Evolutionary studies of *Corynebacteria* by comparative genomics

Yousuke Nishio

Doctor of Philosophy

Department of Genetics
School of Life Science
The Graduate University for Advanced Studies

2005

Acknowledgement

I greatly appreciate the kind direction of Prof. Takashi Gojobori. I feel deep gratitude to Dr. Kazuho Ikeo for general leading and discussion. I thank Professors Hideaki Sugawara, Hiroyuki Araki, Hironori Niki and Toshimichi Ikemura for their useful comments on my work, serving as the members of my supervisory committee. I am grateful to Drs. Kazuhiko Matsui, Shin-ichi Sugimoto and Yoshihiro Usuda for the conduct of a series of my work at Ajinomoto Co., Inc. I acknowledge the contributions of all of the members in the Laboratory for DNA Data Analysis, especially Dr. Yoji Nakamura. And, I am thankful to the secretaries of Gojobori lab, especially Mses. Yuki Katsube, Yoko Ueda and Keiko Okuda for the help with a lot of the administrative work. Finally, I express my appreciation to my wife for the cooperation in this work.

Contents

Abstract

Corynebacterium efficiens is a gram-positive non-pathogenic bacterium previously known as *Corynebacterium thermoaminogenes*. This strain has recently been shown to be a near relative of *Corynebacterium glutamicum* and *Corynebacterium callunae*, both of which are recognized as glutamic acid-producing *Corynebacterium.* The optimal temperature for glutamate production by *C. glutamicum* is around 30ºC, and this microorganism can neither grow nor produce glutamate at 40ºC or above. On the other hand, *C. efficiens* can grow and produce glutamate above 40ºC. The glutamic-acid-producing species of corynebacteria are known to overproduce glutamic acid under a variety of conditions, such as biotin limitation, although the mechanism of this phenomenon remains unclear. Another member of this genus, *Corynebacterium diphtheriae,* is a well-known pathogen that does not produce glutamic acid. The purpose of the present study is to elucidate the mechanism underlying the thermal stability of *C. efficiens* and to investigate the evolutionary processes that are related to the glutamic-acid-overproduction mechanisms in *C. glutamicum* and *C. efficiens* through considering the genome evolution of *Corynebacterium.* In order to describe the mechanism, I conducted a comparative genomics study using a genome-wide comparison of amino-acid substitutions and metabolic pathways using whole genome sequences.

This thesis comprises five chapters. In **chapter 1**, I describe the research background on this study, placing particular emphasis on the relationship between thermostability and fermentation. I noted that the industrial fermentation process could be carried out at a higher temperature; it might be possible to reduce the electric power consumption and carbon dioxide generation.

In **chapter 2**, I describe the thermostability mechanism of *C. efficiens* revealed by the complete genome sequence comparison between *C. efficiens* and *C. glutamicum*. Differences in the growth temperature, protein stability and GC content between *C. efficiens* and *C. glutamicum* can be investigated through comparative genomics using the complete genome sequences of these bacteria. Because these two species are phylogenetically closely related, more than 1,000 orthologous genes with 60–95% amino-acid sequence identity can be compared. Taking an advantage of comparative genomic studies, I found that there was tremendous bias in amino acid substitutions in all orthologous ORFs. Analysis of the direction of the amino acid substitutions suggested that three substitutions from lysine to arginine, serine to alanine, and serine to threonine, are important for the thermostability of the *C. efficiens* proteins. On the basis of these findings, I suggest that the accumulation of these three types of amino acid substitutions correlates with the acquisition of thermostability and is responsible for the greater GC content of *C. efficiens*.

In **chapter 3**, I make an attempt to understand the evolutionary process involved in the ability of amino acid production in *Corynebacterium*. To attain this purpose, I analyzed the differentiation of metabolic pathways based on a comparative genome analysis of high GC Gram-positive bacteria, including *Mycobacterium* and *Streptomyces*. When *Mycobacterium* and *Streptomyces* were used as outgroups, the comparative study suggested that the common ancestor of *Corynebacteria* already possessed almost all of the gene sets necessary for amino acid production. However, *C. diphtheriae* was found to have lost the genes responsible for amino acid production. Moreover, I found that the common ancestor of *C. efficiens* and *C. glutamicum* have acquired some of genes responsible for amino acid production by horizontal gene transfer. Thus, I show that the evolutionary events of gene loss and horizontal gene transfer must have been responsible for functional differentiation in amino acid biosynthesis of the three species of *Corynebacteria.*

In **chapter 4**, I discuss the evolutionary process for glutamic acid overproduction mechanism under the biotin limitation condition in *C. glutamicum.* To attain this purpose, I compared between the biotin biosynthesis related genes in high GC Gram-positive bacteria. I found that the complete biotin biosynthesis pathway was inherited in *C. diphtheriae*, while *C. glutamicum* and *C. efficiens* only possessed an incomplete pathway. Furthermore, the complete biotin biosynthesis pathway in *C. diphtheriae* suggested to be achieved by the horizontal gene transfer. I conclude that this evolutionary event may have affected metabolic regulation in corynebacteria following the loss of the glutamic acid overproduction mechanism in *C. diphtheriae.*

Finally, in **chapter 5**, I describe the summary and the conclusion of the present study. This study acquired significant knowledge of the protein thermostabilization mechanism and evolutionary process for amino acid production mechanism in *Corynebacterium* by conducting whole genome comparisons. I conclude that this study gives significant insight to the evolutionary process of bacterial diversity from view point of genome evolution.

Chapter 1

Introduction

1.1 Fermentation and heat generation

Five primary elements of taste have been described: sweet, sour, salty, bitter and umami. The last of these, umami, was originally discovered in glutamic acid as the source of the flavor of kelp, which is a type of seaweed. Glutamate is an amino acid that occurs naturally in food and is used throughout the world as a seasoning product. Glutamate has been produced by a fermentation process using the Gram-positive bacterium *Corynebacterium glutamicum* for 50 years (Kinoshita et al., 1957, Udaka, 1960). The production of glutamate is increasing and now the global yield exceeds one million tons per year (Kimura, 2002). Improving the production yield of glutamate is important not only economically but also in terms of the environment, as global warming, which might be caused by increased carbon dioxide levels, has become a serious problem. One approach to reducing the level of industrial carbon dioxide production is to improve the efficiency of fermentation. During the general fermentation process, the temperature of the fermenter is increased by heat that is generated by the growth of the bacteria (Fig. 1.1). A chilling unit is therefore used to keep the fermenter at the optimal temperature. Electronic power, which is produced at power plants that generate carbon dioxide, is required to drive the chilling unit. The fermentation process for glutamate production using *C. glutamicum* is usually carried out at around 30 ºC. Hence, if this fermentation process could be carried out at a higher temperature, it might be possible to reduce the associated electric power consumption and carbon dioxide generation (Adachi et al., 2003).

Corynebacterium efficiens was originally isolated and identified as *Corynebacterium thermoaminogenes* by Yamada and Seto (1987). It was subsequently reclassified as a new species, *C. efficiens,* the nearest relatives of which are the glutamic acid-producing species *C. glutamicum* and *Corynebacterium callunae* (Fudou et al., 2002). *C. efficiens,* unlike *C. glutamicum,* can grow and produce glutamic acid at temperatures above 40 ºC (Fudou et al., 2002). This feature of *C. efficiens* could help to reduce carbon dioxide production by reducing the energy needed to drive the chilling units during the fermentation process.

1.2 Protein thermostabilization

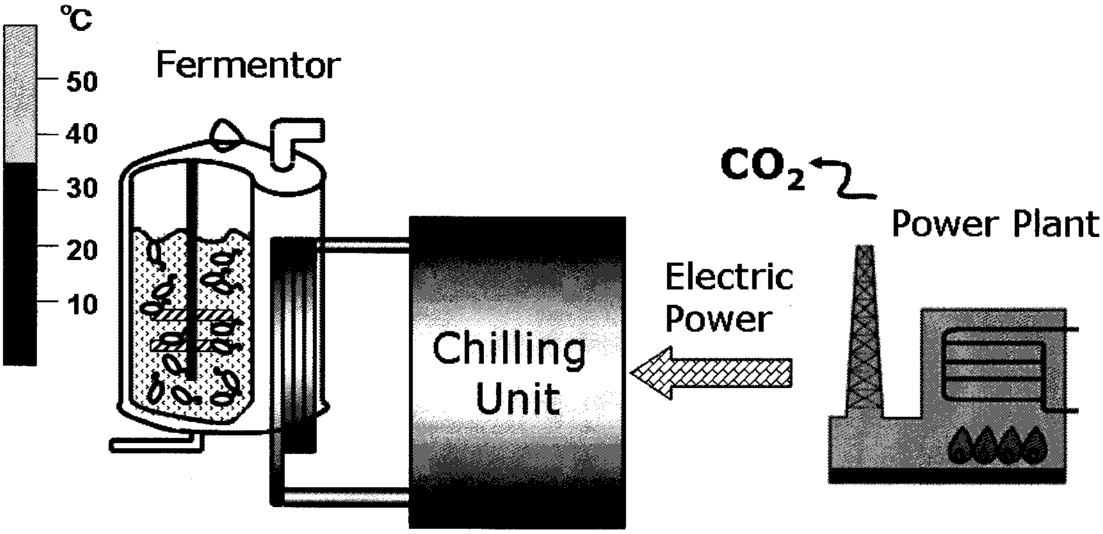
The present study focused on thermostability and, particularly, the various physiological characteristics that can be understood using a comparative approach. We aimed to elucidate the mechanism underlying the thermal stability of *C. efficiens* using a genome-wide comparison of amino-acid substitutions. Our ultimate goal was to identify a general method for protein thermostabilization.

Until now, there have been many excellent studies of protein thermostabilization. Before the complete genome sequences were available, by the comparison of more or less than hundreds of protein sequences from phylogenetically closely related species, the amino acid substitution patterns which might be related to protein thermostabilization were estimated (Haney et al., 1999; McDonald et al., 1999; McDonald, 2001). It has been known that each genome showed the different GC content and codon usage pattern (Grantham et al., 1980) that may contribute to the thermostability in each genome (Musto et al., 2004). The whole genome comparison between *C. efficiens* and *C. glutamicum* will provide attractive topics of protein thermostabilization because of their closely related phylogeny and the difference of growth temperature and GC contents (Fudou et al., 2002).

Nowadays, many microbial complete genome sequences have been determined and one can compare several genome sequences simultaneously. Comparative genomics has been contributed to the studies for protein thermostabilization. Kreil and Ouzounis showed the difference of amino acid pattern between thermophile and mesophile (Kreil and Ouzounis, 2001). Singer and Hickey described that amino acid frequencies contribute to the growth temperature of microbes by showing the correlation between optimal growth temperature and nucleotide frequencies (Singer and Hickey, 2003). They showed statistically significant changes in the frequencies of eight amino acids; there are increases in the proportion of Glu, Ile, Val and Tyr among the thermophiles, while there are significant decreases in Ala, His, Gln and Thr. The previous comparison of a large number of complete genome sequences between thermophiles and mesophiles established the significant difference of optimal temperature for each protein. However, there is a shortcoming in this study. Because of the large phylogenetic distances, many parallel and backward mutations in each protein may be accumulated in their evolutionary process. The comparison between *C. efficiens* and *C. glutamicum* will overcome this difficulty because of their closely related phylogeny. Although the difference in optimal growth temperature is smaller than that in mesophile and thermophile, more than 1,000 orthologous genes with 60–95% amino-acid sequence identity can be compared individually. This is advantageous for our comparative genomic study — previous genome-wide comparisons between thermophilic archaea and mesophilic bacteria have been hindered by the fact that the amino-acid residues did not correspond on a one-to-one basis.

1.3 Evolution of metabolic pathway and amino acid production

It is also important in applied biotechnology studies to understand the relevant metabolic pathways and their evolutionary history. The glutamic-acid-producing species of corynebacteria are known to overproduce glutamic acid under a variety of conditions, such as biotin limitation (Kimura, 2003), although the mechanism of this phenomenon remains unclear. Another member of this genus, *Corynebacterium diphtheriae,* is a well-known pathogen that does not produce glutamic acid. It is therefore of great interest to investigate the evolutionary processes that are related to the glutamic-acid-overproduction mechanisms in *C. glutamicum* and *C. efficiens,* through considering the genome evolution of high GC Gram-positive bacteria. Here I discuss the evolutionary mechanisms involved in the differentiation of metabolic pathways and their regulation, based on a comparative genome analysis of high GC Gram-positive bacteria, including *Mycobacterium* and *Streptomyces.*



Figurel.1. Heat generation of fermentation process

Chapter 2

Thermostability mechanism in *Corynebacterium efficiens*

2.1 Introduction

More than 266 bacterial genomes have already been sequenced and published (http://www.genomesonline.org/, 2005). Although many of these bacteria were pathogens or model organisms, some are of industrial interest (Nelson et al. 2000). *Corynebacterium glutamicum* is a well-known industrial strain widely used for the production by fermentation of various amino acids, such as glutamate and lysine. *Corynebacterium efficiens* is a gram-positive non-pathogenic bacterium previously known as *C. thermoaminogenes*. This strain has recently been shown to be a near relative of *C. glutamicum* and *C. callunae*, both recognized as glutamic acid-producing species (Fudou et al. 2002). The optimal temperature for glutamate production by *C. glutamicum* is around 30ºC, and this microorganism can neither grow nor produce glutamate at 40ºC or above. On the other hand, *C. efficiens* can grow and produce glutamate above 40ºC. Some comparative experimental results are summarised in Table 2.1, showing clearly distinct upper temperature limits for growth. The relative glutamate productivity of *C. glutamicum* by the biotin limitation method (Kimura et al. 1999) was shown to be severely reduced at 37ºC, whereas that of *C. efficiens* was unaffected. The thermostability of *C. efficiens* is a useful trait from an industrial viewpoint as it reduces the considerable cost of cooling needed to dissipate the heat generated during glutamate fermentation.

Many physiological, biochemical, and genetic analyses of *C. glutamicum* have been performed and the genome sequence of *C. glutamicum* ATCC 13032 has been determined by Kyowa Hakko, and is in the public domain. The finding that *C. efficiens* can grow at a temperature 5 ºC higher than *C. glutamicum* and that its guanine plus cytosine (GC) content is 5% higher (Fudou et al. 2002), provides an attractive topic for study by comparative genomics. Experimental data on the thermal stabilities of 11 metabolic enzymes of the two species suggest that many *C. efficiens* proteins are more thermostable than those of *C. glutamicum* (Kimura et al., manuscript in preparation). Furthermore, the two species are closely related phylogenetically, despite the above differences in physiological characteristics. The genome sequence of *Corynebacterium diphtheriae,* a well-known pathogenic strain, has been determined by the Sanger Institute. Because *C. diphtheriae* does not belong to the glutamic acid-producing species, it is useful as a phylogenetic outgroup.

Hyperthermophilic enzymes have been extensively studied (Vieille and Zeikus 2001) and genome-wide comparisons between thermophilic archae and mesophilic bacteria have been reported (Chakravarty and Varadarajan 2000; Kreil and Ouzounis 2001). Thermophilic enzymes are indeed useful for industrial purposes and many examples of protein thermostabilisation have been reported (Vieille and Zeikus 2001). However, the genome-wide amino acid substitutions responsible for the thermal stability of an organism have not been studied. The genome sequences of *C. efficiens* and *C. glutamicum,* permit us to compare mesophiles with different optimal temperatures for growth. The greatest advantage is the opportunity to compare more than 1,000 orthologous genes one by one, because they are so closely related. We have tried here to elucidate the mechanism underlying the thermal stability of *C. efficiens* by a genome-wide comparison of amino acid substitutions, in the hope that such a comparison may indicate a general method for protein thermostabilisation.

2.2 Methods

2.2.1 Genome Sequencing.

The genome of *C. efficiens* JCM 44549 (strain YS-314) was sequenced by the shotgun method (Fleischmann et al. 1995). The end sequences from two pUC118 shotgun libraries, one containing short fragments (0.8-1.2 kb), the other longer fragments (2.0-2.5 kb), were collected. Sequencing reactions were performed on 377 DNA sequencers using dye primer and dye terminator cycle sequencing kits, and M13 universal primers. The data were processed with the Phred/Phrap/Consed package (http://www.phrap.org/) and the assembled sequences, split into 30 kb segments, were re-assembled and edited by Sequencher (GeneCodes, Ann Abor, MI, USA). The details of genome sequencing will be described (Y. Kawarabayasi et. al. manuscript in preparation). Prediction of protein coding regions was performed with the Glimmer 2.0 program under default conditions (Delcher et al. 1999). The sequence, 5'-AAAGAGG-3', was used as Shine-Dalgarno sequence (Amador et al. 1999). The genome sequence itself was used for training.

2.2.2 Informatics.

The genome sequences of *C. glutamicum* ATCC 13032 determined by Kyowa Hakko (European Patent No. 1108790, BA000036 in DDBJ/ EMBL/ GenBank database) and of *Corynebacterium diphtheriae* NCTC13129 by the Sanger Institute (http://www.sanger.ac.uk/Projects/C_diphtheriae/), were used as references. The BLASTP program was used (Altschul et al. 1997) to determine orthologous corynebacterial pairs. Codon usage was examined using cusp programs (http://www.uk.embnet.org/Software/EMBOSS/Apps/cusp.html). The GC contents of ORFs were examined using geecee programs (http://www.uk.embnet.org/Software/EMBOSS/Apps/geecee.html). Window analyses for GC content ((G+C)/(G+A+T+C)) and GC skew ((C-G)/(C+G)) were performed by the windowgc.pl script (Y. Nakamura, unpublished). Stretcher (Myers and Miller 1988) was used for pairwise alignment. Orthologous genes are defined as the best pair of homologues in comparisons between two organisms (Tatusov et al. 1997). GETAREA 1.1 was used to calculate solvent accessible surface areas from PDB files (Fraczkiewicz and Braun 1998). For calculation of various interactions between amino acid residues in a protein, LPCCSU server was employed (Sobolev et al. 1999). tRNA was examined using tRNAscan-SE (Lowe and Eddy, 1997).

2.3 Results

2.3.1 Genome sequence and GC content.

Sequencing was performed by the whole genome shotgun method. Genome size, GC content, tRNA, and the numbers of predicted genes used in this study are shown in Table 2.1 and Supplementary Table 1 for *C. efficiens*, *C. glutamicum*, and *C. diphtheriae*. To gain an overview of corynebacterial genome structure, we compared the GC content (Fig. 2.1), GC skew (Fig. 2.2) and gene order (Fig. 2.3). *C. glutamicum* had a GC content between 50% and 60% in most regions of the chromosome, and its average GC content was 53.8%. On the other hand, the average GC content of *C. efficiens* was 63.1%, higher than *C. glutamicum* over the entire chromosome (Fig. 2.1). This tendency was also clearly displayed by the predicted ORFs (Fig. 2.3A for *C. efficiens*; Fig. 2.3B for *C. glutamicum*). Although the GC content of *C. efficiens* had previously been reported to be 5% higher than that of *C. glutamicum* (Fudou et al. 2002), the whole genome analysis reveals that the true figure is 10%.

C. diphtheriae was used as an outgroup of the glutamic acid producing strains. *C. diphtheriae* showed a window analysis profile of GC content more similar to *C. glutamicum* than to *C. efficiens* (Fig. 2.1 and Fig. 2.3C). This suggests that the ancestral genome structure of corynebacteria may be closer to that of *C. glutamicum* than to that of *C. efficiens*. The GC skew profile supported this hypothesis: whereas *C. glutamicum* (Fig. 2.2A) and the outgroup, *C. diphtheria* (Fig. 2.2C), showed clear GC skew profiles with an inversion point that corresponds to the replication terminus or origin (McLean et al. 1998), *C. efficiens* gave an irregular GC skew profile (Fig. 2.2B). In addition gene order was well conserved (Fig. 2.4) while the GC content of *C. efficiens* was higher than that of *C. glutamicum* and *C. diphtheriae* (Fig. 2.1) (Nakamura et al., 2003). We therefore inferred that the genome structure of the common ancestor was more similar to that of *C. glutamicum* and *C. diphtheriae* than to *C. efficiens,* so that *C. efficiens* must have acquired its thermostability by an increase of GC content after divergence from its sister species.

There was a region of low GC content between 1.8 Mb and 2.0 Mb in *C. glutamicum* (Fig. 2.3B) and another from 1.2 Mb to 1.7 Mb in *C. diphtheriae* (Fig. 2.3C). In these regions the values of GC skew in *C. glutamicum* were under -0.1, whereas in *C. diphtheriae,* they were above -0.1 (Fig.2.2C), pointing to a difference between the two regions. In the comparison of orthologous gene order, prominent gaps between *C. glutamicum* and *C. efficiens* (Fig. 2.4A) and *C. glutamicum* and *C. diphtheriae* (Fig. 2.4B) corresponded to the region of low GC content of *C. glutamicum.* We did not find a similar large gap corresponding to the low GC content region of *C. diphtheriae* (Fig. 2.4B, 2.4C). These results suggest that the low GC content region in *C. glutamicum* was acquired by horizontal gene transfer and the transposase homologues were found in this region (Ikeda and Nakagawa, 2003, Kalinowski et al., 2003). There may be a tendency towards lower GC content in that region in *C. diphtheriae.* Thus in spite of the conserved gene order, there is massive variability in genomic GC content among *Corynebacteria* that may be a strong driving force for evolution.

2.3.2 Codon usage and amino acid composition of ORFs.

The numbers of ORFs extracted by the Glimmer program as a function of GC content were analyzed (Fig. 2.5). The peak of ORF number in *C. efficiens* shifts to higher GC than in *C. glutamicum*. The difference in average GC content between the two micro organisms is directly reflected in the GC content of the ORFs. To investigate the difference in GC content of the ORFs, codon usage and nucleotide substitutions were examined in the gene-coding regions.

The codon usage of *C. efficiens* genes was much more biased than that of *C. glutamicum* (Table 2.2). For example, CTC (Leu) and CTG (Leu) were used more frequently in *C. efficiens,* although the two species had almost the same total number of Leu codons. The most frequently used Asp and Ala codons, GAC (Asp) and GCC (Ala) in *C. efficiens* differed from those in *C. glutamicum*, GAT (Asp) and GCA (Ala), respectively. Thirteen codons are rarely used in the highly expressed genes of *C. glutamicum* (Malumbres et al. 1993). The number of codons per 1000 bases (fraction values) are below 10 in *C. glutamicum*, whereas the number of GGG (Gly) and CGG (Arg) codons exceeds 10 in *C. efficiens* (Table 2.2).

Among the ten most frequently used codons in *C. glutamicum*, 7 have GC in the third position whereas all 10 codons do so in *C. efficiens*. Of 10 rarely used codons, none contains GC in the third position in *C. efficiens* against 3 in *C. glutamicum.* Also it should be noted that only the fraction value of the GGT (Gly) codon, among the codons with AT in the third position, was higher by more than 6 points in *C. efficiens* than in *C. glutamicum*. These findings seem to reflect clearly the higher GC content of *C. efficiens*.

The amino acid composition of the protein coding regions is analysed in Table 2.3. Lys, Asn, Ser, Ile and Phe are more frequently used in *C. glutamicum* than in *C. efficiens*. The increased usage of Arg, Asp, Gly, His, Pro and Val in *C. efficiens* is shown to be statistically significant by z test. The high utilization frequency of Asn, Ile, Phe and Lys in *C. glutamicum* agrees with the tendency of these amino acids to increase with decreasing GC content reported in a statistical analysis of the complete genomes of six thermophilic archaea, two thermophilic bacteria, 17 mesophilic bacteria and two eukaryotic species (Kreil and Ouzounis 2001). On the other hand, the high frequency of Gly and Arg in *C. efficiens* concurs with the view that these amino acid residues increase parallel to rises in GC content.

2.3.3 Base replacement and amino acid substitution.

The orthologous ORFs of *C. glutamicum* and *C. efficiens* were extracted and sorted according to their degree of identity. They were then divided into three groups, a group with identity of more than 95%, another with identity from 60% to 95%, and a third with identity under 60%. More than 95% of the genes belonging to the first group are ribosomal proteins and we did not analyse these proteins because of their anticipated conservative nature. The third group, with identity under 60%, was also omitted, because of the large calculated p-distance value of 0.4 and the need to take account of backward and parallel mutations (Nei and Sudhir 2000). 1,619 orthologous pairs of genes with identity from 60% to 95% (p-distance value 0.2) were used to examine position-specific mutations. Synonymous codon replacement was analysed (Table 2.4), and among the 30 most frequent synonymous substitutions, 26 were changes in the third letter from AT in *C. glutamicum* to GC in *C. efficiens*. The only substitution that involved GC in *C. glutamicum* and AT in *C. efficiens* was from GGC (Gly) in *C. glutamicum* to GGT (Gly) in *C. efficiens*. Among the 30 most frequent non-synonymous substitutions in *C. efficiens,* 27 increased GC content and in 21 of these, GC was in the third position (Table 2.5). Of these 21, 3 substitutions, from Lys to Arg (AAA to CGG, AAA to CGC, and AAG to CGC), involved changes in all three letters. The trend of nucleotide substitutions at each codon position in *C. efficiens* also involved an increase of GC content (Supplementary Table 2).

The amino acid sequences of 1,619 orthologous genes with identity from 60% to 95% were aligned using the pairwise alignment program, Stretcher (Fraczkiewicz and Braun 1998), and the amino acid substitutions obtained were placed in a matrix (Supplementary Table 3). By analysing the differences between the matrix and the transposed matrix, the asymmetric mutations from *C. glutamicum* to *C. efficiens* were extracted (Table 2.6). The results of biased mutations in the two other categories (the groups with identity under 60% and over 95%) differed from those in Table 2.6 (Supplementary Table 3 and 4). Some of the amino acid substitutions in this table have often been observed before, with Leu, Ile, Val, and Met replacing each other (Henikoff and Henikoff 1992). Because the fourth most frequent substitution, from Ile to Val, is commonly observed in situations unrelated to thermostabilisation, the three most frequent substitutions (Lys to Arg, Ser to Thr, Ser to Ala) are the best candidates for stabilising the proteins. Indeed many studies have suggested that the Lys to Arg substitution affects thermal stability (Vieille and Zeikus 2001). If the evolutionary development of the thermal stability of proteins is responsible for the thermostability of *C. efficiens* itself, then the observed amino acid substitutions must be adaptive mutations leading to overall thermostability. In a separate study, the thermal stability of 13 pairs of enzymes on the Glu and Lys biosynthetic pathways in the two species were compared on the basis of the enzymatic activities remaining after heat treatment of crude extracts. In Table 2.7 the numbers of the three kinds of amino acid substitutions within the amino acid sequence are assigned points depending on their directions, and we compare the number of calculated points with the experimental results of enzyme thermal stability (Supplementary Fig.1). Nine out of 13 enzymes, the thermostabilities of which had been measured, agree with the calculated points, 3 can not be determined, and only one does not coincide (Table 2.7). These results suggest that there is a significant correlation between the three kinds of amino acid substitution and the thermal stability of proteins.

2.4 Discussion

There is controversy over whether the first life forms were hyperthermophiles (Woese 1987; Pace 1991; Nisbet and Fowler 1996; Yamagishi et al. 1998) or not (Miller and Lazcano 1995; Forterre 1996; Galtier et al. 1999). As far as we know, among the species belonging to the genus *Corynebacterium*, *C. efficiens* can grow at the highest temperature, and is unique in its ability to produce glutamate above 40 ºC. The main point of interest in relation to the above controversy is whether *C. efficiens* acquired the ability to grow at higher temperature, or whether *C. glutamicum* lost it. On the basis of GC content and GC skew analyses, we concluded that *C. glutamicum* is closer to the common ancestor of the glutamic acid-producing strains, and therefore that *C. efficiens* acquired its thermostability and higher GC content in the course of evolution. To understand the basis of this thermostability, we compared the C. *efficiens* to *C. glutamicum* genomes.

Studies of protein thermostability using genome sequences have generally compared hyperthermophiles or thermophiles, and mesophiles (Chakravarty and Varadarajan 2000; Kreil and Ouzounis 2001). In such cases, the differences in growth temperature are clear, but the amino acid residues do not correspond one-to-one because thermophiles and mesophiles are not close phylogenetically. In this report, we have compared two mesophiles with different optimal temperatures for growth and were able to make a statistical comparison of amino acid residues one by one because of the close phylogenetic relationship. Among asymmetrical amino acid substitutions between *C. glutamicum* and *C. efficiens*, that from Lys to Arg was the most frequent (Table 2.6). This substitution is known to contribute to protein stability. The mechanism of thermostabilization is thought to depend on the resonance stabilisation effect of Arg (Vieille and Zeikus 2001). Thus Arg is assumed to contribute to protein thermostability because it maintains ion pairs more easily. Nevertheless, the Arg/Lys ratios, 2.94 in *C. efficiens*, and 1.61 in *C. glutamicum* are larger than the 2.19 ratio of *Aeropyurum pernix*, which has the highest Arg/Lys ratio of the hyperthermophiles, and a GC content of 53.6% (Kreil and Ouzounis 2001). Thus the substitutions from lysine in *C. glutamicum* to Arg in *C. efficiens* appear to result from the increase of GC content and constitute the basis of protein thermostabilisation.

With regard to the substitutions from Ser in *C. glutamicum* to Ala or Thr in *C. efficiens*, we consider that Ala and Thr can strengthen hydrophobic interaction inside proteins, because Ala and Thr are more hydrophobic in the environment of a protein than Ser (Taylor 1996). McDonald et al. have analyzed the asymmetric amino acid substitution patterns in 229 genes of the bacterial genus *Bacillus* and 99 genes of the archaeal genus *Methanococcus* (McDonald et al. 1999). The differences in GC content in *Bacillus* are similar (*B. stearothermophilus* 52% vs. *B. subtilis* 43.5%) to the difference between *C. efficiens* and *C. glutamicum,* and the asymmetrical amino acid substitution patterns found in *Bacillus* are very similar. However the analysis of *Bacillus* and other works were based on far fewer genes than our analysis and did not confirm orthologous relationships (Haney et al. 1999, McDonald 2001). The two most frequent substitutions found in *Bacillus* were the same as in our analysis (Lys to Arg and Ser to Thr), but, the Ser to Ala substitution found in genus *Corynebacterium* was less evident in *Bacillus*. Nevertheless Wintrode et al. (2001) have reported substitutions from serine to various amino acids in a thermostable subtilisin made by directed evolution, and their findings suggest that mutation from Ser to Ala or Thr may be one of the effective ways to generate thermostable proteins.

The X-ray structure of the diaminopimelate dehydrogenase (Ddh) of *C. glutamicum,* one of the enzymes in our analysis, has been determined (Cirilli et al. 2000). Interestingly, *C. glutamicum* Ddh was found to be more stable than that of *C. efficiens* and the mutations responsible are of great interest. We have tried to identify the most effective of the three amino acid substitutions responsible for the thermostability of *C. glutamicum* Ddh over *C. efficiens*. The amino acid substitution which acts to lower thermostability of *C. glutamicum* Ddh is most probably that of ^{113}Ala, which is replaced by Ser in *C. efficiens*. The Ser residue tends to impair hydrophobic interaction between β-strands whereas the Ala can be effective in bridging strands (Fig. 2.6). It is likely that some but not all of the observed substitutions affect protein stability. To identify the actually effective mutations, actual amino acid substitution experiments and measurements of protein thermostability are needed. Recently, many protein crystal structures have been determined and structure-modeling technology is developing rapidly, so that we may soon be able to predict which mutations among the proposed substitutions increase stability.

An interesting question concerns which event occurred first in evolution, the increase in genomic GC content or the adaptive amino acid substitutions. Due to the close phylogenetic relationship *of C. efficiens* and *C. glutamicum,* this study was focused on only one letter substitutions, and the three substitutions that are not caused by the replacement of the third letter of codons. The one-base substitutions from Lys (AAA and AAG) to Arg (AGA and AGG) and from Ser (TCA, TCC, TCG, and TCT) to Ala (GCA, GCC, GCG, and GCT) are compatible with the increase of GC content in *C. efficiens*. However, the possible replacements from Ser (TCA, TCC, TCG, and TCT) to

Thr (ACA, ACC, ACG, and ACT) are not explained by the GC increase. Thus, the increase in GC content alone cannot predict all three amino acid substitutions thought from the statistical analysis to be involved in thermostabilization.

Table 2.1 Summary of characteristics of corynebacteria

	C. efficiens	*C. glutamicum*	*C. diphtheriae*
Upper temperature limit for growth (ºC)	45	40	-
Glutamate production at 32 ºC (%)[a)]	80	100	-
Glutamate production at 37 ºC (%)	78	40	-
Genome size (bp)	3,147,090	3,309,401	2,488,635
GC content (%)	63.1	53.8	53.5
Number of predicted gene	2,942	3,099	2,320

[a)] Glutamate production in typical experiments using the biotin limitation method as a percent of the production by *C. glutamicum* at 32 ºC.

Table 2.2 Condon usage in *C. glutamicum* and *C. efficiens*

Codon	Amino acid	Rare codon[a]	Fraction value[b]	
			C. glutamicum	*C. efficiens*
GCA	Ala		30.66	12.29
GCC	Ala		27.18	54.41
GCG	Ala		23.15	28.82
GCT	Ala		24.96	8.75
TGC	Cys		4.87	6.33
TGT	Cys		2.66	2.81
GAC	Asp		26.11	34.68
GAT	Asp		32.89	29.26
GAA	Glu		35.50	18.10
GAG	Glu		27.41	42.27
TTC	Phe		22.87	26.53
TTT	Phe		13.78	3.82
GGA	Gly		15.43	12.47
GGC	Gly		33.34	37.00
GGG	Gly	○	6.97	18.19
GGT	Gly		24.44	31.09
CAC	His		14.52	19.77
CAT	His		7.22	8.68
ATA	Ile	○	2.05	1.89
ATC	Ile		33.55	42.38
ATT	Ile		21.48	5.21
AAA	Lys		14.27	5.58
AAG	Lys		20.75	17.38
CTA	Leu	○	5.94	1.78
CTC	Leu		22.05	33.67
CTG	Leu		27.38	47.63
CTT	Leu		17.01	8.15
TTA	Leu	○	5.31	1.38
TTG	Leu		19.99	6.08
ATG	Met		21.90	19.52
AAC	Asn		21.94	18.40
AAT	Asn		11.29	6.26

CCA	Pro		16.80	6.76
CCC	Pro		9.83	20.98
CCG	Pro		10.38	21.56
CCT	Pro		11.26	3.94
CAA	Gln		13.23	3.33
CAG	Gln		20.76	31.64
AGA	Arg	○	2.71	1.94
AGG	Arg	○	3.79	4.73
CGA	Arg	○	6.73	4.07
CGC	Arg		24.54	26.91
CGG	Arg	○	5.13	16.99
CGT	Arg		13.50	12.90
AGC	Ser		10.89	8.91
AGT	Ser	○	5.28	3.76
TCA	Ser	○	8.43	4.01
TCC	Ser		21.01	25.23
TCG	Ser	○	7.71	7.91
TCT	Ser		10.99	2.43
ACA	Thr	○	7.90	4.23
ACC	Thr		32.14	46.33
ACG	Thr		8.96	10.18
ACT	Thr		12.51	3.22
GTA	Val	○	8.41	5.23
GTC	Val		22.22	34.05
GTG	Val		29.01	37.03
GTT	Val		21.03	9.58
TGG	Trp		14.13	13.21
TAC	Tyr		14.40	13.35
TAT	Tyr		7.45	5.00

[a]Rare codons are adapted from Malumbers et al. (1993).

[b]Fraction value represents the number of codons per 1000 bases.

Table 2.3 Amino acid composition of protein coding regions

Amino acid	Number		Ratio (%)[a]		P[b]
	C. glutamicum	*C. efficiens*	*C. glutamicum*	*C. efficiens*	
Ala	107484	122084	10.58	10.44	
Arg	57210	79096	5.63	6.76	***
Asn	33710	28875	3.32	2.47	***
Asp	59858	74866	5.89	6.40	***
Cys	7643	10706	0.75	0.92	
Gln	34477	40943	3.39	3.50	
Glu	63816	70689	6.28	6.04	*
Gly	81344	115628	8.01	9.88	***
His	22050	33308	2.17	2.85	***
Ile	57899	57934	5.70	4.95	***
Leu	99098	115556	9.76	9.88	
Lys	35527	26882	3.50	2.30	***
Met	22217	22860	2.19	1.95	*
Phe	37182	35530	3.66	3.04	***
Pro	48961	62331	4.82	5.33	***
Ser	65246	61183	6.42	5.23	***
Thr	62400	74898	6.14	6.40	*
Trp	15465	14339	1.52	1.23	*
Tyr	22164	21488	2.18	1.84	**
Val	81846	100565	8.06	8.60	***
Total	1015597	1169761			

[a] The ratio is the percentage of the number of a given amino acid to the total number of amino acids.

[b] P is the significant difference level by z test: * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$.

Table 2.4 Top 30 synonymous codon replacement between *C. glutamicum* and *C. efficiens*

C. glutamicum	*C. efficiens*	Value
GAA	GAG	9247
GCA	GCC	7355
ATT	ATC	7063
GCT	GCC	6325
TTG	CTG	5576
CCA	CCG	4236
TTT	TTC	4228
GTT	GTC	3886
CTT	CTG	3868
ACT	ACC	3746
GTT	GTG	3664
CAA	CAG	3644
CCA	CCC	3273
GAT	GAC	3220
TCT	TCC	2744
GCG	GCC	2615
GCA	GCG	2591
CCT	CCG	2476
CTT	CTC	2322
TTG	CTC	2319
CCT	CCC	2295
GCT	GCG	2248
GGC	GGG	1895
GGC	GGT	1856
AAA	AAG	1847
GGA	GGC	1790
CGC	CGG	1570
GTA	GTG	1488
CTA	CTG	1345
GGT	GGG	1321

Value is defined as the difference between the number of amino acid substitutions from *C. glutamicum* to *C. efficiens* and the number of substitutions in the opposite direction.

Table 2.5 Top 30 nonsynonymous codon substitution between *C. glutamicum* and *C. efficiens*

C. glutamicum	*C. efficiens*	Value
GAA	GAC	827
GAT	GAG	803
GTT	ATC	795
ATT	GTG	712
ATT	GTC	588
TCC	GCC	488
GAA	CAG	471
ATG	CTG	422
ATT	CTG	400
AAG	CGG	397
GAA	GCC	385
AAA	CGG	355
AAA	CGC	335
TCT	GCC	329
TCT	ACC	328
CAA	GAG	322
AAC	GAC	304
GCT	TCC	300
GAA	GAT	299
AAG	CGC	296
GCA	ACC	294
GAC	GAG	283
GCA	TCC	274
GCT	ACC	274
ATT	CTC	273
GCA	GAG	258
AAG	AGG	253
TCC	ACC	243
ATC	GTC	232
CTT	ATC	230

Value is defined as the difference between the number of amino acid substitutions from *C. glutamicum* to *C. efficiens* and the number of substitutions in the opposite direction.

Table 2.6 Biased amino acid substitutions in the orthologous genes of *C. glutamicum* and *C. efficiens*

C. glutamicum	*C. efficiens*	Forward	Reverse	Point[a]	G+C change by one-base substitution
Lys	Arg	2855	664	1095.5	AAA->AGA, AAG->AGG
Ser	Ala	3378	2372	503.0	TCA->GCA, TCC->GCC, TCG->GCG, TCT->GCT
Ser	Thr	2623	1723	450.0	
Ile	Val	4332	3585	373.5	ATA->GTA, ATC->GTC, ATT->GTT
Asn	Arg	978	372	303.0	
Gln	Glu	1321	747	287.0	
Ile	Leu	2191	1642	274.5	ATA->CTA, ATC->CTC, ATT->CTT
Ser	Gly	1013	610	201.5	AGC->GGC, AGT->GGT
Lys	Thr	600	235	182.5	AAA->ACA, AAG->ACG
Ala	Pro	1019	656	181.5	

[a] Point is defined as the difference between the number of amino acid substitutions from *C. glutamicum* to *C. efficiens* and the number of substitutions in the opposite direction, divided by two.

All of the asymmetrical amino acid substitutions showed probability of obtaining the observed deviation from 50:50 by chance less than 0.001

Table 2.7 Check of predictions against actual measurements

Entry	Enzyme	Thermostable species	Point	Result
1	2-Oxoglutarate dehydrogenase	*C. efficiens*	0	-
2	Glutamate dehydrogenase	*C. efficiens*	1	Yes
3	Isocitrate lyase	*C. efficiens*	2	Yes
4	Phosphofructokinase	*C. efficiens*	-3	No
5	Fructose-1-phosphate kinase	*C. efficiens*	4	Yes
6	Isocitrate dehydrogenase	*C. efficiens*	4	Yes
7	Aconitase	*C. efficiens*	0	-
8	Phosphoenolpyruvate carboxylase	*C. efficiens*	10	Yes
9	Citrate synthase	*C. efficiens*	3	Yes
10	Aspartate kinase	*C. glutamicum*	-1	Yes
11	Dihydrodipicolinate synthase	*C. efficiens*	0	-
12	Diaminopimelate dehydrogenase	*C. glutamicum*	-2	Yes
13	Diaminopimelate decarboxylase	*C. efficiens*	2	Yes

Point is defined as the difference between the sum of the three kinds of substitutions from *C. glutamicum* to *C. efficiens* (Lys to Arg, Ser to Ala and Ser to Thr) and the sum of the reverse substitutions (Point = {number of (Lys→Arg + Ser→Ala + Ser→Thr)} — {number of (Arg→Lys + Ala→Ser + Thr→Ser)}).

Results are indicated by 1) Yes: when the enzyme from *C. efficiens* was more thermostable and the point is positive, or when the enzyme from *C. glutamicum* was more thermostable and point is negative. 2) -: when the point was zero. 3) No: all other case

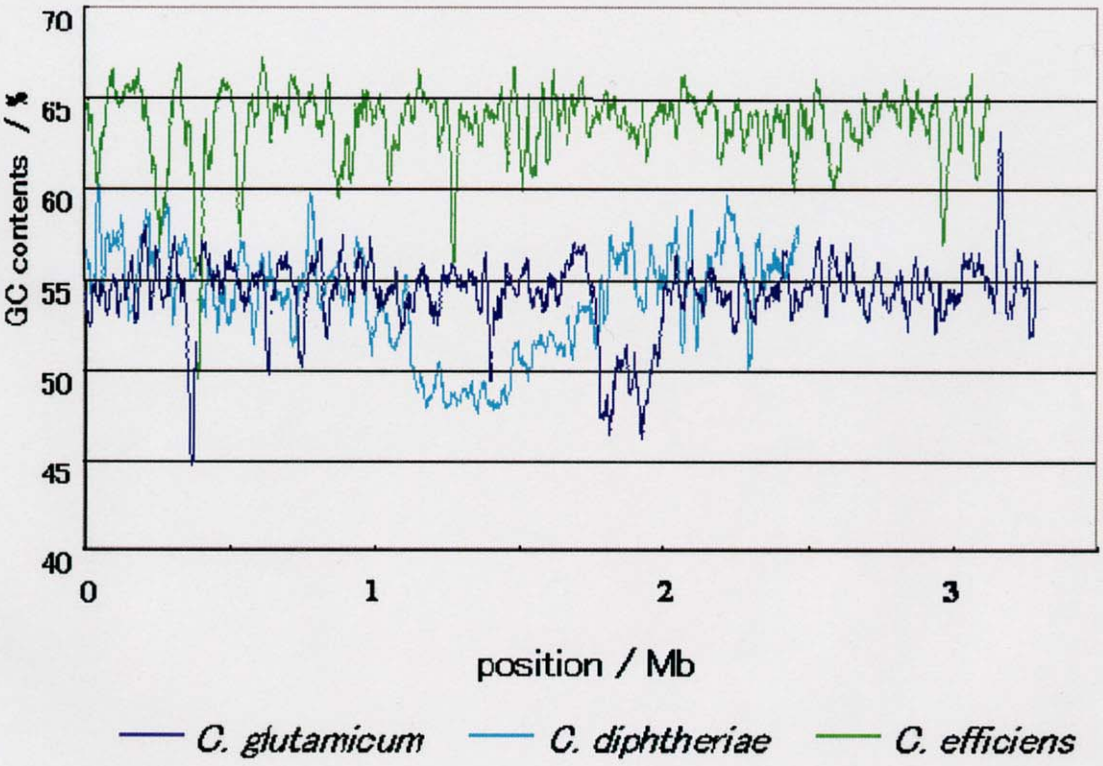


Figure 2.1 GC content of three corynebacterial genomes

Window analysis of GC content performed at 20 kb window size and 1 kb step size. Linear representation of GC content along the chromosome. Green, *C. efficiens*; dark blue, *C. glutamicum*; light blue, *C. diphtheriae*

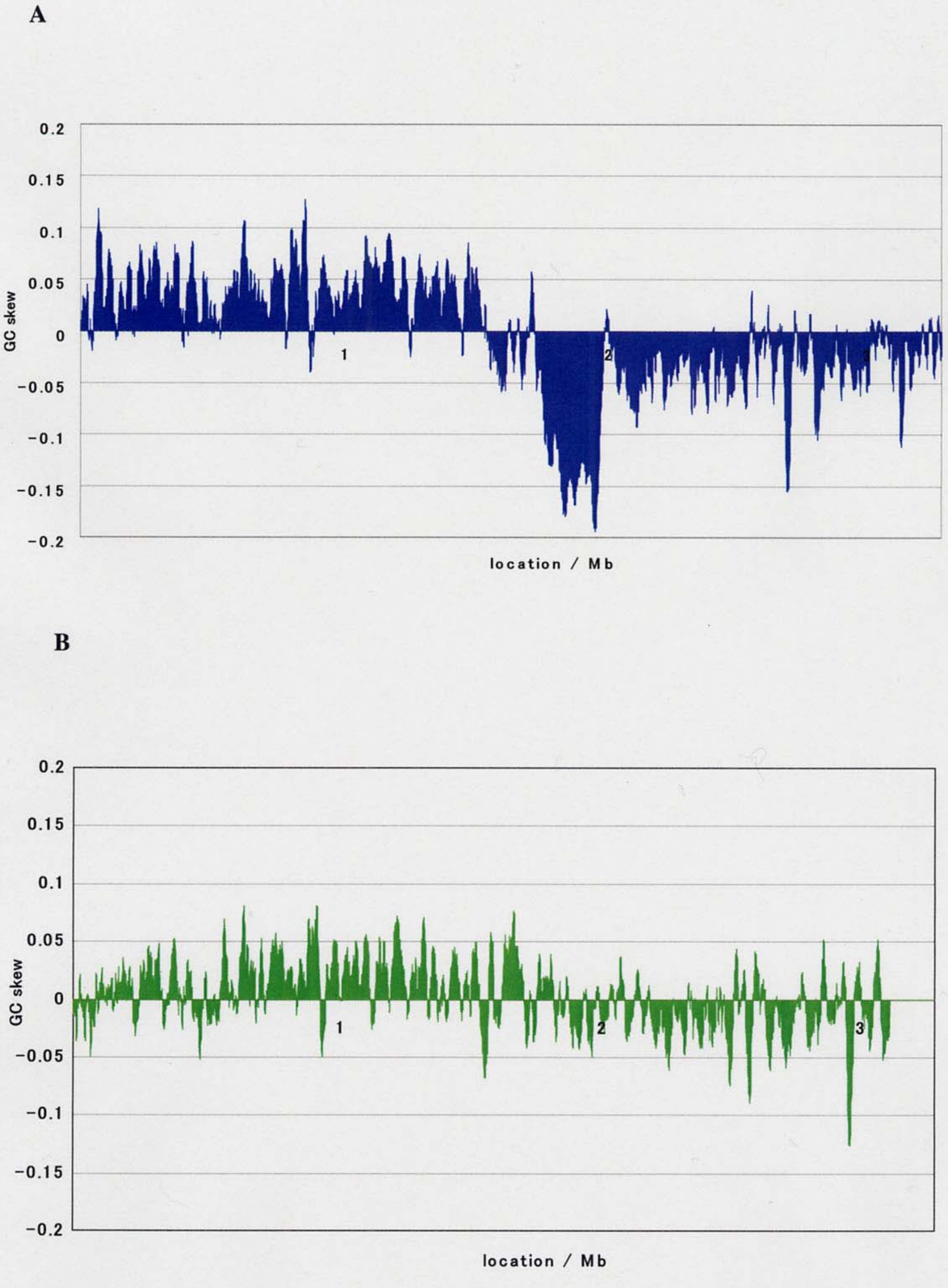

A
0.2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
GC skew
1
2
3
location / Mb
B
0.2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
GC skew
1
2
3
location / Mb

Figure 2.2

C

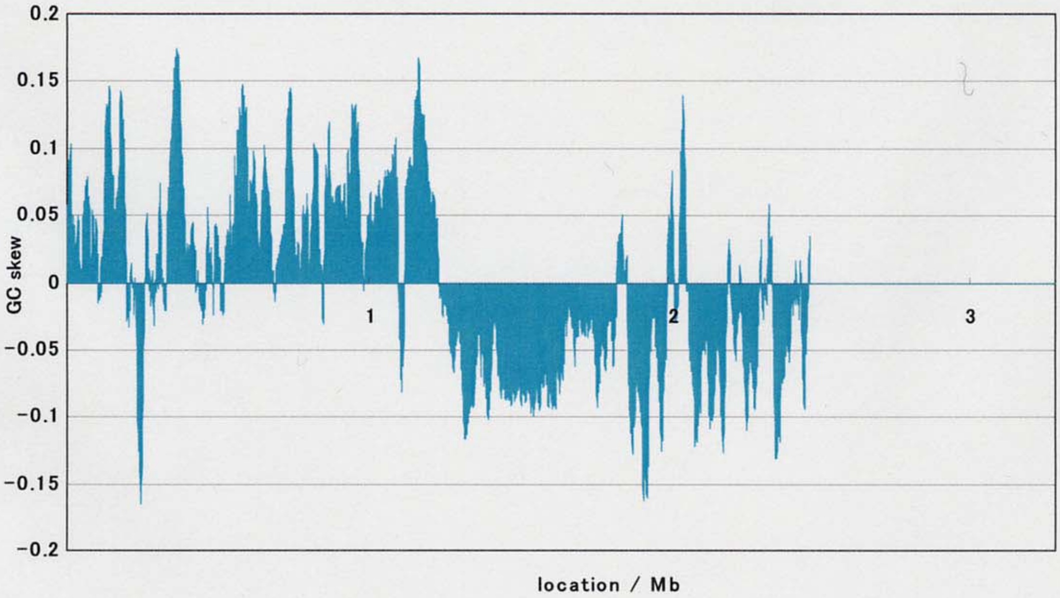


Figure 2.2 (Continued) *C. glutamicum* GC skew of the three corynebacteria

Window analysis of GC skew was performed at 20 kb window size and 1 kb step size. *C. glutamicum* (A), *C. efficiens* (B) and *C. diphtheriae* (C).

A

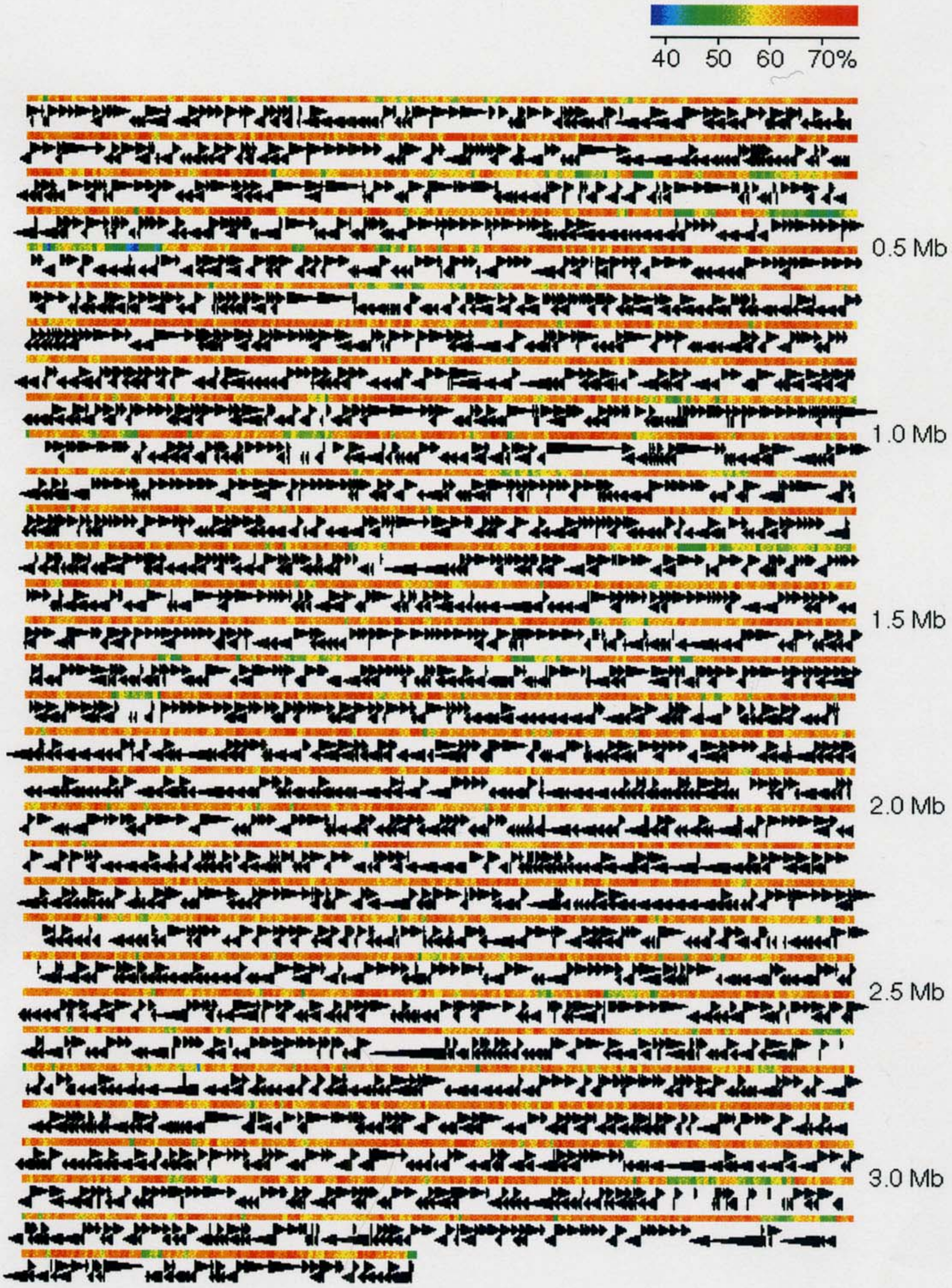


Figure 2.3

B

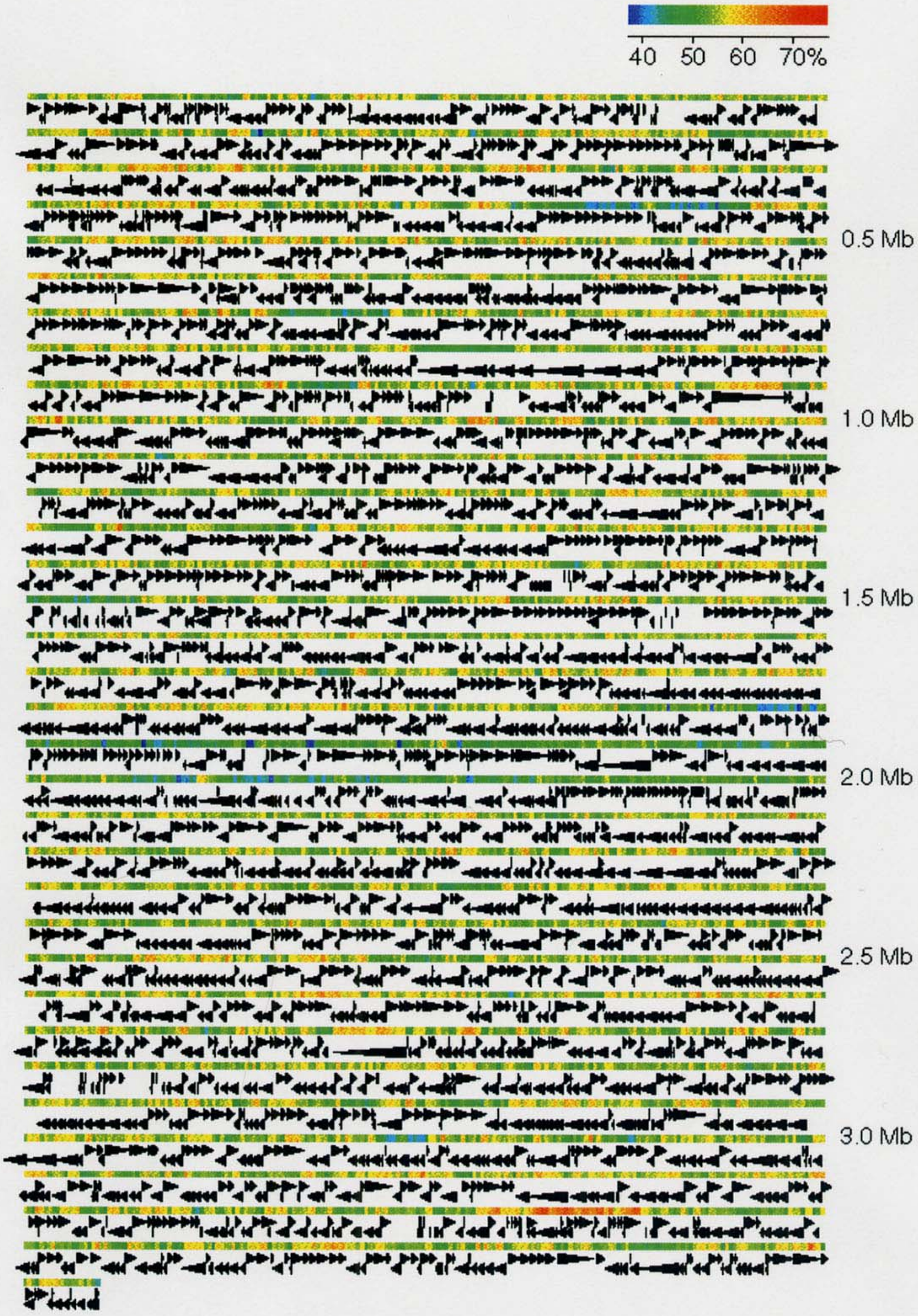


Figure 2.3 (Continued)

C

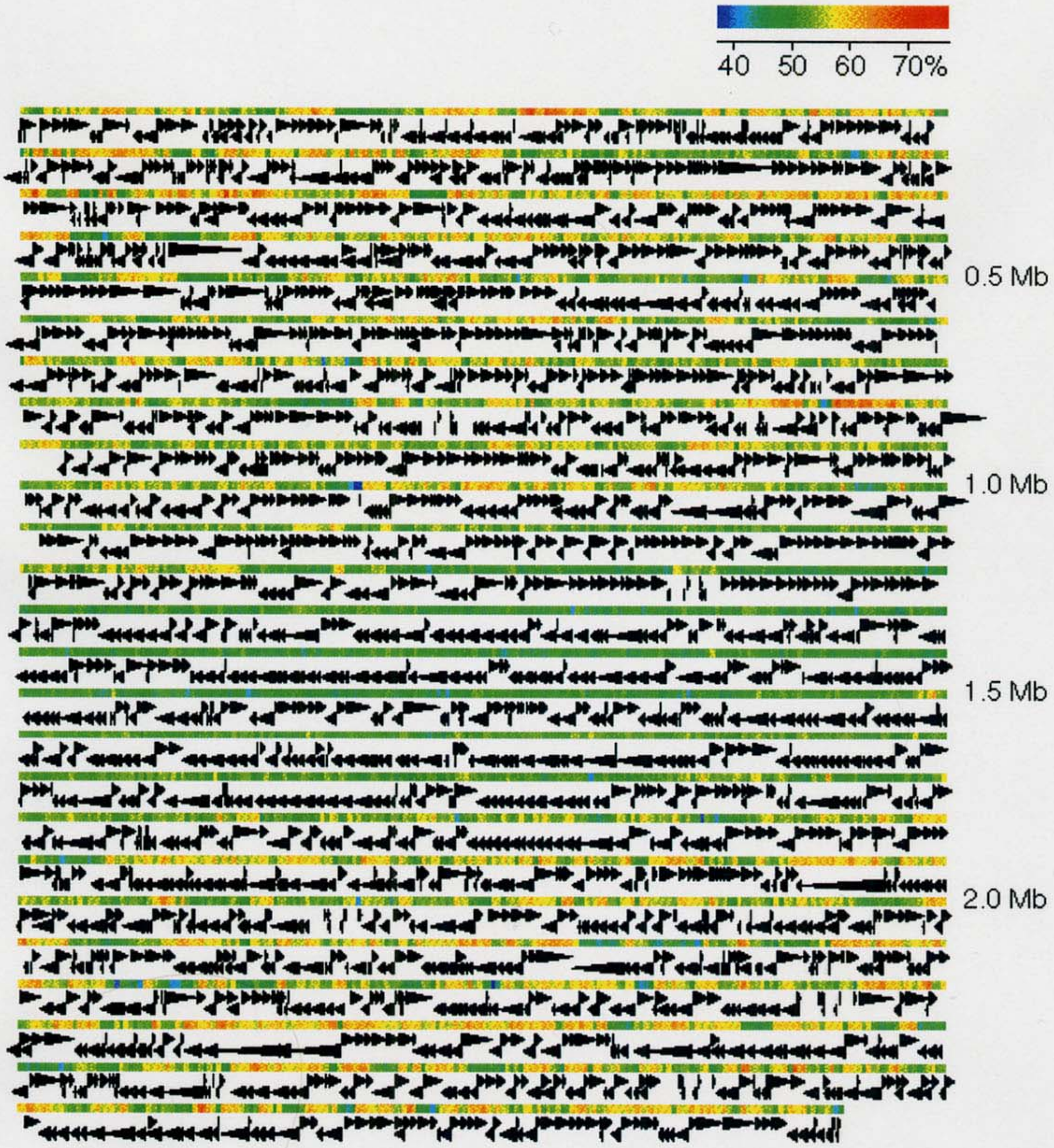
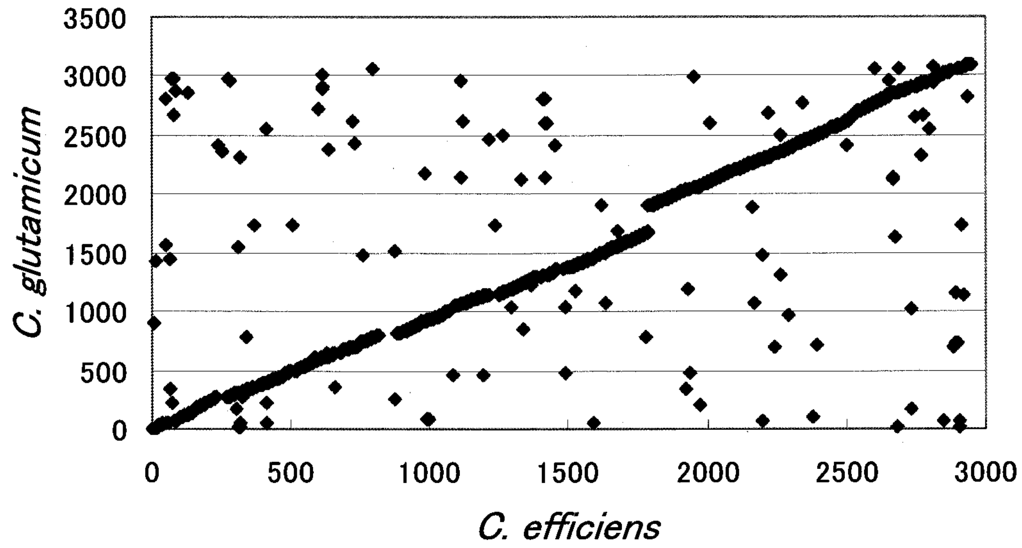


Figure 2.3 (Continued) GC content of three corynebacterial genomes

Predicted locations of ORFs of *C. efficiens* (A), *C. glutamicum* (B) and *C. diphtheriae* (C). The color shows the average GC content per 500 bp from the beginning of the chromosome. Yellow is set at 57%, the average GC content of the three genomes. The arrows give the locations of the ORFs and their direction.

A


3500
3000
2500
2000
1500
1000
500
0
C. glutamicum
0
500
1000
1500
2000
2500
3000
C. efficiens

B

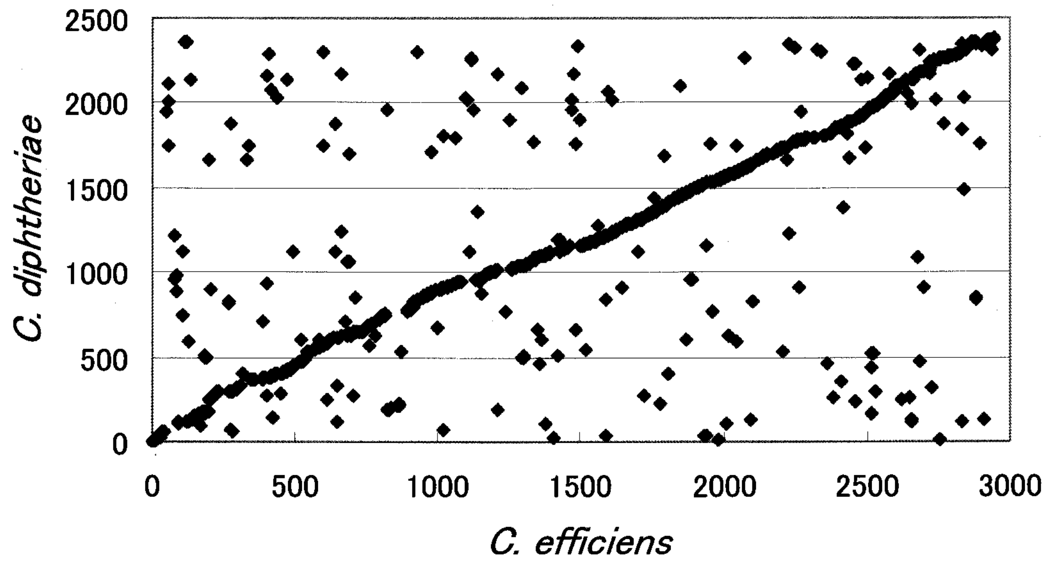

2500
2000
1500
1000
500
0
C. diphtheriae
0
500
1000
1500
2000
2500
3000
C. efficiens

Figure 2.4

C

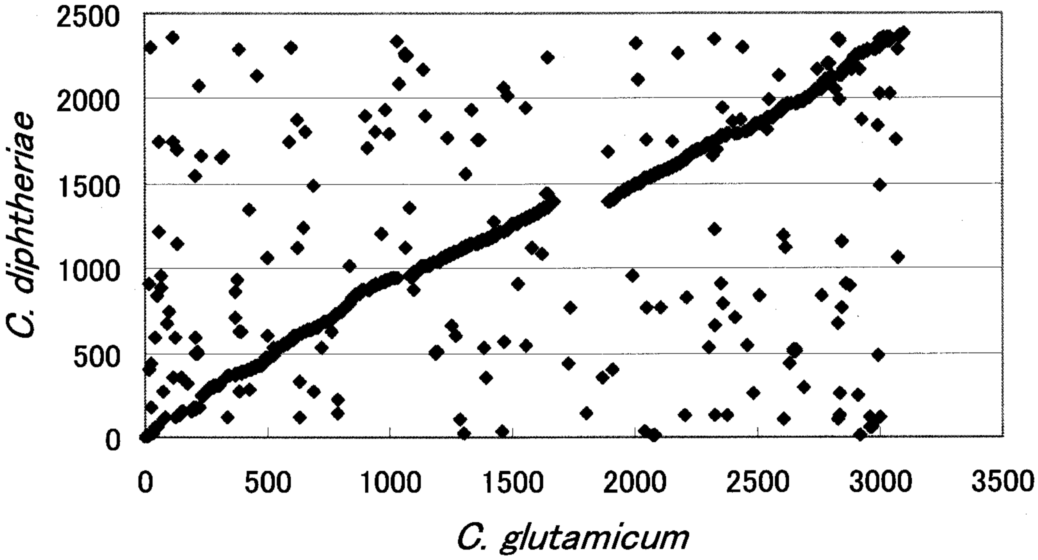


Figure 2.4 (Continued) Comparison of gene order

C. efficiens versus *C. glutamicum* (A), *C. efficiens* versus *C. diphtheriae* (B) and *C. glutamicum* versus *C. diphtheriae* (C). Axes represent the order of the ORFs predicted by the Glimmer program and the numbers represent the ORF ID numbers derived from the genome annotation. For each genome, the *dnaA* gene was the first gene. Dots represent the orthologous ORFs in each pair of species.

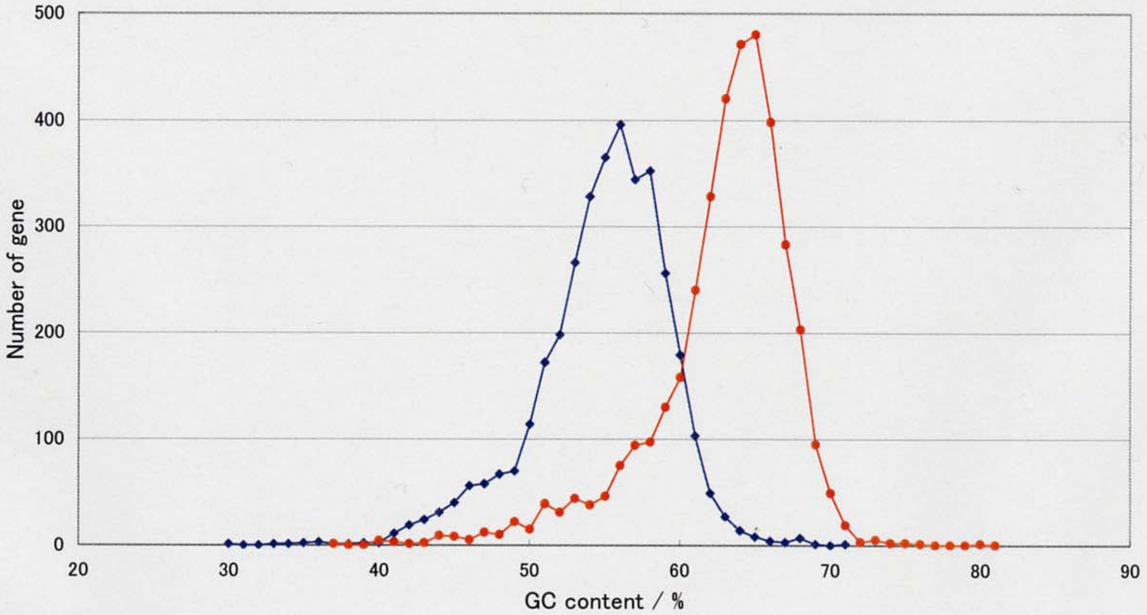


Figure 2.5 The GC content of the ORFs of *C. efficiens* and *C. glutamicum*

Numbers of ORFs are plotted against percentage GC content. Red, *C. efficiens*; blue, *C. glutamicum*.

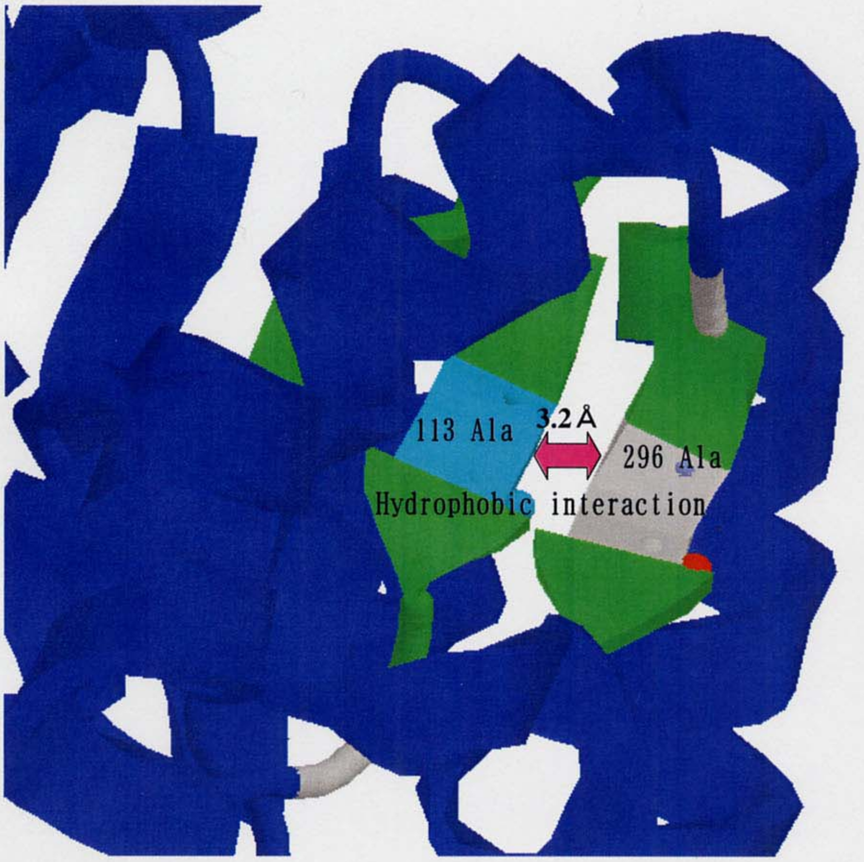


Figure 2.6 Proposed hydrophobic interaction in *C. glutamicum* Ddh

The residue ^{113}Ala is substituted to Ser in *C. efficiens*. This substitution may destroy hydrophobic interaction and destabilize the protein structure. Flat arrows represent β-sheet.

Chapter 3

Evolution of amino acid biosynthesis in *Corynebacteria*

3.1 Introduction

The genomes of several industrially useful bacteria as well as of pathogenic bacteria have been sequenced (Nelson et al., 2000), and one can now compare several genome sequences simultaneously. The reconstruction of metabolic pathways from genome sequences by means of tools such as KEGG (Kanehisa, 1997) and WIT (Overbeek et al., 2000) can provide much useful information on differences in the metabolic pathways of bacteria that lead to variation in growth characteristics and in ability to assimilate different substances. It is especially important in applied technology to understand differences in metabolic pathways and their evolutionary history. Until now, there have been only a few studies of metabolic pathways based on comparison of the genome sequences of closely related species (Marais et al., 1999). The same is true of comparisons of specific pathways in more distantly related microorganisms with the aim of accounting for their differences from a biological and evolutionary point of view (Boucher and Doolittle, 2000; Lange et al., 2000).

We have previously sequenced and annotated the genome of *Corynebacterium efficiens* (Nishio et al., 2003). This bacterium is a close relative of *Corynebacterium glutamicum,* which has been widely used in the industrial production of glutamate, lysine and other amino acids by fermentation. The two species are recognized as glutamic acid-producing *Corynebacteria* (Fudou et al., 2002). The optimal temperature for glutamate production by *C. glutamicum* is around 30ºC, and it does not grow or produce glutamate at 40ºC or above. On the other hand, *C. efficiens* can grow and produce glutamate above 40ºC. On the basis of genome comparisons between these two species, three kinds of amino acid substitutions were suggested to be responsible for the thermostability of *C. efficiens* and the increase of 10% in genome GC contents in *C. efficiens* (Nishio et al., 2003). In addition, the comparative genome sequence analysis suggested that the absence of a RecBCD pathway may have been responsible for suppressing genome shuffling in *Corynebacterium* (Nakamura et al., 2003). One of our research interests is the extent to which the genetic control of amino acid biosynthesis differs between these closely related species. It is well known that *C. glutamicum* overproduces glutamic acid under a variety of conditions such as biotin limitation (Kimura, 2003). We are interested in the evolutionary events responsible for the acquisition of this feature. Furthermore, *C. glutamicum* also overproduces lysine, arginine, threonine, isoleucine, valine, serine, tryptophan, phenylalanine and histidine (Supplementary Table 5; Ikeda, 2003). It is therefore of great interest to investigate the evolutionary processes involved in the acquisition of these productive capabilities. *Corynebacterium diphtheriae* is a well-known pathogen (Collins and Cummins 1986, Graevenitz and Krech 1991) whose genome has been sequenced by the Sanger Center (Cerdeno-Tarraga et al., 2003). Although the main focus of interest in the study of *C. diphtheriae* has been its pathogenicity, we were interested in understanding the evolutionary process of functional differentiation between the amino acid producing species and this pathogenic strain.

From the complete genome sequence data, some of the industrially useful phenotypes are suggested to be acquired by horizontal gene transfer and gene duplication. For example, *Streptomyces* may have acquired many genes for the antibiotics production by gene duplication (Bentley et al., 2002). From the comparison of the complete genome sequences for closely related species, the functional differentiation among species will be clarified. Different phenotypes in closely related species have originated from the difference of the gene contents and regulatory mechanisms of genes. The comparison of complete genome sequences enables us to know the difference of gene content and regulation. To understand the difference of gene contents among *Corynebacteria* should be the first step for the study of a regulatory system of amino acid overproduction. Making the comparison of gene contents among *Corynebacteria* using the complete genome sequences, we conducted our study to understand when the amino acid overproduction system in *C. glutamicum* was acquired. Our analysis showed that the common ancestor of *Corynebacterium* had already possessed almost all the genes needed for the overproduction of amino acids, and that *C. diphtheriae* lost many of these genes. However, the difference in gene contents between glutamic acid-producing *Corynebacteria* and *C. diphtheriae* may account for the amino acid productivity in *C. glutamicum*. Both of *C. efficiens* and *C. glutamicum* acquired several genes that may be important for amino acid production, after their divergence from the common ancestor of the three *Corynebacteria*. Furthermore, the genes controlling amino acid biosynthesis differentiate *C. glutamicum* from *C. efficiens*. In particular, *C. efficiens* possesses a paralogous gene encoding glutamine synthetase I that may be responsible for its differences from *C. glutamicum* in glutamic acid productivity. Our results suggest that gene transfer and gene loss in *Corynebacterium* were responsible for functional differentiation of the three bacterial species; the emergence of features favoring the capacity for amino acid production, and acquirement of pathogenicity against human.

3.2 Materials and Methods

3.2.1 ORF prediction

The ORFs were selected by using Glimmer 2.0 (Delcher et al., 1999). Glimmer used the "open reading frames" with longer than 500bp as the learning data set for the ORF prediction. Then dozens of ORFs that were rejected by Glimmer but those were selected using the learning dataset constructed with ribosomal protein and tRNA synthetase or but with more than 40 score in Smith-Waterman homology search (Smith and Waterman, 1981) against SWISS-PROT database, were added to the original ORF set.

3.2.2 Genome annotation

We gave the following kinds of annotations to *C. efficiens* genome sequence. In the case that protein functions have been demonstrated experimentally in closely related species such as *Corynebacterium glutamicum, Brevibacterium flavum,* or *Corynebacterium ammoniagenesis,* the same database descriptions of products information, gene name and EC number were adopted in *C. efficiens*. When the product information was established in distant species such as *Escherichia coli,* it was treated that the responsible ORF in *C. efficiens* has a putative function. When ORFs had significant homologies against database entries but functions were not clear, they were annotated as conserved hypothetical protein. If ORFs had no homologies against database entries, they were treated as a hypothetical protein. After products information identified, terms were integrated. In the case that there were alternative names in products, the protein name entry in SWISS-PROT was used as the product name. The overlapped ORFs were deleted when they were the hypothetical proteins, or when they had delayed or rejected information in the Glimmer output file. The ORFs shorter than 150 bp were also deleted when they were the hypothetical proteins.

The complete genome sequences of *C. efficiens* (Nishio et al., 2003), *C. glutamicum* (Ikeda and Nakagawa, 2003; Kalinowski et al., 2003), *C. diphtheriae* (Cerdeno-Tarraga et al., 2003), *Mycobacterium tuberculosis* (Cole et al., 1998), *Mycobacterium leprae* (Cole et al., 2001) and *Streptomyces coelicolor* A3(2) (Bentley et al., 2002) were obtained from DDBJ/EMBL/Genbank (accession numbers: BA000035, BA000036, BX248353, AL123456, AL450380 and AL645882, respectively). In the case of *S. coelicolor*, we also used the Web server (http://jic-bioinfo.bbsrc.ac.uk/S.coelicolor/index.html).

3.2.3 Phylogenetic analysis

The BLAST (Altschul et al., 1997) and FASTA (Pearson, 2000) programs were used for database searches, and ClustalW (Thompson et al., 1997) for multiple alignments. Phylogenetic trees were constructed by the neighbor-joining method with p-distance or Kimura's distance (Saitou and Nei, 1987). Estimates of synonymous (Ks) and nonsynonymous (Ka) per sites and standard deviations were calculated using Li's method (Li, 1993) implemented in DAMBE (Xia and Xie, 2001). We also used the Nei and Gojobori (Nei and Gojobori, 1986) method, but it gave virtually the same results.

3.2.4 Comparison of gene contents in *Corynebacterium*

Multiple alignments and phylogenic trees were constructed of the high-GC Gram-positive bacteria, *C. efficiens*, *C. glutamicum*, *C. diphtheriae*, *M. tuberculosis*, *M. leprae* and *S. coelicolor*, using all highly conserved proteins involved in amino acid biosynthesis. Criteria for highly conserved sequences were defined using the FASTA program. The query sequences used in the FASTA program searches were from *C. glutamicum* or *C. efficiens*. The Z-scores of the FASTA program, identities of overlapping regions, and detected sequence lengths were used to establish the highly conserved sequences. All alignments were checked manually. The highly conserved gene pairs which defined above selected as the paralogous gene set. And those phylogentic relations were also checked manually.

3.3 Results

3.3.1 Differences between *C. efficiens* and *C. glutamicum* in genes related to amino acid biosynthesis

To evaluate the evolutionary processes that led to the biological capacity for amino acid production on a large scale, we collected the amino acid sequences of amino acid biosynthetic enzymes and related enzymes from genome annotations of a number of high-GC Gram positive-bacteria. All the phylogenetic trees of these enzymes were compared with a 16S rRNA-based phylogenetic tree (Fig. 3.1). In this phylogenetic tree, *C. diphtheriae* diverged from the common ancestor of *Corynebacteria*, and after that, *C. efficiens* and *C. glutamicum* diverged from the common ancestor of glutamic acid producing *Corynebacteria.* This representative topology of a phylogeny was supported by the phylogenetic trees for most translation/transcription-related genes. We found that only 5 amino acid biosynthesis-related genes possessed their paralogous genes in glutamic acid-producing *Corynebacteria* (Table 3.1 and Supplementary Table 7). The topology of phylogenetic tree for the five genes was shown to be distinctively different from the representative topology of a phylogeny in High GC gram-positive bacteria. Not only by the NJ method (Saitou and Nei, 1987) but also by the maximum likelihood method (Adachi and Hasegawa, 1996), the same topologies were obtained for each of five genes and 16S rRNAs, respectively. Four of the five genes encouraged us to study the genome structure such as gene transfer/duplication/loss in glutamic acid-producing *Corynebacteria* because of their tree topologies, multiple alignments and operon structures. We focused a further analysis on four genes: *trpB* (encoding tryptophan synthase beta chain), *ilvD* (encoding dihydroxy-acid dehydratase), *aroQ* (encoding 3-dehydroquinate dehydratase), and *glnA* (glutamine synthetase I).

In the phylogenetic tree of TrpB, the *C. efficiens* (CE2880) and *C. diphtheriae* (DIP2351) were positioned outside the orthologues of *C. glutamicum* (CE2872, Cglu3034, DIP2360) (Fig. 3.2A). The location on the genome of the paralogue *trpB* (CE2880 and DIP2351) was very close to that of the orthologue *trpB* (CE2872 and DIP2360) in *C. efficiens* and *C. diphtheriae*. From these results, we suggest that gene duplication took place in the common ancestor of the *Corynebacterium*, and that gene loss was responsible for the single copy of this gene in *C. glutamicum*.

We constructed a multiple alignment and a phylogenetic tree of IlvD, again using high-GC Gram-positive bacterial sequences. In the phylogenetic tree, the highly conserved sequence in *Bacillus subtilis,* a low GC Gram-positive bacterium, was used as outgroup (Supplementary Fig. 2 and Fig. 3.2B, respectively). This phylogenetic tree contained two clusters whose topologies were unlike the trees obtained from the 16S rRNA sequences (Fig. 3.1, 3.2B). In the multiple alignment of IlvD, a large insertion was observed between positions 412 and 450 in *C. efficiens* CE1362, *C. glutamicum* Cgl1268, *C. diphtheriae* DIP1096 and *S. coelicolor* SCO3345(Supplementary Fig. 2), and these four sequences were clustered in the phylogenetic tree. A large insertion was also observed in multiple alignment of the dehydratase family (PfamA, ILVD_EDD). It implies that this insertion took place a long time ago, even before the emergence of the common ancestor of high GC gram-positive bacteria.

The phylogenetic tree of AroQ was constructed in the same manner as IlvD and its topology also differed from the 16S rRNA-based phylogenetic tree (Fig. 3.1, 3.2C). *C. efficiens* CE1739, *C. diphtheriae* DIP1342, *C. pseudotuberculosis, M. leprae*

ML0519 and *M. tuberculosis* Rv2537c form a cluster in the phylogenetic tree. *AroQ* in *C. efficiens* CE1739, *C. diphtheriae* DIP1342, *M. leprae* ML0519 and *M. tuberculosis* Rv2537c is part of the *aroCKBQ* operon. Another AroQ cluster was composed of an additional *aroQ* in *C. efficiens* CE0442, *C. glutamicum* Cgl0423 and *S. coelicolor* SCO1961. The additional *aroQ* in *C. efficiens* CE0442 and *aroQ* in *C. glutamicum* Cgl0423 lie next to *aroE* on the chromosome, whereas in *S. coelicolor* SCO1961 there is no nearby aromatic amino acid biosynthesis gene. These results suggest that the evolution of the *aroQ* gene in high-GC Gram-positive bacteria was related to operon organization, and it is curious that *C. efficiens* retained two *aroQ* genes within conserved operon structures.

The phylogenetic tree of GlnA showed that the paralogous GlnA of *C. efficiens* CE2116 was positioned outside that of *C. diphtheriae* DIP1644 (Fig. 3.2D). This result suggests that *glnA* of *C. efficiens* CE2116 was not acquired by gene duplication within its own evolutionary linage (unless it is a pseudogene), but rather by gene duplication in the common ancestor of *Corynebacterium*, or by horizontal gene transfer. To find a more likely explanation, we compared the genome structures of the three *Corynebacteria* (Fig. 3.3). In *C. efficiens* and *C. diphtheriae*, there were additional genes next to orthologous GlnA than in *C. glutamicum.* These additional genes are from CE2105 to CE2116 in *C. efficiens* and DIP1644 to DIP1661 in *C. diphtheriae,* as shown in Fig. 3. These genes were dissimilar at both the DNA and amino acid levels, implying that they were acquired independently in each species. The GC contents of these additional regions were 61.9% in *C. efficiens* and 50.2% in *C. diphtheriae*. The *C. diphtheriae* specific genes are annotated as putative phage-related and antibiotic resistance-related pathogenicity island and showed unusual GC-contents and dinucleotide signature (Cerdeno-Tarraga et al. 2003). This result suggested that the *C. diphtheriae* specific genes were acquired by horizontal gene transfer. On the other hand, the paralogous *ocd* gene (CE2115) encoding ornithine cyclodeaminase and the paralogue *glnA* (CE2116) (Fig. 3.3) were *C. efficiens*-specific genes. The paralogous *ocd* gene (CE2115) was located next to the paralogue *glnA* (CE2116) in *C. efficiens*. The phylogenetic tree of Ocd showed that the paralogous Ocd of *C. efficiens* (CE2115) was positioned outside the orthologous corynebacterial Ocd (CE1700, Cgl1582) (Fig. 3.2E). Moreover, *C. diphtheriae* has lost the *ocd* gene. The orthologous *ocd* gene was not located near the orthologous *glnA* in *C. efficiens* CE2104 and *C. glutamicum* Cgl2214, suggesting that the paralogous *glnA* (CE2116) and paralogous *ocd* (CE2115) genes of *C. efficiens* were not acquired by gene duplication in the common ancestor of *Corynebacterium*: A possible explanation was due to the lack of a RecBCD pathway (Nakamura et al., 2003), genome rearrangement could not take place, and duplicated genes must remain close to where they originate. Another possible explanation was that it was a pseudogene. An analysis of the number of nonsynonymous versus synonymous substitutions showed a larger number of nonsynonymous substitutions in the paralogous *glnA* of *C. efficiens* (CE2116) than in the orthologous corynebacterial *glnA* (CE2104, Cgl2214, DIP1644); however it was not as high as in *Mycobacterium*, and GC content analysis showed that there was no difference in the 2nd position GC content (Tables 3.2, 3.3). If paralogous *glnA* gene was a pseudogene on which there were no functional constraints, a significant difference would be observed in the 2nd position GC content of paralogous gene when comparing with that of orthologous gene. Evidently, the paralogous *glnA* of *C. efficiens* (CE2116) is not a pseudogene, but was acquired by horizontal gene transfer.

3.3.2 Newly acquired genes in amino acid producing species

The genome of *C. diphtheriae* comprises 2,488,635 bp, thus being smaller than those of other high-GC Gram-positive bacteria (Cerdeno-Tarraga et al., 2003). The evolutionary origin of this small genome must have been either massive gene loss in *C. diphtheriae,* or massive gene acquisition in the other high-GC Gram-positive bacteria. To clarify the evolutionary event responsible, we identified the common orthologous genes in the five high-GC Gram-positive bacteria, *C. efficiens*, *C. glutamicum*, *C. diphtheriae*, *M. tuberculosis* and *S. coelicolor*, by the reciprocal best-hit method using BLAST (Mineta et al., 2003), as well as four species excluding one of *Corynebacteria.* There were 748 orthologous genes in the five bacteria, 768 when excluding *C. glutamicum*, 773 when excluding *C. efficiens* and 831 when excluding *C. diphtheriae*. This shows that it is likely that *C. diphtheriae* lost many orthologues that were found in the four other bacteria after it diverged from the common ancestor of the *Corynebacterium*. *C. diphtheriae* has lost many genes present in the sister species; for example, *gltBD, metE, metB, malE, cysH, cysI, cysN* and *cysD* are missing from *C. diphtheriae*, but present in *C. efficiens, C. glutamicum* and the outgroup bacteria (Table 3.1, Supplementary Table 7, and Fig. 3.4).

C. diphtheriae does not possess a paralogous pyruvate kinase (*pyk2*) or phosphoenolpyruvate synthase (*pps*) gene in the anaplerotic pathway, nor an *aroG* encoding 3-deoxy-D-arabinoheptulosonate-7-phosphate synthase in aromatic amino acid biosynthesis, or a diaminopimelate dehydrogenase (*ddh*) gene in lysine biosynthesis. These genes are also absent from the other high-GC Gram-positive bacteria (Supplementary Table 7). There are only two homologues of Pyk2 of *C. glutamicum* and *C. efficiens* among known protein sequences. One is in

Thermosynechococcus elongates, a kind of *Cyanobacterium* and the other in *Arabidopsis thaliana.* In *C. efficiens* and *C. glutamicum, pps* (CE0560 and Cgl0551) and *pps2* (CE0561 and Cgl0552) are adjacent to each other. The N-terminal region of the *pps2* of *Corynebacteria* (CE0561 and Cgl0552) is similar to that of bacterial phosphoenolpyruvate synthase. We found only one species from known protein sequences that has these two homologues in the same arrangement, and they were isolated as putative phenol phosphorylation related genes (Breinig et al., 2000). *Bacillus sphaericus* and *Clostridium tetani* have homologous Ddh sequences at the amino acid level. Together these results suggest that *pyk2, pps, pps2* and *ddh* were acquired by the common ancestor of the amino acid producing species, rather than having been lost in *C. diphtheriae*. There are no homologues of AroG in *Mycobacterium* or *Streptomyces* among known protein sequences: However, other high-GC Gram-positive bacteria, *Actinomycetales*, *Thermobifida fusca* and *Amycolatopsis mediterranei* have highly conserved sequences. We infer that *aroG* was lost in *C. diphtheriae*, *Mycobacterium* and *Streptomyces,* but retained in *C. efficiens* and *C. glutamicum.*

One of the biologically important characteristics in *C. glutamicum* is that it has been known to be a biotin requirement organism (Kimura, 2003). The biotin requirement is also observed in *C. efficiens*. These bacteria lack the complete biotin biosynthesis pathway from pimelate to biotin. Glutamic acid overproduction in *C. glutamicum* is due to the shortage of biotin (Kimura, 2003). Interestingly, *C. diphtheriae* may not be a biotin requiring organism because it possesses the complete biotin biosynthesis pathway. From this reason, it is strongly speculated that *C. diphtheriae* does not possess the glutamic acid overproduction mechanism induced by the biotin limitation. Moreover, in *C. diphtheriae*, DIP1381 encoding 6-carboxyhexanoate—CoA ligase as the first enzyme in biotin biosynthesis, may have been acquired by horizontal gene transfer in *C. diphtheriae* (Table 3.1 and Supplementary Table 7). This is because any other high GC gram-positive bacteria except *C. diphtheriae* did not possess orthologous genes of DIP1381.

3.4 Discussion

Why do the glutamic acid-producing *Corynebacteria* have such a remarkable capacity for producing many different amino acids? To answer this question from an evolutionary point of view, we reconstructed metabolic pathways using the complete genome sequences of high-GC Gram-positive bacteria, and made a detailed comparison of their pathway genes. We first tried to determine whether *C. efficiens* and *C. glutamicum* had acquired the genes necessary for amino acid overproduction. Our analysis suggested that other high-GC Gram-positive bacteria had orthologues for most of the characteristic genes needed for amino acid overproduction in *C. glutamicum* (Vrljic et al., 1996; Kimura et al., 1996; Kimura, 2003; Simic et al., 2001). In a previous study, 2,101 orthologues were identified between *C. efficiens* and *C. glutamicum* (Nakamura et al., 2003). Only 177 orthologues failed to have any homologues in *C. diphtheriae*, *Mycobacterium* and *Streptomyces*. These results suggest that the capacity for overproducing amino acids was inherited from a common ancestor, and that actual overproduction may have emerged in the course of evolution of glutamic acid-producing *Corynebacteria.*

The loss of genes in *C. diphtheriae* may be correlated with its loss of amino acid production capability. Our analysis suggested that *C. diphtheriae* has lost many genes present in the common ancestor and that this is reflected in its genome size. *C. diphtheriae* lacks the genes *gltBD*, *ddh*, *metE* and *metB* whose products encode redundant pathways for glutamate, lysine and methionine biosynthesis in the amino acid producing species (Fig. 3.4). Surprisingly, it has also lost all genes of the sulfur incorporation pathway, suggesting that it cannot synthesize cysteine. The addition of cysteine was critical for toxin production and cell growth of *C. diphtheriae* (Nagarkar et al., 2002), consistent with the absence of the sulfur incorporation pathway.

To estimate what evolutionary events are needed for the capacity for amino acid overproduction in the industrially useful *Corynebacteria* genome, we compared amino acid biosynthesis related genes of *C. efficiens* and *C. glutamicum*. We found that although amino acid biosynthesis pathways were well conserved, the number of paralogues related to amino acid biosynthesis differed (Table 3.1). Our phylogenetic analysis suggested that the paralogues *glnA* (CE2116) (Schulz et al., 2001) and *ocd* (CE2115) of *C. efficiens* were acquired by horizontal gene transfer. If acquisition of *ocd* and *glnA* paralogous genes was made together, then the creation of ammonia recycle pathway can be achieved in *C. efficiens*. Gene transfer may therefore be one of the important factors in the evolution of the amino acid producing species.

Choice for particular genes may also have been important in the evolution of bacterial phenotypes. In the phylogenetic tree of AroQ (Fig. 3.1C), one cluster contains only non-pathogenic bacteria, and another pathogenic bacteria other than *C. efficiens* (the pathogenic cluster). This was the only phylogenetic tree of all the phylogenetic trees for amino acid biosynthesis-related genes in the high-GC Gram-positive bacteria to show that *Corynebacteria* are separated into two clusters of pathogens and non-pathogens. One possible evolutionary explanation is that gene duplication occurred in the common ancestor of the high-GC Gram-positive bacteria and that, as a result of the choice, each species, except *C. efficiens*, lost one of the two *aroQ* genes depending on their phenotypic features. Mutation of the common aromatic amino acid biosynthetic gene for the inhibition of the folic acid biosynthesis is one of the strategies for vaccine development against pathogenic bacteria. In fact, it has been observed that the growth in more than 10 pathogens was attenuated by single mutation of aromatic amino acid biosynthesis related genes (*aro* genes) (Simmons et al., 1997). In *C. pseudotuberculosis*, mutation of *aroQ* weakened its pathogenicity in the mouse (Simmons et al., 1997). Thus, *aroQ* may be related to pathogenicity.

To understand the phylogeny of IlvD, there are two possible evolutionary events; ancient gene duplication or horizontal gene transfer (Fig. 3.2B). Our results suggested that *ilvD* in *C. efficiens* was acquired by an ancient gene duplication rather than horizontal gene transfer. In the case of TrpB, the phylogenetic tree clearly showed that gene duplication had occurred in the common ancestor, and that *C. glutamicum* may have lost the duplicated ORF (Fig. 3.2A). The paralogous *trpB* was located near the orthologous *trpB* in *C. efficiens* and *C. diphtheriae*. This location in the *Corynebacteria* supports the rule that duplicated genes are located next to one another due to the absence of genome rearrangement resulting from the lack of a RecBCD pathway (Nakamura et al., 2003). It has been proposed that the paralogous *trpB* in *C. diphtheriae* is a pseudogene because of the long branch length (Xie et al., 2002). However, persistence of this paralogue in *C. diphtheriae* but not in *C. glutamicum* seems strange because *C. diphtheriae* seems to have lost many genes during its evolution and the selective pressure to discard unnecessary genes appears to have been much higher in its case than in *C. glutamicum*.

Our findings suggested that almost all the genes required for amino acid production already existed in the common ancestor of *Corynebacterium*. We also believe that newly acquired genes in glutamic acid-producing *Corynebacteria* contribute to amino acid overproduction capacity. Actually, *ddh*, one of the newly acquired genes in the amino acid producing species, has been known to contribute to lysine production in *C. glutamicum.* An interesting question is whether the newly acquired and previously unrecognized enzyme phosphoenolpyruvate synthase in *C. efficiens* and *C. glutamicum* contributes to their ability to overproduce amino acids. Previous studies of glutamate and lysine production have not highlighted the existence of this enzyme. For example, Park et al. (1997) did not assume this enzyme in the flux calculation for lysine production in *C. glutamicum.* In *E. coli*, the same enzyme plays an important role in the production of aromatic compounds (Yi et al., 2002), and furthermore, *pps* and its homologue in *T. aromatica* were isolated as phenol-induced proteins (Breinig et al., 2000). In fact, *aroG* encoding 3-deoxy-D-arabino-heptulosonate-7-phosphate synthase, which is on the aromatic amino acid biosynthesis pathway, may have been retained in *C. efficiens* and *C. glutamicum* although it was lost in *C. diphtheriae, Mycobacteria* and *Streptomyces*. As the gene for benzoate 1,2-dioxygenase reductase, which is related to genes for benzoate degradation (CE2306, Cgl2405), was newly acquired in amino acid producing species (Table 3.1), phosphoenolpyruvate synthase may cooperate with that gene in these *Corynebacteria.* Thus, newly acquired genes may also contribute to productivity of amino acids. Small numbers of those genes homologues are found among known protein sequences. Therefore, they may be have been acquired by horizontal gene transfer.

We have now shown differences in gene contents among *Corynebacteria.* It may give us a clue for elucidating the regulatory mechanisms for amino acid overproduction. Although we do not know the regulatory sequences related to glutamic acid production in *C. glutamicum*, the comparison of regulatory regions of glutamate overproduction related-genes between different species may lead to an overview of the regulation for amino acid production mechanism. In *C. glutamicum*, there may be a strong relationship between the attenuation of the 2-oxoglutarate dehydrogenase (ODH) activity and glutamic acid production (Shimizu et al., 2003). One of our interests is the similarity of the regulatory regions among three *Corynebacteria*. The regulatory regions of *odhA* gene encoding ODH were more strongly conserved between *C. efficiens* and *C. glutamicum* than between *C. diphtheriae* and *C. glutamicum* or *C. efficiens* (Supplementary Fig.3). On the other hand, enhanced glutamate dehydrogenase (GDH) activity may not contribute to glutamic acid production (Shimizu et al., 2003). The conservation of regulatory regions for *gdh* genes encoding GDH were almost the same (Supplementary Fig.4). These results are consistent with the previous knowledge of glutamic acid production, suggesting the lack of glutamic acid overproduction mechanism in *C. diphtheriae*. It is also supported by the complete biotin biosynthesis pathway of *C. diphtheriae*. As mentioned earlier, the comparison of regulatory regions among three *Corynebacteria* may be important for studying regulatory systems of amino acids production. In this case, we may have to assume that the important part of regulatory regions is conserved in spite of a difference in the genome GC contents. Note that the genome GC content of *C. efficiens* was 10% higher than that of *C. glutamicum* or *C. diphtheriae* (Nishio et al., 2003).

In this study, we have attempted to analyze the evolutionary process by which the capacity for amino acid overproduction was acquired by glutamic acid-producing *Corynebacteria*. Gene transfer/duplication/loss events in *Corynebacteria* may facilitate the formation of amino acid overproduction mechanisms. Retention of ancestral genes and gain of new genes by horizontal gene transfer may have been the major motive forces in establishing their capability for amino acid overproduction, while gene loss may have resulted in the loss of that capacity by *C. diphtheriae*. We have also found some genes that may be responsible for different amino acid productivity between *C. efficiens* and *C. glutamicum* by comparison and detailed analysis of their genome sequences. Experimental analysis will be needed to clarify the contribution of these genes and their regulatory sequences to the overproduction of amino acids.

Table 3.1 The summary of amino acid biosynthesis related genes examined in this study

product	gene name	*C. efficiens*	*C. glutamicum*	*C. diphtheriae*
6-carboxyhexanoate–CoA ligase				DIP1381
3-dehydroquinate dehydratase	*aroQ*	CE0442 CE1739	Cgl0423	DIP1342
3-deoxy-D-arabino-heptulosonate-7-phosphate synthase	*aroG*	CE1054	Cgl0990	
3-deoxy-D-arabino-heptulosonate-7-phosphate synthase	*aroH*	CE2073	Cgl2178	DIP1616
5-methyltetrahydropteroyltriglutamate–homocysteine methyltransferase	*metE*	CE1209	Cgl1139	
citrate synthase	*gltA*	CE0905	Cgl0829	DIP0785
citrate synthase		CE0718	Cgl0659 Cgl0696	
detergent sensitivity rescuer DtsR	*dtsR*	CE0738	Cgl0708	DIP0658
detergent sensitivity rescuer DtsR homolog	*dtsR2*	CE0737	Cgl0707	DIP0660
diaminopimelate dehydrogenase	*ddh*	CE2498	Cgl2617	
dihydroxy-acid dehydratase	*ilvD*	CE1362 CE2439	Cgl1268	DIP1096
glutamate synthase large subunit	*gltB*	CE0158	Cgl0184	
glutamate synthase small subunit	*gltD*	CE0159	Cgl0185	
glutamine synthetase I	*glnA*	CE2104 CE2116	Cgl2214	DIP1644
lysine exporter protein	*lysE*	CE1357	Cgl1262	DIP1091
malic enzyme	*malE*	CE2839	Cgl3007	
O-acetylhomoserine (thiol)-lyase	*metB*	CE2343	Cgl2446	
phosphoenolpyruvate carboxylase	*ppc*	CE1703	Cgl1585	DIP1122
putative adenosine 5'-phosphosulphate reductase	*cysH*	CE2642	Cgl2816	
putative benzoate 1,2-dioxygenase reductase		CE2306	Cgl2405	
putative ferredoxin–nitrite reductase	*cysI*	CE2644	Cgl2817	
putative ornithine cyclodecarboxylase / cyclodeaminase	*ocd*	CE1700 CE2115	Cgl1582	

subunit 2				
putative sulfate adenylyltransferase subunit 1	*cysN*	CE2640	Cgl2814	
pyruvate carboxylase	*pyc*	CE0709	Cgl0689	DIP0641
pyruvate kinase	*pyk*	CE1989	Cgl2089	DIP1553
threonine export carrier protein	*thrE*	CE2506	Cgl2622	DIP1964
tryptophan synthase beta chain	*trpB*	CE2872 CE2880	Cgl3034	DIP2360 DIP2351

Table 3.2. Average number of Ks, synonymous, and Ka, nonsynonymous, substitution rates and Ks/Ka ratio for glutamine synthetase in *Corynebacterium* and *Mycobacterium*.

Ks ± SE[a]	Cgl2214	DIP1644	CE2116	ML0925	Rv2220
CE2104	0.7071 ± 0.0976	0.9229 ± 0.1295	0.7347 ± 0.1205	1.0973 ± 0.1891	1.0138 ± 0.3565
Cgl2214		0.8088 ± 0.1265	1.1220 ± 0.1900	1.4686 ± 0.2994	1.5947 ± 0.3460
DIP1644			1.6231 ± 0.3330	2.4380 ± 14.8729	1.8361 ± 0.4702
CE2116				1.3250 ± 0.3194	1.1132 ± 0.4601
ML0925					0.5924 ± 0.0719

Ka ± SE[a]	Cgl2214	DIP1644	CE2116	ML0925	Rv2220
CE2104	0.0423 ± 0.0108	0.1185 ± 0.0194	0.2013 ± 0.0248	0.2136 ± 0.0276	0.2188 ± 0.0281
Cgl2214		0.1234 ± 0.0205	0.1946 ± 0.0250	0.2211 ± 0.0279	0.2237 ± 0.0280
DIP1644			0.2304 ± 0.0265	0.2461 ± 0.0293	0.2483 ± 0.0298
CE2116				0.2718 ± 0.0298	0.2704 ± 0.0296
ML0925					0.0410 ± 0.0096

Ks/Ka	Cgl2214	DIP1644	CE2116	ML0925	Rv2220
CE2104	16.70	7.79	3.65	5.14	4.63
Cgl2214		6.56	5.76	6.64	7.13
DIP1644			7.05	9.91	7.40
CE2116				4.87	4.12
ML0925					14.44

[a] standard error

Table 3.3. GC contents of glutamine synthetase in *Corynebacterium* and *Mycobacterium*

	GC %	1stGC %	2ndGC %	3rdGC %
CE2104	62.8	60.9	40.2	86.6
Cgl2214	57.5	60.4	40.6	71.4
DIP1644	55.5	58.3	40.3	68.0
CE2116	60.6	61.2	42.3	82.4
ML0925	57.3	56.5	39.0	76.2
Rv2220	61.2	57.2	39.6	86.8

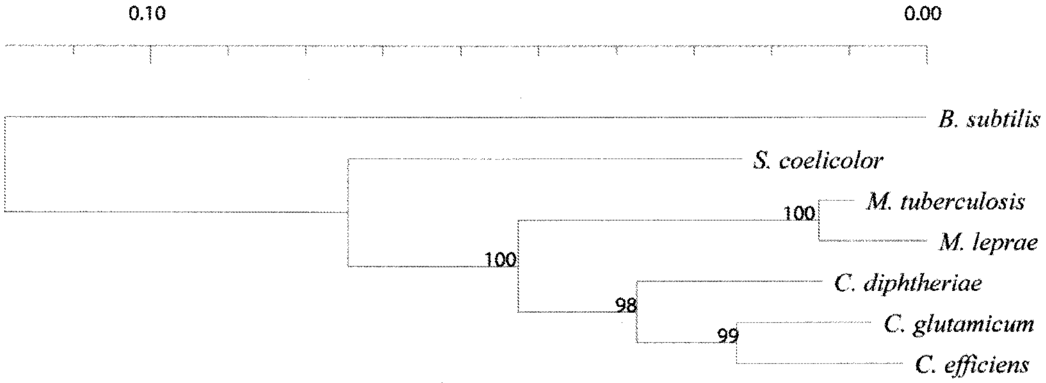


Figure 3.1 Phylogenetic tree of the 16S rRNA sequences of the high-GC Gram-positive bacteria examined in this study. *B. subtilis* was used as outgroup. The tree was constructed by the neighbor-joining method (Saitou and Nei 1987), and numbers indicate bootstrap values for 100 replications.

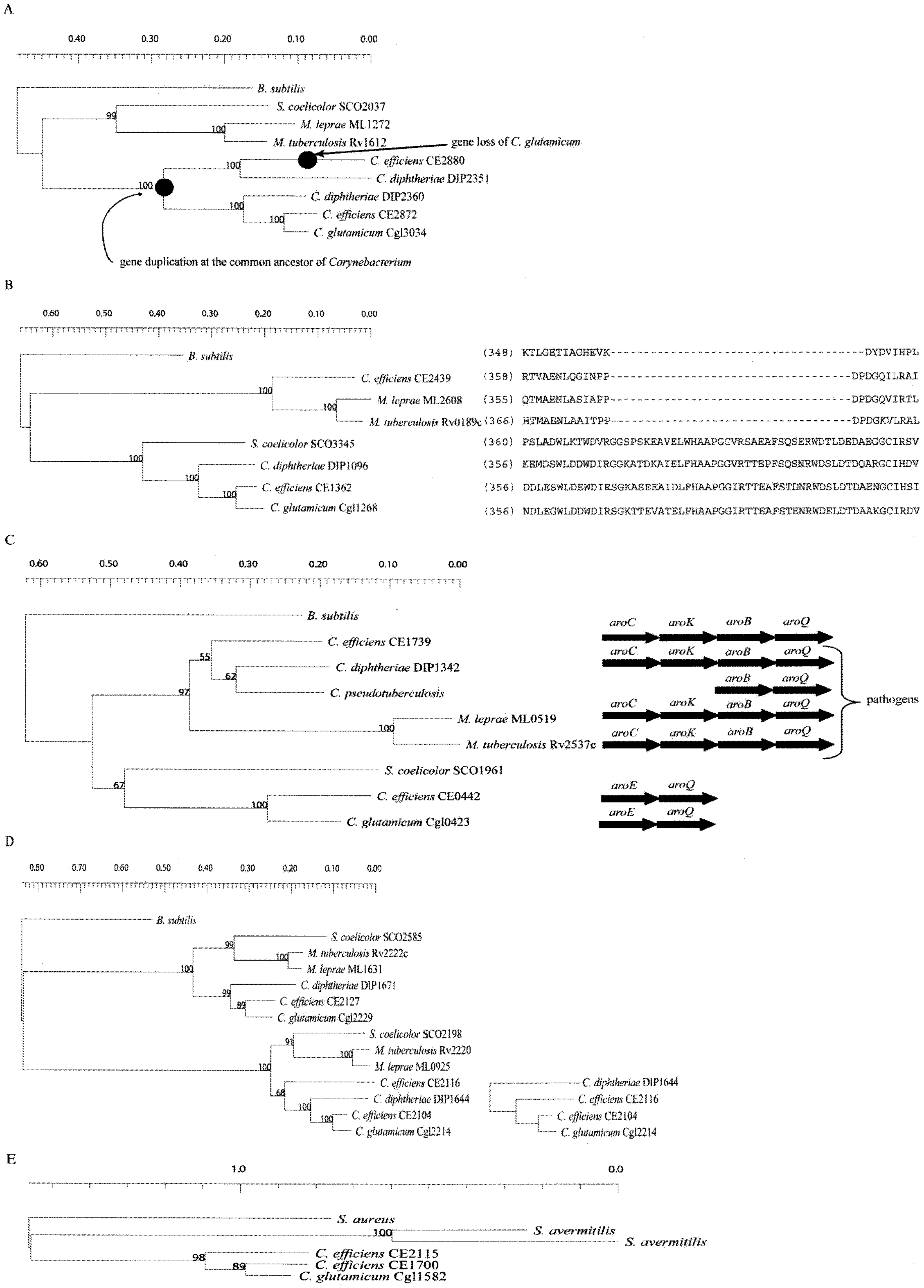

A
0.40
0.30
0.20
0.10
0.00
B. subtilis
S. coelicolor SCO2037
99
M. leprae ML1272
100
M. tuberculosis Rv1612
gene loss of C. glutamicum
100
C. efficiens CE2880
C. diphtheriae DIP2351
100
C. diphtheriae DIP2360
100
C. efficiens CE2872
100
C. glutamicum Cgl3034
gene duplication at the common ancestor of Corynebacterium
B
0.60
0.50
0.40
0.30
0.20
0.10
0.00
B. subtilis
C. efficiens CE2439
100
M. leprae ML2608
100
M. tuberculosis Rv0189c
S. coelicolor SCO3345
100
C. diphtheriae DIP1096
100
C. efficiens CE1362
100
C. glutamicum Cgl1268
(348) KTLGETIAGHEVK------------------------------------DYDVIHPL
(358) RTVAENLQGINPP----------------------------------DPDGQILRAI
(355) QTMAENLASIAPP----------------------------------DPDGQVIRTL
(366) HTMAENLAAITPP----------------------------------DPDGKVLRAL
(360) PSLADWLKTWDVRGGSPSKEAVELWHAAPGCVRSAEAFSQSERWDTLDEDAEGGCIRSV
(356) KEMDSWLDDWDIRGGKATDKAIELFHAAPGGVRTTEPFSQSNRWDSLDTDQARGCIHDV
(356) DDLESWLDEWDIRSGKASEEAIDLFHAAPGGIRTTEAFSTDNRWDSLDTDAENGCIHSI
(356) NDLEGWLDDWDIRSGKTTEVATELFHAAPGGIRTTEAFSTENRWDELDTDAAKGCIRDV
C
0.60
0.50
0.40
0.30
0.20
0.10
0.00
B. subtilis
C. efficiens CE1739
55
C. diphtheriae DIP1342
62
C. pseudotuberculosis
97
M. leprae ML0519
100
M. tuberculosis Rv2537c
S. coelicolor SCO1961
67
C. efficiens CE0442
100
C. glutamicum Cgl0423
aroC
aroK
aroB
aroQ
aroC
aroK
aroB
aroQ
aroB
aroQ
aroC
aroK
aroB
aroQ
aroC
aroK
aroB
aroQ
pathogens
aroE
aroQ
aroE
aroQ
D
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
B. subtilis
S. coelicolor SCO2585
99
100
M. tuberculosis Rv2222c
M. leprae ML1631
100
C. diphtheriae DIP1671
99
89
C. efficiens CE2127
C. glutamicum Cgl2229
S. coelicolor SCO2198
91
100
M. tuberculosis Rv2220
M. leprae ML0925
100
C. efficiens CE2116
68
C. diphtheriae DIP1644
100
100
C. efficiens CE2104
C. glutamicum Cgl2214
C. diphtheriae DIP1644
C. efficiens CE2116
C. efficiens CE2104
C. glutamicum Cgl2214
E
1.0
0.0
S. aureus
100
S. avermitilis
S. avermitilis
98
C. efficiens CE2115
89
C. efficiens CE1700
C. glutamicum Cgl1582

Figure 3.2 Phylogenetic trees of proteins related to amino acid biosynthesis in the high-GC Gram-positive bacteria

(a) tryptophan synthase beta chain (TrpB); TrpB in *B. subtilis* was used as outgroup. (b) dihydroxy-acid dehydratase (IlvD); IlvD in *B. subtilis* was used as outgroup. Sequences in the figure show the region of the multiple alignment which contains the most critical differences (see text). (c) 3-dehydroquinate dehydratase (AroQ): 3-dehydroquinate dehydratase (AroC) in *B. subtilis* was used as outgroup. Arrows in the figure show the operon structure. The complete genome sequence of *C. pseudotuberculosis* was not available. (d) glutamine synthetase (GlnA): GlnA in *B. subtilis* was used as outgroup. The right part of figure shows the tree assuming gene duplication in *C. efficiens*. The position of CE2116 should be positioned inside of *C. glutamicum* Cgl2214. (e) ornithine cyclodeaminase (Ocd): the ornithine cyclodeaminase homologues in *Streptomyces avermitilis* were used as a member of high GC gram-positive bacteria. The ornithine cyclodeaminase homologue in *Staphylococcus aureus* was used as outgroup. The numbers indicate bootstrap values for 100 replications.

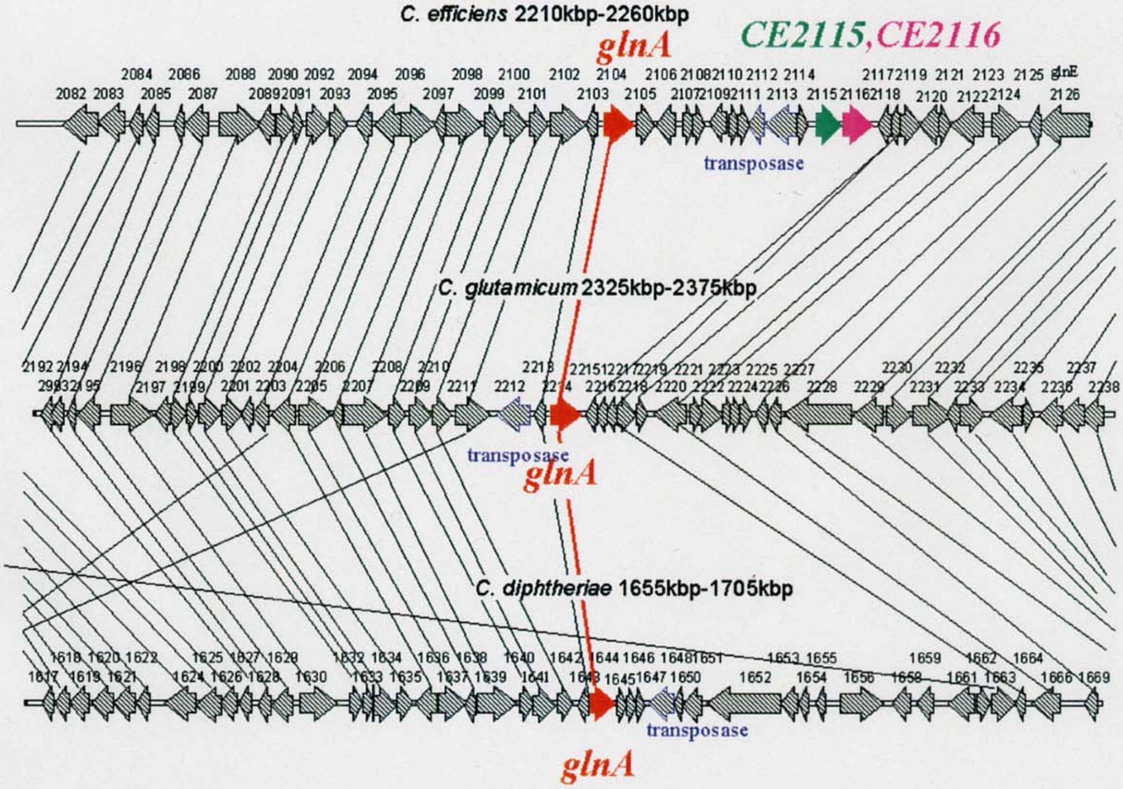


Figure 3.3 ORFs in the *glnA* region of *Corynebacteria*

The numbers correspond to the gene designations in each species. Orthologous genes are connected with lines.

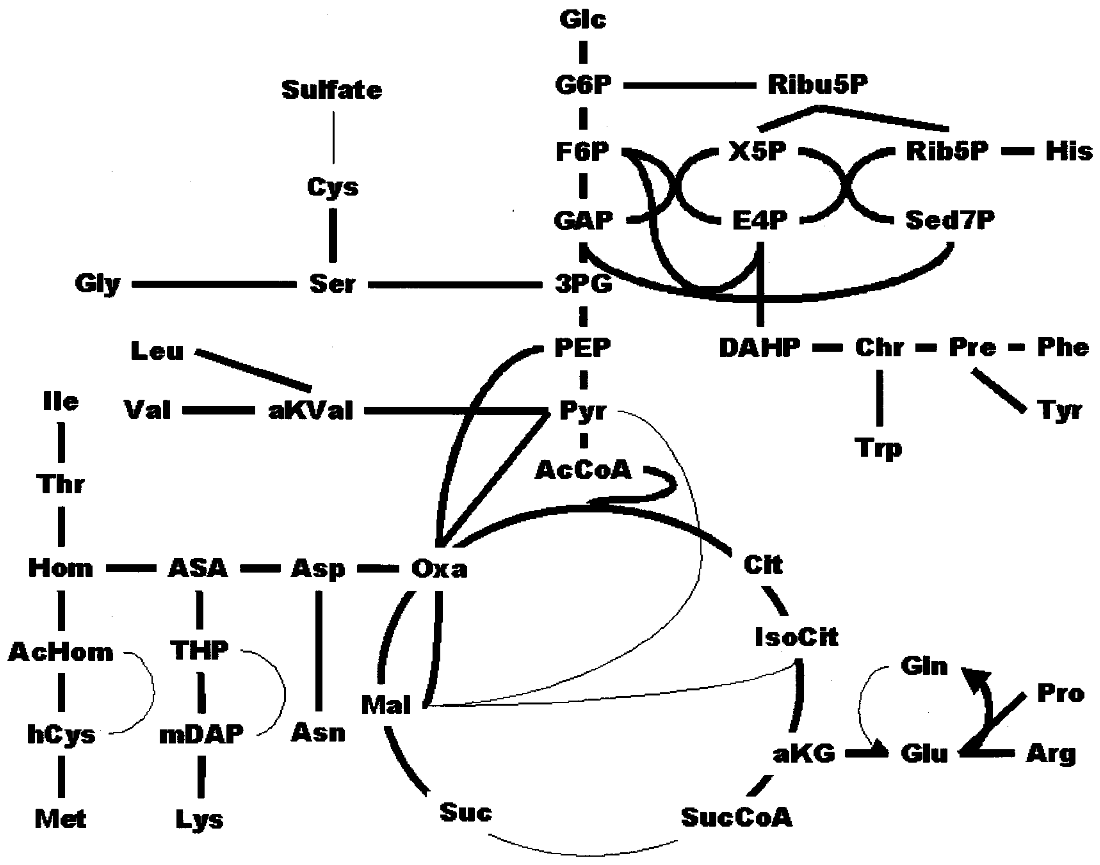


Figure 3.4 The overview of amino acid biosynthesis pathway in *Corynebacteria*

Broad line shows the conserved pathway among three *Corynebacteria*. Narrow line shows the lost pathway in *C. diphtheriae*. Glc: glucose, G6P: glucose-6-phosphate, F6P: fructose-6-phosphate, GAP: glyceraldehyde-3-phosphate, 3PG: 3-phosphoglycerate, PEP: phosphoenolpyruvate, Pyr: pyruvate, AcCoA: acetyl-coenzyme A, Cit: citrate, IsoCit: isocitrate, aKG: alpha-ketoglutarate, SucCoA: succinyl-coenzyme A, Suc: succinate, Mal: maleate, Oxa: oxaloacetic acid, Ribu5P: ribulose-5-phosphate, X5P: xylulose-5-phosphate, Rib5P: ribose-5-phosphate, E4P: erythrose-4-phosphate, Sed7P: sedoheptulose-7-phosphate, His: histidine, DAHP: 3-deoxy-D-arabino-heptulosonate-7-phosphate, Chr: chorismate, Trp: tryptophan, Pre:

prephenate, Phe: phenylalanine, Tyr: tyrosine, Glu: glutamate, Gln: glutamine, Pro: proline, Arg: arginine, Ser: serine, Gly: glycine, Cys: cysteine, aKVal: alpha-ketovaline, Leu: leucine, Val: valine, Ile isoleucine, Thr; threonine, Asp: aspartate, Asn: asparagine, ASA: aspartate-semialdehyde, THP: tetrahydropicolinate, mDAP: meso-diaminopimelate, Lys: lysine, Hom: homoserine, AcHom: acetylhomoserine, hCys: homocysteine, Met: methionine

Chapter 4

Evolutionary significance of *Corynebacterium*

4.1 Genome sequence and fermentation

The genome comparison and its experimental validations have been recognized as the effective way for the breeding and process-development of industrial amino acid production by fermentation method. In *Escherichia coli,* the standard model organism in microbiology, advanced technologies like genome, transcriptome, proteome and metabolome analysis have been used for the study of cell physiology. Especially, DNA array technology are powerful tools not only for the basic research but also for applied technology such as the breeding and process-development of amino acid fermentation (Imaizumi et al., 2005). *C. glutamicum* has been widely used for industrial fermentation of glutamic acid, lysine and other amino acid production. However, fundamental knowledge based on biochemical and genetic analyses in glutamic acid producing coryneform bacteria is less than that in model organisms. To overcome this difficulty, the comparative genome sequence analysis using phylogenetically near relatives may be required. The genome sequences of *C. efficiens* and *C. diphtheriae* have been suitable materials for the comparison with that of *C. glutamicum.* Through the comparison of three corynebacterial genome sequences, we tried to reveal the genome evolution in *Corynebacteria* in the hope of finding an application in applied biotechnology.

4.2 Evolutionary process of protein thermostabilization and organism thermostabilization

Until now, many attempts have been reported to find the motive force for the protein thermostabilization in nature. Chakravarty and Varadarajan showed several factors responsible for protein stability using high quality structural alignment with 9 thermophilic and 21 mesophilic bacterial genomes (Chakravarty and Varadarajan, 2002). They showed that the most remarkable differences of amino acid compositions in proteins between mesophile and thermophile were occurred at protein exposed sites. Most prominent substitution patterns at exposed sites were that noncharged polar (Thr, Ser, Asn, and Gln) residues in mesophiles were replaced either by rigid (Pro), branched nonpolar (Ile or Val), large aromatic (Tyr or Trp), or charged (Lys, Arg, Asp, or Glu) residues in thermophiles. La et al. proposed a motif based analysis method to identify the protein sequences responsible for the protein thermostabilization (La et al., 2003).

Our approach was the comparison of the complete genome sequences of two mesophiles from closely related species. Although there is less certainty factor in the difference of protein optimal temperature, more than 1,000 orthologous genes with 60–95% amino-acid sequence identity can be compared individually. This is advantageous for our comparative genomic study — previous genome-wide comparisons between thermophilic archaea and mesophilic bacteria have been hindered by the fact that the amino-acid residues did not correspond on a one-to-one basis. By comparison of the complete genome sequences of *C. efficiens* and *C. glutamicum*, we identified three kinds of amino acid substitutions responsible for protein thermostabilization: Lys in *C. glutamicum* to Arg in *C. efficiens*, Ser in *C. glutamicum* to Ala or Thr in *C. efficiens*. These kinds of amino acid substitutions were responsible for greater GC contents in *C. efficiens*. McDonald et al. have analyzed the asymmetric amino acid substitution patterns in 229 genes of the bacterial genus *Bacillus* (McDonald et al. 1999). The differences in GC content in *Bacillus* are similar (*B. stearothermophilus* 52% vs. *B. subtilis* 43.5%) to the difference between *C. efficiens* and *C. glutamicum,* and the asymmetrical amino acid substitution patterns found in *Bacillus* are very similar. We believe that greater genome GC content in closely related mesophiles may be one of the general strategies for the acquisition of thermostability in bacteria. The amino acid substitutions pattern that we obtained here seems to be different from the previous comparative genomics study's results using both of mesophile and thermophile, suggesting the reflection of a different time scale of evolution between our study and previous studies. It was suggested that the reason for the increase of genome GC contents in *C. efficiens* was the lack of *mutT* gene in *C. efficiens* (Nakamura et al., 2003). The *mutT* gene has been known to be suppressing the nucleotide mutation from A-T pair to C-G pair in *E. coli* (Horst et al., 1999).

By the results of comparison of the protein thermostability in 13 enzymes, it was suggested that the difference in growth temperature between *C. efficiens* and *C. glutamicum* was due to the difference in protein thermostability. Repeated attention should be paid to the estimation of the mechanism of an organism's thermostability by comparative genomics. One of the important facts is that many thermophilic archaea adopt different types of enzymes from the prokaryotes in the same metabolic pathway. For example, it was shown that glucokinase/phosphofructokinase in *Methanococcus jannaschii* was ADP-dependent. This enzyme was a member of the glycolysis pathway in *M. jannaschii* and its evolutionary origin was different from the phosphofructokinase in prokaryotes (Sakuraba et al., 2002). By adopting those enzymes, the organisms may have achieved total thermostability. We need to analyze not only the asymmetrical amino acid substitution pattern, but also gene loss and gain which may related to thermostability in *C. efficiens*.

It has been known that trehalose accumulation in the cell enhances the thermostabilization of organisms (Canovas et al., 2001). By the comparison of *C. efficiens* and *M. tuberculosis*, it was suggested that *C. efficiens* possessed another trehalose biosynthesis related gene whereas *C. glutamicum* did not. It was a fusion gene of phosphoglucomutase (*pgm*) and a functional unidentified region (Fig. 4.1). Further experimental study of the thermostability in *C. efficiens* considering both of protein thermostabilization and metabolic effect will be required.

4.3 The impact of the complete genome sequence on the evolutionary study and amino acid fermentation in *Corynebacterium*

4.3.1 Breeding of amino acid production strain

Although *C. glutamicum* has been widely used for industrial amino acid production by fermentation method, the fundamental knowledge like the cell physiology in *Corynebacterium* has been less than that of model organisms such as *E. coli* or *B. subtilis*. Recently, three species of corynebacterial genome sequences have been available. These complete genome sequences have been largely contributed to the progress of cell biology and industrial application of *C. glutamicum.*

Historically, the major breeding method for amino acid producing strain in *Corynebacterium* was random mutagenesis. By this method, both useful phenotypes and undesired properties were accumulated in the producing strain. Ohnishi et al. proposed the idea “genome breeding”. They tried to identify the effective mutations for fermentation and reconstruct the production strain only with them (Ohnishi et al., 2002). They identified three kinds of mutations from Lysine production strain; from 59Val in wild strain to 59Ala in lysine producing strain in the homoserine dehydrogenase, from 311Thr to 311Ile in aspartokinase, from 458Pro to 458Ser in pyruvate carboxylase. Finally, they achieved the lysine accumulation of 80 g/l after 27 h at 30 °C and 85 g/l after 28 h at 40 °C with the decrease in final achievement of growth. (Ohnishi et al., 2002, Ohnishi et al., 2003). These results suggested that many mutations accumulated in the current amino acid producing strain may not be necessary for the ideal fermentation process. At the same time, it should be noted that the classification of desired and non-desired mutation from many mutations accumulated in the whole genome might be a difficult task. We identified newly acquired genes at the common ancestor of *C. glutamicum* and *C. efficiens* and some of them showed a significant homology with functional genes in other organisms. For example, *pps* homologous genes were newly acquired genes in glutamic-acid-producing *Corynebacteria*, but their function is poorly characterized in *Corynebacterium* (Nishio et al., 2004). And *pps* gene has been known to be contributed to the effective aromatic amino acid production in *E. coli*. By considering the evolutionary process in *Corynebacteria* with the result of random mutagenesis in the corynebacterial amino acid production strain, it may be easier to extract an effective and novel mutation for amino acid production.

Metabolomic studies will provide us the detailed information for the phenotypic change caused by the random mutagenesis (Raamsdonk et al., 2001). Wittmann and Heinzle analyzed the metabolic flux genealogies of several lysine producing strains in *C. glutamicum* obtained by random mutagenesis (Wittmann and Heinzle, 2002). They showed a clear tendency for the increase of lysine yield with the increase of carbon flux into pentose phosphate pathway and from pyruvate to oxaloacetate. From this result, supply of NADPH and carbon dioxide fixation were suggested to be the important factors for the lysine production in industrial scale. Kromer et al. integrated the metabolome and transcriptome data obtained from the lysine producing strain in *C. glutamicum*, and analyzed them with flux analysis technique (Kromer et al., 2004). By the integration of those data, they estimated the major cause of every flux changing at each sampling points. The flux changes at lysine biosynthesis pathway were regulated at metabolic level. On the other hand, highly regulated gene expressions were found at the major branch points of central metabolism. For example, the expression level of the glucose 6-phosphate dehydrogenase gene was changed about seven folds, in comparison with growth phase and lysine production phase. Glucose 6-phosphate dehydrogenase catalyzes the first step reaction at pentose phosphate pathway. We expect that by combining many omic technologies, a new analytical method for choosing the effective mutation which causes the carbon flux change may be developed. And in near future, the system level comparison of amino acid overproduction in different bacteria will be available, and this knowledge may provide new aspects of evolutional study.

4.3.2 The cell wall biosynthesis in *Corynebacteria*

The study of cell wall structure has been recognized as one of the important subjects for the cell biology in *Corynebacteria.* The most remarkable feature is the presence of mycolic acid which was only observed in the several Gram-positive bacteria including *Mycobacteria, Corynebacteria* and *Nocardia* (Bayan et al., 2003). Mycolic acid is related to formation of the hydrophobic layer of the cell wall. And the hydrophobic layer is related to the drug and substrate permeability (Tzvetkov et al., 2003). It has been known that the structure of the fatty acid biosynthesis gene in *Mycobacteria* and *Corynebacteria* is fusion of subunit for fatty acid chain elongation reactions (Stuible et al., 1996). The basic knowledge of cell wall structure may be also important for the glutamic acid overproduction in *C. glutamicum.* Glutamic acid overproduction by biotin limitation is supposed to be related to fatty acid biosynthesis regulation (Kimura, 2003). Unfortunately, general representation of cell wall biosynthesis pathway is still unclear. Further knowledge related to cell wall biosynthesis has been expected to be obtained by the availability of the complete genome sequence.

Brand et al. extracted the candidates of mycolic acid biosynthesis related genes (*cmt1*, *cmt2*, *cmt3*, *cmt4* and *cmt5*) by mycolyltransferase domain comparison from the complete genome sequence of *C. glutamicum* (Brand et al., 2003). The homologues of these candidates were also observed in the genome sequence of *M. tuberculosis* and *M. leprae*. The mutations to these genes were affected by the amount of trehalose monocorynomycolate and trehalose dicorynomycolate in the cell envelope in *C. glutamicum*. In *Corynebacteria* and *Mycobacteria*, three kinds of trehalose biosynthesis pathway were found on their genome sequences; OtsA/OtsB pathway which synthesizes trehalose from glucose-6-phosphate and UDP glucose, TreY/TreZ pathway which degrades α-1,4-glucan polysaccharides into trehalose, TreS pathway which converts maltose into trehalose (Fig. 4.2). Tzvetkov et al. tried to elucidate the biological meaning of these redundant pathways (Tzvetkov et al., 2003). They found that inactivation of OtsA/OtsB pathway and TreY/TreZ pathway cause serious delay for growth comparing with wild type cells. And the delay was complemented with the addition of trehalose into the medium. Interestingly, the inactivation of OtsA/OtsB pathway and glycogen biosynthesis pathway showed similar effect to that of OtsA/OtsB pathway and TreY/TreZ pathway. In the mutant which could not accumulate trehalose, not only were trehalose monocorynomycolate or trehalose dicorynomycolate not observed but also other sugar and corynomycolic acid ester were not detected. Furthermore, by the analysis of the mutant of mycolyltransferase coding gene (*csp1*), trehalose 6-phosphate was suggested to serve as an acceptor for the freshly synthesized corynomycolic acid. These results suggested the important role of trehalose for cell wall biosynthesis.

Trehalose also works as the compatible solute with proline in *C. glutamicum.* Wolf et al. also investigated the role of three trehalose biosynthetic pathways on osmotic stress (Wolf et al., 2003). They evaluated all of the possible combinations for the trehalose biosynthetic pathways. *ΔotsAΔtreY* and *ΔotsAΔtreSΔtreY* strains showed growth inhibition and the absence of trehalose in cytoplasm under elevated osmolarity condition. In many other organisms, OtsAB pathway played an important role in the response of osmotic stress. However in *C. glutamicum*, OtsAB pathway did not but TreYZ pathway did play a central role for osmoresponsive trehalose biosynthesis. And they found that TreS pathway works for the conversion from trehalose to maltose rather than opposite reaction.

Cell wall biosynthesis or fatty acid biosynthesis has been important topics for not only amino acid fermentation but also drug design. Further comparative genomics and omic studies will promote the progress of this area. The merit for further research into amino acid fermentation is the elucidation for high tolerance of osmotic pressure.

4.3.3 Metabolic regulation

C. glutamicum has been known to assimilate both glucose and acetate for amino acid production (Liebl, 1991). For the growth on acetate, the activation of acetyl kinase (Ack), phosphotransacetylase (Pta) for the synthesis of acetyl-CoA, and glyoxylate shunt (AceA, AceB) as the anaplerotic pathway is required. The activation of these pathways has been shown to be regulated at transcriptional level (Reinscheid et al., 1999; Wendisch et al., 1997). Gerstmeir et al. isolated the transcriptional regulator of these genes by DNA affinity chromatography using proposed cis-regulatory elements (Gerstmeir et al., 2004). The regulatory protein was named as the regulator of acetate metabolism B, RamB. The analysis of *ramB* mutant in *C. glutamicum* showed that RamB negatively affected the expression of *ack*, *pta*, *aceA* and *aceB* genes growth on glucose.

GlxR (glyoxylate bypass regulator) was also isolated as the repressor protein of *aceB* gene expression (Kim et al., 2004). By the analysis of the amino acid sequence, GlxR showed the similarity with CRP in *E. coli*, suggesting the cAMP binding protein. In fact, GlxR was able to complement *E. coli crp* mutant strain (Kim et al., 2004). In the case of *E. coli*, intracellular cAMP concentration was kept low when it was grown on glucose and in high on acetate. However, the observed intracellular cAMP concentration in *C. glutamicum* was different from that in *E. coli*. The intracellular cAMP concentration was lower in acetate medium than in glucose medium for *C. glutamicum*. And multi copies of *glxR* gene in *C. glutamicum* repressed the expression of isocitrate lyase (ICL), malate synthase (MS), acetate kinase (ACK), and isocitrate dehydrogenase (ICDH) on acetate medium (Kim et al., 2004). From these results, the following hypothesis was proposed; cAMP binding GlxR repressed the gene expression of *ack*, *pta*, *aceA* and *aceB* genes when *C. glutamicum* was grown on glucose. And when the carbon source was changed into acetate, the cAMP concentration was decreased and the complex of cAMP-GlxR was not formed. As a result, those gene expressions were induced.

In previous study, acetyl-CoA was assumed to be the inducer molecule of *aceA* and *aceB* gene expression in *C. glutamicum* (Wendisch et al., 1997). Until now, no direct relationship between acetyl-CoA and RamB or GlxR has been shown. Further analysis is required for the clarification of the mechanism of *aceAB* gene activation. Interestingly, there was no glyoxylate shunt in *C. diphtheriae*, but the orthologous gene of *ramB* and *glxR* in *C. glutamicum* were also found in *C. diphtheriae*. These facts support our claims that the common ancestor of the three species of *Corynebacteria* may possess the ability of amino acid biosynthesis and *C. diphtheria* has lost its ability. In *C. glutamicum*, RamB or GlxR may regulate the expression of several genes including corynebacterial orthologs which may be lost in *C. diphtheriae*. This may be one of the reasons that although the pathway has been lost, the regulatory genes have been kept on the genome in *C. diphtheriae*.

4.3.4 Evolutionary process of glutamic acid overproduction mechanism in *Corynebacterium*

Still now, we have not elucidated the molecular mechanism for glutamate production responsible for biotin limitation in *C. glutamicum*. We believe that more detailed comparative genomics study will be required for that purpose. One biologically important characteristic of *C. glutamicum* is the biotin requirement for growth, which is closely related to glutamate overproduction (Kimura, 2003). This biotin requirement was also observed in *C. efficiens*. Both of these bacteria lack the complete biotin-biosynthesis pathway from pimelate to biotin (Fig. 4.3). By contrast, *C. diphtheriae* has the complete biotin-biosynthesis pathway. In addition, DIP1381 encoding 6-carboxyhexanoate-CoA ligase (BioW), which is the first enzyme in biotin biosynthesis, might have been acquired by horizontal gene transfer in this species (Table 4.1). This is suggested by the fact that none of the bacteria that are closely related to *C. diphtheriae* possess orthologues of DIP1381. BirA is a bifunctional protein that exhibits biotin ligase activity and also acts as the DNA-binding transcriptional repressor of the biotin operon, which is conserved in many organisms. The regulatory sequence of BirA might be conserved among many bacteria (Rodionov et al., 2002). However, the corynebacteria have lost the DNA-binding region in the orthologous *birA* gene.

Glutamate overproduction in *C. glutamicum* is induced by a shortage of biotin (Kimura, 2003). However, the regulatory sequences that are associated with the biotin-biosynthesis-related genes and glutamate production remain to be identified. Comparing the regulatory regions of glutamate overproduction-related genes between glutamic-acid-producing and non-producing species might help to elucidate the regulatory mechanism of glutamate production. In *C. glutamicum,* a lack of biotin attenuated the 2-oxoglutarate dehydrogenase complex (ODHC) activity and the initiation of glutamate production simultaneously (Kawahara et al., 1997, Shimizu et al., 2003). By contrast, enhanced glutamate dehydrogenase (GDH) activity might not contribute to glutamate production (Shimizu et al., 2003) and showed no response to biotin limitation (Kawahara et al., 1997). The *odhA* gene, which encodes the OdhA subunit of the ODHC, has a lineage-specific structure in the corynebacteria and mycobacteria (Usuda et al., 1996), while the structure of the *gdh* gene is common among a wide range of bacteria (Bormann et al. 1992). Here we focused on the conservation of the regulatory regions of these genes among the corynebacteria. Although the regulatory sequences of these genes remain to be identified experimentally, it may be possible to discuss the evolutionary process of gene regulation by the window analysis of the identity for the five prime regions of these genes. Because the five prime regions may include the regulatory regions, the conservation in five prime regions was discussed here in place of that in regulatory regions. The five prime regions of the *odhA* gene were more strongly conserved between *C. efficiens* and *C. glutamicum* than between *C. diphtheriae* and either *C. glutamicum* or *C. efficiens* (Fig. 4.4). The regulatory region of the *odhA* gene may be also conserved between *C. efficiens* and *C. glutamicum*. The accumulation of mutations in *C. diphtheriae* might have explained this pattern of conservation. By contrast, the upstream regions of the *gdh* gene were equally conserved among all three species (Fig. 4.4). These results suggest that decreased ODHC activity induced by biotin limitation might be regulated at the gene-expression level. The conservation of the five prime regions in *odhA* gene showed that the regulatory region in *odhA* gene should be conserved between *C. glutamicum* and *C. efficiens*. And this conservation of regulation also suggested that the conservation of regulatory mechanism. Moreover, the loss of the glutamate-overproduction ability in *C. diphtheriae* might have originated with the acquisition of the complete biotin-biosynthesis pathway through horizontal gene transfer. In this case, it is assumed that the important parts of the regulatory regions were conserved, despite the differences in the genome GC content, which was 10% higher in *C. efficiens* than in either *C. glutamicum* or *C. diphtheriae* throughout the genome (Nishio et al., 2003).

We suggested that the ability of various amino acid production was inherited at the common ancestor of *Corynebacterium* and *C. diphtheria* has been lost them on their evolutionary process after divergence from its sister species. If this hypothesis is true, the regulatory sequences related to glutamate production may be conserved between *C. glutamicum* and *C. efficiens*, but not in *C. diphtheriae*. In fact, the five prime regions including regulatory sequence in *odhA* gene was highly conserved between *C. glutamicum* and *C. efficiens*. The gene expression of *odhA* gene was responsible for cell growth, biotin limitation and glutamate production in *C. glutamicum* (Kawahara et al. 1997). And by the comparison of regulatory sequence comparison, we may estimate the gene networks responsible for glutamate production in *Corynebacteria*. The bioinformatics technique for the identification of gene regulatory sequences has not been worked well. The statistical method may need for the evaluation of the biological significance of detected regulatory sequence. To cover the incompletion the regulatory sequence detection and comparison methodology, the integration of transcriptome, proteome and metabolome data with sequence data may be required.

Table 4.1 Biotin biosynthesis genes in high GC Gram-positive bacteria

	bioW	*bioF*	*bioA*	*bioD*	*bioB*
C. glutamicum			Cgl2604	Cgl2605	Cgl0072
C. efficiens			CE1421	CE1420	CE0089
C. diphtheriae	DIP1381	DIP1382	DIP1191	DIP1189 DIP1192	DIP0105 DIP1124
M. tuberculosis		Rv0032 Rv1569	Rv1568	Rv1570	Rv1589
M. leprae		ML1217	ML1216	ML1218	ML1120
S. coelicolor		SCO1243	SCO1245	SCO1246	SCO1124

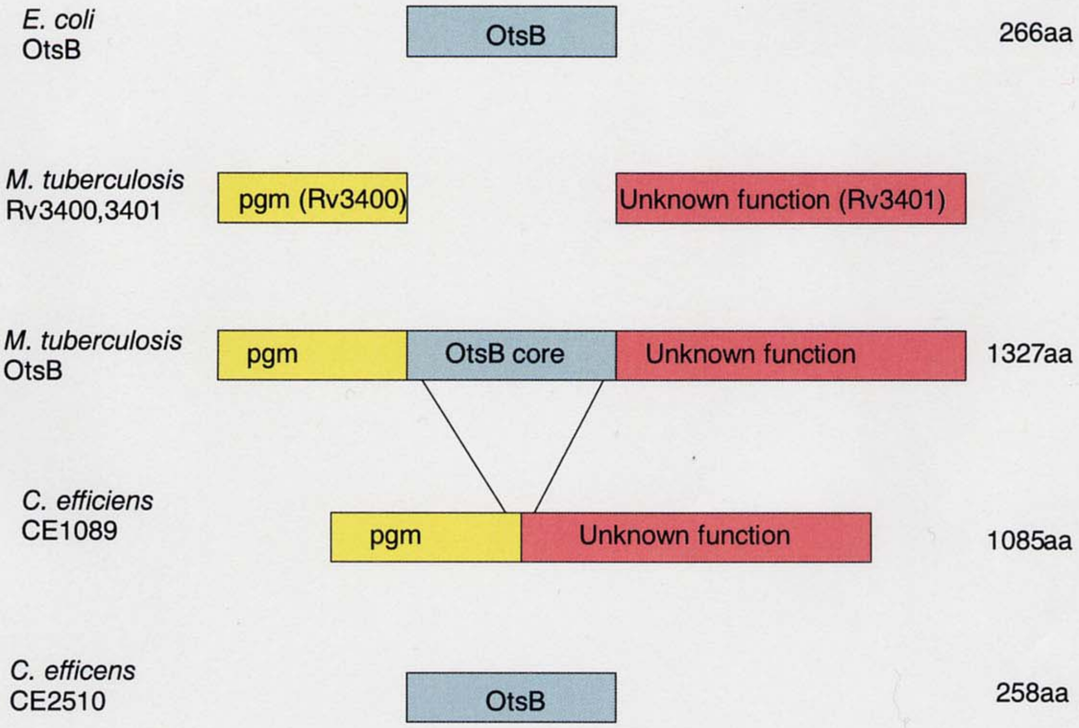


Figure 4.1 Structual alignment of OtsB and its related genes

OtsB: trehalose-6-phosphate phophatase, Pgm: phosphoglucomutase

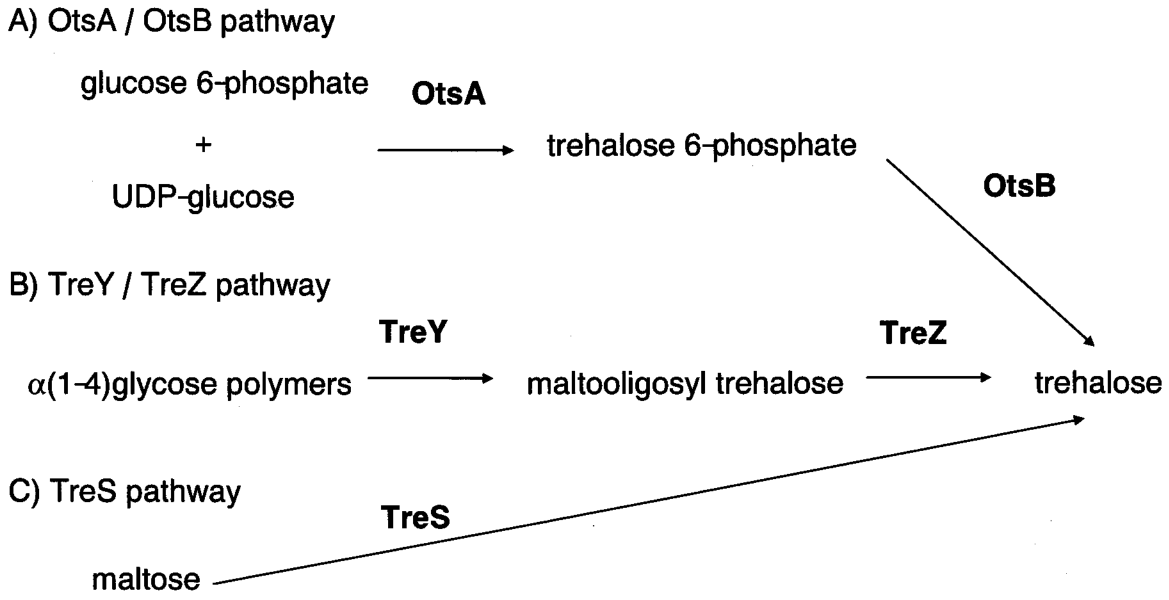


Figure 4.2 Three kinds of trehalose biosynthesis pathways in *Corynebacterium*

OtsA: trehalose 6-phosphate synthase, OtsB: trehalose 6-phosphate phosphatase, TreY: maltooligosyltrehalose synthase, TreZ: maltooligosyltrehalose hydrolase, TreS: trehalose synthase

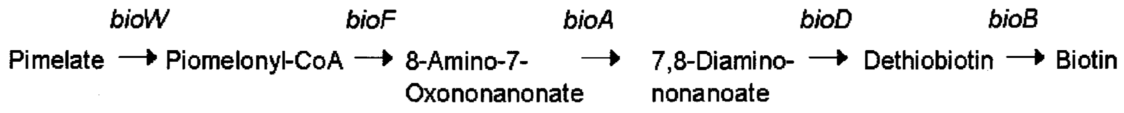


Figure 4.3 Biotin-biosynthesis pathway

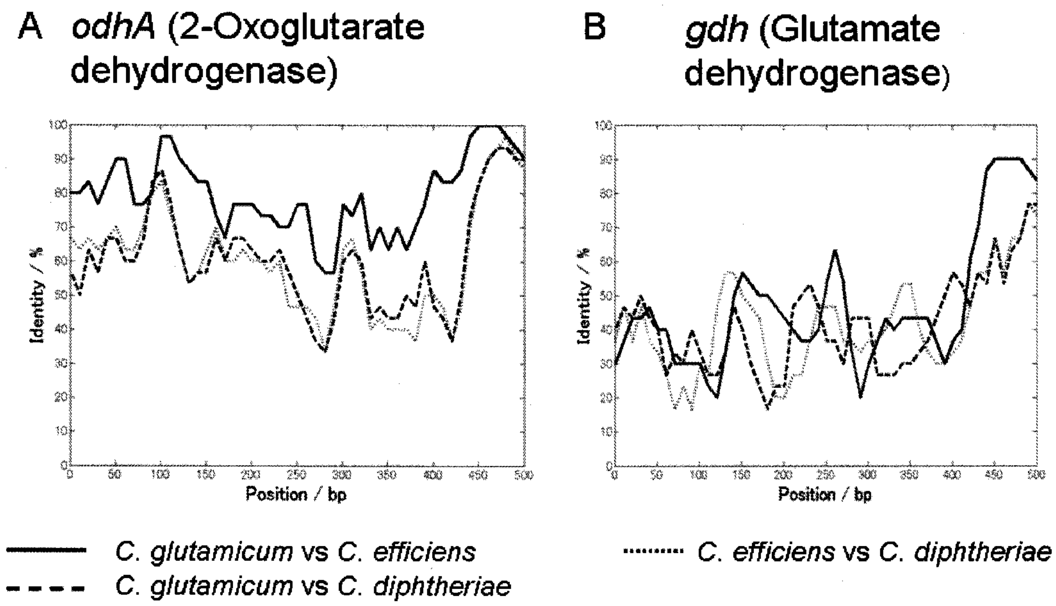


Figure 4.4 Window analyses of the five prime regions

The 500-bp sequence upstream from the start codon of each gene was analyzed according to the 30-bp window size and 10-bp step size. After alignment of the regulatory region plus the coding region, the gaps were removed from the multiple alignment and the identity was calculated. (A) *odhA* gene. (B) *gdh* gene.

Chapter 5

Conclusion

The evolutionary mechanism of protein thermostabilization in *C. efficiens* was elucidated by whole genome comparison between *C. efficiens* and *C. glutamicum.* The difference in GC content between the species was reflected in codon usage and nucleotide substitutions. My comparative genomic study clearly showed that there was tremendous bias in amino acid substitutions in all orthologous ORFs. Analysis of the direction of the amino acid substitutions suggested that three substitutions: from lysine to arginine, serine to alanine, and serine to threonine, are important for the stability of the *C. efficiens* proteins. It is suggested that the accumulation of these three types of amino acid substitutions correlates with the acquisition of thermostability and is responsible for the greater GC content of *C. efficiens*.

Gene loss and horizontal gene transfer were important for the amino acid pathway organization and metabolic regulation in *Corynebacteria.* When *Mycobacterium* and *Streptomyces* were used as outgroups, it was suggested that the common ancestor of *Corynebacteria* already possessed almost all of the gene sets necessary for amino acid production. However, *C. diphtheriae* was found to have lost the genes responsible for amino acid production. Moreover, I found that the common ancestor of *C. efficiens* and *C. glutamicum* have acquired some of genes responsible for amino acid production by horizontal gene transfer. Thus, I concluded that the evolutionary events of gene loss and horizontal gene transfer must have been responsible for functional differentiation in amino acid biosynthesis of the three species of *Corynebacteria*.

By the analysis of genome GC contents and GC skew, it was suggested that the genome structure of the common ancestor of *Corynebacterium* was more similar to that of *C. glutamicum* and *C. diphtheriae* than to *C. efficiens*. On the other hand, by the phylogenetic analysis of 16S rRNA or protein coding genes, it was suggested that *C. glutamicum* and *C. efficiens* were closely related phylogenetically. The comparative genome analysis suggested that after divergence of the common ancestor of *C. glutamicum* and *C. efficiens*, the genome GC contents in *C. efficiens* was increased. Thus the discrepancy of the evolutionary process between phylogenetically and genome structure was generated. This study showed the process of genome evolution through the gene loss, gene duplication and horizontal gene transfer after divergence from the common ancestor of *Corynebacteria*. These evolutionary events were related to the acquisition of protein thermostability in *C. efficiens* and the loss of glutamic acid productivity in *C. diphtheriae*.

In conclusion, the differentiation of the metabolic pathways among the corynebacteria appears to have been caused by dynamic genome evolution involving not only amino-acid substitutions but also gene loss and gene gain. This comparative genomics study indicates that dynamic genome evolution within the corynebacteria is associated with the major biological features of each species: that is, glutamate overproduction in *C. glutamicum*, thermostability in *C. efficiens* and pathogenesis in *C. diphtheriae*. Further comparative studies, particularly of gene expression and metabolic regulation, will help to realize an ecological fermentation process. I concluded that this study showed the evolutionary process of bacterial diversity from view point of genome evolution.

Reference

Adachi, J., Hasegawa, M. 1996. MOLPHY version 2.3: programs for molecular phylogenetics based on maximum likelihood. Computer Science Monographs 28. Institute Statistical Mathematics, Tokyo.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs *Nucleic Acids Res.* **25**: 3389-3402

Amador, E., Castro, J.M., Correia, A., Martin, J.F. 1999. Structure and organization of the *rrnD* operon of '*Brevibacterium lactofermentum*': analysis of the 16S rRNA gene *Microbiology* **145**: 915-924

Bayan, N., Houssin, C., Chami, M., Leblon, G. 2003. Mycomembrane and S-layer: two important structures of *Corynebacterium glutamicum* cell envelope with promising biotechnology applications. *J. Biotechnol.* **104**: 55-67.

Bentley, S. D., Chater, K. F., Cerdeno-Tarraga, A. M., et al. (40 co-authors) 2002. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* **417**: 141-147.

Brand, S., Niehaus, K., Puhler, S., Kalinowski, J., 2003. Identification and functional analysis of six mycolyltransferase genes of *Corynebacterium glutamicum* ATCC 13032: the genes *cop1*, *cmt1*, and *cmt2* can replace each other in the synthesis of trehalose dicorynomycolate, a component of the mycolic acid layer of the cell envelope. *Arch. Microbiol.* **180**: 33–44

Bormann E. R., Eikmanns B. J., Sahm H., 1992, Molecular analysis of the *Corynebacterium glutamicum gdh* gene encoding glutamate dehydrogenase. *Mol*

Microbiol. **6**: 317-26.

Boucher, Y., Doolittle, W. F., 2000. The role of lateral gene transfer in the evolution of isoprenoid biosynthesis pathways. *Mol. Microbiol.* **37**: 703-716.

Breinig, S., Schiltz, E., Fuchs, G. J., 2000. Genes involved in anaerobic metabolism of phenol in the bacterium *Thauera aromatica. J. Bacteriol.* **182**: 5849-5863.

Canovas, D., Fletcher, S. A., Hayashi, M., Csonka, L. N. 2001. Role of trehalose in growth at high temperature of *Salmonella enterica* serovar Typhimurium. *J Bacteriol.* **183**: 3365-3371.

Cerdeno-Tarraga, A. M., Efstraitou, A., Dover, L. G., et al., (26 co-authors) 2003. The complete genome sequence and analysis of *Corynebacterium diphtheriae* NCTC13129. *Nucleic Acids Res.* **31**: 6516-6523.

Chakravarty, S., Varadarajan, R. 2000. Elucidation of determinants of protein stability through genome sequence analysis. *FEBS Letters* **470**: 65-69

Chakravarty, S., Varadarajan, R. 2002. Elucidation of Factors Responsible for Enhanced Thermal Stability of Proteins: A Structural Genomics Based Study. *Biochemistry* **41**:8152-8161

Cirilli, M., Scapin, G., Sutherland, A., Vederas, J.C., Blanchard, J.S. 2000. The three-dimensional structure of the ternary complex of *Corynebacterium glutamicum* diaminopimelate dehydrogenase-NADPH-L-2-amino-6-methylene-pimelate. *Protein Sci.* **9**: 2034-2037

Cole, S. T., Brosch, R., Parkhill, J., et al. (39 co-authors) 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**: 537-544.

Cole, S. T., Eiglmeier, K., Parkhill, J., et al. (41 co-authors) 2001. Massive gene decay in the leprosy bacillus. *Nature* **409**: 1007-1011.

Collins, M. D., Cummins, C. S. 1986. Genus *Corynebacterium*. Vol. 2 Pp. 1266-1766 *in* P. H. A. Sneath eds. Bergey's Manual of Systematic Bacteriology. Baltimore: Williams & Willkins.

Delcher, A.L., Harmon, D., Kasif, S., White, O., Salzberg, S.L. 1999. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27**: 4636-4641

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M., et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**: 496-512

Forterre, P. 1996. A hot topic: The origin of hyperthermophiles. *Cell* **85**: 789-792

Fraczkiewicz, R., Braun, W. 1998. Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *J. Comp. Chem.* **19**: 319-333

Fudou, R., Jojima, Y., Seto, A., Yamada, K., Kimura, E., Nakamatsu, T., Hiraishi, A., Yamanaka, S. 2002. *Corynebacterium efficiens* sp. nov., a glutamic-acid-producing species from soil and vegetables. *Int. J. Syst. Evol. Microbiol.* **52**: 1127-1131

Galtier, N., Taurasse, N., Gouy, M. 1999. A nonhyperthermophilic common ancestor to extant life forms. *Science* **283**: 220-221

Gerstmeir, R., Cramer, A., Dangel, P., Schaffer, S., Eikmanns, B. J. 2004. RamB, a novel transcriptional regulator of genes involved in acetate metabolism of *Corynebacterium glutamicum*. *J. Bacteriol.* **186**: 2798-2809.

Graevenitz, A. V., Krech, T. 1991. The Genus *Corynebacterium*-Medical. vol.2 Pp.1173-1187 *in* A. Balows, H. G. Trüper, M. Dworkin, W. Harder, and K. H.

Schleicer, eds. The Prokaryotes, 2nd edition. Springer-Verlag, New York.

Haney, P.J., Badger, J.H., Buldak, G.L., Reich, C.I., Woese, C.R., Olsen, G.J. 1999. Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic *Methanococcus* species. *Proc. Natl. Acad. Sci. USA* **96**: 3578-3583

Henikoff, S., Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **89**: 10915-10919

Horst, J.P., Wu, T.H., Marinus, M.G., 1999. *Escherichia coli* mutator genes. *Trends Microbiol.* **7**: 29– 36.

Ikeda, M. 2003. Amino acid production processes. Pp. 1–35 in R. Faurie and J. Thommel, eds. Adv. Biochem. Eng. Biotechnol., vol 79. Microbial production of l-amino acids. Springer, Berlin Heidelberg New York.

Ikeda, M., Nakagawa, S. 2003. The *Corynebacterium glutamicum* genome: features and impacts on biotechnological processes. *Appl. Microbiol. Biotechnol.* **62**: 99-109.

Imaizumi, A., Takikawa, R., Koseki, C., Usuda, Y., Yasueda, H., Kojima, H., Matsui, K., Sugimoto, S., 2005, Improved production of l-lysine by disruption of stationary phase-specific *rmf* gene in *Escherichia coli. J. Biotechnol.* **117**: 111-118.

Kalinowski, J., Bathe, B., Bartels, D., et al. (27 co-authors) 2003. The complete *Corynebacterium glutamicum* ATCC 13032 genome sequence and its impact on the production of L-aspartate-derived amino acids and vitamins. *J. Biotechnol.* **104**: 5-25.

Kanehisa, M. 1997. A database for post-genome analysis. *Trends Genet.* **13**: 375-376.

Kawahara, Y., Takahashi-Fuke, K., Shimizu, E., Nakamatsu, T., and Nakamori, S., 1997, Relationship between the glutamate production and the activity of 2-oxoglutarate dehydrogenase in *Brevibacterium lactofermentum. Biosci. Biotechnol. Biochem.* **61**: 1109-1112.

Kim, H.J., Kim, T.H., Kim, Y., Lee, H.S. 2004. Identification and characterization of *glxR*, a gene involved in regulation of glyoxylate bypass in *Corynebacterium glutamicum. J Bacteriol.* **186**: 3453-3460.

Kimura, E., Abe, C., Kawahara, Y., Nakamatsu, T. 1996. Molecular cloning of a novel gene, *dtsR*, which rescues the detergent sensitivity of a mutant derived from *Brevibacterium lactofermentum. Biosci. Biotechnol. Biochem.* **60**: 1565-1570.

Kimura, E., Yagoshi, Y., Kawahara, Y., Ohsumi, T., Nakamatsu, T., Tokuda, H. 1999. *Corynebacterium glutamicum* triggered by a decrease in the level of a complex comprising DtsR and a biotin-containing subunit. *Biosci. Biotechnol. Biochem.* **63**: 1274-1278

Kimura, E. 2003. Metabolic engineering of glutamate production. Pp. 37-57 in R. Faurie and J. Thommel, eds. Adv. Biochem. Eng. Biotechnol., vol 79. Microbial production of l-amino acids. Springer, Berlin Heidelberg New York.

Kinoshita S, Udaka S, Shimono M. 1957. Studies on the amino acid fermentation. Part I. Production of L-glutamic acid by various microorganisms. *J. Gen. Appl. Microbiol.* **3**: 193-205.

Kreil, D.P., Ouzounis, C.A. 2001. Identification of thermophilic species by the amino acid compositions deduced from their genomes. *Nucleic Acids Res.* **29**: 1608-1615

Kromer, J.O., Sorgenfrei, O., Klopprogge, K., Heinzle, E., Wittmann, C. 2004. In-depth profiling of lysine-producing *Corynebacterium glutamicum* by combined analysis of the transcriptome, metabolome, and fluxome. *J Bacteriol.* **186**:1769-1784.

La, D., Silver, M., Edgar, R.C., Livesay, D. R. 2003. Using Motif-Based Methods in

Multiple Genome Analyses: A Case Study Comparing Orthologous Mesophilic and Thermophilic Proteins. *Biochemistry* **42**: 8988-8998.

Lange, B. M., Rujan, T., Martin, W. Croteau, R. 2000. Isoprenoid biosynthesis: the evolution of two ancient and distinct pathways across genomes. *Proc. Natl. Acad. Sci. USA* **97**: 13172-13177.

Li, W. H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.* **36**: 96-99.

Liebl, W. 1991. *The genus Corynebacterium—nonmedical,* In: Balows, A., Tru¨per, H. G., Dworkin, M., Harder, W. and Schleifer, K. H. (ed.), *The procaryotes*, vol. 2. Springer, New York, N.Y. p. 1157–1171.

Lowe, T.M., Eddy, S.R. 1997. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucl. Acids Res.* **25**: 955-964.

Malumbres, M., Gil, J.A., Martin, J.F. 1993. Codon preference in corynebacteria. *Gene* **134**: 15-24

Marais, A., Mendz, G. L., Hazell, S. L., Megraud, F. 1999. Metabolism and genetics of *Helicobacter pylori*: the genome era. *Microbiol. Mol. Biol. Rev.* **63**: 642-674.

McDonald, J.H., Grasso, A.M., Rejto, L.K. 1999. Patterns of temperature adaptation in proteins from *Methanococcus* and *Bacillus*. *Mol. Biol. Evol.* **16**: 1785-1790

McDonald, J.H. 2001. Patterns of temperature adaptation in proteins form the bacteria *Deinococcus radiodurans* and *Thermus thermophilus*. *Mol. Biol. Evol.* **18**: 741-749

McLean, M.J., Wolfe, K.H., Devine, K.M. 1998. Base Composition Skews, Replication Orientation, and Gene Orientation in 12 Prokaryote Genomes. *J. Mol. Evol.* **47**: 691-696

Miller, S.L., Lazcano, A. 1995. The origin of life - did it occur at high temperatures? *J.*

Mol. Evol. **41**: 689-692

Mineta, K., Nakazawa, M., Cebria, F., Ikeo, K., Agata, K., Gojobori, T. 2003. Origin and evolutionary process of the CNS elucidated by comparative genomics analysis of planarian ESTs. *Proc. Natl. Acad. Sci. USA* **100**: 7666-7671.

Musto H, Naya H, Zavala A, Romero H, Alvarez-Valin F, Bernardi G. 2004. Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. *FEBS Lett.* **573**: 73-77

Myers, E.W., Miller, W. 1988. Optimal alignments in linear space. *Comput. Appl. Biosci.* **4**: 11-17

Nagarkar, P. P., S. D. Ravetkar and M. G. Watve. 2002. The amino acid requirements of *Corynebacterium diphtheriae* PW 8 substrain CN 2000. *J. Appl. Microbiol.* **92**: 215-220.

Nakamura, Y., Nishio, Y., Ikeo, K., Gojobori, T. 2003. The genome stability in *Corynebacterium* species due to lack of the recombinational repair system. *Gene* **317**: 149-155.

Nei, M., Gojobori, T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**: 418-426.

Nei, M., Sudhir, K. 2000. *Molecular Evolution and Phylogenetics*, Oxford University Press, New York, pp. 17-31.

Nelson, K.E., Paulsen, I.T., Heidelberg, J.F., Fraser, C.M. 2000. Status of genome projects for nonpathogenic bacteria and archaea. *Nat. Biotechnol.* **18**: 1049-1054

Nisbet, E.G., Fowler, C.M.R. 1996. Some linked it hot. *Nature* **382**: 404-405

Nishio, Y., Nakamura, Y., Kawarabayasi, Y., Usuda, Y., Kimura, E., Sugimoto, S., Matsui, K., Yamagishi, A., Kikuchi, H., Ikeo, K. Gojobori, T. 2003. Comparative complete genome sequence analysis of the amino acid replacements responsible for the thermostability of *Corynebacterium efficiens*. *Genome Res.* **13**: 1572-1579.

Nishio, Y., Nakamura, Y., Usuda, Y., Sugimoto, S., Matsui, K., Kawarabayasi, Y., Kikuchi, H., Gojobori, T., Ikeo, K. 2004. Evolutionary process of amino acid biosynthesis in *Corynebacterium* at the whole genome level. *Mol Biol Evol.* **21**, 1683-1691.

Ohnishi, J., Mitsuhashi, S., Hayashi, M., Ando, S., Yokoi, H., Ochiai, K., Ikeda, M. 2002, A novel methodology employing *Corynebacterium glutamicum* genome information to generate a new L-lysine-producing mutant. *Appl Microbiol Biotechnol.* **58** :217-23.

Ohnishi, J., Hayashi, M., Mitsuhashi, S., Ikeda, M. 2003, Efficient 40ºC fermentation of l-lysine by a new *Corynebacterium glutamicum* mutant developed by genome breeding. *Appl. Microbiol. Biotechnol.* **62**: 69-75

Overbeek, R., Larsen, N., Pusch, G. D., D'Souza, M., Selkov , Jr., E., Kyrpides, N., Fonstein, M., Maltsev N., Selkov, E. 2000. WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.* **28**:123-125.

Pace, N.R. 1991. Origin of life - facing up to the physical setting. *Cell* **65**: 531-533

Park, S. M., Shaw-Reid, C., Sinskey, A. J., Stephanopoulos, G. 1997. Elucidation of anaplerotic pathways in *Corynebacterium glutamicum* via 13C-NMR spectroscopy and GC-MS. *Appl. Microbiol. Biotechnol.* **47**: 430-440.

Pearson, W. R. 2000. Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.* **132**: 185-219.

Raamsdonk, L. M., Teusink, B., Broadhurst, D., Zhang, N., Hayes, A., Walsh, M. C., Berden, J. A., Brindle, K. M., Kell, D. B., Rowland, J. J., Westerhoff, H. V., van Dam, K., Oliver, S.G. 2001. A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nat. Biotechnol.* **19**: 45-50.

Reinscheid, D. J., Schnicke, S., Rittmann, D., Zahnow, U., Sahm, H., Eikmanns, B. J. 1999. Cloning, sequence analysis, expression and inactivation of the *Corynebacterium glutamicum pta-ack* operon encoding phosphotransacetylase and acetate kinase. *Microbiology* **145**:503–513.

Rodionov, D. A., Mironov, A. A., and Gelfand, M. S., 2002, Conservation of the biotin regulon and the BirA regulatory signal in Eubacteria and Archaea, *Genome Res.* 12: 1507-1516.

Saitou, N., Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**: 406-425.

Sakuraba, H., Yoshioka, I., Koga, S., Takahashi, M., Kitahama, Y., Satomura, T., Kawakami, R., Ohshima, T. 2002, ADP-dependent glucokinase/phosphofructokinase, a novel bifunctional enzyme from the hyperthermophilic archaeon *Methanococcus jannaschii. J. Biol. Chem.* **277**: 12495-12498

Schulz, A. A., Collett, H. J., Reid, S. J. 2001. Nitrogen and carbon regulation of glutamine synthetase and glutamate synthase in *Corynebacterium glutamicum* ATCC 13032. *FEMS Microbiol. Lett.* **205**: 361-367.

Shimizu, H., Tanaka, T., Nakato, A., Nagahisa, K., Kimura, E., Shioya, S. 2003, Effects of the changes in enzyme activities on metabolic flux redistribution around the 2-oxoglutarate branch in glutamate production by *Corynebacterium glutamicum.*

Bioprocess Biosyst. Eng. **25**: 291-298.

Simic, P., Sahm, H., Eggeling, L. 2001. L-threonine export: use of peptides to identify a new translocator from *Corynebacterium glutamicum. J. Bacteriol.* **183**: 5317-5324.

Simmons C. P., Hodgson, A. L., Strugnell, R. A. 1997. Attenuation and vaccine potential of *aroQ* mutants of *Corynebacterium pseudotuberculosis. Infect. Immun.* **65**: 3048-3056.

Singer, G.A.C, Hickey, D.A. 2003. Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. *Gene* **317**: 39-47.

Smith, T., Waterman, M. S., 1981. Identification of common molecular subsequences. *J. Mol. Biol.* **147**: 195–197

Sobolev, V., Sorokine, A., Prilusky, J., Abola, E.E., Edelman, M. 1999. Automated analysis of interatomic contacts in proteins. *Bioinformatics* **15**: 327-332

Stuible, H. P., Wagner, C., Andreou, I., Huter, G., Haselmann, J., Schweizer, E. 1996. Identification and functional differentiation of two type I fatty acid synthases in *Brevibacterium ammoniagenes*, *J. Bacteriol.* **178**: 4787-4793.

Tatusov, R.L., Koonin, E.V., Lipman, D.J. 1997. A Genomic perspective on protein families. *Science* **278**: 631-637

Taylor, W.R. 1986. The classification of amino acid conservation. *J. Theor. Biol.* **119**: 205-218

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., Higgins, D. G. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25**: 4876-4882.

Tzvetkov, M., Klopprogge, C., Zelder, O., Liebl, W. 2003. Genetic dissection of trehalose biosynthesis in *Corynebacterium glutamicum*: inactivation of trehalose production leads to impaired growth and an altered cell wall lipid composition. *Microbiology*. **149**: 1659-73.

Udaka S. 1960. Screening method for microorganisms accumulating metabolites and its use in the isolation of *Micrococcus glutamicus*. *J. Bacteriol.* 79: 754-755.

Usuda, Y., Tujimoto, N., Abe, C., Asakura, Y., Kimura, e., Kawahara, Y., Kurahashi, O., and Matsui, H. 1996. Molecular cloning of the *Corynebacterium glutamicum* ('*Brevibacterium lactofermentum*' AJ12036) *odhA* gene encoding a novel type of 2-oxoglutarate dehydrogenase. *Microbiology* **142**: 3347-3354.

Vieille, C., Zeikus, G.J. 2001. Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. *Microbiol. Mol. Biol. Rev.* **65**: 1-43

Vrljic, M., Sahm, H., Eggeling, L. 1996. A new type of transporter with a new type of cellular function: L-lysine export from *Corynebacterium glutamicum*. *Mol. Microbiol.* **22**: 815-826.

Wendisch, V.F., Spies, M., Reinscheid, D.J., Schnicke, S., Sahm, H., Eikmanns, B. J. 1997. Regulation of acetate metabolism in *Corynebacterium glutamicum*: transcriptional control of the isocitrate lyase and malate synthase genes. *Arch. Microbiol.* **168**:262□269.

Wintrode, P.L., Miyazaki, K., Arnold, F.H. 2001. Patterns of adaptation in a laboratory evolved thermophilic enzyme. *Biochim. Biophys. Acta* **1549**: 1-8

Wittmann, C., Heinzle, E. 2002. Genealogy profiling through strain improvement by using metabolic network analysis: metabolic flux genealogy of several generations of lysine-producing corynebacteria. *Appl Environ Microbiol.* **68**: 5843-59.

Woese, C.R. 1987. Bacterial evolution. *Microbiol. Rev.* **51**: 221-271

Wolf, A., Kramer, R., Morbach, S. 2003. Three pathways for trehalose metabolism in *Corynebacterium glutamicum* ATCC13032 and their significance in response to osmotic stress. *Mol Microbiol.* **49**: 1119-1134.

Xia, X., Xie, Z. 2001. DAMBE: Data analysis in molecular biology and evolution. *J. Hered.* **92**: 371-373.

Xie, G., Forst, C., Bonner, C., Jensen, R. A. 2002. Significance of two distinct types of tryptophan synthase beta chain in bacteria, archaea and higher plants. *Genome Biology* **3**: 0004.1-0004.13.

Yamagishi, A., Kon, T., Takahashi, G., Oshima, T. 1998. *From the common ancestor of all living organisms to protoeukaryotic cell* In: Wiegel, J., Adams, M.W.W. (eds.) *The Keys to Molecular Evolution and the Origin of Life?* Taylor & Francis, London, pp. 287-295

Yi, J., Li, K., Draths, K. M., Frost, J. W. 2002. Modulation of phosphoenolpyruvate synthase expression increases shikimate pathway product yields in *E. coli. Biotechnol. Prog.* **18**: 1141-1148.

Supplementary Tables and Figures

Supplementary Table 1. tRNA gene in *Corynebacteria*

Isotype	anticodon	*C. glutamicum*	*C. efficiens*	*C. diphtheriae*
Arg	ACG	2	2	2
	GCG			
	CCG	1	1	1
	TCG			
	CCT	1	1	1
	TCT	1	1	1
Leu	AAG			
	GAG	2	2	1
	CAG	1	1	1
	TAG	1	1	1
	CAA	1	1	1
	TAA	1	1	1
Ser	AGA			
	GGA	1	1	1
	CGA	1	1	1
	TGA	1	1	1
	ACT			
	GCT	1	1	1
Ala	AGC			
	GGC	1	2	1
	CGC			
	TGC	3	1	3
Gly	ACC			
	GCC	3	3	3
	CCC	1	1	1
	TCC	1	1	1
Pro	AGG			
	GGG	1	1	1
	CGG	1	1	1
	TGG	1	1	1
Thr	AGT			
	GGT	2	1	1
	CGT	1	1	1
	TGT	1	1	1
Val	AAC			
	GAC	2	2	2
	CAC	1	1	1
	TAC	1	1	1
Asn	ATT			
	GTT	2	1	1
Asp	ATC			
	GTC	2	2	2

Cys	ACA			
	GCA	1	1	1
Gln	CTG	2	1	1
	TTG	1	1	
Glu	CTC	3	3	2
	TTC	1	1	1
His	ATG			
	GTG	1	1	1
Ile	AAA			
	GAT	2	1	2
	TAT			
Lys	CTT	2	2	1
	TTT	1	1	1
Met	CAT	4	4	3
Phe	AAA			
	GAA	1	1	1
SelCys	TCA			
Trp	CCA	1	1	1
Tyr	ATA			
	GTA	1	1	1
Supres	CTA			
	TTA			

Supplementary Table2 Codon replacement between *C. effciens* and *C. glutamicum*

C. glutamicum -> *C. efficiens*

C. efficiens -> *C. glutamicum*

	TTT	TTC	TTA	TTG	TCT	TCC	TCA	TCG	TAT	TAC	TAA	TAG	TGT	TGC	TGA	TGG	CTT	CTC	CTA	CTG	CCT	CCC	CCA	CCG	CAT	CAC	CAA	CAG	CGT	CGC	CGA	CGG	ATT	ATC	ATA	ATG	ACT	ACC	ACA	ACG	AAT	AAC	AAA	AAG	AGT	AGC	AGA	AGG	GTT	GTC	GTA	GTG	GCT	GCC	GCA	GCG	GAT	GAC	GAA	GAG	GGT	GGC	GGA	GGG
TTT	607	456	4	15	2	0	1	1	21	22	0	0	4	0	0	10	22	10	0	17	0	1	0	0	2	2	2	3	2	0	0	0	12	7	1	9	3	4	1	1	0	2	1	0	1	1	0	0	8	4	3	6	2	2	2	4	0	2	1	0	2	1	1	0
TTC	4684	11550	25	119	14	21	6	4	199	411	0	0	9	12	0	54	90	131	32	140	5	5	4	2	18	33	9	6	3	10	2	1	66	96	5	67	8	25	4	4	6	14	1	8	2	6	2	1	49	54	10	38	17	15	27	17	6	5	3	5	4	12	11	2
TTA	5	3	65	51	2	0	1	1	0	0	0	0	0	0	0	0	32	24	18	43	0	1	0	0	1	0	1	1	0	0	0	3	4	6	1	8	0	1	1	0	1	0	2	0	0	0	0	0	2	1	3	5	0	0	1	0	1	0	2	0	0	0	0	0
TTG	24	26	118	669	2	2	1	6	1	0	0	0	0	3	0	7	283	349	108	669	6	0	5	2	1	4	6	7	3	2	1	4	42	48	3	68	6	10	5	2	0	1	4	4	0	1	1	1	19	9	4	50	8	4	7	8	0	1	5	5	1	2	1	1
TCT	4	2	2	4	234	187	93	48	1	2	0	0	1	1	0	0	2	2	1	0	13	4	4	2	1	4	4	8	0	4	1	0	1	3	0	1	18	17	7	5	5	10	3	2	10	10	0	1	1	0	0	1	25	14	17	10	7	5	7	6	0	3	2	3
TCC	14	26	3	15	2931	7693	1564	1399	6	14	0	0	9	19	0	4	11	10	2	15	36	47	47	27	24	31	54	61	14	26	6	9	16	22	2	20	177	504	90	100	79	160	54	78	118	403	2	4	24	27	6	36	376	399	395	289	104	88	99	92	58	101	41	16
TCA	2	7	3	1	224	389	336	167	1	0	0	0	2	4	0	2	1	3	3	5	7	5	16	4	0	4	7	10	3	3	4	1	1	5	1	3	21	37	17	11	12	11	7	17	20	31	0	5	0	5	3	4	27	36	43	35	13	10	17	8	10	6	3	0
TCG	2	3	0	4	694	1234	508	853	2	2	0	0	2	10	0	1	2	7	1	2	16	13	18	9	5	9	14	16	4	13	4	5	4	7	1	6	43	103	39	55	13	29	19	21	49	108	0	5	6	4	3	10	88	97	139	111	26	30	46	23	15	18	13	4
TAT	121	120	3	4	4	4	0	2	1164	1389	0	0	3	4	0	11	1	4	0	8	1	2	0	0	11	31	2	8	2	9	0	1	5	3	1	4	0	4	1	1	7	7	4	3	1	0	0	1	2	2	2	5	0	2	4	4	2	3	1	1	0	2	0	0
TAC	186	456	6	16	5	17	3	2	1908	6746	0	0	1	5	0	22	11	15	7	18	4	0	4	3	38	102	4	11	2	5	5	4	8	13	0	11	4	13	4	4	13	24	4	5	3	7	0	0	4	6	3	12	7	6	11	3	13	12	5	3	0	4	2	0
TAA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TAG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TGT	6	4	1	3	5	8	6	5	5	3	0	0	473	441	0	1	2	2	0	6	1	1	0	0	1	0	1	1	2	4	1	2	2	2	0	0	1	8	2	1	1	6	0	2	6	8	0	0	5	5	0	7	7	6	7	7	1	0	1	0	3	3	1	0
TGC	5	5	1	5	7	24	5	9	2	6	0	0	585	1583	0	9	5	2	2	4	1	0	0	0	1	4	2	2	4	8	5	2	5	4	1	5	6	17	1	6	2	7	6	2	4	28	1	0	11	11	1	5	11	21	17	16	1	0	3	0	7	11	1	3
TGA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TGG	33	51	5	26	3	8	2	5	21	35	0	0	0	0	0	7645	8	9	6	17	3	2	1	0	2	4	5	5	5	10	1	6	4	15	0	17	1	4	2	5	1	5	7	7	0	4	0	7	6	4	1	16	5	9	6	7	2	2	1	1	9	5	2	3
CTT	18	25	57	223	2	5	0	0	1	0	0	0	0	0	0	1	762	340	107	338	3	0	1	1	4	2	4	2	6	2	0	1	44	57	2	27	4	4	0	0	1	0	2	3	0	2	0	2	12	16	5	16	7	3	2	6	2	3	4	3	1	2	1	0
CTC	114	197	562	2668	7	12	5	5	10	14	0	0	1	6	0	13	2662	5617	684	3788	7	12	8	4	13	19	26	25	17	27	6	5	293	560	15	287	19	51	10	11	4	13	16	10	1	11	2	5	131	179	45	196	36	58	43	34	5	4	16	18	11	9	5	3
CTA	3	8	26	87	0	0	0	0	0	1	0	0	0	0	0	1	74	62	76	120	0	2	1	2	1	0	0	1	1	1	0	1	12	19	3	6	4	2	1	1	0	1	2	1	0	0	1	1	0	5	1	8	1	2	1	2	1	1	0	1	0	0	0	0
CTG	182	270	994	6245	14	21	12	10	11	27	0	0	6	9	0	37	4206	4998	1465	10653	4	7	11	12	16	22	33	90	18	33	10	13	427	640	36	748	26	61	16	29	8	18	39	41	5	13	1	6	189	176	86	362	64	45	69	64	11	13	31	28	7	18	10	3
CCT	2	2	1	4	15	5	7	10	1	3	0	0	0	0	0	0	9	1	1	2	400	156	345	149	3	1	9	8	5	3	1	1	4	3	1	1	13	12	3	3	6	5	9	6	4	5	1	0	8	3	3	2	17	16	15	18	7	3	17	11	2	5	3	3
CCC	11	9	7	19	52	106	51	40	1	4	0	0	0	1	0	3	18	15	6	17	2451	2653	3516	1893	13	25	54	51	18	29	9	6	15	14	4	17	41	88	30	18	16	33	25	45	19	30	3	6	21	20	16	32	94	120	124	81	46	56	98	86	14	32	15	7
CCA	2	4	2	5	16	12	18	5	1	2	0	0	0	1	0	0	2	3	1	3	376	243	934	349	0	6	9	13	2	3	1	0	3	3	2	4	9	15	9	9	4	5	6	6	4	2	0	1	7	4	3	5	18	16	34	22	9	12	16	18	2	6	2	3
CCG	8	8	5	10	43	65	50	40	1	3	0	0	0	3	0	3	14	10	6	22	2625	1767	4585	2769	15	17	50	52	10	21	13	6	14	17	1	12	35	71	30	22	17	44	35	45	10	40	1	6	22	21	15	37	112	97	149	127	55	39	100	71	14	20	18	11
CAT	15	18	2	6	11	12	3	4	32	41	0	0	1	0	0	2	13	6	1	7	1	1	4	3	802	1261	48	50	15	36	11	8	5	7	1	6	8	20	4	3	54	50	17	19	7	8	1	6	3	3	2	3	10	7	10	4	21	23	25	20	5	4	1	1
CAC	36	59	6	18	18	38	17	13	74	165	0	0	2	1	0	15	16	36	2	21	8	17	9	6	1588	5870	96	166	41	113	24	23	8	17	4	14	16	57	10	13	84	227	54	72	13	33	2	9	9	10	4	19	24	41	38	28	81	66	73	43	9	29	8	4
CAA	0	1	4	3	5	7	7	5	1	2	0	0	0	0	0	1	3	2	2	6	3	2	5	2	11	14	313	266	6	11	7	3	3	7	1	4	4	18	7	4	12	15	32	31	1	5	2	3	0	2	1	5	12	5	8	7	15	12	38	24	3	4	2	1
CAG	11	13	9	36	43	100	40	38	5	12	0	0	3	3	0	3	48	44	17	56	30	22	37	27	101	175	3910	8625	63	156	50	36	11	32	3	60	61	134	47	47	78	171	242	423	27	69	15	20	34	29	24	47	132	110	129	91	149	112	600	431	21	37	18	14
CGT	4	5	1	10	15	16	7	8	7	4	0	0	1	5	0	4	10	10	0	10	4	5	13	5	33	59	64	90	2980	3311	674	356	8	9	0	13	16	37	10	11	24	32	161	247	20	19	139	230	10	11	3	10	27	30	28	20	16	8	35	18	14	18	8	5
CGC	11	8	4	32	26	44	24	17	2	19	0	0	6	9	0	11	18	45	12	33	10	17	20	12	56	149	153	196	3486	8741	1264	696	12	20	3	16	28	65	22	16	30	106	394	536	19	80	273	428	12	18	5	30	38	54	58	37	34	21	89	64	22	33	13	10
CGA	1	1	1	0	0	4	2	7	2	1	0	0	0	3	0	1	1	5	3	4	0	2	3	3	7	11	13	16	235	371	234	76	2	1	2	2	4	4	2	5	5	7	42	47	1	5	39	72	1	2	0	3	5	3	4	4	2	0	8	4	2	3	8	2
CGG	8	8	12	16	25	43	11	16	4	6	0	0	1	9	0	17	12	17	10	23	12	6	21	9	46	58	136	217	1175	2266	657	717	7	17	5	18	33	50	17	31	34	46	367	453	20	39	152	344	19	13	8	21	46	42	43	41	19	19	82	84	13	26	8	16
ATT	10	12	2	21	1	1	2	1	1	1	0	0	2	0	0	1	21	20	7	27	0	2	1	1	2	1	1	4	1	0	0	0	873	762	23	25	3	8	4	5	3	4	5	3	1	4	1	0	103	45	17	67	7	7	6	4	0	0	5	3	0	1	1	0
ATC	71	149	62	248	7	28	10	3	9	12	0	0	7	9	0	4	287	420	79	424	9	4	5	2	4	17	15	21	8	10	5	3	7825	14860	212	302	55	198	33	43	10	31	28	26	3	20	4	3	930	1051	277	1072	42	68	64	54	9	8	25	20	9	11	6	3
ATA	4	1	3	6	0	1	1	0	0	0	0	0	0	0	0	1	5	5	9	4	0	1	1	1	0	0	2	4	0	0	0	0	66	85	36	18	0	4	5	0	3	0	1	0	0	1	0	0	10	11	4	16	0	1	1	3	4	1	3	0	0	0	0	0
ATG	40	66	43	261	13	10	13	15	5	9	0	0	2	6	0	9	175	207	51	326	4	9	6	7	7	10	32	47	5	12	1	1	157	246	22	10607	36	86	18	39	6	19	32	42	8	16	3	5	59	71	28	182	33	42	32	35	11	9	25	22	5	7	4	3
ACT	0	1	2	1	24	18	6	7	1	2	0	0	1	2	0	0	1	4	3	1	7	4	6	2	1	3	3	10	2	1	2	0	7	13	2	11	275	266	86	76	16	21	7	17	15	8	0	3	11	6	1	11	31	13	13	15	8	5	16	5	3	5	2	0
ACC	20	44	15	46	345	747	219	190	9	12	0	0	15	14	0	11	50	70	16	73	47	39	50	43	47	88	132	185	40	121	37	19	101	207	20	108	4012	14246	1694	2366	184	372	162	261	165	439	13	21	204	215	87	267	324	377	382	244	173	154	281	222	43	96	33	25
ACA	2	3	2	4	15	37	18	19	0	1	0	0	0	0	0	3	6	4	3	5	5	11	11	9	4	6	11	18	4	7	5	2	12	9	6	13	181	394	230	135	9	16	26	25	18	26	2	2	15	15	7	18	18	31	44	17	21	14	18	11	9	7	0	3
ACG	5	9	5	14	46	83	35	57	2	3	0	0	2	2	0	0	8	12	1	13	9	5	9	11	10	9	26	43	11	19	6	9	18	40	6	45	544	1180	355	766	27	38	42	76	44	57	2	4	30	37	19	57	42	67	64	84	34	21	52	46	8	16	3	4
AAT	3	3	0	3	15	28	10	8	4	2	0	0	0	1	0	0	2	1	2	5	4	0	4	3	27	26	17	25	6	15	2	2	7	6	2	3	24	42	14	6	1097	1556	36	35	32	42	1	5	2	1	2	5	21	15	21	13	132	61	48	18	13	22	13	6
AAC	4	8	2	4	48	113	25	31	4	21	0	0	0	2	0	0	6	7	3	7	10	5	7	3	53	111	59	89	21	46	15	10	9	14	1	13	70	171	39	19	2147	8017	109	144	49	197	2	7	14	19	4	12	44	45	64	48	303	255	112	84	49	99	46	17
AAA	0	1	3	5	8	9	6	2	0	1	0	0	0	0	0	0	1	6	2	4	4	6	4	1	5	14	48	47	24	59	28	12	3	4	1	9	17	18	7	8	27	42	1096	1087	8	15	11	20	5	1	3	4	6	12	17	14	11	13	37	28	7	7	7	0
AAG	2	5	4	10	24	47	14	28	3	5	0	0	1	3	0	1	9	13	5	15	20	1	22	10	23	31	94	228	98	240	47	56	11	12	4	33	33	101	23	29	65	146	2934	7924	6	44	15	55	15	13	5	26	61	61	74	46	43	50	160	127	13	25	14	5
AGT	1	2	1	2	40	59	29	32	0	3	0	0	2	5	0	2	1	2	0	3	3	3	5	2	4	7	7	10	9	5	5	4	2	1	1	4	34	50	24	14	45	56	13	17	419	503	4	2	3	3	1	5	21	15	12	21	25	11	12	12	23	35	6	4
AGC	3	5	3	7	116	269	77	75	2	6	0	0	8	24	0	4	6	4	2	2	7	7	9	7	11	19	17	34	19	39	11	7	4	18	1	3	66	229	38	42	73	200	36	41	495	1788	5	10	17	12	2	9	57	64	52	44	55	60	50	47	54	117	47	23
AGA	0	0	0	4	0	2	2	1	1	2	0	0	0	0	0	0	0	0	0	2	1	2	3	0	1	1	8	6	32	73	35	21	2	2	1	3	3	4	6	2	2	3	64	49	8	8	84	32	2	0	1	3	3	3	2	1	2	3	13	2	2	1	2	1
AGG	4	4	1	4	4	8	7	5	0	2	0	0	1	0	0	9	2	2	2	6	3	3	3	3	9	9	18	37	161	289	113	110	4	3	0	9	11	34	13	10	13	20	180	308	9	8	60	247	5	4	2	8	16	15	14	18	9	3	29	22	6	7	7	3
GTT	9	9	4	13	3	5	7	2	6	1	0	0	0	2	0	0	16	19	9	19	2	3	3	0	3	0	6	4	1	4	0	1	109	135	8	22	28	41	7	6	4	9	9	5	3	2	0	2	1134	593	226	580	33	29	37	18	5	4	15	14	9	4	3	4
GTC	34	85	33	100	15	38	12	12	2	24	0	0	6	13	0	9	96	155	30	150	12	6	14	11	8	14	23	37	8	15	5	2	633	1283	40	111	56	226	34	42	14	30	23	36	7	23	2	8	4479	6452	1135	4456	159	199	195	123	24	16	86	44	13	27	13	5
GTA	4	2	4	7	5	3	4	0	0	0	0	0	2	0	0	1	8	9	6	9	2	0	3	0	0	0	9	4	0	1	2	0	41	64	8	14	5	13	5	3	1	1	4	4	1	1	0	1	235	161	231	275	9	10	19	12	1	3	10	2	0	2	3	0
GTG	45	70	52	226	24	43	14	27	11	11	0	0	9	24	0	17	130	166	48	286	20	13	20	16	12	17	31	42	5	23	5	12	779	1175	62	224	83	213	59	84	15	29	30	34	10	29	1	5	4244	4026	1763	8601	174	200	216	209	22	17	104	63	26	17	15	13
GCT	3	8	1	6	66	76	27	24	1	2	0	0	4	3	0	0	10	6	2	14	14	10	20	7	2	9	16	19	5	9	1	2	9	13	0	16	44	50	16	8	12	20	25	27	11	25	1	5	26	19	8	32	1203	560	880	479	27	27	52	36	26	34	20	14
GCC	27	44	14	56	343	887	249	207	13	19	0	0	9	34	0	17	38	60	14	64	78	71	132	60	28	60	137	162	33	83	15	15	64	145	9	83	188	482	104	106	64	176	123	170	75	191	5	13	218	221	101	256	6885	8852	8295	5375	204	195	443	327	207	321	133	64
GCA	8	11	6	14	64	121	59	44	1	2	0	0	2	6	0	5	8	19	5	18	28	17	29	16	3	15	31	33	5	12	5	7	13	22	1	19	55	88	35	35	19	28	28	46	24	38	1	6	43	40	29	59	1223	940	2264	1037	48	43	107	91	27	37	30	12
GCG	21	17	12	31	170	307	126	150	5	1	0	0	3	13	0	7	25	23	10	38	48	42	83	42	13	25	59	83	14	24	7	16	32	71	7	54	65	171	67	98	30	62	60	105	39	85	8	10	110	96	50	200	2727	2760	3628	3831	96	82	231	185	110	124	78	37
GAT	1	4	3	3	51	63	27	28	4	5	0	0	0	1	0	1	2	5	1	3	22	17	28	15	33	39	37	78	10	21	4	7	9	6	2	4	46	83	20	23	211	339	58	57	20	82	5	6	9	9	7	15	80	62	59	52	7597	4155	662	504	66	110	46	28
GAC	9	3	2	5	55	119	49	42	3	8	0	0	0	2	0	0	3	7	3	6	27	26	45	19	24	58	79	107	13	30	3	6	6	11	1	11	47	140	40	35	235	559	80	102	33	101	1	11	24	13	11	21	112	131	112	77	7375	9031	1083	863	88	166	55	30
GAA	3	2	3	4	22	45	21	20	0	0	0	0	1	4	0	2	4	9	7	7	11	10	14	12	11	22	123	129	13	21	6	4	9	8	3	15	21	61	26	19	32	70	76	64	11	31	0	4	18	14	16	27	50	58	80	63	363	256	4115	2570	12	22	15	14
GAG	5	6	6	17	70	171	61	83	5	9	0	0	1	2	0	1	18	14	2	22	38	33	77	41	34	78	346	519	24	58	19	19	21	30	0	22	82	195	53	75	95	176	163	288	36	89	10	18	60	46	27	71	233	231	349	224	1307	1146	11817	10229	59	94	48	48
GGT	7	6	5	2	31	54	22	24	4	1	0	0	5	6	0	5	4	5	1	9	6	10	17	4	10	12	36	37	17	21	9	6	2	7	2	10	17	45	17	15	59	98	31	24	61	126	5	7	14	14	5	11	105	93	112	88	138	89	74	49	6158	6403	2002	602
GGC	7	10	3	5	33	84	22	20	1	5	0	0	3	18	0	7	4	11	1	12	10	14	16	9	13	29	33	42	16	34	8	5	8	13	1	10	19	64	14	21	65	162	39	41	31	188	1	5	15	15	5	14	100	125	130	96	130	175	87	63	4547	10071	3054	813
GGA	1	2	0	0	9	9	3	11	1	0	0	0	1	2	0	6	3	1	0	0	6	5	6	2	1	2	8	16	3	9	3	4	4	3	0	1	3	15	8	2	16	33	14	8	10	25	14	2	0	1	1	6	28	36	44	26	30	22	34	21	852	1264	1270	272
GGG	9	7	5	7	20	46	17	24	3	6	0	0	1	4	0	13	7	8	0	10	9	6	12	13	14	12	43	49	10	25	8	9	6	9	1	11	17	35	13	18	48	59	52	42	31	97	5	10	7	4	6	20	59	74	92	93	119	94	134	116	1923	2708	1377	856

Supplementary Table 3 Amino acid replacement using orthologous genes with identity from 60% to 95% between *C. glutamicum* and *C. efficiens*

C. efficiens -> *C. glutamicum*

		Ala	Cys	Ala	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr
C. glutamicum -> *C. efficiens*	Ala	50566	96	648	1244	90	1316	146	259	283	463	134	241	1019	475	536	2372	1763	1634	28	37
	Cys	73	3068	3	9	22	37	4	19	4	25	7	4	5	6	34	86	38	55	0	12
	Asp	679	2	27982	3055	13	794	187	20	114	38	13	747	237	283	140	447	427	91	2	24
	Glu	1428	3	3113	28520	9	560	170	53	348	103	42	255	406	1096	460	419	641	321	5	13
	Phe	135	20	18	15	17176	49	126	245	7	868	100	18	41	26	52	75	78	259	82	876
	Gly	1265	28	581	314	28	43863	62	28	69	64	18	263	163	100	227	610	247	142	19	12
	His	149	5	149	147	54	97	9466	23	72	78	14	221	74	299	443	120	156	54	5	178
	Ile	385	15	35	66	187	56	38	24516	35	2191	414	33	69	57	94	91	435	4332	18	31
	Lys	578	8	284	593	11	248	157	54	12938	123	73	312	188	727	2855	320	600	136	14	16
	Leu	494	35	44	105	596	92	124	1642	72	48877	1095	39	167	219	306	126	356	1626	71	95
	Met	160	4	13	25	71	28	19	331	36	1170	10591	15	30	57	58	32	153	410	15	15
	Asn	396	13	1321	373	22	557	416	51	273	43	28	12726	123	273	333	708	678	88	5	48
	Pro	656	6	199	235	15	148	51	21	71	76	20	36	25111	122	156	297	254	121	5	15
	Gln	528	6	296	1121	20	248	342	42	415	190	74	192	240	13003	978	237	420	157	8	26
	Arg	301	26	113	195	23	238	295	37	664	173	23	137	139	372	29928	229	322	92	27	32
	Ser	3378	109	682	659	58	1013	173	81	215	139	67	623	623	353	525	23062	2623	264	22	48
	Thr	1603	40	415	510	50	313	126	353	235	263	182	386	395	313	423	1723	26670	897	12	27
	Val	1485	48	101	265	174	129	50	3585	65	1529	366	52	197	136	182	177	994	38327	25	38
	Trp	27	8	3	4	67	29	17	6	0	58	9	0	6	3	40	12	13	26	7589	32
	Tyr	43	16	20	16	650	22	307	22	9	63	13	30	16	21	47	38	31	49	58	11115

Supplementary Table 4 Amino acid replacement using orthologous genes with identity more than 95% between *C. glutamicum* and *C. efficiens*

C. efficiens -> *C. glutamicum*

		Ala	Cys	Ala	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr
C. glutamicum -> *C. efficiens*	Ala	848	0	2	3	0	2	0	0	2	0	0	5	6	0	0	5	7	5	0	0
	Cys	0	47	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
	Asp	2	0	525	18	0	0	0	0	0	0	0	1	1	2	0	1	1	1	0	0
	Glu	2	0	20	676	0	0	0	0	2	0	1	0	0	6	0	1	2	0	0	0
	Phe	0	0	0	0	260	0	0	1	0	1	0	0	0	0	0	0	0	1	0	4
	Gly	3	0	1	0	0	799	0	0	0	0	0	1	0	1	0	1	0	0	0	0
	His	0	0	0	1	0	0	160	0	0	1	0	1	0	1	1	0	0	0	0	0
	Ile	0	0	0	0	1	0	0	573	0	6	0	0	0	0	0	0	0	14	0	0
	Lys	2	0	0	1	0	0	0	0	509	0	0	0	1	8	5	0	2	0	0	0
	Leu	0	0	0	0	0	0	0	3	0	757	5	0	0	1	0	0	0	1	0	2
	Met	0	0	0	0	0	0	0	1	0	2	203	0	0	0	0	0	1	1	0	0
	Asn	0	0	5	2	0	3	0	0	0	0	0	303	0	1	2	1	6	1	0	0
	Pro	0	0	0	0	0	1	0	0	1	0	0	1	384	0	1	1	0	0	0	0
	Gln	0	0	0	1	0	0	0	0	3	2	0	0	0	344	2	1	1	1	0	0
	Arg	0	0	0	0	0	0	2	0	4	0	0	0	0	1	705	1	0	0	0	0
	Ser	14	0	0	0	0	5	1	0	0	0	0	2	3	0	1	427	8	0	0	0
	Thr	5	0	0	0	0	0	0	1	0	0	1	1	0	1	1	13	477	3	0	0
	Val	3	0	0	0	0	0	0	20	1	5	2	0	1	0	0	0	2	782	0	0
	Trp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	66	0
	Tyr	0	0	0	0	4	0	1	0	0	0	0	0	0	0	0	0	0	0	0	228

Supplementary Table 5 Amino acid replacement using orthologous genes with identity under 60% between *C. glutamicum* and *C. efficiens*

C. efficiens -> *C. glutamicum*

C. glutamicum -> *C. efficiens*

	Ala	Cys	Ala	Glu	Phe	Gly	Hs	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr
Ala	8062	60	281	432	84	773	92	204	78	413	93	102	462	199	308	870	928	823	24	43
Cys	84	613	5	9	16	112	11	18	2	44	5	5	14	7	43	54	28	52	4	6
Asp	292	8	4196	880	11	387	140	26	43	48	10	262	179	141	142	251	233	64	8	11
Glu	532	6	992	3711	22	293	126	38	110	82	17	102	233	457	322	261	345	175	9	18
Phe	138	13	25	23	2642	75	81	207	10	765	80	17	44	15	61	84	90	220	66	308
Gly	645	39	269	170	28	7594	60	49	40	121	18	125	168	82	227	303	176	148	24	11
His	117	10	149	82	34	85	1524	24	26	62	8	74	53	150	219	78	106	46	16	83
Ile	291	22	38	40	163	79	35	3318	18	1188	190	21	75	39	88	85	251	1756	29	32
Lys	249	7	185	296	18	130	82	34	1183	83	34	98	123	253	861	156	234	77	6	12
Leu	411	47	58	96	435	134	72	885	33	8982	441	43	190	103	254	125	272	1087	59	76
Met	97	5	9	27	63	34	18	153	15	498	1283	13	28	33	51	43	97	233	11	16
Asn	187	9	563	193	10	249	185	24	81	37	16	1547	93	117	217	264	288	76	11	24
Pro	323	6	129	147	17	147	52	34	29	95	13	25	4770	82	115	187	213	111	11	15
Gln	262	6	195	480	15	136	139	34	140	117	48	102	139	1916	475	174	248	84	14	23
Arg	239	27	88	163	18	307	161	59	236	133	30	79	118	221	4575	149	222	90	22	30
Ser	1313	43	378	337	58	725	103	78	85	175	64	219	374	210	340	3654	1044	208	20	36
Thr	833	21	239	270	41	228	68	200	72	215	83	141	252	132	257	695	4097	459	17	32
Val	786	33	63	137	133	153	49	1252	34	995	221	26	144	87	106	132	432	5980	33	39
Trp	28	3	2	11	29	72	19	14	2	71	10	0	8	10	76	21	21	35	1630	30
Tyr	56	6	42	21	263	17	156	38	11	95	15	23	15	29	64	46	49	57	47	1582

Supplementary Table 6 Glutamic acid and lysine productivities in *Corynebacteria*.

Species	Amino acid	relative productivities	conditions
C. glutamicum	glutamic acid	100[a)]	32 ºC
C. glutamicum	glutamic acid	40	37 ºC
C. efficiens	glutamic acid	80	32 ºC
C. efficiens	glutamic acid	78	37 ºC
C. glutamicum	lysine	100[b)]	31.5 ºC, AEC^r, Ala
C. efficiens	lysine	25	43 ºC, AEC^r

[a] Glutamate production in typical experiments using the biotin limitation method as a percent of the production by *C. glutamicum* at 32 ºC (Nishio et al., 2003). [b] Lysine productivities were expressed as a percent of the production by *C. glutamicum* (Ikeda, 2003).

Supplementary Table 7 Amino acid biosynthesis related genes in high GC gram-positive bacteria

Product	gene name	*C. efficiens*	*C. glutamicum*	*C. diphtheriae*	*M. leprae*	*M. tuberculosis*	*S. coelicolor*
2-isopropylmalate synthase	*leuA*	CE0216	Cgl0248	DIP0266	ML2324	Rv3710	
2-oxoglutarate dehydrogenase E1 component	*odhA*	CE1190	Cgl1129	DIP1002	ML1095	Rv1248c	SCO5281
3-dehydroquinate dehydratase	*aroQ*	CE0442 CE1739	Cgl0423	DIP1342	ML0519	Rv2537c	SCO1961
3-dehydroquinate synthase	*aroB*	CE1740	Cgl1621	DIP1343	ML0518	Rv2538c	SCO1494
3-deoxy-D-arabinoheptulosonate-7-phosphate synthase	*aroG aroH*	CE1054 CE2073	Cgl0990 Cgl2178	DIP1616	ML0896	Rv2178c	SCO2115 SCO3210
3-isopropylmalate dehydrogenase	*leuB*	CE1383	Cgl1286	DIP1105	ML1691	Rv2995c	SCO5522
acetylglutamate kinase	*argB*	CE1528	Cgl1396	DIP1169	ML1408	Rv1654	SCO1578
acetylornithine aminotransferase	*argD*	CE1529	Cgl1397	DIP1170	ML1409	Rv1655	SCO1577
aconitate hydratase	*acn*	CE1661	Cgl1540	DIP1283	ML1814	Rv1475c	SCO5999
adenosylmethionine—8-amino-7-oxononanoate transaminase	*bioA*	CE1421	Cgl2604	DIP1191	ML1216	Rv1568	SCO1245
anthranilate phosphoribosyltransferase	*trpD*	CE2870	Cgl3032	DIP2354	ML0883	Rv2192c	SCO3212 SCO2417
anthranilate synthase							SCO2117
anthranilate synthase component I	*trpE*	CE2868	Cgl3029	DIP2352	ML1269	Rv1609	SCO2043 SCO3214
anthranilate synthase component II	*trpG*	CE2869	Cgl3031	DIP2353	ML0015	Rv0013	SCO3220 SCO3851
argininosuccinate lyase	*argH*	CE1533	Cgl1401	DIP1174	ML1413	Rv1659	SCO1570
argininosuccinate synthetase	*argG*	CE1532	Cgl1400	DIP1173	ML1412	Rv1658	SCO7036
aspartate-semialdehyde dehydrogenase	*asd*	CE0221	Cgl0252	DIP0279	ML2322	Rv3708c	SCO2640
aspartokinase	*ask*	CE0220	Cgl0251	DIP0277	ML2323	Rv3709c	SCO3615

ATP phosphoribosyltransferase	*hisG*	CE1634	Cgl1504	DIP1256	ML1310	Rv2121c	SCO1438
biotin synthase	*bioB*	CE0089	Cgl0072	DIP0105 DIP1124	ML1120	Rv1589	SCO1124
branched-chain amino acid aminotransferase	*ilvE*	CE2095	Cgl2204	DIP1636	ML0866	Rv2210c	SCO5523
chorismate synthase	*aroC*	CE1743	Cgl1623	DIP1345	ML0516	Rv2540c	SCO1496
citrate synthase		CE0718	Cgl0696 Cgl0659			Rv1131	
citrate synthase	*gltA*	CE0905	Cgl0829	DIP0785	ML2130	Rv0896	SCO2736
cyclase HisF	*hisF*	CE1994	Cgl2094	DIP1558	ML1263	Rv1605	SCO2048
cystathionine beta-lyase	*metC*	CE2211	Cgl2309	DIP1736	ML1794	Rv0075 Rv2294	SCO0731
dethiobiotin synthase	*bioD*	CE1420	Cgl2605	DIP1189 DIP1192	ML1218	Rv1570	SCO1246
diaminopimelate dehydrogenase	*ddh*	CE2498	Cgl2617				
dihydrodipicolinate reductase	*dapB*	CE1866	Cgl1973	DIP1466		Rv2773c	SCO5739
dihydrodipicolinate synthase	*dapA*	CE1864	Cgl1971	DIP1464	ML1513	Rv2753c	SCO5744 SCO1912
dihydrolipoamide dehydrogenase	*lpd*	CE0383	Cgl0366	DIP0368	ML2387	Rv0462	SCO2180
dihydrolipoamide dehydrogenase		CE0708	Cgl0688	DIP0645		Rv3303c Rv2713	SCO4919
dihydrolipoamide dehydrogenase						Rv0794c	SCO0884 SCO3443 SCO3460
dihydrolipoamide dehydrogenase						Rv0794c	SCO0884 SCO3443 SCO3460

EPSP synthase	*aroA*	CE0779	Cgl0764	DIP0706	ML0792	Rv3227	SCO5212 SCO6819
fructose-bisphosphate aldolase	*fda*	CE2601	Cgl2770	DIP2094	ML0286	Rv0363c	SCO3649
fumarate dehydrogenase subunit A						Rv1552	SCO5106
fumarate dehydrogenase subunit B						Rv1553	SCO5106
fumarate dehydrogenase subunit C						Rv1554	SCO5108
fumarate dehydrogenase subunit D						Rv1555	
glutamate 5-kinase	*proB*	CE2265	Cgl2356	DIP1777	ML1464	Rv2439c	SCO4958
glutamate N-acetyltransferase	*argJ*	CE1527	Cgl1395	DIP1168	ML1407	Rv1653	SCO1579
glutamate synthase large subunit	*gltB*	CE0158	Cgl0184		ML0061	Rv3859c	SCO2026
glutamate synthase small subunit	*gltD*	CE0159	Cgl0185		ML0062	Rv3858c	SCO2025 SCO1977
glutamate-5-semialdehyde dehydrogenase	*proA*	CE2260	Cgl2354	DIP1776	ML1458	Rv2427c	SCO2587
glutamine amidotransferase	*hisH*	CE1997	Cgl2097	DIP1561	ML1260	Rv1602	SCO2051
glutamine synthetase I	*glnA*	CE2104 CE2116	Cgl2214	DIP1644	ML0975	Rv2220	SCO2198
glutamine synthetase II	*glnA2*	CE2127	Cgl2229	DIP1671	ML1631	Rv2222c	SCO2585
glutamte dehydrogenase	*gdh*	CE1982	Cgl2079	DIP1547			SCO4683
glutamyl-tRNA(Gln) amidotransferase subunit A		CE1345	Cgl1247	DIP1080	ML1702	Rv3011c	SCO5499
glutamyl-tRNA(Gln) amidotransferase subunit B		CE1351	Cgl1259	DIP1089	ML1700	Rv3009c	SCO5501
glutamyl-tRNA(Gln) amidotransferase subunit C		CE1344	Cgl1246	DIP1079	ML1703	Rv3012c	SCO5498
glyceraldehyde-3-phosphate dehydrogenase		CE1008	Cgl0937	DIP0892			SCO7040
glyceraldehyde-3-phosphate dehydrogenase	*gap*	CE1706	Cgl1588	DIP1310	ML0570	Rv1436	SCO1947

							SCO7511
GTP-dependent phosphoenolpyruvate carboxykinase	*pck*	CE2691	Cgl2863	DIP2180	ML2624	Rv0211	SCO4979
histidinol dehydrogenase	*hisD*	CE2003	Cgl2102	DIP1566	ML1257	Rv1599	SCO2054
histidinol-phosphate aminotransferase	*hisC2*	CE0193	Cgl0218	DIP0178		Rv3772	SCO3944
histidinol-phosphate aminotransferase	*hisC1*	CE2002	Cgl2101	DIP1565	ML1258	Rv1600	SCO2053
homoserine o-acetyltransferase	*metX*	CE0678	Cgl0652	DIP0623	ML0682	Rv3341	
imidazoleglycerol-phosphate dehydratase	*hisB*	CE2001	Cgl2100	DIP1564	ML1259	Rv1601	SCO2052
indole-3-glycerol phosphate synthase/N-(5'-phospho-ribosyl)anthranilate isomerase	*trpC*	CE2871 CE1991	Cgl3033 Cgl2091	DIP2355 DIP1555	ML1271	Rv1611	SCO2039 SCO3211
isocitrate lyase	*aceA*	CE2232	Cgl2331			Rv0467	SCO0982
ketol-acid reductoisomerase	*ilvC*	CE1367	Cgl1273	DIP1100	ML1694	Rv3001c	SCO5514 SCO7154
LtsA protein / asparagine synthase		CE2088	Cgl2196	DIP1630	ML0874	Rv2201	SCO0386
malate dehydrogenase	*mdh*	CE2285	Cgl2380	DIP1787	ML1091	Rv1240	SCO4827
malate synthase	*masZ*	CE2231	Cgl2329		ML2069	Rv1837c	
malate:quinone oxidoreductase	*mqo*	CE1894	Cgl2001	DIP1492		Rv2852c	
malic enzyme	*malE*	CE2839	Cgl3007				SCO5261 SCO2951
N-acetylglutamate-5-semialdehyde dehydrogenase	*argC*	CE1526	Cgl1394	DIP1167	ML1406	Rv1652	SCO1580
NADP-dependent isocitrate dehydrogenase	*icd*	CE0682	Cgl0664	DIP0631	ML2672	Rv0066c	SCO7000
O-acetylhomoserine (thiol)-lyase	*metB*	CE2343	Cgl2446		ML2394	Rv1079	SCO1808
O-acetylhomoserine sulfhydrylase	*metY*	CE0679	Cgl0653	DIP0630		Rv3340	
ornithine carbamolytransferase	*argF*	CE1530	Cgl1398	DIP1171	ML1410	Rv1656	SCO5976

phosphoenolpyruvate carboxylase	*ppc*	CE1703	Cgl1585	DIP1122	ML0578		SCO3127
phosphoglycerate kinase	*pgk*	CE1705	Cgl1587	DIP1309	ML0571	Rv1437	SCO1946
phosphoribosyl-ATP pyrophosphatase	*hisE*	CE1635	Cgl1505	DIP1257	ML1309	Rv2122c	SCO1439
phosphoribosylformimino-5-amino-1-phosphoribosyl-4-imidazolecarboxamide isomerase	*hisA*	CE1996	Cgl2096	DIP1560	ML1261	Rv1603	SCO2050
prephenate dehydratase	*pheA*	CE2732	Cgl2899	DIP2246	ML0078	Rv3838c	SCO3962
PTS enzyme I		CE1826	Cgl1933	DIP1428			SCO1391
PTS glucose-specific IIABC		CE1458	Cgl1360	DIP1151			
putative 3-isopropylmalate dehydratase large subunit	*leuC*	CE1427	Cgl1315	DIP1127	ML1685	Rv2988c	SCO5553
putative 3-isopropylmalate dehydratase small subunit	*leuD*	CE1428	Cgl1316	DIP1128	ML1684	Rv2987c	SCO5554
putative 5-methyltetrahydrofolate--homocysteine methyltransferase	*metH*	CE1637	Cgl1507	DIP1259	ML1307	Rv2124c	SCO1657
putative 5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase	*metE*	CE1209	Cgl1139		ML0961	Rv1133c	SCO0985
putative 6-carboxyhexanoate-CoA ligase	*bioW*			DIP1381			
putative 6-phosphofructokinase		CE1348	Cgl1250	DIP1088	ML1701	Rv3010c	SCO2119 SCO5426 SCO1214
putative 6-phosphofructokinase		CE1828 CE1825	Cgl1935 Cgl1932	DIP1430		Rv2029c	SCO3197 SCO4283
putative 6-phosphogluconate dehydrogenase		CE1588	Cgl1452	DIP1213	ML1369	Rv1844c	SCO0975
putative 6-phosphogluconolactonase		CE1698	Cgl1578	DIP1306	ML0579	Rv1445c	SCO1939

putative 8-amino-7-oxononanoate synthase	*bioF*			DIP1382	ML1217	Rv0032 Rv1569	SCO1243
putative acetolactate synthase large subunit	*ilvB*	CE1365	Cgl1271	DIP1098	ML1696	Rv3003c	SCO5512
putative acetolactate synthase small subunit	*ilvN*	CE1366	Cgl1272	DIP1099	ML1695	Rv3002c	SCO5513
putative adenosine 5'-phosphosulphate reductase	*cysH*	CE2642	Cgl2816			Rv2392	SCO6100
putative aminotransferase	*dapC*	CE1161	Cgl1103	DIP0974	ML1488	Rv1178	SCO5136
putative aspartate aminotransferase	*aspB*	CE2489	Cgl2599	DIP1929		Rv3565	
putative aspartate aminotransferase	*aspC*	CE2661	Cgl2844	DIP2136	ML2502	Rv0337c	SCO6222 SCO4984
putative cyclohexadienyl dehydrogenase	*tyrA*	CE0195	Cgl0226	DIP0245	ML2472	Rv3754	SCO3221
putative cysteine synthase	*cysM*	CE1418	Cgl2136			Rv0848	SCO0992
putative cysteine synthase	*cysK*	CE2446	Cgl2562	DIP1890	ML0839	Rv2334	
putative cysteine synthase	*cysM*					Rv1336	SCO2910
putative cysteine synthase (putative cystathionine beta-synthase)	*cysM2*				ML2396	Rv1077	SCO3077
putative D-3-phosphoglycerate dehydrogenase	*serA*	CE1379	Cgl1284	DIP1104	ML1692	Rv2996c	SCO5515
putative diaminopimelate decarboxylase	*lysA*	CE1277	Cgl1180	DIP1035	ML1128	Rv1293	SCO5353
putative diaminopimelate epimerase	*dapF*	CE1837	Cgl1943	DIP1442	ML0996	Rv2726c	SCO5793
putative dihydrolipoamide acyltransferase		CE2098	Cgl2207	DIP1639	ML0861	Rv2215	SCO2181
putative dihydroxy-acid dehydratase	*ilvD*	CE1362 CE2439	Cgl1268	DIP1096	ML2608	Rv0189c	SCO3345 SCO1888 SCO1176
putative enolase		CE1042	Cgl0974	DIP0917	ML0255	Rv1023	SCO3096
putative ferredoxin—nitrite reductase	*cysI*	CE2644	Cgl2817			Rv2391	SCO6102
putative fructose-1,6-bisphosphatase		CE1075	Cgl1019	DIP0939	ML1946	Rv1099c	SCO5047

putative fumarate hydratase		CE1071	Cgl1010	DIP0938	ML1947	Rv1098c	SCO5042
putative glucose-6-phosphate 1-dehydrogenase		CE1696	Cgl1576	DIP1304		Rv1447c	SCO1937
		CE0542				Rv1121	SCO6661
putative glucose-6-phosphate isomerase		CE0927	Cgl0851	DIP0832	ML0150	Rv0946c	SCO1942
							SCO6659
putative homoserine dehydrogenase	*thrA*	CE1289	Cgl1183	DIP1036	ML1129	Rv1294	SCO5354
putative homoserine kinase	*thrB*	CE1290	Cgl1184	DIP1037	ML1131	Rv1296	SCO5356
putative phosphoenolpyruvate synthase		CE0560	Cgl0551				
putative phosphoglycerate mutase		CE0423	Cgl0402	DIP0389	ML2441	Rv0489	SCO4209
putative phosphoglycerate mutase		CE2254	Cgl2350	DIP1773	ML1452	Rv2419c	SCO2576
putative phosphoglycerate mutase		CE2731	Cgl2898	DIP2245	ML0079	Rv3837c	SCO1666
				DIP1678			
putative phosphoglycerate mutase						Rv3214	SCO2806
							SCO6218
putative phosphoribosyl-AMP cyclohydrolase	*hisI*	CE1993	Cgl2093	DIP1557	ML1264	Rv1606	SCO2044
putative phosphoserine aminotransferase	*serC*	CE0903	Cgl0828	DIP0784	ML2136	Rv0884c	SCO4366
putative phosphoserine phosphatase	*serB2*	CE2417	Cgl2522	DIP1863	ML1727	Rv3042c	SCO3077
putative phosphoserine phosphatase	*serB1*	CE0434	Cgl0415	DIP0398	ML2424	Rv0505c	SCO3322
putative pyruvate dehydrogenase E1 component		CE2143	Cgl2248	DIP1687	ML1651	Rv2241	SCO2371
							SCO7124
							SCO2183
putative ribulose-phosphate 3-epimerase		CE1717	Cgl1598	DIP1320	ML0554	Rv1408	SCO1464
putative serine O-acetyltransferase	*cysE*	CE2447	Cgl2563	DIP1891	ML0838	Rv2335	
putative succinate dehydrogenase subunit A		CE0387	Cgl0371	DIP0371			SCO0923
putative succinate dehydrogenase subunit B		CE0388	Cgl0372	DIP0372			SCO0922
putative succinate dehydrogenase subunit C		CE0386	Cgl0370	DIP0370			SCO0924

putative succinyl-CoA synthetase alpha subunit		CE2449	Cgl2565		ML0156	Rv0952	SCO4809 SCO6586
putative succinyl-CoA synthetase beta subunit		CE2451	Cgl2566		ML0155	Rv0951	SCO4808 SCO6585
putative sugar phosphate isomerase (rpi)		CE2318	Cgl2423	DIP1796	ML1484	Rv2465c	SCO2627 SCO1224 SCO0579
putative sulfate adenylate transferase subunit 1	*cysN*	CE2640	Cgl2814			Rv1286	SCO6097
putative sulfate adenylate transferase subunit 2	*cysD*	CE2641	Cgl2815			Rv1285	SCO6098
putative transaldolase		CE1695	Cgl1575	DIP1303	ML0582	Rv1448c	SCO6663 SCO1936
pyrroline-5-carboxylate reductase	*proC*	CE0430	Cgl0410	DIP0394	ML2430	Rv0500	SCO3337
pyruvate carboxylase	*pyc*	CE0709	Cgl0689	DIP0646		Rv2967c	SCO0546
pyruvate kinase		CE1989	Cgl2089	DIP1553	ML1277	Rv1617	SCO5423 SCO2014
pyruvate kinase	*pyk*	CE2752	Cgl2910				
serine hydroxymethyltransferase	*glyA*	CE1058	Cgl0996	DIP0932	ML1953	Rv1093 Rv0070c	SCO5470
shikimate 5-dehydrogenase	*aroE*	CE0443 CE1194 CE1745	Cgl0424 Cgl1132 Cgl1629	DIP1006 DIP1347	ML0515	Rv2552c	SCO1498
shikimate kinase	*aroK*	CE1741	Cgl1622	DIP1344	ML0517	Rv2539c	SCO1495
succinate dehydrogenase subunit A					ML0697	Rv3318	SCO4856
succinate dehydrogenase subunit B					ML0696	Rv3319	SCO4855
succinate dehydrogenase subunit C					ML0699	Rv3316	SCO4858
succinate dehydrogenase subunit D					ML0698	Rv3317	SCO4857

succinyl-diaminopimelate desuccinylase	*dapE*	CE1166	Cgl1109	DIP0982	ML1059	Rv1202	SCO5139
tetrahydropicolinate succinylase	*dapD*	CE1163	Cgl1106	DIP0979	ML1058	Rv1201c	SCO1916
		CE1165	Cgl1108	DIP0981			
threonine dehydratase	*ilvA*	CE2026	Cgl2127	DIP1579	ML1209	Rv1559	SCO4962
							SCO0821
							SCO7292
threonine synthase	*thrC*	CE2122	Cgl2220	DIP1666		Rv1295	SCO2241
transketolase	*tkt*	CE1694	Cgl1574	DIP1302	ML0583	Rv1449c	SCO1935
							SCO6497
							SCO6663
triose-phosphate isomerase	*tpi*	CE1704	Cgl1586	DIP1308	ML0572	Rv1438	SCO1945
tryptophan synthase alpha chain	*trpA*	CE2873	Cgl3035	DIP2361	ML1273	Rv1613	SCO2036
tryptophan synthase beta chain	*trpB*	CE2872	Cgl3034	DIP2360	ML1272	Rv1612	SCO2037
		CE2880		DIP2351			

(A)

```
C. glutamicum  1    MLQLGLRHNQPTTNVTVDKTKLNKPSRSKEKRRVPAVSSASTFGQ 45
                                                        :||||||||
C. efficiens   1                                        MSSASTFGQ 9

C. glutamicum  46   NAWLVDEMFQQFQKDPKSVDKEWRELFEAQGGP...NTTPATTEA 87
                    ||||||||||||:|||:|||||||||||:||||    |||| ||
C. efficiens   10   NAWLVDEMFQQFKKDPQSVDKEWRELFESQGGPQAEKATPATPEA 54

C. glutamicum  88   QPSAPKES.........AKPAPKAAPA...AKAAPRVETKPADKT 120
                    : :|  :|         | || | |||   ||||| |:   | |
C. efficiens   55   KKAASSQSSTSGQSTAKAAPAAKTAPASAPAKAAP.VKQNQASKP 98

C. glutamicum  121  APKAKESSVPQQPKLPEPGQTPIRGIFKSIAKNMDISLEIPTATS 165
                    | ||||| : :   :|||| ||:||||||||||||:|||:|||||
C. efficiens   99   AKKAKESPLSKPAAMPEPGTTPLRGIFKSIAKNMDLSLEVPTATS 143

C. glutamicum  166  VRDMPARLMFENRAMVNDQLKRTRGGKISFTHIIGYAMVKAVMAH 210
                    |||||||||||||||||||||||||||||||||||||||||||||
C. efficiens   144  VRDMPARLMFENRAMVNDQLKRTRGGKISFTHIIGYAMVKAVMAH 188

C. glutamicum  211  PDMNNSYDVIDGKPTLIVPEHINLGLAIDLPQKDGSRALVVAAIK 255
                    ||||||||::||||:|:||||||||||||||||||||||||||||
C. efficiens   189  PDMNNSYDIVDGKPSLVVPEHINLGLAIDLPQKDGSRALVVAAIK 233

C. glutamicum  256  ETEKMNFSEFLAAYEDIVARSRKGKLTMDDYQGVTVSLTNPGGIG 300
                    ||||| ||:|| ||||:|||| ||||||||||||:|||||||||
C. efficiens   234  ETEKMTFSQFLEAYEDVVARSRVGKLTMDDYQGVTISLTNPGGIG 278

C. glutamicum  301  TRHSVPRLTKGQGTIIGVGSMDYPAEFQGASEDRLAELGVGKLVT 345
                    ||||:||||||||||||||||||||||||||||||||||||||||
C. efficiens   279  TRHSIPRLTKGQGTIIGVGSMDYPAEFQGASEDRLAELGVGKLVT 323

C. glutamicum  346  ITSTYDHRVIQGAVSGEFLRTMSRLLTDDSFWDEIFDAMNVPYTP 390
                    ||||||||||||| |||||||||:|| ||:||| ||: |||||||
C. efficiens   324  ITSTYDHRVIQGAESGEFLRTMSQLLVDDAFWDHIFEEMNVPYTP 368
```

```
C. glutamicum  391   MRWAQDVPNTGVDKNTRVMQLIEAYRSRGHLIADTNPLSWVQPGM 435
                     ||||||:|||||||||||||||||||||||||||||||| ||||||
C. efficiens   369   MRWAQDLPNTGVDKNTRVMQLIEAYRSRGHLIADTNPLPWVQPGM 413

C. glutamicum  436   PVPDHRDLDIETHNLTIWDLDRTFNVGGFGGKETMTLREVLSRLR 480
                     ||||||||||||| ||:|||||||:||||||||||||||||||||
C. efficiens   414   PVPDHRDLDIETHGLTLWDLDRTFHVGGFGGKETMTLREVLSRLR 458

C. glutamicum  481   AAYTLKVGSEYTHILDRDERTWLQDRLEAGMPKPTQAEQKYILQK 525
                     ||||||||||||||||||||||||||||||||||| |||||||||
C. efficiens   459   AAYTLKVGSEYTHILDRDERTWLQDRLEAGMPKPTAAEQKYILQK 503

C. glutamicum  526   LNAAEAFENFLQTKYVGQKRFSLEGAEALIPLMDSAIDTAAGQGL 570
                     |||||||||||||||||||||||||||:|||||||||||||||||
C. efficiens   504   LNAAEAFENFLQTKYVGQKRFSLEGAESLIPLMDSAIDTAAGQGL 548

C. glutamicum  571   DEVVIGMPHRGRLNVLFNIVGKPLASIFNEFEGQMEQGQIGGSGD 615
                     |||||||||||||||||||||||||||||||||||||||||||||
C. efficiens   549   DEVVIGMPHRGRLNVLFNIVGKPLASIFNEFEGQMEQGQIGGSGD 593

C. glutamicum  616   VKYHLGSEGQHLQMFGDGEIKVSLTANPSHLEAVNPVMEGIVRAK 660
                     ||||||||| |||||||||||||||||||||||||||:|||||||
C. efficiens   594   VKYHLGSEGTHLQMFGDGEIKVSLTANPSHLEAVNPVVEGIVRAK 638

C. glutamicum  661   QDYLDKGVDGKTVVPLLLHGDAAFAGLGIVPETINLAKLRGYDVG 705
                     || |||| || |||||||||||||||||||||||||| |||||||
C. efficiens   639   QDILDKGPDGYTVVPLLLHGDAAFAGLGIVPETINLAALRGYDVG 683

C. glutamicum  706   GTIHIVVNNQIGFTTTPDSSRSMHYATDYAKAFGCPVFHVNGDDP 750
                     |||||||||||||||||||||||||||| ||||||||||||||||
C. efficiens   684   GTIHIVVNNQIGFTTTPDSSRSMHYATDCAKAFGCPVFHVNGDDP 728

C. glutamicum  751   EAVVWVGQLATEYRRRFGKDVFIDLVCYRLRGHNEADDPSMTQPK 795
                     |||||||||||||||||||||||||:|||||||||||||||||||
C. efficiens   729   EAVVWVGQLATEYRRRFGKDVFIDLICYRLRGHNEADDPSMTQPK 773
```

```
C. glutamicum   796    MYELITGRETVRAQYTEDLLGRGDLSNEDAEAVVRDFHDQMESVF 840
                       ||||||||::||| |||||||||||| ||||||||||||||||||
C. efficiens    774    MYELITGRDSVRATYTEDLLGRGDLSPEDAEAVVRDFHDQMESVF 818

C. glutamicum   841    NEVKEGGKKQAEAQTGITGSQKLPHGLETNISREELLELGQAFAN 885
                       ||||| |||| : ||||||||:|  ||:|||:||||:|||||| |
C. efficiens    819    NEVKEAGKKQPDEQTGITGSQELTRGLDTNITREELVELGQAFVN 863

C. glutamicum   886    TPEGFNYHPRVAPVAKKRVSSVTEGGIDWAWGELLAFGSLANSGR 930
                       ||||| ||||||||||||  ||||||||||||||:|||||| |||
C. efficiens    864    TPEGFTYHPRVAPVAKKRAESVTEGGIDWAWGELIAFGSLATSGR 908

C. glutamicum   931    LVRLAGEDSRRGTFTQRHAVAIDPATAEEFNPLHELAQSKGNNGK 975
                       |||||||||||||||||||||||| |||||||||||||:||  ||
C. efficiens    909    LVRLAGEDSRRGTFTQRHAVAIDPNTAEEFNPLHELAQAKG.GGK 952

C. glutamicum   976    FLVYNSALTEYAGMGFEYGYSVGNEDSIVAWEAQFGDFANGAQTI 1020
                       |||||||||||||||||||||||| |:::|:|||||||||||||||
C. efficiens    953    FLVYNSALTEYAGMGFEYGYSVGNPDAVVSWEAQFGDFANGAQTI 997

C. glutamicum   1021   IDEYVSSGEAKWGQTSKLILLLPHGYEGQGPDHSSARIERFLQLC 1065
                       ||||:||||||||||| :|||||||||||||||||||||||||||
C. efficiens    998    IDEYISSGEAKWGQTSSVILLLPHGYEGQGPDHSSARIERFLQLC 1042

C. glutamicum   1066   AEGSMTVAQPSTPANHFHLLRRHALSDLKRPLVIFTPKSMLRNKA 1110
                       ||||||:|||:||||:|||||||||  :|||||:|||||||||||
C. efficiens    1043   AEGSMTIAQPTTPANYFHLLRRHALGKMKRPLVVFTPKSMLRNKA 1087

C. glutamicum   1111   AASAPEDFTEVTKFQSVINDPNVADAAKVKKVMLVSGKLYYELAK 1155
                       | ||||:|||||:|:|||:|||||||:||||:|| |||:||||||
C. efficiens    1088   ATSAPEEFTEVTRFKSVIDDPNVADASKVKKIMLCSGKIYYELAK 1132

C. glutamicum   1156   RKEKDGRDDIAIVRIEMLHPIPFNRISEALAGYPNAEEVLFVQDE 1200
                       ||||| |||||||||||||||||||: :|  |||||||:||||||
C. efficiens    1133   RKEKDNRDDIAIVRIEMLHPIPFNRLRDAFDGYPNAEEILFVQDE 1177
```

```
C. glutamicum  1201   PANQGPWPFYQEHLPELIPNMPKMRRVSRRAQSSTATGVAKVHQL 1245
                      ||||| ||||||||| ||  |  |||:|||.|||||||:|||| :
C. efficiens   1178   PANQGAWPFYQEHLPNLIEGMLPMRRISRRSQSSTATGIAKVHTI 1222

C. glutamicum  1246   EEKQLIDEAFEA                                  1257
                      |:::|:|:|| |
C. efficiens   1223   EQQKLLDDAFNA                                  1234
```

Supplementary Figure 1.

(B)

```
C. glutamicum   1                      MTVDEQVSNYYDMLLKRNAGEPEFHQ 26
                                       ||||||||||||||||||||||||||
C. efficiens    1   MKFHCKFTCPRCRDGNVEFMTVDEQVSNYYDMLLKRNAGEPEFHQ 45

C. glutamicum   27  AVAEVLESLKIVLEKDPHYADYGLIQRLCEPERQLIFRVPWVDDQ 71
                    |||||||||||||||||||||||||||||||||||||||||||| 
C. efficiens    46  AVAEVLESLKIVLEKDPHYADYGLIQRLCEPERQLIFRVPWVDDN 90

C. glutamicum   72  GQVHVNRGFRVQFNSALGPYKGGLRFHPSVNLGIVKFLGFEQIFK 116
                    |||||||||||||||||||||||||||||||||||||||||||||
C. efficiens    91  GQVHVNRGFRVQFNSALGPYKGGLRFHPSVNLGIVKFLGFEQIFK 135

C. glutamicum   117 NSLTGLPIGGGKGGSDFDPKGKSDLEIMRFCQSFMTELHRHIGEY 161
                    |||||||||||||||||||||:|||||||||||||||||||||||
C. efficiens    136 NSLTGLPIGGGKGGSDFDPKGKSELEIMRFCQSFMTELHRHIGEY 180

C. glutamicum   162 RDVPAGDIGVGGREIGYLFGHYRRMANQHESGVLTGKGLTWGGSL 206
                    ||||||||||||||||||||||||||:||||||||||||||||||
C. efficiens    181 RDVPAGDIGVGGREIGYLFGHYRRLANQHESGVLTGKGLTWGGSL 225

C. glutamicum   207 VRTEATGYGCVYFVSEMIKAKGESISGQKIIVSGSGNVATYAIEK 251
                    |||||||:| |||| |||||:||:: |:|:|||||||||||||:|
C. efficiens    226 VRTEATGFGTVYFVQEMIKAEGETLEGKKVIVSGSGNVATYAIQK 270

C. glutamicum   252 AQELGATVIGFSDSSGWVHTPNGVDVAKLREIKEVRRARVSVYAD 296
                     ||||| |:||||||||| |||||||||||||||||||||| |||
C. efficiens    271 VQELGAVVVGFSDSSGWVSTPNGVDVAKLREIKEVRRARVSSYAD 315

C. glutamicum   297 EVEGATYHTDGSIWDLKCDIALPCATQNELNGENAKTLADNGCRF 341
                    ||||| ||||||||||  ||||||||||||:|:||:|||||||||
C. efficiens    316 EVEGAEYHTDGSIWDLTADIALPCATQNELDGDNARTLADNGCRF 360

C. glutamicum   342 VAEGANMPSTPEAVEVFRERDIRFGPGKAANAGGVATSALEMQQN 386
                    |||||||||||||::||||| : ||||||||||||||||||||||
C. efficiens    361 VAEGANMPSTPEAIDVFRERGVLFGPGKAANAGGVATSALEMQQN 405
```

```
C. glutamicum  387   ASRDSWSFEYTDERLQVIMKNIFKTCAETAAEYGHENDYVVGANI 431
                     |||||||||||||||  |||||||:||:|| ||||| :|||||||
C. efficiens   406   ASRDSWSFEYTDERLHRIMKNIFKSCADTAKEYGHEKNYVVGANI 450

C. glutamicum  432   AGFKKVADAMLAQGVI                              447
                     ||||||||||||||||
C. efficiens   451   AGFKKVADAMLAQGVI                              466
```

Supplementary Figure 1. (Continued)

(C)

```
C. glutamicum   1     MSNVGKPRTAQEIQQDWDTNPRWNGITRDYTADQVADLQGSVIEE 45
                      ||||| ||||||||||||||||||||||||||:|||:|||||:||
C. efficiens    1     MSNVGTPRTAQEIQQDWDTNPRWNGITRDYTAEQVAELQGSVVEE 45

C. glutamicum   46    HTLARRGSEILWDAVTQEGDGYINALGALTGNQAVQQVRAGLKAV 90
                      ||||:||:|||||||: ||| ||||||||||||||||||||||||
C. efficiens    46    HTLAKRGAEILWDAVSAEGDDYINALGALTGNQAVQQVRAGLKAV 90

C. glutamicum   91    YLSGWQVAGDANLSGHTYPDQSLYPANSVPSVVRRINNALLRSDE 135
                      |||||||||||||:||||||||||||||||:|||||||||||:||
C. efficiens    91    YLSGWQVAGDANLAGHTYPDQSLYPANSVPNVVRRINNALLRADE 135

C. glutamicum   136   IARTEGDTSVDNWVVPIVADGEAGFGGALNVYELQKAMIAAGAAG 180
                      ||| |||||||||:|||||||||||||||||||||| || |||||
C. efficiens    136   IARVEGDTSVDNWLVPIVADGEAGFGGALNVYELQKGMITAGAAG 180

C. glutamicum   181   THWEDQLASEKKCGHLGGKVLIPTQQHIRTLNSARLAADVANTPT 225
                      |||||||||||||||||||||||||||||||||||||||||||||
C. efficiens    181   THWEDQLASEKKCGHLGGKVLIPTQQHIRTLNSARLAADVANTPT 225

C. glutamicum   226   VVIARTDAEAATLITSDVDERDQPFITGERTAEGYYHVKNGLEPC 270
                      ||||||||||||||||||||||:|||||||||||||||| |||||
C. efficiens    226   VVIARTDAEAATLITSDVDERDRPFITGERTAEGYYHVKPGLEPC 270

C. glutamicum   271   IARAKSYAPYADMIWMETGTPDLELAKKFAEGVRSEFPDQLLSYN 315
                      |||||||||||||||||||||||||||||||||||||||||||||
C. efficiens    271   IARAKSYAPYADMIWMETGTPDLELAKKFAEGVRSEFPDQLLSYN 315

C. glutamicum   316   CSPSFNWSAHLEADEIAKFQKELGAMGFKFQFITLAGFHSLNYGM 360
                      |||||||||||||||||||||||||||||||||||||||||||||
C. efficiens    316   CSPSFNWSAHLEADEIAKFQKELGAMGFKFQFITLAGFHSLNYGM 360

C. glutamicum   361   FDLAYGYAREGMTSFVDLQNREFKAAEERGFTAVKHQREVGAGYF 405
                      |||||||||||| :|||||||||||||||||||||||||||||||
C. efficiens    361   FDLAYGYAREGMPAFVDLQNREFKAAEERGFTAVKHQREVGAGYF 405
```

```
C. glutamicum   406     DQIATTVDPNSSTTALKGSTEEGQFH          431
                        | ||||||||||||||||||||||||
C. efficiens    406     DTIATTVDPNSSTTALKGSTEEGQFH          431
```

Supplementary Figure 1. (Continued)

(D)

```
C. glutamicum   1     MEDMRIATLTSGGDCPGLNAVIRGIVRTASNEFGSTVVGYQDGWE 45
                      |  ||||||||||||||||||||||||||||||||||||||||||
C. efficiens    1     MGAMRIATLTSGGDCPGLNAVIRGIVRTASNEFGSTVVGYQDGWE 45

C. glutamicum   46    GLLGDRRVQLYDDEDIDRILLRGGTILGTGRLHPDKFKAGIDQIK 90
                      ||| |||||||||||||||||||||||||||||||||:|||||:|
C. efficiens    46    GLLADRRVQLYDDEDIDRILLRGGTILGTGRLHPDKFRAGIDQVK 90

C. glutamicum   91    ANLEDAGIDALIPIGGEGTLKGAKWLSDNGIPVVGVPKTIDNDVN 135
                      ||| ||||||||||||||||||||||:||||||||||||||||||
C. efficiens    91    ANLADAGIDALIPIGGEGTLKGAKWLADNGIPVVGVPKTIDNDVN 135

C. glutamicum   136   GTDFTFGFDTAVAVATDAVDRLHTTAESHNRVMIVEVMGRHVGWI 180
                      |||||||||:||:|||||:||||||||||||||||||||||||||
C. efficiens    136   GTDFTFGFDSAVSVATDAIDRLHTTAESHNRVMIVEVMGRHVGWI 180

C. glutamicum   181   ALHAGMAGGAHYTVIPEVPFDIAEICKAMERRFQMGEKYGIIVVA 225
                      ||||||||||||||||||||||:|||| |||||||||||||||||
C. efficiens    181   ALHAGMAGGAHYTVIPEVPFDISEICKRMERRFQMGEKYGIIVVA 225

C. glutamicum   226   EGALPREGTMELREGHIDQFGHKTFTGIGQQIADEIHVRLGHDVR 270
                      |||||:||||||||| :||||||||||||||||||:| |||||||
C. efficiens    226   EGALPKEGTMELREGEVDQFGHKTFTGIGQQIADEVHRRLGHDVR 270

C. glutamicum   271   TTVLGHIQRGGTPTAFDRVLATRYGVRAARACHEGSFDKVVALKG 315
                      ||||||||||||||||||||||||||||||||||| |: ||||||
C. efficiens    271   TTVLGHIQRGGTPTAFDRVLATRYGVRAARACHEGQFNTVVALKG 315

C. glutamicum   316   ESIEMITFEEAVGTLKEVPFERWVTAQAMFG               346
                      | | ||:|:|||||||:|| |||||||||||
C. efficiens    316   ERIRMISFDEAVGTLKKVPMERWVTAQAMFG               346
```

Supplementary Figure 1. (Continued)

(E)

```
C. glutamicum    1    MIITFTPNPSIDSTLSLGEELSRGSVQRLDSVTAVAGGKGINVAH  45
                      ||:| ||||||||||:|| ||:|| ||||:|||||||||||||||
C. efficiens     1    MIVTLTPNPSIDSTLALGGELTRGEVQRLESVTAVAGGKGINVAH  45

C. glutamicum   46    AVLLAGFETLAVFPAGKLDPFVPLVRDIGLPVETVVINKNVRTNT  90
                      || ||| :|:|:||||:|||||||||:|| |:|||:|  ||||||
C. efficiens    46    AVFLAGVDTVALFPAGRLDPFVPLVREIGFPIETVLIPTNVRTNT  90

C. glutamicum   91    TVTEPDGTTTKLNGPGAPLSEQKLRSLEKVLIDALRPEVTWVVLA  135
                      |:||||||||||||||||||:  :  ||: |:|||  :|||||||
C. efficiens    91    TITEPDGTTTKLNGPGAPLSQAHVTRLERTLVDALSADVTWVVLA  135

C. glutamicum  136    GSLPPGAPVDWYARLTALIHSARPDVRVAVDTSDKPLMALGESLD  180
                      ||||||||:|||:|||||:| | |  |||||||| ||  ||  ||
C. efficiens   136    GSLPPGAPLDWYSRLTALVHRACPGARVAVDTSDAPLQELGRHLD  180

C. glutamicum  181    TPGAAPNLIKPNGLELGQLANTDGEELEARAAQGDYDAIIAAADV  225
                       |||||||||||| |||||   ||:||| ||  |||  ||  | |
C. efficiens   181    EPGAAPNLIKPNGRELGQLVGVDGQELEDRARGGDYAPIIDCATV  225

C. glutamicum  226    LVNRGIEQVLVTLGAAGAVLVNAEGAWTATSPKIDVVSTVGAGDC  270
                      || ||||||||||| ||||||| |||| :::|:|  |||||||||
C. efficiens   226    LVERGIEQVLVTLGEAGAVLVNTEGAWVSSTPEITAVSTVGAGDC  270

C. glutamicum  271    ALAGFVMARSQKKTLEESLLNAVSYGSTAASLPGTTIPRPDQLAT  315
                      ||||||:||::  :: :|:|:||:||| | :|||||||||||: |
C. efficiens   271    ALAGFVLARTRGLSIPDSVLHAVAYGSAATALPGTTIPRPDQITT  315

C. glutamicum  316    AGATVTQVKGLKESA                                330
                       || | :
C. efficiens   316    KGAEVKKAV                                      324
```

Supplementary Figure 1. (Continued)

(F)

```
C. glutamicum    1     MAKIIWTRTDEAPLLATYSLKPVVEAFAATAGIEVETRDISLAGR 45
                       |||||||||||||||||||||||||||||||||||||||||||||
C. efficiens     12    MAKIIWTRTDEAPLLATYSLKPVVEAFAATAGIEVETRDISLAGR 56

C. glutamicum    46    ILAQFPERLTEDQKVGNALAELGELAKTPEANIIKLPNISASVPQ 90
                       ||||| ::| |:||| :||||||||||||||||||||||||||||
C. efficiens     57    ILAQFADQLPEEQKVSDALAELGELAKTPEANIIKLPNISASVPQ 101

C. glutamicum    91    LKAAIKELQDQGYDIPELPDNATTDEEKDILARYNAVKGSAVNPV 135
                       ||||:||||:||||:||  |       ||   || || || ||||
C. efficiens     102   LKAAVKELQEQGYDLPEYED......AKD...RYAAVIGSNVNPV 137

C. glutamicum    136   LREGNSDRRAPIAVKNFVKKFPHRMGEWSADSKTNVATMDANDFR 180
                       |||||||||||:||||||||||||||||||||||||||| |:|||
C. efficiens     138   LREGNSDRRAPVAVKNFVKKFPHRMGEWSADSKTNVATMGADDFR 182

C. glutamicum    181   HNEKSIILDAADEVQIKHIAADGTETILKDSLKLLEGEVLDGTVL 225
                        ||||:|:| || | |||:|||||||:||||| ||:|||:||| :
C. efficiens     183   SNEKSVIMDEADTVVIKHVAADGTETVLKDSLPLLKGEVIDGTFI 227

C. glutamicum    226   SAKALDAFLLEQVARAKAEGILFSAHLKATMMKVSDPIIFGHVVR 270
                       ||||||||||:|| ||| ||||||||:|||||||||||||||:||
C. efficiens     228   SAKALDAFLLDQVKRAKEEGILFSAHMKATMMKVSDPIIFGHIVR 272

C. glutamicum    271   AYFADVFAQYGEQLLAAGLNGENGLAAILSGLESLDNGEEIKAAF 315
                       ||||||:||||||||||||||||||||| :||: |||| ||||||
C. efficiens     273   AYFADVYAQYGEQLLAAGLNGENGLAAIYAGLDKLDNGAEIKAAF 317

C. glutamicum    316   EKGLEDGPDLAMVNSARGITNLHVPSDVIVDASMPAMIRTSGHMW 360
                       :||||:||||||||||:||||||||||||:|||||||||||| ||
C. efficiens     318   DKGLEEGPDLAMVNSAKGITNLHVPSDVIIDASMPAMIRTSGKMW 362

C. glutamicum    361   NKDDQEQDTLAIIPDSSYAGVYQTVIEDCRKNGAFDPTTMGTVPN 405
                       ||||| || ||:|||||||||||||||||||||||||||||||||
C. efficiens     363   NKDDQTQDALAVIPDSSYAGVYQTVIEDCRKNGAFDPTTMGTVPN 407
```

```
C. glutamicum   406   VGLMAQKAEEYGSHDKTFRIEADGVVQVVSSNGDVLIEHDVEAND 450
                      |||||||||||||||||||||||| ||||:||||||||||||  |
C. efficiens    408   VGLMAQKAEEYGSHDKTFRIEADGKVQVVASNGDVLIEHDVEKGD 452

C. glutamicum   451   IWRACQVKDAPIQDWVKLAVTRSRLSGMPAVFWLDPERAHDRNLA 495
                      |||||| ||||||||||||| |:||||||||||||| |||||||
C. efficiens    453   IWRACQTKDAPIQDWVKLAVNRARLSGMPAVFWLDPARAHDRNLT 497

C. glutamicum   496   SLVEKYLADHDTEGLDIQILSPVEATQLSIDRIRRGEDTISVTGN 540
                      :|||||||||||||||||||||||||| :||||||||||||||||
C. efficiens    498   TLVEKYLADHDTEGLDIQILSPVEATQHAIDRIRRGEDTISVTGN 542

C. glutamicum   541   VLRDYNTDLFPILELGTSAKMLSVVPLMAGGGLFETGAGGSAPKH 585
                      |||||||||||||||||||||||||||||||||||||||||||||
C. efficiens    543   VLRDYNTDLFPILELGTSAKMLSVVPLMAGGGLFETGAGGSAPKH 587

C. glutamicum   586   VQQVQEENHLRWDSLGEFLALAESFRHELNNNGNTKAGVLADALD 630
                      |||| |||||||||||||||||||||||||   ||||||||||||
C. efficiens    588   VQQVIEENHLRWDSLGEFLALAESFRHELNTRNNTKAGVLADALD 632

C. glutamicum   631   KATEKLLNEEKSPSRKVGEIDNRGSHFWLTKFWADELAAQTEDAD 675
                      :||||||||||||||||||||||||||||  :|||||| |||||:
C. efficiens    633   RATEKLLNEEKSPSRKVGEIDNRGSHFWLATYWADELANQTEDAE 677

C. glutamicum   676   LAATFAPVAEALNTGAADIDAALLAVQGGATDLGGYYSPNEEKLT 720
                      || ||||||||||  ||||||||:  ||   ||||||:|::|| :
C. efficiens    678   LAETFAPVAEALNNQAADIDAALIGEQGKPVDLGGYYAPSDEKTS 722

C. glutamicum   721   NIMRPVAQFNEIVDALKK                            738
                       |||||| ||||:|:|||
C. efficiens    723   AIMRPVAAFNEIIDSLKK                            740
```

Supplementary Figure 1. (Continued)

(G)

```
C. glutamicum   1    MELTVTESKNSFNAKSTLEVGDKSYDYFALSAVPGMEKLPYSLKV 45
                     |||||||||||||||||||:||:|||||:|||||||||||||||||
C. efficiens    1    MELTVTESKNSFNAKSTLQVGEKSYDYYALSAVPGMEKLPYSLKV 45

C. glutamicum   46   LGENLLRTEDGANITNEHIEAIANWDASSDPSIEIQFTPARVLMQ 90
                     ||||||||||||||| ||||||||||||:||||||||||||||||
C. efficiens    46   LGENLLRTEDGANITEEHIEAIANWDASADPSIEIQFTPARVLMQ 90

C. glutamicum   91   DFTGVPCVVDLATMREAVAALGGDPNDVNPLNPAEMVIDHSVIVE 135
                     ||||||||||||||||||  |||||: |||||||||||||||||
C. efficiens    91   DFTGVPCVVDLATMREAVKTLGGDPDKVNPLNPAEMVIDHSVIVE 135

C. glutamicum   136  AFGRPDALAKNVEIEYERNEERYQFLRWGSESFSNFRVVPPGTGI 180
                     |||||||| ||||||||||||||||||||||:|||||||||||||
C. efficiens    136  AFGRPDALEKNVEIEYERNEERYQFLRWGSEAFSNFRVVPPGTGI 180

C. glutamicum   181  VHQVNIEYLARVVFDNEGLAYPDTCIGTDSHTTMENGLGILGWGV 225
                     ||||||||||||||||:||||||||||||||||||||||||||||
C. efficiens    181  VHQVNIEYLARVVFDNDGLAYPDTCIGTDSHTTMENGLGILGWGV 225

C. glutamicum   226  GGIEAEAAMLGQPVSMLIPRVVGFKLTGEIPVGVTATDVVLTITE 270
                     |||||||||||||||||||||||||||||||||||||||||||||
C. efficiens    226  GGIEAEAAMLGQPVSMLIPRVVGFKLTGEIPVGVTATDVVLTITE 270

C. glutamicum   271  MLRDHGVVQKFVEFYGSGVKAVPLANRATIGNMSPEFGSTCAMFP 315
                     ||||||||||||||||:|||:||||||||||||||||||||||||
C. efficiens    271  MLRDHGVVQKFVEFYGNGVKSVPLANRATIGNMSPEFGSTCAMFP 315

C. glutamicum   316  IDEETTKYLRLTGRPEEQVALVEAYAKAQGMWLDEDTVEAEYSEY 360
                     ||||| ||||||||||||:||||||||||||||::|  |||||||
C. efficiens    316  IDEETIKYLRLTGRPEEQIALVEAYAKAQGMWLEQDAPEAEYSEY 360

C. glutamicum   361  LELDLSTVVPSIAGPKRPQDRILLSEAKEQFRKDLPTYTDDAVSV 405
                     ||||||||||||||||||||||||||||||||:||  ||:| | |
C. efficiens    361  LELDLSTVVPSIAGPKRPQDRILLSEAKEQFREDLKAYTNDPVQV 405
```

```
C. glutamicum   406   DTSIPATRMVNEGGGQPEGGVEADNYNASWAGSGESLATGAEGRP 450
                      | |||| || |||| ||    : ||||||| | ||| |  |||||
C. efficiens    406   DQSIPAKRMANEGGFQPGSTSDLDNYNASWPGEGESAAANAEGRP 450

C. glutamicum   451   SKPVTVASPQGGEYTIDHGMVAIASITSCTNTSNPSVMIGAGLIA 495
                      | |||| ||||||||||||||||||||||||||||||||||||||
C. efficiens    451   SNPVTVVSPQGGEYTIDHGMVAIASITSCTNTSNPSVMIGAGLIA 495

C. glutamicum   496   RKAAEKGLKSKPWVKTICAPGSQVVDGYYQRADLWKDLEAMGFYL 540
                      ||||||||||||||||||||||||||||||||||||||||:||||
C. efficiens    496   RKAAEKGLKSKPWVKTICAPGSQVVDGYYQRADLWKDLEALGFYL 540

C. glutamicum   541   SGFGCTTCIGNSGPLPEEISAAINEHDLTATAVLSGNRNFEGRIS 585
                      |||||||||||||||||||| ||||||| ||||||||||||||||
C. efficiens    541   SGFGCTTCIGNSGPLPEEISEAINEHDLAATAVLSGNRNFEGRIS 585

C. glutamicum   586   PDVKMNYLASPIMVIAYAIAGTMDFDFENEALGQDQDGNDVFLKD 630
                      |||||||||||||||||||||||||||||||||||||||||||||
C. efficiens    586   PDVKMNYLASPIMVIAYAIAGTMDFDFENEALGQDQDGNDVFLKD 630

C. glutamicum   631   IWPSTEEIEDTIQQAISRELYEADYADVFKGDKQWQELDVPTGDT 675
                      |||||||||:||| |||||||||||||||||||||||||:|:| |
C. efficiens    631   IWPSTEEIEETIQAAISRELYEADYADVFKGDKQWQELDIPSGKT 675

C. glutamicum   676   FEWDENSTYIRKAPYFDGMPVEPVAVTDIQGARVLAKLGDSVTTD 720
                      |||||||||||||||||||  ||  ||||: ||||||||||||||
C. efficiens    676   FEWDENSTYIRKAPYFDGMTAEPQPVTDIENARVLAKLGDSVTTD 720

C. glutamicum   721   HISPASSIKPGTPAAQYLDEHGVERHDYNSLGSRRGNHEVMMRGT 765
                      ||||||||||||||||||| ||||| |||||||||||||||||||
C. efficiens    721   HISPASSIKPGTPAAQYLDAHGVERQDYNSLGSRRGNHEVMMRGT 765

C. glutamicum   766   FANIRLQNQLVDIAGGYTRDFTQEGAPQAFIYDASVNYKAAGIPL 810
                      ||||||||||||||||||||||||| |||||||| |||| |||||
C. efficiens    766   FANIRLQNQLVDIAGGYTRDFTQEGGPQAFIYDACVNYKEAGIPL 810
```

```
C. glutamicum   811   VVLGGKEYGTGSSRDWAAKGTNLLGIRAVITESFERIHRSNLIGM 855
                      ||| |||||||||||||||||||||:|||||||||||||||||||
C. efficiens    811   VVLAGKEYGTGSSRDWAAKGTNLLGVRAVITESFERIHRSNLIGM 855

C. glutamicum   856   GVVPLQFPAGESHESLGLDGTETFDITGLTALNEGETPKTVKVTA 900
                      |||||||| |||||||||||||||||||||||||| |||||||||
C. efficiens    856   GVVPLQFPEGESHESLGLDGTETFDITGLTALNEGTTPKTVKVTA 900

C. glutamicum   901   TKENGDVVEFDAVVRIDTPGEADYYRHGGILQYVLRQMAAS     941
                      |||||: |||||||||||||||||:||||||||||||||||
C. efficiens    901   TKENGEKVEFDAVVRIDTPGEADYFRHGGILQYVLRQMAAS     941
```

Supplementary Figure 1. (Continued)

(H)

```
C. glutamicum  1      MTDFLRDDIRFLGQILGEVIAEQEGQEVYELVEQARLTSFDIAKG 45
                      | : ||||||:||:||||||:||||  |:||||:|| ||||||||
C. efficiens   1      MNELLRDDIRYLGRILGEVISEQEGHHVFELVERARRTSFDIAKG 45

C. glutamicum  46     NAEMDSLVQVFDGITPAKATPIARAFSHFALLANLAEDLYDEELR 90
                       |||||||:|| || |  |||:||||:||||||||||||:|   |
C. efficiens   46     RAEMDSLVEVFAGIDPEDATPVARAFTHFALLANLAEDLHDAAQR 90

C. glutamicum  91     EQALDAGDTPPDSTLDATWLKLNEGNVGAEAVADVLRNAEVAPVL 135
                      ||||:::|:  |||||:|||:||::  ||:  || |:||| |||||
C. efficiens   91     EQALNSGEPAPDSTLEATWVKLDDAGVGSGEVAAVIRNALVAPVL 135

C. glutamicum  136    TAHPTETRRRTVFDAQKWITTHMRERHALQSAEPTARTQSKLDEI 180
                      ||||||||||||||||| ||  | ||| | :    ||||||||:|
C. efficiens   136    TAHPTETRRRTVFDAQKHITALMEERHLLLALPTHARTQSKLDDI 180

C. glutamicum  181    EKNIRRRITILWQTALIRVARPRIEDEIEVGLRYYKLSLLEEIPR 225
                      |:|||||||||||||||||||||||||:|||||||||||| ||||
C. efficiens   181    ERNIRRRITILWQTALIRVARPRIEDEVEVGLRYYKLSLLAEIPR 225

C. glutamicum  226    INRDVAVELRERFGEGVPLKPVVKPGSWIGGDHDGNPYVTAETVE 270
                      || || |||  |||  :|   :|:|||||||||||||:||||||
C. efficiens   226    INHDVTVELARRFGGDIPTTAMVRPGSWIGGDHDGNPFVTAETVT 270

C. glutamicum  271    YSTHRAAETVLKYYARQLHSLEHELSLSDRMNKVTPQLLALADAG 315
                      |:|||||||||||| :|||:|||||||||||| :: :|  |||||
C. efficiens   271    YATHRAAETVLKYYVKQLHALEHELSLSDRMNVISDELRVLADAG 315

C. glutamicum  316    HNDVPSRVDEPYRRAVHGVRGRILATTAELIGEDAVEGVWFKVFT 360
                       ||:|||||||||||:||:|||:||||| ||||:|||| ||| ||
C. efficiens   316    QNDMPSRVDEPYRRAIHGMRGRMLATTAALIGEEAVEGTWFKTFT 360

C. glutamicum  361    PYASPEEFLNDALTIDHSLRESKDVLIADDRLSVLISAIESFGFN 405
                      ||    ||  |   :| ||| |:| :||||||::| ||::|||||
C. efficiens   361    PYTDTHEFKRDLDIVDGSLRMSRDDIIADDRLAMLRSALDSFGFN 405
```

```
C. glutamicum  406   LYALDLRQNSESYEDVLTELFERAQVTANYRELSEAEKLEVLLKE 450
                     ||:|||||||: :||||||||  ||   ||| |:|||||::|::|
C. efficiens   406   LYSLDLRQNSDGFEDVLTELFATAQTEKNYRGLTEAEKLDLLIRE 450

C. glutamicum  451   LRSPRPLIPHGSDEYSEVTDRELGIFRTASEAVKKFGPRMVPHCI 495
                     | :||||||||  :||| |:||||||  |:|||:|||| ||||||
C. efficiens   451   LSTPRPLIPHGDPDYSEATNRELGIFSKAAEAVRKFGPLMVPHCI 495

C. glutamicum  496   ISMASSVTDVLEPMVLLKEFGLIAANGDNPRGTVDVIPLFETIED 540
                     |||||||||:||||||||||||| ||| || |:||||||||||:|
C. efficiens   496   ISMASSVTDILEPMVLLKEFGLIRANGKNPTGSVDVIPLFETIDD 540

C. glutamicum  541   LQAGAGILDELWKIDLYRNYLLQRDNVQEVMLGYSDSNKDGGYFS 585
                     || |||||:||| |||||||| ||||||||||||||||||||||:
C. efficiens   541   LQRGAGILEELWDIDLYRNYLEQRDNVQEVMLGYSDSNKDGGYFA 585

C. glutamicum  586   ANWALYDAELQLVELCRSAGVKLRLFHGRGGTVGRGGGPSYDAIL 630
                     ||||||||||:||||||   |||||||||||||||||||||||||
C. efficiens   586   ANWALYDAELRLVELCRGRNVKLRLFHGRGGTVGRGGGPSYDAIL 630

C. glutamicum  631   AQPRGAVQGSVRITEQGEIISAKYGNPETARRNLEALVSATLEAS 675
                     |||:|||:|:||:||||||||||||||:|||||||||||||||||
C. efficiens   631   AQPKGAVRGAVRVTEQGEIISAKYGNPDTARRNLEALVSATLEAS 675

C. glutamicum  676   LLDVSELTDHQRAYDIMSEISELSLKKYASLVHEDQGFIDYFTQS 720
                     |||  || : :||: || |||||| ::|:|||||| ||| |||||
C. efficiens   676   LLDDVELPNRERAHQIMGEISELSFRRYSSLVHEDPGFIQYFTQS 720

C. glutamicum  721   TPLQEIGSLNIGSRPSSRKQTSSVEDLRAIPWVLSWSQSRVMLPG 765
                     |||||||||||||||||||||::||||||||||||||||||||||
C. efficiens   721   TPLQEIGSLNIGSRPSSRKQTNTVEDLRAIPWVLSWSQSRVMLPG 765

C. glutamicum  766   WFGVGTALEQWIGEGEQATQRIAELQTLNESWPFFTSVLDNMAQV 810
                     |||||||| :|||||| | :|||||| ||  ||||||||||||||
C. efficiens   766   WFGVGTALREWIGEGEGAAERIAELQELNRCWPFFTSVLDNMAQV 810
```

```
C. glutamicum   811   MSKAELRLAKLYADLIPDTEVAERVYSVIREEYFLTKKMFCVITG 855
                      |||||||||:|||||||| |||:|:|  |  ||||||:||| |||
C. efficiens    811   MSKAELRLARLYADLIPDREVADRIYETIFGEYFLTKEMFCTITG 855

C. glutamicum   856   SDDLLDDNPLLARSVQRRYPYLLPLNVIQVEMMRRYRKGDQSEQV 900
                      | ||||||| |||||: |:|||||||||||||||||| ||:   |
C. efficiens    856   SQDLLDDNPALARSVRSRFPYLLPLNVIQVEMMRRYRSGDEGTAV 900

C. glutamicum   901   SRNIQLTMNGLSTALRNSG                           919
                       |||:||||||||||||||
C. efficiens    901   PRNIRLTMNGLSTALRNSG                           919
```

Supplementary Figure 1. (Continued)

(I)

```
C. glutamicum   1    MFERDIVATDN 11
                                                        |||:|||:||
C. efficiens    1    METFVSRNNILAARDASDLVIESGDLPQPGGTDKKFEREIVASDN 45

C. glutamicum   12   NKAVLHYPGGEFEMDIIEASEGNNGVVLGKMLSETGLITFDPGYV 56
                     |||||||||||||| | :|:|||:||:||||||||||:|||||||
C. efficiens    46   NKAVLHYPGGEFEMGIKQATEGNSGVILGKMLSETGLVTFDPGYV 90

C. glutamicum   57   STGSTESKITYIDGDAGILRYRGYDIADLAENATFNEVSYLLING 101
                     ||||||||||||||||||||||||||||||||||||||||||| |
C. efficiens    91   STGSTESKITYIDGDAGILRYRGYDIADLAENATFNEVSYLLIKG 135

C. glutamicum   102  ELPTPDELHKFNDEIRHHTLLDEDFKSQFNVFPRDAHPMATLASS 146
                     |||||:|||||||||||||||||||||||||||||||||||||||
C. efficiens    136  ELPTPEELHKFNDEIRHHTLLDEDFKSQFNVFPRDAHPMATLASS 180

C. glutamicum   147  VNILSTYYQDQLNPLDEAQLDKATVRLMAKVPMLAAYAHRARKGA 191
                     ||||||||||||:||||||||||||||||||||||||||||||||
C. efficiens    181  VNILSTYYQDQLDPLDEAQLDKATVRLMAKVPMLAAYAHRARKGA 225

C. glutamicum   192  PYMYPDNSLNARENFLRMMFGYPTEPYEIDPIMVKALDKLLILHA 236
                     ||||||||||||||||||||||||||||:||||||||||||||||
C. efficiens    226  PYMYPDNSLNARENFLRMMFGYPTEPYEVDPIMVKALDKLLILHA 270

C. glutamicum   237  DHEQNCSTSTVRMIGSAQANMFVSIAGGINALSGPLHGGANQAVL 281
                     |||||||||||||||||||||||||||||||||||||||||||||
C. efficiens    271  DHEQNCSTSTVRMIGSAQANMFVSIAGGINALSGPLHGGANQAVL 315

C. glutamicum   282  EMLEDIKSNHGGDATEFMNKVKNKEDGVRLMGFGHRVYKNYDPRA 326
                     ||||:| :| |||||:|||:||||| |||||||||||||||||||
C. efficiens    316  EMLEEIAAN.GGDATDFMNRVKNKEKGVRLMGFGHRVYKNYDPRA 359

C. glutamicum   327  AIVKETAHEILEHLGGDDLLDLAIKLEEIALADDYFISRKLYPNV 371
                     ||||:|||||||||||| |||||:||||||| |||||||||||||
C. efficiens    360  AIVKDTAHEILEHLGGDPLLDLALKLEEIALNDDYFISRKLYPNV 404
```

```
C. glutamicum   372   DFYTGLIYRAMGFPTDFFTVLFAIGRLPGWIAHYREQLGAAGNKI 416
                      ||||||||||||||||||||||||||||||||||||||   | ||
C. efficiens    405   DFYTGLIYRAMGFPTDFFTVLFAIGRLPGWIAHYREQLADPGAKI 449

C. glutamicum   417   NRPRQVYTGNESRKLVPREER                         437
                      |||||:|||  :||::|||||
C. efficiens    450   NRPRQIYTGETARKIIPREER                         470
```

Supplementary Figure 1. (Continued)

(J)

```
C. glutamicum   1     MALVVQKYGGSSLESAERIRNVAERIVATKKAGNDVVVVCSAMGD 45
                      :||||||||||||||||||||||||||||||||||||||||||||
C. efficiens    8     VALVVQKYGGSSLESAERIRNVAERIVATKKAGNDVVVVCSAMGD 52

C. glutamicum   46    TTDELLELAAAVNPVPPAREMDMLLTAGERISNALVAMAIESLGA 90
                      ||||||:||||||||||||||||||||||||||||||||||||||
C. efficiens    53    TTDELLDLAAAVNPVPPAREMDMLLTAGERISNALVAMAIESLGA 97

C. glutamicum   91    EAQSFTGSQAGVLTTERHGNARIVDVTPGRVREALDEGKICIVAG 135
                      |||||||||||||||||||||||||||||||||||||||||||||
C. efficiens    98    EAQSFTGSQAGVLTTERHGNARIVDVTPGRVREALDEGKICIVAG 142

C. glutamicum   136   FQGVNKETRDVTTLGRGGSDTTAVALAAALNADVCEIYSDVDGVY 180
                      |||||||||||||||||||||||||||||| ||||||||||||||
C. efficiens    143   FQGVNKETRDVTTLGRGGSDTTAVALAAALGADVCEIYSDVDGVY 187

C. glutamicum   181   TADPRIVPNAQKLEKLSFEEMLELAAVGSKILVLRSVEYARAFNV 225
                      ||||||||||||||:||||||||||||||||||||||||||||||
C. efficiens    188   TADPRIVPNAQKLERLSFEEMLELAAVGSKILVLRSVEYARAFNV 232

C. glutamicum   226   PLRVRSSYSNDPGTLIAGSMEDIPVEEAVLTGVATDKSEAKVTVL 270
                      |:||||||||||||||||||||||:||||||||||||||||||||
C. efficiens    233   PMRVRSSYSNDPGTLIAGSMEDIPMEEAVLTGVATDKSEAKVTVL 277

C. glutamicum   271   GISDKPGEAAKVFRALADAEINIDMVLQNVSSVEDGTTDITFTCP 315
                      || ||||||||||||||||||||||||||||||||||||||||||
C. efficiens    278   GIPDKPGEAAKVFRALADAEINIDMVLQNVSSVEDGTTDITFTCP 322

C. glutamicum   316   RSDGRRAMEILKKLQVQGNWTNVLYDDQVGKVSLVGAGMKSHPGV 360
                      |||| ||||:|||:| ||:||||||||||||||||||||||||||
C. efficiens    323   RSDGPRAMELLKKMQQQGDWTNVLYDDQVGKVSLVGAGMKSHPGV 367

C. glutamicum   361   TAEFMEALRDVNVNIELISTSEIRISVLIREDDLDAAARALHEQF 405
                      ||||||||||||||:|||||||||||||||||||| :|:||||:|
C. efficiens    368   TAEFMEALRDVNVNVELISTSEIRISVLIREDDLDKSAKALHEKF 412
```

```
C. glutamicum   406     QLGGEDEAVVYAGTGR          421
                        ||||::|| |||||||
C. efficiens    413     QLGGDEEATVYAGTGR          428
```

Supplementary Figure 1. (Continued)

(K)

```
C. glutamicum   1    MSTGLTAKTGVEHFGTVGVAMVTPFTESGDIDIAAGREVAAYLVD 45
                     ||||||||||||||||||||||||||||||:|:|||||:||:|||
C. efficiens    4    MSTGLTAKTGVEHFGTVGVAMVTPFTESGDLDVAAGREIAAHLVD 48

C. glutamicum   46   KGLDSLVLAGTTGESPTTTAAEKLELLKAVREEVGDRAKLIAGVG 90
                      |:|:|:|||||||||| | |||| |||||||||||||||||| |
C. efficiens    49   NGVDALILAGTTGESPTVTTAEKLTLLKAVREEVGDRAKLIAGAG 93

C. glutamicum   91   TNNTRTSVELAEAAASAGADGLLVVTPYYSKPSQEGLLAHFGAIA 135
                     |||||:||||||| |  ||||||||||||||||||||: ||  ||
C. efficiens    94   TNNTRSSVELAEAFAEVGADGLLVVTPYYSKPSQEGLVRHFTEIA 138

C. glutamicum   136  AATEVPICLYDIPGRSGIPIESDTMRRLSELPTILAVKDAKGDLV 180
                      ||::|||||||||||||||||||:|||||||||||:||||||:|
C. efficiens    139  QATDLPICLYDIPGRSGIPIESDTIRRLSELPTILAMKDAKGDVV 183

C. glutamicum   181  AATSLIKETGLAWYSGDDPLNLVWLALGGSGFISVIGHAAPTALR 225
                     ||  ||:|||||||||||||||||||||||||||||||||| |||
C. efficiens    184  AAAPLIEETGLAWYSGDDPLNLVWLALGGSGFISVIGHAAPNALR 228

C. glutamicum   226  ELYTSFEEGDLVRAREINAKLSPLVAAQGRLGGVSLAKAALRLQG 270
                     ||||||||||| ||||||| |||||||||||||||:|||||||||
C. efficiens    229  ELYTSFEEGDLARAREINATLSPLVAAQGRLGGVSMAKAALRLQG 273

C. glutamicum   271  INVGDPRLPIMAPNEQELEALREDMKKAGVL               301
                     ||||||||||:|||||||| || ||||||||
C. efficiens    274  INVGDPRLPIVAPNEQELEDLRADMKKAGVL               304
```

Supplementary Figure 1. (Continued)

(L)

```
C. glutamicum  1    MTNIRVAIVGYGNLGRSVEKLIAKQPDMDLVGIFSRRATLDTKTP 45
                    |: || |||||||||:|||||| :||||:|||||||| |||| ||
C. efficiens   1    MSKIRAAIVGYGNLGKSVEKLIVQQPDMELVGIFSRRDTLDTDTP 45

C. glutamicum  46   VFDVADVDKHADDVDVLFLCMGSATDIPEQAPKFAQFACTVDTYD 90
                    ||:||: :||  |||:|||||||||||||||| || |||||||||
C. efficiens   46   VFNVAETEKHTGDVDLLFLCMGSATDIPEQAPGFAAFACTVDTYD 90

C. glutamicum  91   NHRDIPRHRQVMNEAATAAGNVALVSTGWDPGMFSINRVYAAAVL 135
                    ||||||||||||:||| |||||::|:|||||||||||||| ||:|
C. efficiens   91   NHRDIPRHRQVMDEAARAAGNVSVVATGWDPGMFSINRVYGAALL 135

C. glutamicum  136  AEHQQHTFWGPGLSQGHSDALRRIPGVQKAVQYTLPSEDALEKAR 180
                    |:|||||||||||||||||||||| ||:||||||||||||||||||
C. efficiens   136  ADHQQHTFWGPGLSQGHSDALRRIDGVEKAVQYTLPSEDALEKAR 180

C. glutamicum  181  RGEAGDLTGKQTHKRQCFVVADAADHERIENDIRTMPDYFVGYEV 225
                    ||||  |||||||||||||||  ::||||||:|||| ||||||||
C. efficiens   181  RGEAEGLTGKQTHKRQCFVVAPESEHERIENEIRTMADYFVGYEV 225

C. glutamicum  226  EVNFIDEATFDSEHTGMPHGGHVITTGDTGGFNHTVEYILKLDRN 270
                    ||||||||||||||||||||||||||||||||:||||| ||||||
C. efficiens   226  EVNFIDEATFDSEHTGMPHGGHVITTGDTGGFHHTVEYTLKLDRN 270

C. glutamicum  271  PDFTASSQIAFGRAAHRMKQQGQSGAFTVLEVAPYLLSPENLDDL 315
                    |||||||||||||||:|:|: ||:|||||||||||||||  ||||
C. efficiens   271  PDFTASSQIAFGRAAYRLKEAGQAGAFTVLEVAPYLLSPTPLDDL 315

C. glutamicum  316  IARDV                                         320
                    |||||
C. efficiens   316  IARDV                                         320
```

Supplementary Figure 1. (Continued)

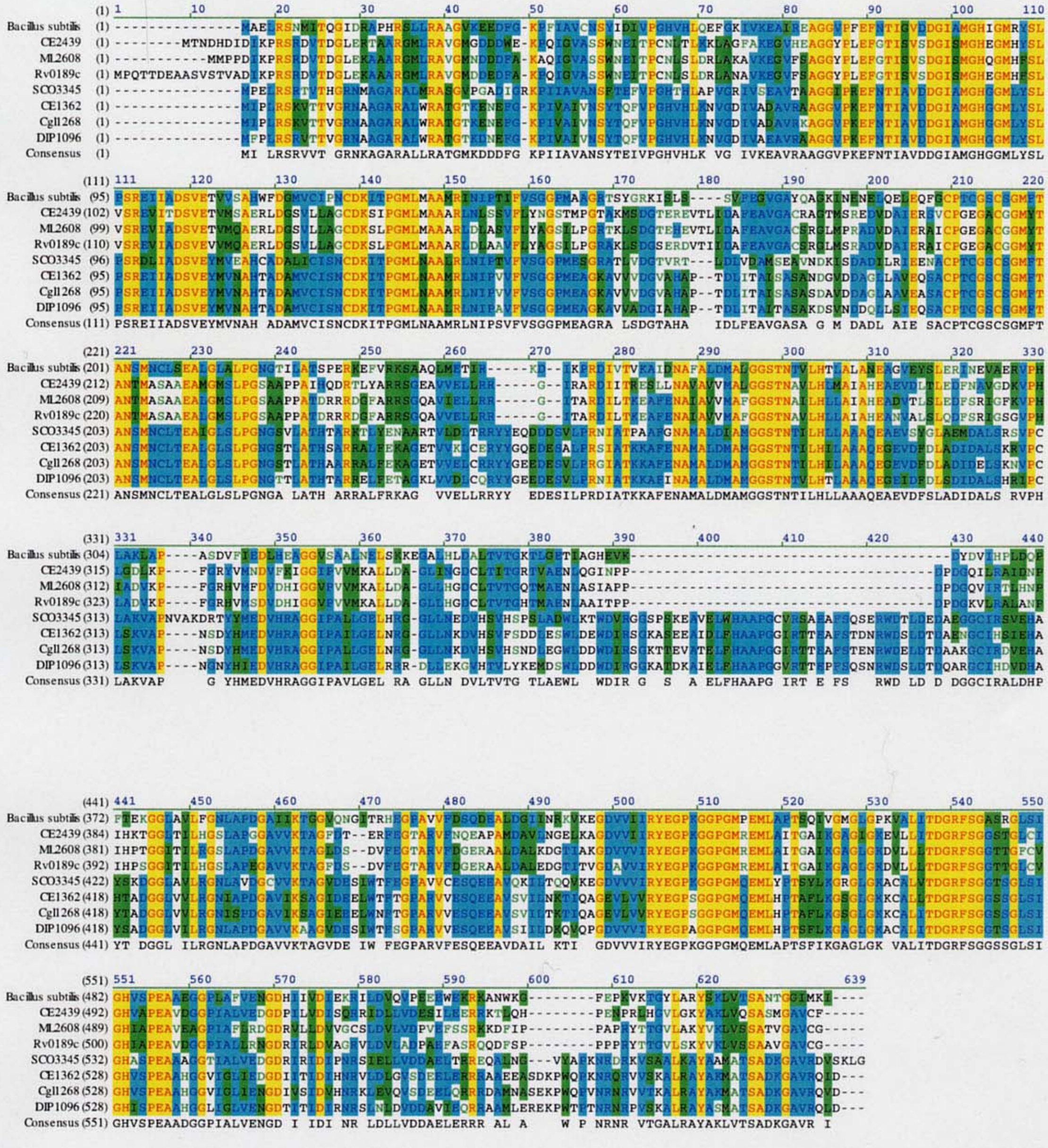
(M)

```
C. glutamicum  1                  MATVENFNELPAHVWPRNAVRQEDGVVTVAG 31
                                      : |||||||||||||||||||||||||
C. efficiens   1    MTAETETGIPGVPGTQAADQFNELPAHVWPRNAVRQEDGVVTVAG 45

C. glutamicum  32   VPLPDLAEEYGTPLFVVDEDDFRSRCRDMATAFGGPGNVHYASKA 76
                    |||||||||||||||||||||||:||||||:|||||  |||||||
C. efficiens   46   VPLPDLAEEYGTPLFVVDEDDFRARCRDMASAFGGPDRVHYASKA 90

C. glutamicum  77   FLTKTIARWVDEEGLALDIASINELGIALAAGFPASRITAHGNNK 121
                    ||:||:|||||||||:||||| ||||||||| ||  |||||||||
C. efficiens   91   FLSKTVARWVDEEGLSLDIASENELGIALAADFPGERITAHGNNK 135

C. glutamicum  122  GVEFLRALVQNGVGHVVLDSAQELELLDYVAAGEGKIQDVLIRVK 166
                       |||| |:| :||||||||||||||||:||||||:| ||||||
C. efficiens   136  DASFLRACVRNNLGHVVLDSAQELELLDYIAAGEGKVQPVLIRVK 180

C. glutamicum  167  PGIEAHTHEFIATSHEDQKFGFSLASGSAFEAAKAANNAENLNLV 211
                    |||||||||||||||||||||||||||:||:||:|| ||||| ||
C. efficiens   181  PGIEAHTHEFIATSHEDQKFGFSLASGAAFDAARAAVNAENLELV 225

C. glutamicum  212  GLHCHVGSQVFDAEGFKLAAERVLGLYSQIHSELGVALPELDLGG 256
                    |||||||||||||:|| ||||||| |||:|| |||| | ||||||
C. efficiens   226  GLHCHVGSQVFDAQGFSLAAERVLELYSRIHDELGVTLAELDLGG 270

C. glutamicum  257  GYGIAYTAAEEPLNVAEVASDLLTAVGKMAAELGIDAPTVLVEPG 301
                    ||||||||||||||| ||| |||||||| ||||||:|||||||||
C. efficiens   271  GYGIAYTAAEEPLNVVEVAHDLLTAVGKTAAELGIEAPTVLVEPG 315

C. glutamicum  302  RAIAGPSTVTIYEVGTTKDVHVDDDKTRRYIAVDGGMSDNIRPAL 346
                    ||||||||||:||||| ||| |||: |||||:|||||||||||||
C. efficiens   316  RAIAGPSTVTVYEVGTIKDVDVDDETTRRYISVDGGMSDNIRPAL 360

C. glutamicum  347  YGSEYDARVVSRFAEGDPVSTRIVGSHCESGDILINDEIYPSDIT 391
                    ||:|||||||||| ||:  :||:|||||||||||||:  |||||
C. efficiens   361  YGAEYDARVVSRFTEGETTNTRVVGSHCESGDILINEATYPSDIH 405
```

```
C. glutamicum  392   SGDFLALAATGAYCYAMSSRYNAFTRPAVVSVRAGSSRLMLRRET 436
                     :|| |||||||||||||||||||| ||||||||||:::|||||||
C. efficiens   406   TGDLLALAATGAYCYAMSSRYNAFARPAVVSVRAGAAKLMLRRET 450

C. glutamicum  437   LDDILSLEA                                     445
                     ||||||||
C. efficiens   451   LDDILSLEV                                     459
```

Supplementary Figure 1. (Continued) Piarwise alignment of thirteen enzyme sequences between *C. glutamicum* and *C. efficiens*.

(A) 2-Oxoglutarate dehydrogenase, (B) Glutamate dehydrogenase, (C) Isocitrate lyase, (D) Phosphofructokinase, (E) Fructose-1-phosphate kinase, (F) Isocitrate dehydrogenase, (G) Aconitase, (H) Phosphoenolpyruvate carboxylase, (I) Citrate synthase, (J) Aspartate kinase, (K) Dihydropicolinate synthase, (L) Diaminopimelate dehydrogenase, (M) Diaminopimelate decarboxylase. Amino acid substitutions from K in *C. glutamicum* to R in *C. efficiens*, S in *C. glutamicum* to A in *C. efficiens*, and S in *C. glutamicum* to T in *C. efficiens* showed with boxes, and the opposite amino acid substitutions in alignments showed with half-tone dot meshings.



Supplementary Figure 2 Multiple alignment of IlvD

This alignment was used for the phylogenetic tree of IlvD shown in Fig. 3.2b.

S-44

```
                  1                                                                                                  100
CE1190      (1)   ----------------------------------------------------------------------------------------------------
Cgl1129     (1)   AAGCACACTTGTTTAGTGGAAGCATCGCCGACAACATTGGCTACGGATGCAGGGAGGCGTCGACAAGCAAAATCGAAGCGGCAGCACGCCGCGTCGGAGC
CDIP1002    (1)   ----------------------------------------------------------------------------------------------------
                  101                                                                                                200
CE1190      (1)   ---------CATCGCCGCGATCCCGGGGGGATTCAACCACCCCGTCGGTGAACGCGGACGCGGCCTGTCCTCCGGGCAGCGGCAGCTGATCGCTCTGGCG
Cgl1129   (101)   CTTAAACGCCATCGCCGCCATCCCTGATGGTTTCAACCATCAAGTCGGTGAACGCGGGCGCAACCTGTCATCCGGACAGCGCCAACTGATCGCGCTGGCG
DIP1002     (1)   --------------------------GCGGCTTCCGCGCCACTGTTGGCGAACGCGGCCAAGGGTTATCTTCAGGACAACGTCAGCTCATTGCCTTGGCA
                  201                                                                                                300
 CE1190    (92)   CGCGCCGAGCTCATCGAACCGGTGATCATGCTTCTCGACGAGGCCACCTCCACCCTCGACCCCGCCACCGAGACGGTCATCCTCAACGCCTCCGACCGGG
Cgl1129   (201)   CGCGCCGAACTTATCGAGCCTTCCATCATGCTTCTCGACGAAGCCACCTCCACCCTCGACCCCGCCACCGAAGCCGTTATCCTCAACGCCTCCGATCGAG
DIP1002    (75)   CGAGCAGAGATGATGAAGCCAGAAATCTTGCTTCTCGACGAAGCCACCGCAACGCTTGATCCTGCAACCGAAAAAACGATCTTGTCTGCCGCCGAACGGC
                  301                                                                                                400
 CE1190   (192)   TCACCCGGAACCGCACGAGCGTGATCGTCGCGCACCGGCTCGCCACCGCTAGCCGGGCCGACCGGATCATCGTGGTTGACGGGGGACGTATCATCGAGGA
Cgl1129   (301)   TCACTAAGGGACGCACCAGCATCATCGTCGCGCACCGCTTGGCAACCGCTAAAAGGGCCGACCGTATTCTTGTTGTTGAACAAGGACGTATCATTGAGGA
DIP1002   (175)   TCACGCAAACACGCACCTCGGTCATTGTTGCCCACCGATTAGCCACCGCCGCGAAAGCAGATCGGATACTTGTGATTGCGAACGGGGCCGTCGTTGAAGA
                  401                                                                                                500
 CE1190   (292)   TGGTTCCCACGATGAACTTCTGGGAGCGAATGGAACCTACGCAACAATGTGG------------CATTTAGTAGGGTGACA----------GGATATTTT
Cgl1129   (401)   CGGATCTCACGACGCGTTGTTGTCTGCTAACGGCACCTACGCCCGCATGTGG------------CATTTAATGGCCTGACA----------CGTTATTTT
DIP1002   (275)   TGGCGACCATGCAAGCCTACGCACTTATGGGGGTATTTACGCCACAATGTGGGCACACGGCGAACAAGAAATCCCGCGATAAAGGCGCGTACAATAGGTG
                  501                                                                                                600
 CE1190   (370)   AGGAAAGACTGTTACCAAAAGG-TGCTAATACTGGGGTGCTAGGTCC----CCGCGACCGGA--ACCAGCGTTA—CAGTGGATAAAATAAAGCCCATTT
Cgl1129   (479)   TAGGAGAACTGTCAACAAATTAATGCTACAACTGGGGCT-TAGGCATAAT-CAGCCAACG----ACCAACGTTA—CAGTGGATAAAACAAAGCTCAATA
DIP1002   (375)   AAATCCCACTGTGCACGTATGG-GGAGGCGATTTTCTTTCTCGGCAGTTCACAGTTGGAGGAAGAAAACCGATAGCCTGTAGTGAAGCTATTACAGTGTG
                  601                                                                                                700
 CE1190   (461)   AGAACCCTCAACAAG-----CAAGGAAAAGAGGCGAGTACCTGCCGTGAGCAGCGCTAGTACTTTCGGCCAGAACGCGTGGCTGGTGGATGAGATGTTCC
Cgl1129   (571)   A--ACCCTCAAGAAG-----CAAGGAAAAGAGGCGAGTACCTGCCGTGAGCAGCGCTAGTACTTTCGGCCAGAATGCGTGGCTGGTAGACGAGATGTTCC
DIP1002   (474)   AATAGACGTTAGAAATCTCACAAAGAAATGAGGCGAGCACCTGCAGTGAGCAGCGCTAGTACTTTCGGCCAGAACGATTGGCTGGTAGACGAGATGTTCC
```

```
                  701                                                                                                800
 CE1190    (556)  AGCAGTTCAAGAAGGACCCCCAGTCCGTGGACAAGGAATGGAGAGAGCTCTTCGAGTCTCAGGGGGGTCCCCAGGCTGAAAAGGCTACCCCCGCCACCCC
Cgl1129    (664)  AGCAGTTCCAGAAGGACCCCAAGTCCGTGGACAAGGAATGGAGAGAACTCTTTGAGGCGCAGGGGGGAC--C-------AAATACTACCCCCGCTACAAC
DIP1002    (574)  AGCAGTTCCAAAAGGATCCGCAGTCCGTAGATAAGGAATGGCGCGACCTTTTCGAGAAGCAGGGTGCCCCGAGCACACCGGGAACTGAGGCTAAGAACAC
```

Supplementary Figure 3 Multiple alignment of regulatory region for *odhA* gene in *Corynebacteria*

Boxes showed the proposed start codon for *odhA* gene in each genome sequence.

```
                       1                                                                                                  100
 CE1982          (1)   CCGGAGGACACCCCGGGGGAGAAGGGTGGCACCGGGGTGCTCGTGG-CGCTGGGCGCGCTGATGGCTCAGCGCAGAGGGGCGTTGTCCTAGAAACTCTAT
CGL2079          (1)   CTCAATTGTGGCCAGGTTATATAACCAGTCAGTCAACTGGTCTCATTCGCTGGTCGGATGAAT—TTAATTAAAGAAGAGACTTCATGCAGTTACCGCGC
DIP1547          (1)   TGAGGTGGTAGC--GGTGGATTTAAAAGAATCTGGAGCGATCACAATC-TTGAACGTGCGATTG---GGGTGGGGCATACCTATCCTTGTGCATTTTAGG
                       101                                                                                                200
 CE1982        (100)   CGTCCGAGGGTGTGCGTTCGGCAA—CCGGCGGGCCCAGCGACGTTCGCGGGACAGTGGTATTAATACCAGTGGGGGCACCGGTTTTATCTCGATGAGCG
CGL2079         (99)   GTTTTGGCGATACAAAATTGATAAACCTAAAGAAATTTTCAAACAATTTTAATTCTTTGTGGTCATATCTGTGCGA-CACTGCCATAAT-TGAACGTG-A
DIP1547         (95)   GTGATAGTAGTAGTCGAATTGTGGA-GTATTGATCCTAAACATTGTACGGGGACGCTTGTAG--ATACCAAGCCGA-CAAGATTTCTAC-TATTTGATTG
                       201                                                                                                300
 CE1982        (198)   GGATCTGTCAGCTCGGGAGTCGTTTCACAAGGAGGAGGGT-TCGGGGTGTGAACCCGCTGGCTGGAAGTGTGAAATTTTCCACATTGTGGT-CATATCG
CGL2079        (196)   GCATTTACCAGCCTAAATGCC----CGCAGTGAGTTAAGTCTCAAAGCAAGAAGTTGCTCTTTAGG-GCATCCGTAGTTTAAAACTATTA-----ACCG
DIP1547        (190)   AAACCTCCTGATTTTTGGAAGACTTTTGGATTTTTGAGGCTATCGGTCAAAAATTT--TTTTTAGC-CTCAAGGGTTTCACTATTATCGCAGCTAAATG
                       301                                                                                                400
 CE1982        (295)   TCATGGGACTGACATAATCGGACGTGAGCATTGGCCTGCCGCT--CTGGCCCTTGTGAGTCAACTCTCATGGTCGAGAGTTGCTCTTTAGGGCC-CGCGT
CGL2079        (285)   T--TAGGTATGACA-AGCCGG---TTGATGTGAACGCAGTTTT--TAAAAGTTTCAGGATCAGATTTTTCA--CAGGCATTTTGCTCCAGCAAA-CGCCT
DIP1547        (286)   CATTTATATCTACGCGTAGGGGAGTGGGTTGGGGAAAGATTTTGGTTAACTTTCGATAGTTCAATAGGAATTTTGTTTTTGCTTGTGTAGAGTCTCACTT
                       401                                                                                                500
 CE1982        (392)   GGTTTAAAAC-TATTAACCGTTAGGTATAACAAGCCGCGCCCCTCGGTGTAGTTGAAATTTCA-TTGCAAATTCACCTGCCCGCGGTGCCGAGATGGGAA
CGL2079        (374)   AGGATGTACA-TGGTGCCC-TCAATGGGAACCACCAACATCACTAAATG-GCCCAGGTACACA-CTTTAAAATCGTGCGCGCATGCAGCCGAGATGGGAA
DIP1547        (386)   GCTGGATAACACATCGGCCGAAAACGGACATAGCTTAAGCGGCTAGATGGCTGCAGGGGCATAGTCCTTAAGTCCGGAGACCACGTTCCTGAGG-AGGTA
                       501                                                                                                600
 CE1982        (490)   CGTTGAATTCATGACTGTAGATGAGCAGGTCTCCAACTACTACGACATGCTGCTGAAGCGCAACGCCGGGGAACCTGAGTTCCACCAGGCTGTCGCGGAG
CGL2079        (470)   CGAGGAAATCATGACAGTTGATGAGCAGGTCTCTAACTATTACGACATGCTTCTGAAGCGCAATGCTGGCGAGCCTGAATTTCACCAGGCAGTGGCAGAG
DIP1547        (485)   CG—AAAATGTCGCCTATCGATGAGAAGGTACAGGGCTACTACGAGCTGCTTTTGAAGCGAAACCCTGCGGAGCCGGAATTCCACCAGGCAGTTAACGAA
                       601                                                                                                700
 CE1982        (590)   GTTCTCGAATCTCTGAAGATCGTCCTGGAGAAGGACCCGCACTACGCCGACTACGGTCTGATCCAGCGTCTCTGCGAACCGGAACGCCAGCTGATCTTCC
CGL2079        (570)   GTTTTGGAATCTTTGAAGATCGTCCTGGAAAAGGACCCTCATTACGCTGATTACGGTCTCATCCAGCGCCTGTGCGAGCCTGAGCGTCAGCTCATCTTCC
DIP1547        (583)   GTCCTTGACTCTCTGAAAATTGTTTTGGAAAAGGATCCTCACTACGCGGACTACGGCTTGATTCAGCGCTTGTGTGAGCCTGAGCGCCAGCTTATGTTCC
```

Supplementary Figure 4. Multiple alignment of regulatory region for *gdh* gene in *Corynebacteria*

Boxes showed the proposed start codon for *gdh* gene in each genome sequence.