.089	0.084±0.0072	0.36±0.010
	D	-	-	-	14±2.1	0.47±0.14	-0.92±0.23	0.04±0.049[a]

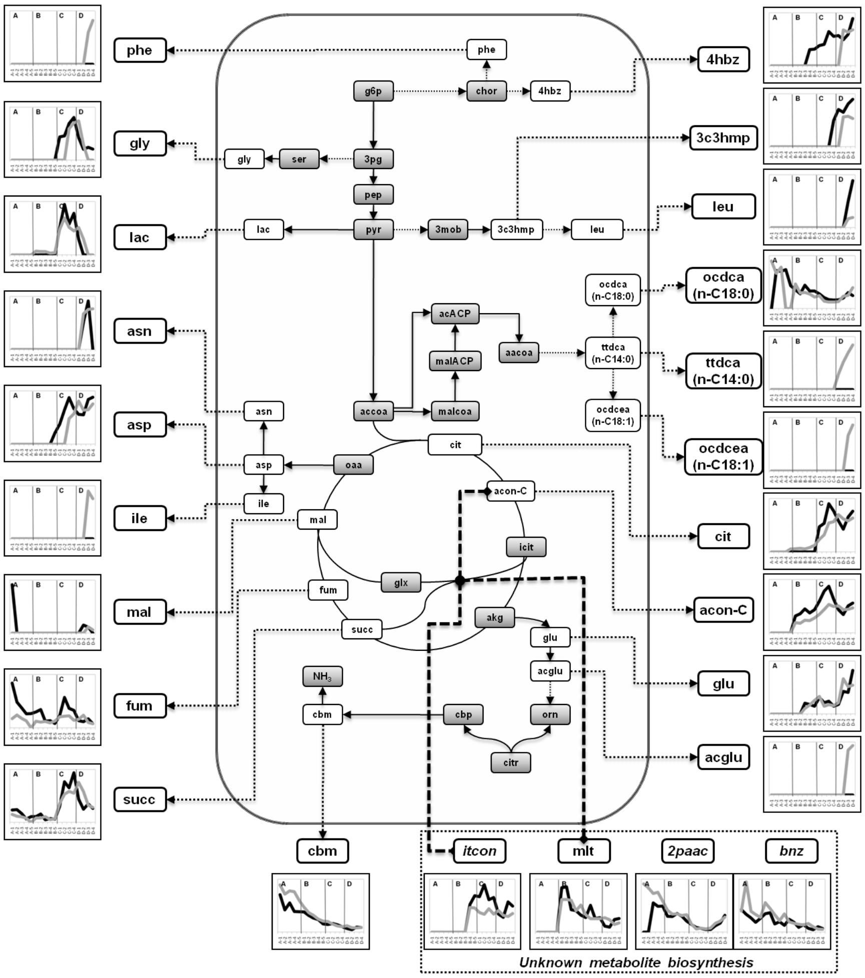
Specific growth rates (μ), AcGFP1 production (q_{AcGFP1}), glucose uptake (q_{Gluc}) and acetate formation (q_{Ac}).

[a]These parameters were calculated for biomass growth under acetate consumption.

6.4.2 METABOLIC FOOTPRINT ANALYSIS

Gas chromatography–mass spectrometry (GC-MS) has been widely used in the analysis of a large number of compounds such as amino acids, sugars, and organic acids. In this work, a GC-MS-based analytical plataform established by Villas-Bôas' group (Smart KF *et al.*, 2010) was used to detect amino and nonamino organic acids secreted by *E. coli* cells during fed-batch cultures. Samples harvested at different time points in the fermentation growth phases were analyzed and the relative concentrations for each detected metabolite were determined. Analytical data was further examined to address any changes in the metabolic footprints resulting from the alteration of culture conditions and to verify if the *rel*A gene mutation could actually influence the metabolic behaviour of this recombinant *E. coli* strain.

A wide variety of metabolites was detected in the extracellular medium, including fatty, amino and organic acids (see Figure 6.2) that ultimately resulted from the metabolic activities of cells at the various tested conditions. These metabolites were found to be mainly involved in the central carbon metabolism, including the tricarboxylic acid cycle (TCA cycle), the biosynthesis of amino and fatty acids as well as other energy generating metabolic reactions.



Abbv.	Metabolite names	Abbv.	Metabolite names	Abbv.	Metabolite names	Abbv.	Metabolite names
2paac	2-phenylglycine	asn	Asparagine	glx	Glyoxylate	NH_3	Ammonia
3c3hmp	2-Isopropylmalate	asp	Aspartate	gly	Glycine	oaa	Oxaloacetate
3mob	2-oxoisovalerate	bnz	Benzoate	icit	Isocitrate	ocdca	Octadecanoate
3pg	3-Phospho-D-glycerate	cbm	Carbamate	ile	Isoleucine	ocdcea	Octadecenoate
4hbz	4-hydroxybenzoate	cbp	Carbamoyl-phosphate	itcon	Itaconate	orn	Ornithine
aacoa	Acetoacetyl-CoA	chor	Chorismate	lac	Lactate	pep	Phosphoenolpyruvate
acACP	Acetyl-ACP	cit	Citrate	leu	Leucine	phe	Phenylalanine
accoa	Acetyl-CoA	citr	Citrulline	malACP	Malonyl-ACP	pyr	Pyruvate
acglu	Acetyl-L-glutamate	fum	Fumarate	malcoa	Malonyl-CoA	ser	Serine
acon-C	*cis*-Aconitate	g6p	D-Glucose 6-phosphate	mal	Malate	succ	Succinate
akg	2-Oxoglutarate	glu	Glutamate	mlt	Malonate	ttdca	Tetradecanoate

Figure 6.2. Schematic diagram of the *E. coli* metabolic map involving the metabolites secreted into the extracellular medium during recombinant fed-batch cultivations.

The accumulation (or assimilation) of metabolites was evaluated along four phases: the pre-IPTG induction phase (A), the post-IPTG induction phase (B), the post-nutrient upshift phase (C) and the glucose limitation phase (D). Graphs represent the relative concentration levels of each metabolite in the metabolic footprint along the fermentation process for the W3110 strain (dark line) and the Δ*rel*A mutant (light gray line) cultures. Bold dashed lines connecting metabolites and biochemical reactions indicate known inhibitory effects on those reaction-associated enzymes. Grey boxes represented in the metabolic map indicate other metabolites that participate in the metabolism, but were not detected in these experiments. The dashed square lists the detected metabolites with unknown biosynthetic reactions in *E. coli*.

However, we detected metabolites such as itaconate, malonate, 2-phenylglycine and benzoate that could not be linked directly to any known metabolic pathway of *E. coli,* according to public databases, such as KEGG (Kanehisa M and Goto S, 2000) and EcoCyc (Keseler IM *et al.*, 2009), and genome-scale metabolic models, like *i*AF1260 model of *E. coli* K12 (Feist AM *et al.*, 2007). However, these metabolites are known to participate in metabolic reactions of other organisms (Kim YS, 2002; Willke T and Vorlop KD, 2001). Therefore, resemblances between enzyme-coding genes from other organisms that produce these metabolites with the *E. coli* genome were investigated, but it was not possible to establish with confidence their participation as reactants/products in *E. coli* reactions. Yet, itaconate and malonate have been described as *in vitro* inhibitors of a key enzyme of the *E. coli* anaplerotic metabolism, i.e.; isocitrate lyase, the first enzyme of the glyoxylate cycle (Hoyt JC *et al.*, 1988), indicating that at least these metabolites are likely to be produced during *E coli* growth.

Principal components analysis (PCA) was performed to investigate whether the samples from different fermentation phases could be distinguished based on their metabolic footprints profiles and to determine the significant metabolic differences between the W3110 and Δ*rel*A mutant *E. coli* strains (Figure 6.3).

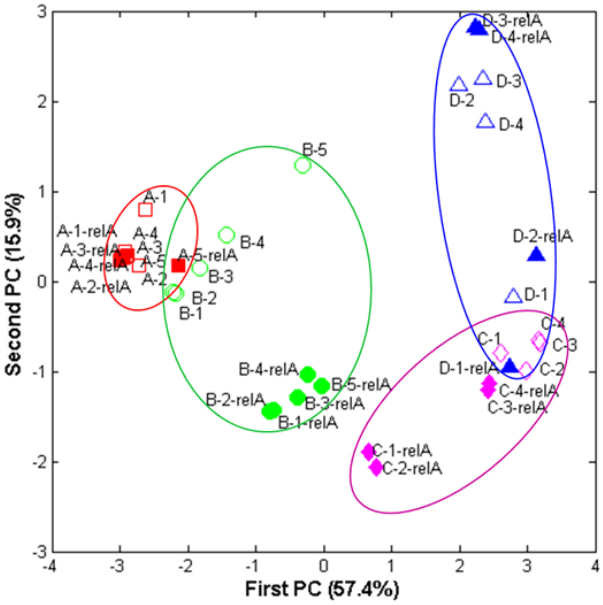


Figure 6.3. Principal component analysis (PCA) 2-dimensional projection of samples from fed-batch cultures of *E. coli* grown at different conditions based on mass fragment profiles of extracellular metabolites analysed by GC-MS.

Each sample is represented by a letter that designates the cultivation phase: samples withdrawn during the pre-induction phase [A] represented by *squares*, the post-induction phase [B] by *circles*, after growth upshift [C] by *diamonds* and after nutrient downshift [D] by *triangles*; and a number that indicates the sampling sequence in each cultivation phase. Samples from the W3110 culture are represented by opened symbols, whereas the Δ*rel*A mutant culture samples are depicted by full symbols.

As represented in Figure 6.2, the relative accumulation of metabolites varies during the fed-batch experiments as the culturing conditions are changed. Some of these variations were different in the two recombinant cultures, indicating that the W3110 and Δ*rel*A mutant *E. coli* strains might respond differently to those experimental conditions. Accordingly, metabolites that presented similar profiles are expected to participate in metabolic activities that were stimulated in both *E. coli* cultures. Table 6.2 shows the Pearson's correlation coefficients (*r*) with *p*-values using a Student's *t* distribution for testing the hypothesis of no correlation against the alternative, which considers the existence of significant correlations between the metabolite profiles produced by the W3110 and Δ*rel*A mutant strains at similar experimental conditions.

Table 6.2. Correlations between metabolite profiles of two *E. coli* strains, W3110 and Δ*rel*A mutant, during fed-batch fermentation phases: (A) pre-induction phase; (B) IPTG post-induction phase; (C) growth upshift phase; and (D) nutrient downshift phase.

Pearson correlation coefficients (*r*) are given with their *p*-values. Underlined *p*-values indicate metabolite profiles that are significantly correlated with a 95% confidence level. See metabolite abbreviations in Figure 6.2.

	A		B		C		D	
	r	*p*-value	*r*	*p*-value	*r*	*p*-value	*r*	*p*-value
2paac	-0.805	0.100	0.182	0.770	0.938	0.062	0.969	0.031
3c3hmp					(b)		0.338	0.662
4hbz			(b)		(b)		0.280	0.720
acglu							(a)	
acon-C			0.810	0.097	0.677	0.323	0.190	0.810
asn							0.515	0.485
asp			(b)		0.678	0.322	-0.086	0.914
bnz	0.349	0.565	0.748	0.811	0.860	0.140	-0.353	0.647
cbm	0.839	0.075	0.972	0.006	0.576	0.424	-0.236	0.764
cit	(a)		(a)		0.941	0.059	-0.046	0.954
fum	0.484	0.409	-0.089	0.887	0.301	0.699	0.063	0.937
glu			0.766	0.131	0.845	0.155	0.576	0.424
gly					0.989	0.011	0.864	0.424
ile							(a)	
itcon			0.983	0.003	-0.607	0.393	-0.418	0.582
lac			(a)		0.696	0.304	0.829	0.171
leu							0.954	0.046
mal	(b)						0.120	0.880
mlt			0.656	0.229	-0.290	0.710	-0.876	0.124
ocdca	-0.342	0.573	0.161	0.795	0.958	0.042	0.539	0.461
ocdcea							(a)	
phe							(a)	
succ	0.710	0.179	-0.035	0.956	0.834	0.166	0.551	0.449
ttdca							(a)	

[a] Undetected in the W3110 culture.

[b] Undetected in the Δ*rel*A mutant culture.

PCA analysis (Figure 6.3), on the other hand, shows that the metabolic footprints from the pre-induction phase (phase A) are identical for both cultures, but during the production of heterologous protein, especially during fermentation phases B to D, these strains behaved very differently. When looking at the Pearson correlation coefficients it is evident that during fermentation phases B to D the amount of metabolites that did not present significant correlated profiles is large. Only a few metabolites were found to be significantly correlated to each other. For example, in the last phase only 2-phenylglycine (2paac) and leucine (leu) profiles were significantly correlated with 95% of confidence. The other seventeen metabolites detected in both cultures, at these conditions, presented profiles that were not considered significantly correlated. In fact, some of these metabolites presented negative correlations, evidencing opposing tendencies regarding their accumulation in the extracellular medium. Moreover, some metabolites were only detected in one of the fed-batch cultures, which indicates that *E. coli* strains were at different metabolic states. For example, the accumulation of isoleucine (ile), phenylalanine (phe), acetyl-L-glutamate (acglu), octadecenoate (ocdcea) and tetradecanoate (ttdca) was only detected in the fed-batch culture with the Δ*rel*A mutant strain during the last fermentation phase.

Although the distinct metabolic states between the Δ*rel*A mutant and W3110 strain may result from changes associated with the genetic perturbation (i.e. *rel*A gene mutation), we cannot ignore the intrinsic metabolic variability that is common to most metabolic systems. As reported previously (Steuer R *et al.*, 2003), intrinsic metabolic fluctuations may arise because organisms are never in the same exact metabolic stage, even when growing in the same conditions, and small differences in enzyme concentrations may also affect metabolite concentrations. Thus, to better understand these differences, the metabolic footprints from each culture were further examined by clustering analysis (Figures 6.4 and 6.5).

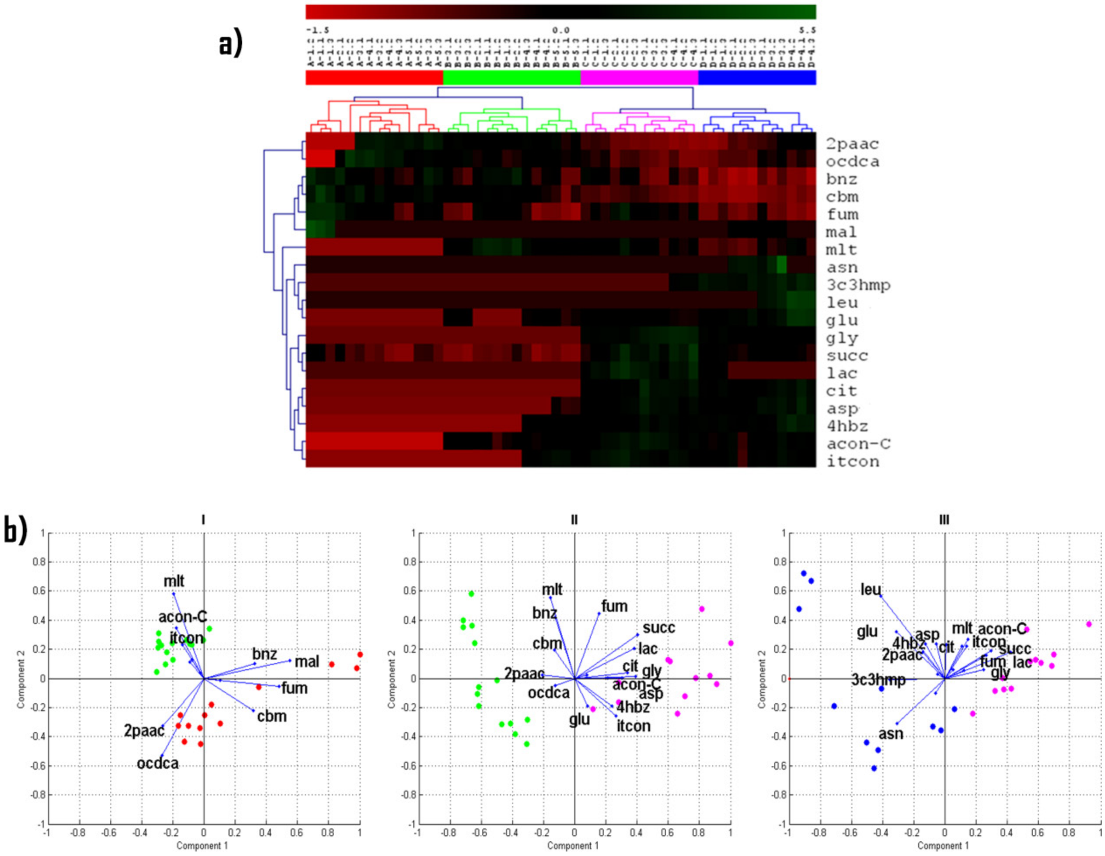


Figure 6.4. Analysis of the metabolic footprint profiles obtained from fed-batch aerobic culture of *E. coli* W3110.

(a) Hierarchical clustering (HCL) distinguished four data classes (vertical clusters), corresponding to extracellular samples taken during each fermentation phase, and several metabolite clusters (horizontal clusters) that characterized the metabolic footprint profiles during the recombinant fed-batch cultivation. (b) Principal component analysis (PCA) was performed to determine the most significant metabolic changes in the extracellular medium when growth conditions were changed: (I) induction of the heterologous protein expression; (II) nutritional upshift; and (III) glucose downshift. Principal component coefficients (represented by metabolite vectors) depicted metabolites that contributed the most to discriminate between sample clusters defined by the vertical clusters in the HCL analysis.

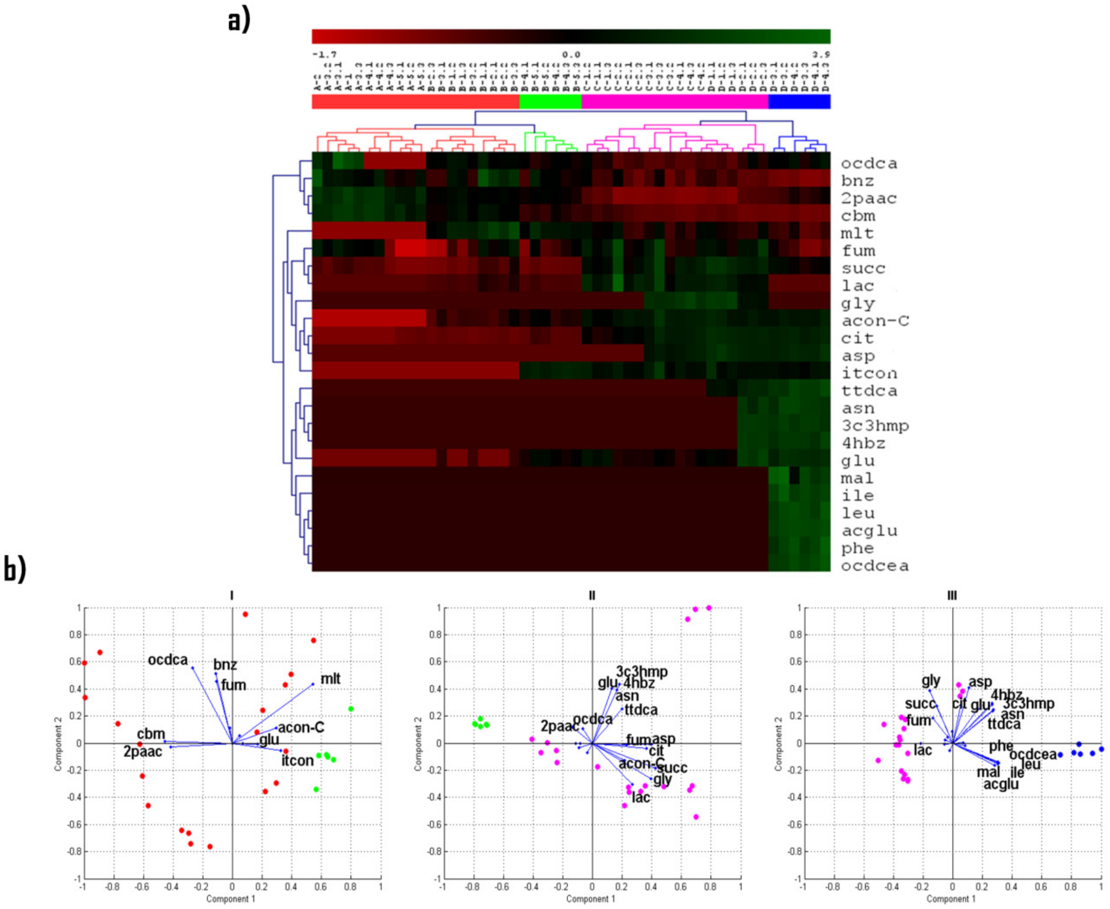


Figure 6.5. Analysis of the metabolic footprint profiles obtained from fed-batch aerobic culture of *E.coli* Δ*rel*A mutant strain.

(a) Hierarchical clustering (HCL) distinguished four sample clusters (vertical clusters) and several metabolite clusters (horizontal clusters). Here, sample clusters did not correspond to each fermentation phases, but illustrate data classes that are characterized by different metabolic states defined by the metabolite footprints detected during the recombinant fed-batch cultivation. (b) Principal component analysis (PCA) was performed to determine the most significant metabolic changes in the extracellular samples classified in each sample cluster. Principal component coefficients (represented by metabolite vectors) depicted metabolites that contributed the most to discriminate between sample clusters.

Hierarchical clustering (HCL) was used to evaluate the metabolic footprints produced by cells during the fed-batch fermentations. As shown in Figure 6.4a, for W3310, samples from the same fermentation phase were clustered together, which indicates that those samples have similar metabolite profiles. Metabolite clusters evidenced the association between metabolite patterns generated along the fermentation process. The starting hypothesis is that metabolites that show a similar variation are related, a relation that conveys information about their proximity or function within the metabolic map. For example, the metabolite profiles of citrate (cit), aspartate (asp), 4-hydroxybenzoate (4hbz), *cis*-aconitate (acon-C) and itaconate (itcon) were custered at these experimental conditions, meaning that in the W3110 *E. coli* culture these metabolites follow a same pattern.

Similarly, samples from the recombinant Δ*rel*A mutant grouped into four major clusters (Figure 6.5a). However, in this particular case, samples were clustered differently. Samples from phase A clustered together, but samples taken immediately after IPTG induction (triplicate samples from B-1 to B-3) clustered together with samples from phase A. The following samples from phase B (i.e. samples B-4 and B-5) clustered separatedly. This indicates that changes in the metabolic footprint profiles were not immediate after the IPTG induction, as observed in the W3110 culture. Only in the late stage of this fermentation phase that changes in metabolite profile were detectable. Samples taken immediately after nutrient downshift (triplicate samples D-1 and D-2) clustered with samples from phase C, which indicates that these samples have metabolic profiles more similar to samples from phase C than from to those from phase D. This can be due to the fact that, at this sampling time, cells were still consuming the excess of glucose accumulated during phase C (see Figure 6.1). The other samples from phase D clustered separatedly. In this fed-batch culture, metabolites were not clustered in the same groups as observed in the W3110 *E. coli* culture, which reflect the existence of important differences in the metabolic footprints produced by the Δ*rel*A mutant cells. For example, the 4-hydroxybenzoate (4hbz) is, at these conditions, clustered with asparagine (asn), 2-isopropylmalate (2paac) and tetradecanoate (ttdca).

Besides the observed differences between the metabolic footprints produced at particular growth conditions, differences between the set of metabolic variables that better characterized each phase (i.e. principal component coefficients represented by vectors) were also explored. Thus, in the next subsections details are exposed on the key metabolites involved in the alteration of the metabolic footprints when cells were: (I) induced to express the heterologous protein; (II) submitted to nutritional upshift; and (III) submitted to glucose-limited conditions.

I - Impact of the IPTG induction on the metabolic footprint. Figure 6.4b shows that the metabolites with the highest positive coefficients in PC2 (malonate (mlt), itaconate (itcon) and *cis*-aconitate (acon-C)) had their levels increased after IPTG induction in the *E. coli* W3110 culture. Similarly, these metabolites also presented high positive coefficients in PC1of samples from Δ*rel*A mutant culture (Figure 6.5b). In addition, itaconate and *cis*-aconitate were clustered using a Pearson correlation metrics, as represented in the HCL diagrams (Figure 6.4a and Figure 6.5a) presenting a strong association between their levels along both fed-batch fermentations.

II - The impact of the nutrient upshift on the metabolic footprint. Figure 6.4b shows that glycine (gly), succinate (succ), lactate (lac), citrate (cit), aspartate (asp), 4-hydroxybenzoate (4hbz), *cis*-aconitate (acon-C) and itaconate (itcon) are the most significant variables that explain the differences projected in PC1 (positive coefficients), which are the differences between growth phases B and C of the *E. coli* W3110 culture. In fact, as illustrated in Figure 6.4a they were grouped in the same metabolite cluster and represent those that were highly accumulated after the nutrient upshift. Similarly, positive coefficients for PC1 and negative for PC2 of the Δ*rel*A mutant culture (Figure 6.5b) exposed these metabolites as the most relevant variables characterizing the samples after nutrient upshift, except for the hydroxybenzote (4hbz) and itaconate (itcon).

III – The impact of the glucose limitation on the metabolic footprint. In the W3110 *E. coli* culture (Figure 6.4b), negative coefficients projected in the PC1, corresponding to leucine (leu), asparagine (asn), 2-isopropylmalate (3c3hmp) and glutamate (glu), showed the largest differences after nutrient downshift. These metabolites were also clustered by HCL (Figure 6.4a). The principal coefficients of metabolic footprints from the Δ*rel*A mutant culture (Figure 6.5b) depicted glycine (gly), succinate (succ) and lactate (lac) as adjacent vectors corresponding to metabolites that were immediately assimilated after glucose limitation.

6.5 DISCUSSION

The metabolic footprint analysis is supported on the basis that cells can secrete metabolites to the extracellular medium during growth and/or in response to environmental changes (Arana I *et al.*, 2004; Lin HY *et al.*, 2004; Srinivasan S *et al.*, 1998). Furthermore, cells may activate a variety of efflux transporters that work like metabolic relief valves or defensive support to survive an antagonistic environment (Van Dyk TK *et al.*, 2004). In the first case, an increase in extracellular metabolites would be associated with an increase in the intracellular concentration of those compounds, while in the second case there would be a gradient that has to be maintained by the cells to achieve specific purposes. In general, the variety and level of the secreted metabolites reflect the metabolic state of the cell and, therefore, may be considered the closest indicator of the phenotype (Fiehn O, 2002; Kell DB *et al.*, 2005; Raamsdonk LM *et al.*, 2001; Villas-Boas SG *et al.*, 2005). Considering this, and the fact that the metabolic impact of recombinant protein expression in the host cells is still not well-understood, the metabolic footprints of the recombinant W3110 and Δ*rel*A mutant *E. coli* cells grown at different experimental conditions were analysed.

The metabolic responses of *E. coli* cells were evaluated by measuring some physiological parameters, such as the cellular growth and acetate formation (Figure 6.1 and Table 6.1). Results corroborated previous works (Bentley WE *et al.*, 1990; Bonomo J and Gill RT, 2005; Harcum SW, 2002), indicating that the decrease on the cellular growth is the major consequence of the metabolic burden on the recombinant host cells due to higher demands of energy and amino acids. The drainage of energy and biosynthetic precursors associated to the expression of foreign proteins imposes severe changes in the metabolic activity of cells and, as a result, reduces the cellular growth. As shown, upon IPTG induction, the specific growth rate of the W3110 and Δ*rel*A mutant *E. coli* strains decreased 50% and 32%, respectively. The Δ*rel*A mutant strain seems to be less affected, which can be explained by the failure to stimulate the RelA-dependent stringent response, a stress response that has been proposed (Andersson L *et al.*, 1996; Haddadin FT and Harcum SW, 2005; Harcum SW and Bentley WE, 1999) to occur when there is a lack of intracellular amino acids associated with the additional requirements for the production of recombinant products. Ultimately, this stress response may induce a decrease of cellular growth and protein production. When increasing the nutrient availability, cells can generate sufficient metabolic and energetic resources for the formation of the heterologous protein, as well as for growth-associated processes, and as a result no physiological differences between cultures were observed. However, it was expected that at nutrient deprived conditions (phase D), the physiological responses of the two strains would be, at some extent, distinct. Besides the substrate uptake rates, the estimated physiological parameters did not show significant differences between the W3110 and Δ*rel*A mutant *E. coli* cultures. Cellular growth was rapidly arrested and formation of the heterologous protein decreased to similar levels, which did not allow to deduce any fundamental alterations in the cellular metabolism caused by the single gene mutation. Nevertheless, the analysis of extracellular metabolites was invaluable to determine the main consequences on the physiology of *E. coli* cells derived from the *rel*A mutation and the experimental conditions.

As previously reported (Arana I *et al.*, 2004; Lin HY *et al.*, 2004), *E. coli* cells secrete metabolites according to the adjustments needed in the cellular metabolism to cope with different physiological demands. A small, but considerable number of metabolites characterized the metabolic footprints produced during the fed-batch processes and their level fluctuations were inspected to estimate their relationship with the intracellular metabolic changes. Metabolic footprints allowed not only to discriminate between samples withdrawn at different fermentation phases, but also to disclose the main metabolic changes that were not evidenced by the physiological characterization of the fed-batch cultures. As illustrated in Figure 6.4a, samples from the W3110 *E. coli* fed-batch culture were discriminated according to the fermentation phases, which defined clusters with characteristic metabolic properties. In contrast, samples from the Δ*rel*A mutant *E. coli* culture (Figure 6.5a) were not equally clustered. When cells were induced to express the heterologous protein (I) or submitted to glucose-limiting conditions (III), the metabolic footprints observed before and immediately after these experimental shifts were equivalent. Samples taken immediately after IPTG induction (i.e. samples from B-1 to B-3) were clustered with samples from phase A and samples withdrawn after ceasing the glucose feeding (i.e. samples D-1 and D-2) were clustered with samples from phase C, indicating that the metabolic changes were more significant at the late stage of this fermentation phase. Nevertheless, it is noteworthy that, when cells were submitted to nutritional upshift, there were clear metabolic changes in the footprints that established a separate cluster (i.e. samples from phase C were clearly isolated from the previous phase).

Metabolites were clustered according to their relative concentration profiles measured along the fermentation phases and revealed the existence of some correlations between metabolites. Although most metabolites presented unrelated profiles, some were consistently grouped in both experiments. For example, glycine (gly), *cis*-aconitate (acon-C), citrate (cit), aspartate (asp) and itaconate (itcon) were clustered together in both HCL analyses, which indicates that these metabolites can be, at some degree, interrelated. Indeed, citrate and *cis*-aconitate are neighbours on the metabolic network (i.e. *cis*-aconitate is an intermediate in the enzymatic isomerization of citrate to isocitrate in the TCA cycle) and itaconate and *cis*-aconitate are both enzymatic inhibitors of the first enzymatic reaction of the glyoxylate shunt (i.e. isocitrate lyase). Nevertheless, it should be emphasized that metabolite correlations that were consistent between the W3110 and Δ*rel*A mutant experiments, should be examined with care. Metabolite correlations that arise from changes that influence a large number of cellular functions (e.g. RelA-dependent responses) are the hardest to interpret in terms of the underlying biochemical network. Therefore, to further understand these observed “associations” between the metabolite levels and to investigate their changes according to the environmental conditions (e.g. IPTG induction), metabolic footprints from each culture were independently analysed.

Regarding IPTG induction of heterologous protein expression in the W3110 *E. coli* fed-batch culture, a set of metabolites were depicted as key elements characterizing the metabolic changes associated with this experimental transition. Three carboxylic acids (malonate, itaconate and *cis*-aconitate) were secreted into the culture broth (Figure 6.4b) suggesting that they were released from the cells to prevent any inhibitory effects on the activity of the isocitrate lyase. As mentioned before, *cis*-aconitate and itaconate, as well as malonate (Hoyt JC *et al.*, 1988), are inhibitors of this enzyme, controlling the activity of the first reaction of the glyoxylate shunt that was reported (Wittmann C *et al.*, 2007) as serving an anaplerotic function in the cell during heterologous protein production. Since several TCA intermediates are withdrawn from the TCA cycle as amino acid precursors and need to be replenish, these anaplerotic reactions are central to balance the intracellular levels of TCA metabolites, fulfilling this way the additional biosynthetic requirements associated with the formation of the heterologous protein. This could be a plausible assumption, given that induced recombinant cells undergo severe metabolic burden counteracted by the activation of anaplerotic reactions, and the simultaneous secretion of the abovementioned inhibitors may indicate a key point at which the metabolic regulation has changed.

The Δ*rel*A mutant strain response to the IPTG induction was also manifested by the accumulation of the same enzymatic inhibitors: malonate, *cis*-aconitate and itaconate (see Figure 6.5b). Yet, HCL analysis showed that the metabolic footprints produced immediately after IPTG induction did not discriminate these samples from the previous fermentation phase. Moreover, the metabolic footprints generated after IPTG induction were distinct when compared with the W3110 culture and those differences were sufficient to distinguish samples from the two strains (Figure 6.3). Also, most metabolites did not present significant correlated patterns (i.e. *p*-values below 0.05) between the two cultures during phase B and others showed to have a negative correlation, such as succinate and fumarate (Table 6.2). Although some of the changes on the metabolite levels could be associated with the intrinsic variability of this kind of experiments, the overall metabolic patterns and the accumulation of certain metabolites only detected in the Δ*rel*A mutant culture (e.g. acetyl-L-glutamate (acglu)) may result from changes on the cellular metabolism that were not equivalent in the two *E. coli* cultures.

Despite the extensive knowledge on basic aspects such as the changes of growth rates with nutrient concentrations (Ferenci T, 1999; Hua Q *et al.*, 2004; Lendenmann U and Egli T, 1998; Marr AG, 1991; Tweeddale H *et al.*, 1998), information on the effects of the nutritional upshift during recombinant processes on the metabolic footprint is still scarce. In general, studies are focused in the secretion of acetate, not only because it retards growth and inhibits protein formation, but also because it represents a deviation of carbon that might otherwise be used to generate energy and precursors for biosynthetic purposes (Eiteman MA and Altman E, 2006; Suarez DC and Kilikian BV, 2000; Van de Walle M and Shiloach J, 1998). In this study, a glucose feeding upshift was applied to increase the specific growth rate during heterologous protein production and the metabolic footprints were analysed. Numerous metabolites were immediately accumulated in the extracellular medium of the W3110 culture after glucose availability increased, such as glycine (gly), 4-hydroxibenzoate (4hbz), lactate (lac), citrate (cit), *cis*-aconitate (acon-C), itaconate (itcon), succinate (succ), fumarate (fum) and aspartate (asp) (Figure 6.4b). The increasing metabolic activity associated with a high rate of glucose consumption resulted in the accumulation of most intracellular metabolites, including TCA intermediaries, amino acids and amino acid precursors. Moreover, acetate was also accumulated during this fermentation phase resulted from the glucose overflow metabolism. It was reported (Majewski RA and Domach MM, 1990) that in the presence of excess glucose, the carbon flux through glycolysis exceeds the capacity of the TCA cycle and acetate is accumulated. The same set of metabolites, except the 4-hydroxibenzoate and itaconate, were also accumulated during the Δ*rel*A fermentation process at these conditions, which indicate that metabolic adjustments induced by the nutrient upshift resulted in similar alterations in the metabolic footprints from both strains. Although some metabolite patterns revealed significant differences (Figure 6.3), in general, these were less significant than in any other phase. The accumulation of lactate at these conditions was found intriguing. Since the conversion of pyruvate to lactate in *E. coli* is usually exclusively induced at anaerobic conditions (Clark DP, 1989; Tarmy EM and Kaplan NO, 1968), the presence of this by-product implies that the internal accumulation of pyruvate due to the metabolism overflow overrides any other mechanism known to control the activity of the lactate dehydrogenase (LdhA) enzyme under aerobic conditions (Jiang GR *et al.*, 2001) or, for some reason, local oxygen deficiencies during the *E. coli* fed-batch process have triggered the ArcAB system and other genes involved in the mixed acid fermentation pathway (Xu B *et al.*, 1999). The latest assumption seems improbable, as good mixing conditions and dissolved oxygen values above 30% were maintained inside the reactor.

Finally, the metabolic responses to nutritional stress associated with the restriction of glucose feeding during recombinant *E. coli* processes were also evaluated. It is remarkable the amount of metabolites found to be secreted after glucose downshift. Besides the decreasing levels of metabolites that can serve as carbon sources for *E. coli* (e.g. acetate), the accumulation of unexpected metabolites, like amino acids (e.g. leucine (leu), asparagine (asn), L-glutamate (glu) and aspartate (asp)) and amino acid derivatives (e.g. 2-isopropylmalate (3c3hmp)), indicate that the cells accumulated these biosynthetic precursors as a consequence of the reduced activity of the protein translation machinery, as estimated by the experimental AcGFP1 formation rates (Table 6.1). It is acknowledged that under nutrient starvation *E. coli* cells entail complex protective processes that ultimately manage the cellular metabolism to sustain cellular maintenance and viability (Matin A, 1991; Neubauer P *et al.*, 1995; Nystrom T, 1999). Apparently the RelA-dependent response is involved in these metabolic readjustments, as the metabolic footprints exhibited by the wild-type and Δ*rel*A mutant *E. coli* cultures were divergent (Figure 6.3). Major differences were found in the citrate (cit), fumarate (fum), malate (mal) and aspartate (asp) profiles, which indicate that the fine-tuning of the TCA fluxes after glucose depletion and subsequent acetate assimilation were differently coordinated. Furthermore, when comparing the metabolite patterns of both strains after nutrient downshift, we found that acetate utilization by *E. coli* cells has a strong effect on the accumulation of several metabolites, especially in the Δ*rel*A mutant strain. Metabolites such as phenylalanine (phe), isoleucine (ile), leucine (leu), acetyl-L-glutamate (acglu) and octadecenoate (ocdcea) were accumulated in the medium of the Δ*rel*A mutant culture shortly after acetate consumption has started (data shown in Appendix A). Under these conditions, acetate is converted to acetyl coenzyme A (accoA) at the expense of ATP which, in turn, is mainly catabolised via glyoxylate cycle that serves as anaplerotic reactions (Brown TD *et al.*, 1977). This shift in the utilization of carbon sources involves the activation of various cellular processes, including the synthesis of new catabolic enzymes and the activation of substrate-specific transport systems. It seems that the metabolic imbalance caused by these metabolic activities coupled with the additional formation of heterologous protein is the basis for the accumulation of several metabolites, including amino acids, that was more critical in the Δ*rel*A strain. During glucose starvation the translational apparatus, as well as cellular growth, are limited via transcriptional control of several growth-associated genes, like ribosomal operons. The ppGpp-stringent control has been related to this cellular response to nutritional deprivation that redirect the RNA polymerase transcriptional activity from stable RNA (ribosomal and transfer RNA) synthesis to stress-related genes, in particular genes that have protective functions. In the absence of the RelA activity, the ppGpp accumulation is limited and the translational apparatus stays unaffected. This seems to be the main cause for the only slight reduction of heterologous protein production rate in the Δ*rel*A mutant. However, while protein production seems to be unaffected by nutrient starvation, bacterial growth was arrested in this strain, probably by action of responses mediated by other stress proteins. These facts apparently generated an imbalance in the mutant strain's metabolism (probably due to differences in amino acid composition between heterologous and average *E. coli* proteins) that was evidenced by the accumulation of some amino and fatty acids. For example, the synthesis of fatty acids is known to be inhibited during glucose starvation, which did not seem to happen in the mutant strain. However, since these compounds had not been used for biomass production, there was an accumulation of octadecenoate (ocdcea) and tetradecanoate (ttdca) in the Δ*rel*A mutant strain. Moreover, isoleucine (ile), acetyl-L-glutamate (acglu) and phenylalanine (phe) were also accumulated at these conditions by the Δ*rel*A mutant strain. It seems evident that the failure to accumulate ppGpp at these conditions allowed a continuous production of heterologous protein (although slightly reduced) resulting in an unbalanced drainage of precursors. Some metabolites were gradually replenished while others (less required for the production of AcGFP1) were over-accumulated. The fact that most metabolic resources were probably redirected to the formation of AcGFP1, may explain the higher protein synthesis rate and lower biomass yield observed for the Δ*rel*A mutant strain when compared to the W3110 strain. It is clear that the *rel*A mutation influences the natural cellular responses to nutritional downshifts, which can delay, or even suppress, *E. coli* survival and resistance.

6.6 CONCLUSIONS

Although the characterization of metabolites in *E. coli* culture broths has been performed using various detection methods, such analyses are mostly confined to specific metabolites and have not been done in a global scale. For example, the secretion of acetate during aerobic *E. coli* fermentations is regularly measured, because it is considered a major obstacle to enhanced heterologous protein production. However, the present work reveals that the complexity of the generated metabolic footprints at different culture conditions is much higher than what has been admitted. As demonstrated, *E. coli* secretes a vast array of metabolites that participate in a wide range of metabolic pathways. Although the metabolism in *E. coli* has been studied more intensively than in any other bacterium, only recently it has become clear that targeted studies do not provide an accurate picture of the cellular metabolism. A typical metabolomic approach is expected to generate new knowledge from the comprehensive analysis of the metabolome and the distinctive metabolic patterns produced at different environmental and genetic conditions.

Metabolic footprints resulted from the IPTG-induction of heterologous protein expression have shown that there is a rapid accumulation of unexpected metabolites in the culture broth. The secretion of the isocitrate lyase inhibitors suggest that the anaplerotic glyoxylate shunt was activated to replenish the TCA intermediaries engaged in the additional formation of heterologous protein. Moreover, the detection of compounds unknown to participate in *E. coli* metabolism reinforces the importance of unbiased analytical approaches in research.

When cells are exposed to conditions of nutrient-excess, the uncoupling of maximum glucose uptake rates and the TCA fluxes results in a metabolic overflow with consequent accumulation of metabolites. Not only acetate and lactate were secreted at these conditions, but also TCA intermediaries (e.g. fumarate, succinate, citrate and *cis*-aconitate) and some amino acids (e.g. aspartate and glycine). Assuming that the demand for metabolic resources for cellular growth and heterologous protein production is exceeded, the secretion of these metabolites can be understood as a metabolic relief required to avoid adverse effects from the imbalanced cellular metabolism (Kell DB *et al.*, 2005).

At nutrient-limited conditions (i.e. when glucose feeding was discontinued), some of these metabolites (e.g. acetate) were assimilated by the cells as carbon and energy sources. These metabolic activities coupled with the additional formation of heterologous protein may have resulted into severe rearrangements in the cellular metabolism that led to the secretion of amino acids like: phenylalanine, asparagine, isoleucine and leucine and the acetyl-L-glutamate. Once again, these metabolic imbalances were more pronounced in the Δ*rel*A strain, which fails to trigger RelA-dependent processes to respond to nutrient deprivation.

The metabolic flexibility exposed by the alterations in the metabolic footprints, evidenced that cells entail diverse cellular mechanisms to sense and rapidly counteract the adverse environmental conditions. This stringent behaviour was prevalent in the W3110 strain, while the Δ*rel*A strain showed some difficulties to cope immediately with the metabolic imbalances caused by the formation of heterologous protein. However, at nutrient-excess conditions, when none of cellular processes dependent on the activity of RelA are triggered, the metabolic behaviour of both strains was not significantly divergent. It is evident that, although some disadvantages might have been indicated concerning the metabolic behaviour of the Δ*rel*A strain (e.g. failure to manage metabolic imbalances), the enhanced production rate of heterologous protein represents a major benefit.

Metabolic footprinting, more than most other analytical strategies, is a rapid and non-invasive analysis, representing a powerful approach for the characterization of phenotypes and the distinction of specific metabolic states due to environmental or genetic alterations. Nevertheless, metabolic footprints are just a shallow representation of the metabolic state of cells and the full understanding of the underlying mechanisms controlling these metabolic imbalances caused by the heterologous protein production, require further inspection of key metabolites (e.g. metabolites that are important nodes in the metabolic network) or the combination with other experimental strategies (e.g. gene expression and proteomics).

6.7 REFERENCES

1. Aldor IS *et al* (2005) Proteomic profiling of recombinant *Escherichia coli* in high-cell-density fermentations for improved production of an antibody fragment biopharmaceutical. *Applied and Environmental Microbiology* 71 (4):1717-1728.

2. Allen J *et al* (2003) High-throughput classification of yeast mutants for functional genomics using metabolic footprinting. *Nature Biotechnology* 21 (6):692-696.

3. Allen J *et al* (2004) Discrimination of modes of action of antifungal substances by use of metabolic footprinting. *Applied and Environmental Microbiology* 70 (10):6157-6165.

4. Andersson L *et al* (1996) Impact of plasmid presence and induction on cellular responses in fed batch cultures of *Escherichia coli*. *Journal of Biotechnology* 46 (3):255-263.

5. Arana I *et al* (2004) Relationships between *Escherichia coli* cells and the surrounding medium during survival processes. *Antonie Van Leeuwenhoek International Journal of General and Molecular Microbiology* 86 (2):189-199.

6. Artsimovitch I *et al* (2004) Structural basis for transcription regulation by alarmone ppGpp. *Cell* 117 (3):299-310.

7. Bentley WE *et al* (1990) Plasmid-encoded protein - The principal factor in the metabolic burden associated with recombinant bacteria. *Biotechnology and Bioengineering* 35 (7):668-681.

8. Bonomo J and Gill RT (2005) Amino acid content of recombinant proteins influences the metabolic burden response. *Biotechnology and Bioengineering* 90 (1):116-126.

9. Brown TD, Jones-Mortimer MC, and Kornberg HL (1977) The enzymic interconversion of acetate and acetyl-coenzyme A in *Escherichia coli*. *Journal of General Microbiology* 102 (2):327-336.

10. Chatterji D, Fujita N, and Ishihama A (1998) The mediator for stringent control, ppGpp, binds to the beta-subunit of *Escherichia coli* RNA polymerase. *Genes to Cells* 3 (5):279-287.

11. Chatterji D and Ojha AK (2001) Revisiting the stringent response, ppGpp and starvation signaling. *Current Opinion in Microbiology* 4 (2):160-165.

12. Chou CP (2007) Engineering cell physiology to enhance recombinant protein production in *Escherichia coli*. *Applied Microbiology and Biotechnology* 76 (3):521-532.

13. Clark DP (1989) The fermentation pathways of *Escherichia coli*. *Fems Microbiology Reviews* 63 (3):223-234.

14. Durfee T *et al* (2008) Transcription profiling of the stringent response in *Escherichia coli*. *Journal of Bacteriology* 190 (3):1084-1096.

15. Eiteman MA and Altman E (2006) Overcoming acetate in *Escherichia coli* recombinant protein fermentations. *Trends in Biotechnology* 24 (11):530-536.

16. Feist AM *et al* (2007) A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Molecular Systems Biology* 3:121.

17. Ferenci T (1999) Regulation by nutrient limitation. *Current Opinion in Microbiology* 2:208-213.

18. Fiehn O (2002) Metabolomics - the link between genotypes and phenotypes. *Plant Molecular Biology* 48 (1-2):155-171.

19. Glick BR (1995) Metabolic Load and Heterologous Gene-Expression. *Biotechnology Advances* 13 (2):247-261.

20. Gnoth S *et al* (2008) Control of cultivation processes for recombinant protein production: a review. *Bioprocess and Biosystems Engineering* 31 (1):21-39.

21. Haddadin FT and Harcum SW (2005) Transcriptome profiles for high-cell-density recombinant and wild-type *Escherichia coli*. *Biotechnology and Bioengineering* 90 (2):127-153.

22. Harcum SW (2002) Structured model to predict intracellular amino acid shortages during recombinant protein overexpression in *E. coli*. *Journal of Biotechnology* 93 (3):189-202.

23. Harcum SW and Bentley WE (1999) Heat-shock and stringent responses have overlapping protease activity in *Escherichia coli*. Implications for heterologous protein yield. *Applied Biochemistry and Biotechnology* 80 (1):23-37.

24. Hoyt JC *et al* (1988) *Escherichia coli* Isocitrate Lyase - Properties and Comparisons. *Biochimica et Biophysica Acta* 966 (1):30-35.

25. Hua Q *et al* (2004) Analysis of gene expression in *Escherichia coli* in response to changes of growth-limiting nutrient in chemostat cultures. *Applied and Environmental Microbiology* 70 (4):2354-2366.

26. Jain V, Kumar M, and Chatterji D (2006) ppGpp: Stringent response and survival. *Journal of Microbiology* 44 (1):1-10.

27. Jiang GR, Nikolova S, and Clark DP (2001) Regulation of the ldhA gene, encoding the fermentative lactate dehydrogenase of *Escherichia coli*. *Microbiology* 147 (Pt 9):2437-2446.

28. Jishage M *et al* (2002) Regulation of σ factor competition by the alarmone ppGpp. *Genes & Development* 16 (10):1260-1270.

29. Jurgen B *et al* (2000) Monitoring of genes that respond to overproduction of an insoluble recombinant protein in *Escherichia coli* glucose-limited fed-batch fermentations. *Biotechnology and Bioengineering* 70 (2):217-224.

30. Kanehisa M and Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* 28:27-30.

31. Kell DB *et al* (2003) Metabolic footprinting: a high-throughput, high-information approach to cellular characterisation and functional genomics. *Yeast* 20:S335.

32. Kell DB *et al* (2005) Metabolic footprinting and systems biology: The medium is the message. *Nature Reviews Microbiology* 3 (7):557-565.

33. Keseler IM *et al* (2009) EcoCyc: a comprehensive view of *Escherichia coli* biology. *Nucleic Acids Research* 37 (Database issue):D464-D470.

34. Kim YS (2002) Malonate metabolism: Biochemistry, molecular biology, physiology, and industrial application. *Journal of Biochemistry and Molecular Biology* 35 (5):443-451.

35. Lendenmann U and Egli T (1998) Kinetic models for the growth of *Escherichia coli* with mixtures of sugars under carbon-limited conditions. *Biotechnology and Bioengineering* 59 (1):99-107.

36. Lin HY *et al* (2004) Change of extracellular cAMP concentration is a sensitive reporter for bacterial fitness in high-cell-density cultures of *Escherichia coli*. *Biotechnology and Bioengineering* 87 (5):602-613.

37. Magnusson LU, Farewell A, and Nystrom T (2005) ppGpp: a global regulator in *Escherichia coli*. *Trends in Microbiology* 13 (5):236-242.

38. Majewski RA and Domach MM (1990) Simple constrained-optimization view of acetate overflow in *E. coli*. *Biotechnology and Bioengineering* 35:732-738.

39. Marr AG (1991) Growth-Rate of *Escherichia coli*. *Microbiological Reviews* 55 (2):316-333.

40. Matin A (1991) The molecular basis of carbon-starvation-induced general resistance in *Escherichia coli*. *Molecular Microbiology* 5 (1):3-10.

41. Neubauer P *et al* (1995) Response of guanosine tetraphosphate to glucose fluctuations in fed-batch cultivations of *Escherichia coli*. *Journal of Biotechnology* 43 (3):195-204.

42. Nystrom T (1999) Starvation, cessation of growth and bacterial aging. *Current Opinion in Microbiology* 2 (2):214-219.

43. Pope GA *et al* (2007) Metabolic footprinting as a tool for discriminating between brewing yeasts. *Yeast* 24 (8):667-679.

44. Raamsdonk LM *et al* (2001) A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nature Biotechnology* 19 (1):45-50.

45. Roberts JW (2009) Promoter-specific control of *E. coli* RNA polymerase by ppGpp and a general transcription factor. *Genes & Development* 23 (2):143-146.

46. Rocha I and Ferreira EC (2002) On-line simultaneous monitoring of glucose and acetate with FIA during high cell density fermentation of recombinant *E. coli*. *Analytica Chimica Acta* 462 (2):293-304.

47. Rocha I, Veloso ACA, Carneiro S, Costa R, Ferreira EC (2008) Implementation of a specific rate controller in a fed-batch *E. coli* fermentation.17th World Congress The International Federation of Automatic Control

48. Saeed AI *et al* (2003) TM4: A free, open-source system for microarray data management and analysis. *Biotechniques* 34 (2):374-378.

49. Seo JH, Kang DG, and Cha HJ (2003) Comparison of cellular stress levels and green-fluorescent-protein expression in several *Escherichia coli* strains. *Biotechnology and Applied Biochemistry* 37 (Pt 2):103-107.

50. Smart KF et al.Analytical platform for metabolome analysis of microbial cells using gas chromatography-mass spectrometry (GC-MS). Nature Protocols (in press)

51. Srinivasan S *et al* (1998) Extracellular signal molecule(s) involved in the carbon starvation response of marine Vibrio sp. strain S14. *J Bacteriol* 180 (2):201-209.

52. Srivatsan A and Wang JD (2008) Control of bacterial transcription, translation and replication by (p)ppGpp. *Current Opinion in Microbiology* 11 (2):100-105.

53. Stein SE (1999) An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *Journal of the American Society for Mass Spectrometry* 10 (8):770-781.

54. Steuer R *et al* (2003) Observing and interpreting correlations in metabolomic networks. *Bioinformatics* 19 (8):1019-1026.

55. Suarez DC and Kilikian BV (2000) Acetic acid accumulation in aerobic growth of recombinant *Escherichia coli*. *Process Biochemistry* 35 (9):1051-1055.

56. Tarmy EM and Kaplan NO (1968) Kinetics of *Escherichia coli* B D-lactate dehydrogenase and evidence for pyruvate-controlled change in conformation. *Journal of Biological Chemistry* 243 (10):2587-&.

57. Toulokhonov II, Shulgina I, and Hernandez VJ (2001) Binding of the transcription effector ppGpp to *Escherichia coli* RNA polymerase is allosteric, modular, and occurs near the N terminus of the beta '-subunit. *Journal of Biological Chemistry* 276 (2):1220-1225.

58. Tweeddale H, Notley-McRobb L, and Ferenci T (1998) Effect of slow growth on metabolism of *Escherichia coli*, as revealed by global metabolite pool ("metabolome") analysis. *J Bacteriol* 180 (19):5109-5116.

59. Van de Walle M and Shiloach J (1998) Proposed mechanism of acetate accumulation in two recombinant *Escherichia coli* strains during high density fermentation. *Biotechnology and Bioengineering* 57 (1):71-78.

60. Van Dyk TK *et al* (2004) Characterization of the *Escherichia coli* AaeAB efflux pump: A metabolic relief valve? *Journal of Bacteriology* 186 (21):7196-7204.

61. Villas-Boas SG *et al* (2003) Simultaneous analysis of amino and nonamino organic acids as methyl chloroformate derivatives using gas chromatography-mass spectrometry. *Analytical Biochemistry* 322 (1):134-138.

62. Villas-Boas SG *et al* (2008) Phenotypic characterization of transposon-inserted mutants of *Clostridium proteoclasticum* B316(T) using extracellular metabolomics. *Journal of Biotechnology* 134 (1-2):55-63.

63. Villas-Boas SG *et al* (2005) High-throughput metabolic state analysis: the missing link in integrated functional genomics of yeasts. *Biochemical Journal* 388:669-677.

64. Villas-Boas SG *et al* (2006) Extracellular metabolomics: A metabolic footprinting approach to assess fiber degradation in complex media. *Analytical Biochemistry* 349 (2):297-305.

65. Willke T and Vorlop KD (2001) Biotechnological production of itaconic acid. *Applied Microbiology and Biotechnology* 56 (3-4):289-295.

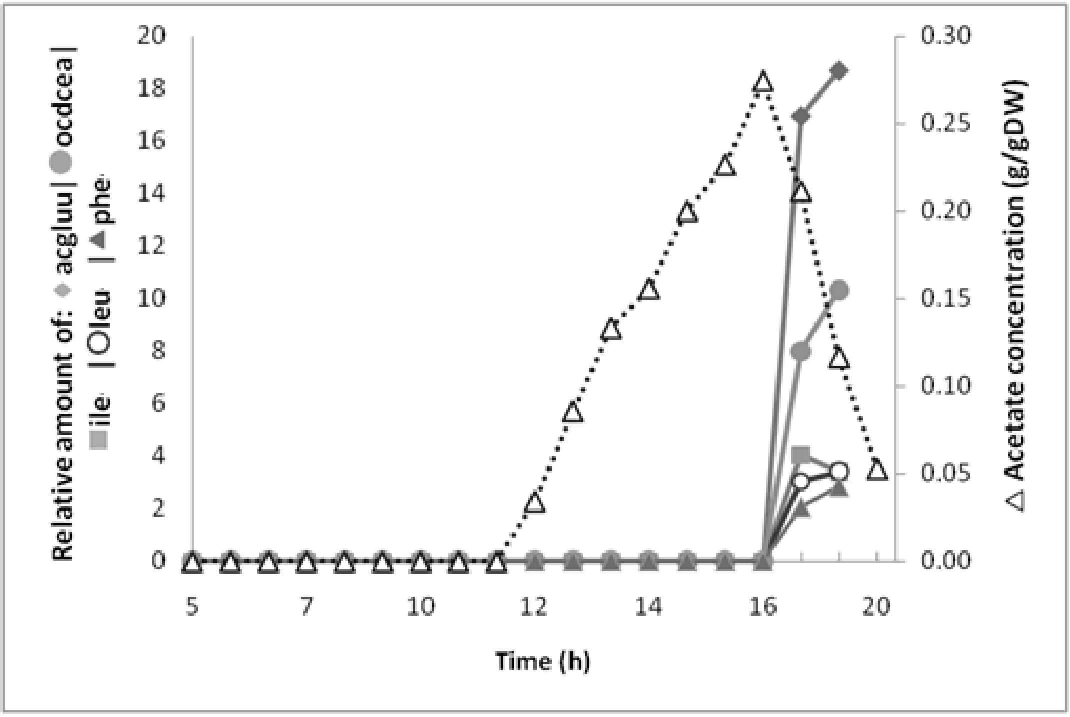
66. Wittmann C *et al* (2007) Response of fluxome and metabolome to temperature-induced recombinant protein synthesis in *Escherichia coli*. *Journal of Biotechnology* 132 (4):375-384.

67. Wu J and Xie J (2009) Magic spot: (p) ppGpp. *Journal of Cellular Physiology* 220 (2):297-302.

68. Xu B *et al* (1999) Glucose overflow metabolism and mixed-acid fermentation in aerobic large-scale fed-batch processes with *Escherichia coli*. *Applied Microbiology and Biotechnology* 51 (5):564-571.

6.8 APPENDIX A

Details on the accumulation of several metabolites that occurred immediately after the acetate consumption by the Δ*rel*A mutant strain.



6.9 Appendix B

Relative concentrations of metabolites measured by GC/MS during during fed-batch fermentation phases: (A) pre-induction phase; (B) IPTG post-induction phase; (C) growth upshift phase; and (D) nutrient downshift phase.

		A-1	A-2	A-3	A-4	A-5	B-1	B-2	B-3	B-4	B-5	C-1	C-2	C-3	C-4	D-1	D-2	D-3
2paac	RelA	82.10	75.10	65.67	67.59	52.79	37.11	30.71	31.03	28.26	29.51	13.58	5.47	6.55	5.20	6.61	14.86	18.65
	W3110	0.00	0.00	49.69	43.87	44.11	26.69	26.49	31.71	27.22	18.53	12.96	7.65	6.50	4.41	4.37	10.73	15.37
3c3hmp	RelA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	10.11	10.32
	W3110	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	8.93	12.66	11.69	14.58
4hbz	RelA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	40.49	38.02
	W3110	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	18.35	18.77	22.83	24.33	38.82	33.31	34.12	29.50	35.67
acglu	RelA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	16.95
	W3110	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
acon-C	RelA	0.00	0.00	0.00	0.00	0.00	19.12	18.08	23.79	23.05	28.02	34.88	43.76	53.36	49.32	50.67	43.43	44.07
	W3110	0.00	0.00	0.00	0.00	0.00	31.63	35.66	44.77	37.26	43.55	59.27	78.10	87.36	63.29	43.52	34.19	50.55
asn	RelA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	10.76	12.11
	W3110	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	8.23	14.62
asp	RelA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	335.71	399.53	566.06	446.41	452.35
	W3110	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	147.47	261.84	449.27	611.85	464.45	392.30	385.79	587.31
bnz	RelA	99.27	237.23	97.72	82.85	69.65	139.22	111.20	81.75	53.54	85.61	48.21	41.70	36.75	33.40	40.36	25.45	13.45
	W3110	106.67	84.15	66.99	46.07	64.69	86.83	91.33	38.06	49.03	31.50	60.48	46.45	22.07	34.67	17.45	10.63	25.02
cbm	RelA	6242.49	5025.13	5461.03	5344.66	4218.20	3301.16	2557.74	2212.88	1711.03	1455.14	1382.78	1043.24	813.88	704.99	826.24	585.38	541.48
	W3110	4956.60	2926.65	3867.89	2722.77	2705.70	2674.03	2198.67	1847.86	1669.44	1087.78	993.32	1127.34	977.15	701.19	479.61	319.42	657.91
cit	RelA	0.00	0.00	0.00	0.00	78.98	80.51	93.38	77.93	115.41	170.56	328.04	448.05	760.24	790.86	980.42	865.33	812.91
	W3110	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	662.70	748.55	1267.47	1075.61	814.33	596.28	916.43
fum	RelA	60.78	84.10	61.30	45.39	0.00	43.06	38.16	42.07	32.37	44.99	71.97	75.96	79.21	49.98	57.14	83.07	35.66
	W3110	293.16	169.46	130.45	92.17	94.17	104.08	120.45	43.32	46.69	28.32	80.69	199.34	128.11	115.33	46.85	40.18	59.20
glu	RelA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	14.76	23.82	35.72	46.53	20.70	24.80	21.98	29.85	93.21	72.56
	W3110	0.00	0.00	0.00	0.00	0.00	0.00	0.00	28.21	25.36	20.60	43.76	29.91	33.89	20.53	25.15	59.57	56.84
gly	RelA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	17.81	21.07	22.24	8.85	0.00
	W3110	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	13.12	13.67	20.71	24.11	12.65	6.05	7.25
ile	RelA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	4.05
	W3110	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
itcon	RelA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	6.51	6.41	5.15	4.70	6.21	4.75	5.70	4.44	4.10
	W3110	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	7.30	9.63	9.37	12.52	8.58	9.13	4.75	3.90	7.99
lac	RelA	0.00	0.00	0.00	0.00	0.00	48.06	44.89	44.77	23.15	29.12	331.81	485.40	362.85	314.29	360.00	208.28	0.00
	W3110	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	360.20	660.66	405.34	546.27	205.78	0.00	0.00
leu-L	RelA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.03
	W3110	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	8.84
mal-L	RelA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	28.85
	W3110	228.62	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	37.40	27.88
mlt	RelA	0.00	0.00	0.00	0.00	0.00	43.00	43.10	42.08	28.24	35.71	22.92	17.66	17.71	17.59	19.00	16.22	9.90
	W3110	0.00	0.00	0.00	0.00	0.00	58.70	59.74	29.47	28.39	22.32	19.24	35.54	17.98	18.93	7.86	6.59	16.54
ocdca	RelA	35.98	26.64	31.30	0.00	0.00	18.57	11.76	12.19	11.02	8.80	9.79	5.50	4.93	4.97	5.28	9.40	9.97
	W3110	0.00	28.70	28.14	29.49	17.32	14.12	11.18	17.61	13.25	13.26	11.12	8.32	6.83	6.41	6.68	7.94	12.01
ocdcea	RelA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	8.00
	W3110	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
phe	RelA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.08
	W3110	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
succ	RelA	91.13	70.38	62.28	46.94	0.00	48.06	49.37	64.37	31.34	41.08	215.78	320.51	351.03	331.99	453.26	348.24	194.09
	W3110	139.98	144.66	99.95	38.27	66.04	63.99	94.94	27.75	41.78	35.10	183.20	460.51	392.42	561.66	262.73	149.03	207.66
ttdca	RelA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	17.39	31.05	41.79
	W3110	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

CHAPTER 7

CONCLUSIONS AND OUTLOOK

"...stress emerges when there is a demand-capability imbalance."

By HI McCubbina and JM Patterson in The Family Stress Process:
The Double ABCX Model of Adjustment and Adaptation,
Marriage & Family Review, 1983; 6(1):7-37

This thesis investigated the extent of metabolic changes induced by recombinant protein production in *E. coli* cells and, more specifically, the influence of the stringent response during recombinant bioprocesses. Metabolomics approaches based on a GC/MS method were at the core of these studies and supported the identification of potential metabolic bottlenecks that may be behind some hindering phenomena during recombinant bioprocesses.

To capture the metabolic behaviour of *E. coli* cells when overexpressing recombinant proteins, a hybrid modelling approach was developed (see Chapter 3). The main purpose of this modelling approach was to estimate, at a systems-level perspective, the degree of metabolic burden imposed by recombinant biosynthetic processes and to predict the induction of stress-responsive events, in particular the ppGpp-induced stringent response triggered by intracellular amino acids shortages. As exemplified, the withdrawn of amino acids for recombinant proteins formation normally exceeds the biosynthetic capacities of the *E. coli* cells, a phenomenon aggravated by the differences in the amino acid composition of recombinant and biomass proteins. As reported in previous studies, drainage of amino acids may result in the sudden induction of the RelA enzyme activity, which synthesizes ppGpp above basal levels (Chatterji D and Ojha AK, 2001; Goldman E and Jakubowski H, 1990; Rojiani MV *et al.*, 1989; Wendrich TM *et al.*, 2002). The pleiotropic effects of this global regulator have been described (Chang DE *et al.*, 2002; Durfee T *et al.*, 2008; Traxler MF *et al.*, 2006), but besides inducing amino acid biosynthetic genes there is little knowledge about its effects on the overall cellular metabolism.

Therefore, the RelA activity was studied both in recombinant and non-recombinant cultures applying a metabolomics approach. First, it was fundamental to understand the influence of the RelA activity in the *E. coli* metabolism at different growth conditions. Chapter 4 presented a metabolic profiling analysis of *E. coli* W3110 and the isogenic Δ*rel*A mutant cells performed to characterize the activity of this enzyme under different growth conditions. Apparently, the *rel*A mutation affects the metabolic behaviour of *E. coli* strains, in particular at lower dilution rates. For example, it was observed that *E. coli* cells lacking the *rel*A gene presented a “relaxed” phenotype that may have led to important shortages in certain metabolites (e.g. amino acids). Moreover, it was suggested that (directly or indirectly) the RelA enzyme might be involved in the synthesis of ppGpp when cells are grown in nutrient-deprived conditions, which potentiates the expression of several stress-response genes, namely the transcriptional regulator cAMP receptor protein (Crp) that governs catabolite repression (Traxler MF *et al.*, 2006). This points to the idea that RelA-mediated ppGpp synthesis is fundamental for the coordination of several metabolic responses to cope with nutritional deprivations.

After exploring the influence of the *rel*A mutation in the *E. coli* W3110 cells' metabolism, it was proposed to investigate if this gene mutation can actually bring any productivity advantages during recombinant processes, as it has been previously reported (Dedhia N *et al.*, 1997; Sanden AM *et al.*, 2003). First, to isolate the impact of recombinant biosynthetic processes from the RelA-mediated metabolic responses to potential amino acid shortages, a chemostat culture with *E. coli* W3110 Δ*rel*A [pTRC-His-AcGFP1] was performed (Chapter 5). The metabolic profiles of host cells were evaluated to investigate the metabolic impact of biosynthetic activities required to maintain the plasmid DNA and recombinant protein expression. It was found that plasmid burden did significantly change the metabolic profiles of the host cells. It was also evident that plasmid burden caused the shortage of certain biosynthetic products, namely amino acids, generating a metabolic imbalance that might also be responsible for the accumulation of some by-products, like acetate. Second, to inspect the metabolic rearrangements associated with the *rel*A mutation and the simultaneous production of recombinant protein, fed-batch cultures with recombinant *E. coli* W3110 and Δ*rel*A cells were performed (Chapter 6). Although some studies have shown that cells that carry a *rel*A mutation are able to produce higher amounts of recombinant protein compared to cells with the stringent (*rel*A$^+$) phenotype (Harcum SW and Bentley WE, 1999), those did not explore the impact of this mutation in the metabolism of host cells, which may ultimately impair cellular survival. A different metabolomic approach, i.e. metabolic footprinting, was devised to estimate the metabolic responses engendered by these cells during the recombinant bioprocess. The metabolic footprints from recombinant *E. coli* W3110 and Δ*rel*A cultures were compared and it was observed that "relaxed" phenotypes (i.e. Δ*rel*A cells) are less effective to stimulate certain metabolic changes, as demonstrated by the delay in the metabolic shift after IPTG-induction. Metabolic profiles of chemostat cultures with non-recombinant *E. coli* W3110 and Δ*rel*A cells were studied in Chapter 4. It was suggested that "relaxed" phenotypes might be inefficient when inducing certain metabolic activities that are required to adjust the cells' metabolism, in particular those that are dependent on the ppGpp-mediated regulation, like anaplerotic functions. Moreover, metabolic footprints revealed that metabolic imbalances were less pronounced in the *E. coli* cells carrying the *rel*A gene. Although the productivity levels were higher in the *E. coli* Δ*rel*A cells, the unexpected accumulation of various metabolites in the extracellular medium, most likely due to the inability to cope with metabolic imbalances caused by recombinant processes, indicates that this cellular system needs to be fine-tuned during recombinant bioprocesses. An option could be the supplementation of the culture medium with amino acids, but it must be acknowledged that though it is a straightforward alternative it may induce severe metabolic constraints, since the excess of some amino acids can stimulate the inhibition of certain amino acid biosynthetic pathways, creating an imbalance in the metabolism. Thus, to enhance the productivity of recombinant proteins, it is crucial to reach a finer balance between strain improvement strategies and culturing conditions. This is the primary goal of biosystems engineering in recombinant bioprocesses to meet the increasing demands of industry.

Altogether, the major outcomes of this thesis contributed significantly to understand the metabolic impact caused by the recombinant protein production and the participation of the RelA enzyme to mediate a series of metabolic adjustments to cope with such demands. It is expected that this information will further contribute to improve the proposed model for the stringent response (Chapter 2). Although modelling approaches are still reliant on bioinformatic and mathematical developments needed to incorporate omics data and, most importantly to manage the complexity of dynamic processes to be modelled, these findings imply that major metabolic changes are stimulated during the RelA-mediated stringent response and, most likely, are closely related with recombinant-induced metabolic imbalances. Ultimately, it is envisioned to create a modelling framework that represents the metabolic behaviour of recombinant cells after ppGpp synthesis has been induced. The integration of new dynamic descriptions that represent these ppGpp-mediated processes in the metabolic network of *E. coli* will improve model predictability when analysing the behaviour of recombinant systems. As previously stated (Hoppe A *et al.*, 2007; Ow DS *et al.*, 2009) flux distributions predicted by FBA are hypothetical and strongly dependent on the cellular objective, and it has been demonstrated that under certain environmental or genetic conditions (e.g. recombinant microorganisms) this approach does not provide the best description of the underlying physiological state of the cells. Therefore, the extension of FBA with other modelling approaches is foreseen to increase the reliability of FBA results and assess new metabolic bottlenecks that must be considered in such systems. For instance, anaplerotic pathways need to be further inspected as they seem to be implicated both in responses to metabolic burden caused by recombinant processes and in the stringent response. Eventually, fluxomics and proteomics data will expose more details about these metabolic activities, providing more detailed models of the *E. coli* metabolism useful for the simulation of recombinant processes.

REFERENCES

1. Chang DE, Smalley DJ, and Conway T (2002) Gene expression profiling of *Escherichia coli* growth transitions: an expanded stringent response model. *Molecular Microbiology* 45 (2):289-306.

2. Chatterji D and Ojha AK (2001) Revisiting the stringent response, ppGpp and starvation signaling. *Current Opinion in Microbiology* 4 (2):160-165.

3. Dedhia N *et al* (1997) Improvement in recombinant protein production in ppGpp-deficient *Escherichia coli*. *Biotechnology and Bioengineering* 53 (4):380-386.

4. Durfee T *et al* (2008) Transcription profiling of the stringent response in *Escherichia coli*. *Journal of Bacteriology* 190 (3):1084-1096.

5. Goldman E and Jakubowski H (1990) Uncharged tRNA, protein synthesis, and the bacterial stringent response. *Molecular Microbiology* 4 (12):2035-2040.

6. Harcum SW and Bentley WE (1999) Heat-shock and stringent responses have overlapping protease activity in *Escherichia coli*. Implications for heterologous protein yield. *Applied Biochemistry and Biotechnology* 80 (1):23-37.

7. Hoppe A, Hoffmann S, and Holzhutter HG (2007) Including metabolite concentrations into flux balance analysis: thermodynamic realizability as a constraint on flux distributions in metabolic networks. *BMC Systems Biology* 1:23.

8. Ow DS *et al* (2009) Identification of cellular objective for elucidating the physiological state of plasmid-bearing *Escherichia coli* using genome-scale *in silico* analysis. *Biotechnology Progress* 25 (1):61-67.

9. Rojiani MV, Jakubowski H, and Goldman E (1989) Effect of variation of charged and uncharged tRNA(Trp) levels on ppGpp synthesis in *Escherichia coli*. *Journal of Bacteriology* 171 (12):6493-6502.

10. Sanden AM *et al* (2003) Limiting factors in *Escherichia coli* fed-batch production of recombinant proteins. *Biotechnology and Bioengineering* 81 (2):158-166.

11. Traxler MF, Chang DE, and Conway T (2006) Guanosine 3 ',5 '-bispyrophosphate coordinates global gene expression during glucose-lactose diauxie in *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America* 103 (7):2374-2379.

12. Wendrich TM *et al* (2002) Dissection of the mechanism for the stringent factor RelA. *Molecular Cell* 10 (4):779-788.