Evolutionary studies of metabolic networks on the basis of genome information

Tsuyoshi Tanaka

Doctor of Philosophy

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

2003 (School Year)

Acknowledgements

I wish to express my sincere gratitude to my supervisor, Professor Takashi Gojobori for his continuous guidance and encouragement during all the stages of this work. I thank Professors Yoshio Tateno, Toshimichi Ikemura, Naruya Saito, Fumiaki Yamao, and Associate Professor Yoko Satta for their useful comments on my work, serving as the members of my supervisory committee. I wish to express my appreciation to Associate Professor Kazuho Ikeo and Assistant Professor Yoshiyuki Suzuki for their valuable advices. I would appreciate the support and encouragement of all of my colleagues and friends at National Institute of Genetics. Finally, I would like to dedicate this thesis to my parents, Yasuo and Masako, and other family members, Kiyoko and Satoshi.

Contents

Acknowledgements

Contents

Abstract

Chapter 1 Introduction.....1

1.1 Complete genome sequence data.....1

1.2 Metabolic network as an example of biological network for the evolutionary study 2

1.3 Evolutionary study of the metabolic networks in the viewpoint of gain and loss of genes.....4

Chapter 2 Evolution of metabolic networks by gain and loss of enzymatic reaction in eukaryotes.....8

2.1 Introduction.....8

2.2 Materials and methods.....11

2.2.1 Database and species-ER matrix.....11

2.2.2 Estimation of an ER set comprising metabolic networks in the ancestor.....13

2.3 Results.....15

2.3.1 Comparison of the ER sets between different species.....15

2.3.2 The numbers of gains and losses of ERs.....17

2.3.3 Evolution of lipid and complex lipid metabolisms by gain of ERs in the vertebrate lineage.....18

2.4 Discussion.....20

Chapter 3 Evolution of Vitamin B_6 (Pyridoxine) metabolism by gain and loss of genes.....38

3.1 Introduction.....38

3.2 Materials and methods.....41

3.2.1 Genes related to the VB6 metabolism.....41

3.2.2 Phylogenetic tree.....42

3.2.3 Estimation of the gene set of the VB6 metabolism in the ancestor.....42

3.2.4 The order of the losses among the genes.....43

3.3 Results.....44

3.3.1 Comparison of the gene sets among species.....44

3.3.2 Distribution of the genes in the three domains of life.....44

3.3.3 Estimation of the losses of genes.....45

3.3.4 Losses of *SNO* and *SNZ* in animal lineages.....46

3.3.5 Correlation of losses of genes among PLP biosynthetic pathways.....47

3.4 Discussion.....48

Chapter 4 Conclusion.....68

References.....70

Supplementary table 1

Supplementary table2

Supplementary results1

Supplementary results2

Abstract

A variety of biological molecules play important roles in maintaining life through their intermingling networks. The metabolic network is one of them that function as the interactions between substrates and enzymes. In other words, the metabolic network is composed of the enzymatic reactions (ERs) in which one or more than one enzymes catalyze the reaction of the pertinent substrates. Since the metabolism is a basal system for maintaining life of all organisms, any changes of the metabolic networks must have greatly affected the organismic evolution. In this thesis, I have studied the evolution of the metabolic networks using the complete genome sequences and genealogical relationships among species. Based on the complete gene sets of the species studied, I examined whether a particular enzyme encoded by its gene existed in the species of interest. In particular, I conducted a comparative analysis of the metabolic networks among the species using the set of genes. First, I investigated the evolutionary process of the metabolic networks focusing on the gain and loss of ERs, because a single gene often functioned in more than one ER in the metabolic networks. Next, taking the pathways of vitamin B_6 (VB6) metabolism as an example, I systematically estimated the gain and loss of the genes during evolution of the species in order to elucidate the evolutionary process of the metabolic networks. In this examination, I used genes instead of ERs because I was able to identify directly the genes encoding for particular enzymes that were involved with the corresponding ERs in the pathways of VB6 metabolism.

In Chapter 1, I have given an overview of my evolutionary study of the metabolic

networks using the complete genome sequences. Here, I have also described the motivation and purposes of the present study.

In Chapter 2, I conducted comparative studies of ERs in the metabolic networks among the 6 eukaryotic species whose complete genome sequences were determined. For prokaryotes, it has been known that many gene losses had occurred during evolution. For eukaryotes, however, the evolutionary events of the gain and loss of genes in the metabolic networks are unknown, because no systematic studies have been conducted so far. The aim of this chapter is to examine how often gains and losses of enzymatic reactions (ERs) have occurred during the evolution of metabolic networks in eukaryotes, and how these evolutionary events have affected phenotypic traits of the organisms. As a result, I found that the losses of ERs had occurred more frequently than the gains during the evolutionary diversification of metabolic networks in different lineages of eukaryotic species. However, the vertebrate lineage after the separation from *Drosophila melanogaster* showed a remarkable increase in the number of ER gains compared with that of ER losses. In particular, 41% of ER gains were deeply involved with the lipid and complex lipid metabolisms. Because some products of these two metabolisms function as hormones, I concluded that ER gains of the two metabolisms accelerated the development of hormonal signal transduction for the elaborated regulation of physiological system during the vertebrate evolution.

In Chapter 3, in order to understand the evolutionary process of the metabolic networks in more details, I focused upon the VB6 metabolism as an example. The group of VB6, particularly pyridoxal 5'-phosphate (PLP), functions as a cofactor of diverse enzymes in the amino acid metabolisms. Most unicellular organisms and plants can biosynthesize PLP using any one of the three known PLP biosynthetic

pathways; the *de novo* pathway, the salvage pathway and the fungi-type pathway. On the other hand, animals such as insects and mammals have to take it as nourishment because there is a deficiency in the VB6 metabolism in animal lineages. To understand the evolutionary diversification and breakdown of the VB6 metabolism from the viewpoint of gain and loss of the genes, I conducted a comparative analysis of the sets of the genes involved in the VB6 metabolism among the 122 species, including prokaryotes, whose genome sequences were completely determined. In this study, I discussed directly the gain and loss of genes instead of ERs. As a result, I found that any of 10 genes examined was lost more than once in the evolutionary lineages of the 122 species examined. I have also made the following three findings in the evolution of VB6 biosynthesis: (1) A breakdown of the fungi-type pathway occurred at least three times independently in some of animal lineages, (2) the *de novo* pathway was established by generation of the *pdxB* gene in gamma-proteobacteria, and (3) a particular order of gene losses in the PLP biosynthetic pathways was evolutionarily conserved among lineages of the different species. These findings strongly suggest that an evolutionary process of the vitamin B_6 metabolism had been quite dynamic through the events of gain and loss of the genes during evolution of the 122 species examined.

Finally, in Chapter 4, I described the conclusions of the present studies. In particular, I have come to the conclusion that the study of the gain and loss of ERs provides us with profound insight into the understanding of the evolutionary process of metabolic networks. Moreover, I have concluded that the gain and loss of ERs not only played important roles in evolutionary diversification of the metabolic networks, but also greatly affected the whole evolutionary process from prokaryotes to eukaryotes.

Chapter 1

Introduction

1.1 Complete genome sequence data

Since the complete genome sequence of *Haemophilus influenzae Rd* was obtained in 1995 (Fleischmann *et al.* 1995), the available number of the completely sequenced genomes has been still increasing. As of November in 2003, we have 150 species whose compete genome sequences are stored in the Genome Information Broker (GIB), the database collecting the complete genome sequences at DDBJ/NIG (the DNA Data Bank of Japan at National Institute of Genetics) (Fumoto *et al.* 2002). The contents of the 150 species are 128 eubacteria, 16 archaebacteria and 6 eukaryotes. Using the complete genome sequence data, we can conduct a comparative study of the genome sequences among the different species in order to know the existence or the absence of particular genes in the species of interest (Huynen *et al.* 1999). Conducting the large-scale comparative analysis of the complete genome sequence data, I studied the evolutionary process of the metabolic networks from the viewpoint of gains and losses of the genes.

1.2 Metabolic network as an example of biological network for the evolutionary study

I am very much interested in the evolution of the biological networks. In particular, as I mentioned in the previous section, the complete genome sequences provided me with a unique opportunity for conducting the evolutionary study of biological networks because the dataset of all possible proteins encoded by the genes becomes available for large-scale comparisons.

Table 1-1 shows the types of the biological networks which are composed by the interactions between the protein and the other molecules. In the network of the transcriptional regulation, for example, the interaction between the protein called a transcriptional factor and the DNA segment called a regulatory region is mainly needed for transcriptional regulation (Lee *et al.* 2002). In the metabolic network, as another example, low-molecular chemical compounds called "substrates" and proteins called "enzymes" interact with each other (Schuster *et al.* 2000).

In my study, I decided to take the metabolic networks as a typical example of biological network in order to understand their evolutionary process. This is mainly because the metabolic networks have been well studied, so that enzymatic reactions (ERs) can be relatively easily identified by the gene prediction conducted on the complete genome sequence data. The metabolic network is also called the metabolic pathway. In this thesis, I use the term "network" in which various molecules just interact with each other and I ignore direction of the ERs in which the certain substrate is converted to the product by the particular enzyme. On the other hand, when I consider the direction of the interaction, I call that network as a "pathway".

The comparison of metabolic networks among different species has also suggested that both gains and losses of the genes encoding enzymes have often occurred in prokaryotes during evolution (Tatusov *et al.* 1996). For eukaryotes, however, there is no systematic study of evolutionary events of the gain and loss of genes in the metabolic networks. Thus, the following questions are immediately raised: How have the metabolic networks evolved in eukaryotes? To answer this important question, we have to detect, first, the evolutionary changes in the metabolic networks from the viewpoint of gains and losses of ERs.

1.3 Evolutionary study of the metabolic networks in the viewpoint of gains and losses of genes

The metabolism is the basal system to maintain the life from the unicellular organisms to the multicellular organisms. The evolution of the metabolic networks must have affected greatly the life system of the species. Because of this nature, the metabolic networks are thought to be evolutionarily conserved among the species. However, it has been known that there are differences in the metabolic networks among the species, and therefore it is of particular interest to conduct the comparison of the metabolic networks between different species of organisms (Bono *et al.* 1998; Peregrin-Alvarez *et al.* 2003). For conducting this line of study, I would also note that there are various databases which are useful and available freely (Table 1-2).

Taking these advantages of studying the metabolic networks, I conducted comparative analyses of the metabolic networks between different species on the basis of the genome sequence information. The most characteristic point of the present study is to examine the genomic changes of the metabolic networks by estimating the gain or loss of genes and enzymatic reactions (ERs). In particular, I estimated the events of gains or losses of genes and ERs in evolutionary lineages of the species by combining the genomic information with the genealogical relationships among the species. Based upon these results, I discussed the evolutionary process of the metabolic networks in the following chapters.

In this thesis, I utilized the two databases, KEGG and ExPASy database. I could gather the dataset of the entire gene sets from the prokaryotes to eukaryotes are included in the KEGG databases. Moreover, by connecting the KEGG and ExPASy databases, I

predicted the function of the gene more exactly than the use of the one database. When I evaluated the effect of the gain and loss of genes and ERs against the phenotypic traits of the species examined, the categorized 104 metabolic networks in KEGG database could be useful. The other three databases also have the numerous dataset of the metabolic networks, however, I could not utilize them. In the case of BioCyc, the dataset concerned with *E. coli* was massive, however, the dataset of eukaryotes was less enough than that of KEGG and ExPASy. The dataset of the function of the gene in the UM-BBD database is linked to the KEGG database. Therefore, I did not need to the UM-BBD database for the function of the gene. Even though the WIT database contains the original set of the orthologous genes mainly for the bacteria, total number of species in the WIT is less than that of KEGG and ExPASy. Therefore, I also did not use this database instead of the KEGG and ExPASy database.

Table 1-1. Networks in the living organism

Component (The type of interaction)	Example of major network
Protein - DNA	Transcriptional regulation
Protein – RNA	Translation
Protein - Protein	Signal transduction
Protein - Low-molecular compound	Metabolic network
	Signal transduction

Table 1-2. Database related to the metabolism

Database	URL	Ref.
KEGG	http://www.genome.ad.jp/kegg/	(Kanehisa *et al.* 2002)
ExPASy	http://kr.expasy.org/	(Gasteiger *et al.* 2003)
BioCyc	http://biocyc.org/	(Karp *et al.* 2002)
UM-BBD	http://umbbd.ahc.umn.edu/	(Ellis *et al.* 2003)
WIT	http://wit.mcs.anl.gov/WIT2/	(Overbeek *et al.* 2000)

Chapter 2

Evolution of metabolic networks by gain and loss of enzymatic reaction in eukaryotes

2.1 Introduction

Metabolisms of the living system are complex networks of physico-chemical processes, most of which are catalyzed by enzymes. The complete genome sequences allow us to obtain the comprehensive data set of genes encoding enzymes in the metabolic networks (Tweeddale *et al.* 1998; Covert *et al.* 2001). Moreover, even if experimental studies have not been conducted in a species, it has become possible that we reconstruct the metabolic networks on the basis of the prediction of the gene and its function from the complete genome sequence data (Gaasterland and Selkov 1995). Using these approaches, the comparative studies of metabolic networks have been conducted in prokaryotes (Huynen *et al.* 1999; Forst and Schulten 2001). These studies suggest that for the prokaryotic species, both the gain and loss of genes have often occurred during the evolution of metabolic networks (Tatusov *et al.* 1996). In particular, the loss of genes has frequently occurred in parasitic and symbiotic bacteria (Shigenobu *et al.* 2000; Akman *et al.* 2002; Oshima *et al.* 2003). Although models have been proposed to explain the evolutionary mechanism of metabolic networks by the gain of enzymes (Horowitz 1945; Jensen 1976; Schmidt *et al.* 2003), the comparative studies of bacterial genomes have shown that it is not sufficient to consider only the gain of genes for explaining the evolution of metabolic networks in the

prokaryotic lineage (Himmelreich *et al.* 1996; Cole *et al.* 2001).

In the case of eukaryotes, there are a few reports that the loss of genes might have occurred in the metabolic pathway such as a vitamin biosynthetic pathway (Smirnoff 2001). However, there is no systematic study of evolutionary events of the gain and loss of ER in the metabolic networks of eukaryotes, particularly in multicellular organisms. Thus, the following questions are immediately raised to understand the evolutionary process of the metabolic networks in eukaryotes: How have the metabolic networks evolved and how were they diversified during the evolution of the three kingdoms of eukaryotes? Moreover, it is also a question of interest to ask how the diversification of the metabolic networks affected phenotypic traits of organisms.

To answer the above-mentioned questions, I conducted comparative studies of the metabolic networks among 6 eukaryotic species whose genome sequences were completely determined. Using the gene set predicted from the complete genome sequence, I examined whether a particular gene does exist or not in the species.

In this study, I made the following assumptions. First, I consider that a metabolic network in a species is composed of a set of ERs, and I used this set of ERs to compare the metabolic networks among the 6 species. Because a single gene may be involved with more than one ER, my attention has been paid to ERs rather than enzymes or genes themselves in this study. Second, if the two genes encoded the homologous protein sequences, they are assumed to share the same gene function that is represented by a single ER. Thus, I conducted the comparative studies of the ER set in metabolic networks between different species, on the basis of the prediction of gene function through the sequence homology. Finally, I assumed that for a given metabolic network, all the sets of ERs for the 6 eukaryotic species must have been derived from

the single common ancestor (Lazcano and Miller 1999). In other words, I ignored the effect of any possible horizontal gene transfer and parallel evolution against the result of this study. In the case of the horizontal gene transfer, it was mainly reported in prokaryotes (Lawrence JG. 2002), and in eukaryotes, especially *H. sapiens*, the estimated number of horizontal transferred genes is small (Salzberg *et al.* 2001). Therefore, I considered that the gain of ER by horizontal gene transfer had occurred rarely in the eukaryotic lineage. On the other hand, the parallel evolution did not affect to the result of this study, because independent generations of homologous sequences between different species are considered to occur rarely. From these reasons, I considered that the effects of those kinds of evolutionary events can be very small in the present analysis. Therefore, I considered that same ERs shared by the species generated by only one evolutionary event during the evolution of the metabolic networks when I estimated the ERs based on the sequence homology. I, then, estimated the ancestral ER set from the comparison of the 6 extant species on the basis of genealogical relationship. Then, I identified the gain and loss events of ERs from the comparison of the ER sets between the ancestral and extant species.

My comparative studies have shown drastic differences of metabolic networks among the species examined. I have also attempted to make answers to the question which lineage has had the change of metabolic networks during evolution, from the viewpoint of the relationship between the differentiation of the metabolic networks and phenotypic traits of organisms. Thus, this study would give us an important clue in understanding the evolutionary processes of the metabolic networks in each lineage.

2.2 Materials and Methods

2.2.1 Database and species-ER matrix

The present analysis of metabolic networks based on sequence information of enzymes requires the access to adequate databases. Kyoto Encyclopedia of Genes and Genomes (KEGG) provides both an online map of metabolic networks and the annotation of enzymatic reactions for species of interest (Kanehisa *et al.* 2002). In the KEGG database, the function prediction for an enzyme encoded by a gene was updated daily by both computational and manual approaches using sequential and experimental evidence. For identifying the set of ERs for a given metabolic network, I used a total of 94,000 protein sequences reported, in the KEGG database, for the 6 eukaryotic species (*Homo sapiens, Drosophila melanogaster, Caenorhabdits elegans, Saccharomyces cerevisiase, Schizosaccharomyces pombe,* and *Arabidopsis thaliana*) whose complete geneome sequences have been determined, as of 20th April 2003. Although the 94,000 protein sequences contain gene products for both metabolic and non-metabolic enzymes, curation is relatively well-conducted for a larger part of genes encoding metabolic enzymes in these protein sequences of the KEGG database.

Expert Protein Analysis System (ExPASy) proteomics server of the Swiss Institute of Bioinformatics (SIB) also provides the database of the functional categorization and its relevant literature of enzymes as well as analytical tools dedicated to the study of proteomics (Gasteiger *et al.* 2003). For identifying ERs in the metabolic networks from protein sequence families, I also used the ExPASy databases in order to ensure the validity of functional identification of ERs. This is because a metabolic network in a species is assumed to be composed of a set of ERs that is used

for making comparison of the metabolic networks among different species.

For making the homologous protein sequence families on the basis of the predicted genes in the complete genome sequence data, I used the blastp homology search (Altschul *et al.* 1990).

First, I classified 94,000 protein sequences into the homologous sequence families, setting up the threshold of E-value to be 10^{-5}. Because the homologous sequence families, of course, depend on the threshold of E-value, I examined four different E-values for the threshold: 10^{-100}, 10^{-50}, 10^{-10}, and 10^{-5}. When I took 10^{-100} and 10^{-50}, the number of sequences in one sequence family became so small that identification of ERs was virtually impossible. On the other hand, whichever 10^{-10} or 10^{-5} is taken, the numbers of sequences in the family were almost the same. Because 10^{-5} is often used for homology search for gene prediction in the genome studies (Read *et al.* 2003; Galagan *et al.* 2003), I took 10^{-5} as the threshold of E-value in the present study. As mentioned above, I identified a total of 751 ERs in 104 different kinds of metabolic networks based on the KEGG and ExPASy databases, obtaining 751 corresponding protein families from the homology search of predicted gene sequences in the complete genomes.

Second, in order to know how many ERs a given species possesses in its relevant metabolic networks, I examined whether each of the 751 ERs existed in the metabolic networks of a species of interest. In practice, if a species has a particular protein sequence that is contained in the family corresponding to an ER, I decided that the species had the ER.

Third, I created a dataset of ERs for the out-group species of eukaryotes, because I had to estimate an ER set of the ancestral species using the out-group species (See the

next section). As the dataset of ERs in the out-group species, I used the dataset of bacterial ERs that was included in the "enzyme" dataset provided by the KEGG database. In practice, I constructed the dataset of the out-group ER set that was made by taking a union of all ER sets among all the 112 bacterial species examined because I did not deal with the gain and loss of ERs in the bacterial lineage in this study.

Fourth, for the ER sets of the 6 eukaryotic species together with the out-group ER set, I formed a matrix of 7 x 751 whose element was n_{ij} (i = 1,2,.., and 7, and j = 1,2, .., and 751), where n_{ij} is represented by the following two states: n_{ij} = 1 when the i-th species had at least one gene for the j-th ER, and n_{ij} = 0 when the i-th species had no gene for the j-th ER. Let us call this matrix as the ER matrix. Thus, the ER matrix was used for the comparison of the set of ERs among species.

2.2.2 Estimation of an ER set comprising metabolic networks in the ancestor

I used a phylogenetic tree that was based on the study of Baldauf *et al.* (2000) to estimate the ER sets of the ancestor of the eukaryotes. For estimating the ancestral set of ERs, I used the ER matrix comparing the 0-1 state elements as mentioned above. There are five internal-nodes in the phylogenetic tree of the 6 eukaryotes. As the first step, I estimated the state (0 or 1) at the immediately common ancestral node of the neighboring two species in the tree by comparing between the states of the two species (Figure 2-1). If two states were the same, namely (0,0) or (1,1), I assumed that the ancestor has the same state, meaning that nothing had changed (Patterns 1 and 2 in Figure 2-1). If two states were different, namely (0,1) or (1,0), I examined the states for all of species that were located in the outside of the two comparing species in the

tree and out-group species. Let me denote a group of those species by species group C. If at least one species in group C have the state of "1", I estimated the ancestral state to be "1" (Pattern 3 in Figure 2-1). If all species of group C have the state of "0", I estimated the ancestral state to be "0" (Pattern 4 in Figure 2-1). In this way, I estimated the states at all internal-nodes for a given ER. Thus, for a given species or ancestral node, the number of existing ERs has been easily obtained by the sum of all the states (0 or 1) over all possible ERs. Thus, the number of existing ERs can be compared between different species. Note that my estimations are based upon the assumption that a single ER was acquired only once during evolution of the 6 eukaryotes in this model. As a result, the number of ERs composing the metabolic networks in the common ancestor of all the 6 eukaryotes was 622.

A total of 104 kinds of metabolic networks were classified into 11 categories according to the definitions given in the KEGG database. For a given category, I estimated how many ERs exist for each of all the 6 species.

2.3 Results

2.3.1 Comparison of the ER sets between different species

As shown in Table 2-1, the number of ERs varied considerably with species. In fact, the number of ERs was 417 in *S. pombe* at the smallest, and 578 in *H. sapiens* at the largest among the 6 species examined. When I counted the number of ERs that were commonly shared among all the 6 species, it was only 226, being 36% of the estimated 622 ERs of the common ancestor of the 6 eukaryotes. Thus, I found that the ERs must have undergone dynamic changes in the number of ERs during the evolutionary diversification of eukaryotic species.

When I compared the number of ERs for each of the 6 eukaryotes with that of the common ancestor of the 6 eukaryotes, I made a surprisingly interesting observation that the ancestor had the number of ERs larger than any of all the species examined. This result strongly suggests that the number of ERs decreased in the metabolic networks for all the lineages during evolution, implying that the losses of ERs occurred more frequently than the gains for all evolutionary lineages. Thus, this feature appears to be common between prokaryotes and eukaryotes.

When I focused on the gain and loss of ERs in each evolutionary lineage towards the extant species from the common ancestors, the total number of gains and losses of ERs varied considerably with evolutionary lineages; from 114 in *A. thaliana* at the smallest to 302 in *D. melanogaster* at the largest (Table 2-1). Interestingly, I observed that the number of losses of ER was more than the number of gains of ER in any of all evolutionary lineages examined. This reinforces my observation that the losses of ER occurred more frequently during the evolution of metabolic networks than the gains of

ER among the 6 eukaryotic species.

Next, I studied the distribution of ERs for the 104 different kinds of metabolic networks in the 6 eukaryotic species as well as the common ancestor of the 6 species (Table 2-2). In the 104 metabolic networks, there are only three networks that were not changed at all during evolution from the common ancestor to all the 6 extant species; the metabolic networks of ATP synthesis, 1,2-Dichloroethane degradation, and phospholipid degradation. It suggests that most of the metabolic networks had experienced one or more than one changes of gains or losses in the number of ERs in evolutionary lineages. It again indicates a dynamic feature of evolutionary changes in metabolic networks.

Interestingly enough, there are metabolic networks in which species-specific changes in the number of ERs were observed. For example, in the networks of valine-leucine-isoleucine biosynthesis, lysine biosynthesis and phenylalanine-tyrosine-tryptophan biosynthesis, the numbers of ERs were observed to have changed drastically only in the animal lineages of *H. sapiens*, *D. melanogaster* and *C. elegans*. For these three biosynthesis pathways, on the other hand, few changes were observed in the plant or yeast lineages of *A. thaliana*, *S. cerevisiae* and *S. pombe*. Moreover, it is of particular interest to note that the pathway of flavonoids, stilbene and lignin biosynthesis showed substantial increase in the number of ERs only in the plant lineage of *A. thaliana*. Because flavonoids are essential for plant reproduction (Winkel-Shirley 2002), the gains of ERs could have contributed to the establishment of the plant-specific system of the reproduction. Thus, such kinds of metabolic networks may have played an important role in characterization of species during evolutionary diversification.

My observation in which the number of ERs involved with the amino acid biosynthesis decrease in the animal lineage is supported by the fact that most species of plants and yeasts are known to be capable of biosynthesizing the amino acids whereas the animals such as human are required to obtain the so-called essential amino acids that are not produced by themselves.

2.3.2 The numbers of gains and losses of ERs

As mentioned above, the distribution of the number of ERs over the 104 metabolic networks were species-dependent as long as the 6 eukaryotes examined were concerned. This is also a strong indication that gains and losses, had taken place considerably in each lineage of the 6 species during evolution. When I analyzed the difference of the metabolic networks between the 6 eukaryotes based on the existence of ERs in detail, I discovered the metabolic networks of which I explained the difference between species by the following four patterns, conservation, alternative conservation, ER losses and ER gains, even if it was hard to categorize almost metabolic networks into each pattern (See the supplementary results1). When I counted the numbers of losses and gains separately, I found that a total numbers of 675 losses and 129 gains of ERs occurred in metabolic networks during evolution of the 6 eukaryotic species (Figure 2-2). It follows that 1.1 losses and 0.2 gains per ER, on the average, had occurred since the 622 ERs of the common ancestor diverged out to the 6 extant eukaryotes. Thus, the event of losses had occurred approximately five times more often than that of gains. Moreover, I found that 37% of 525 ERs, namely 193 ERs, had more than one loss or both of gain and loss of ERs in all the evolutionary lineages of the 6 species (Figure 2-3). It suggests that about one forth of ERs underwent several gains

or losses independently during the evolution of metabolic networks.

According to the KEGG database, the metabolic networks examined in the present study were separated into 11 different categories of biological function (Table 2-3). As shown in Table 2-3, it is interesting to see that the numbers of ER losses were larger than the numbers of ER gains in any of all categories. However, the differences in the numbers between ER losses and ER gains were very large, depending upon categories. I would point out that for the amino acid metabolism, both numbers of ER losses and gains were larger than those for the other categories. It indicates that ERs in the amino acid metabolism are quite flexible in the evolutionary change, probably because necessary amino acids can be acquired from the foods or symbiotic bacteria. On the other hand, in the metabolisms of complex carbohydrates and complex lipid, the numbers of ER gains were larger, compared with those of the other categories. This will be discusses later. These results strongly suggest that the likelihood of occurrences of the gains and losses of ERs depends upon metabolic category.

2.3.3 Evolution of lipid and complex lipid metabolisms by gain of ERs in the vertebrate lineage

I found that the evolutionary lineage toward the vertebrate showed the largest number of gains among the 6 species. In fact, this lineage had undergone 107 gains of ERs, which corresponds to 83% of the estimated number of ER gains. This is an amazingly large number of gains. More interestingly, I found that 46 out of 107 gains of ERs occurred in the lipid and complex lipid metabolisms, which corresponds to 43% of the total number of gains (Table 2-4). In particular, in the two metabolisms, a total of 17 ER gains out of 46 had occurred in the vertebrate lineage after the divergence

from *D. melanogaster*. Because only one ER loss had occurred in the same lineage, the number of ER gains is outstandingly high.

There are several explanations to be considered for the extraordinary gains of ERs in the vertebrate lineage. However, the most possible explanation is that the ER gains in the lipid and complex lipid metabolisms must have contributed to the evolutionary formation of an exquisite system of signal transduction in the vertebrate. This explanation is actually supported by the fact that the products of the two metabolisms were often involved deeply with the regulation system of signal transduction (Yamashita *et al.* 1999; Unger and Orci 2002; Kolesnick and Fuks 2003). In particular, the products of three metabolic networks such as the prostaglandin and leukotriene metabolism, the C21-steroid hormone metabolism and the androgen and estrogen metabolism, were known to work as hormones or autacoids in vertebrates (Huber and Gruber 2001; Funk 2001; Dubey *et al.* 2002). Hormones and autacoids are functionally very important for maintenance of the complex system of physiological regulations (Yamashita *et al.* 1999). When I looked at the prostaglandin and leukotriene metabolism, 6 gains of ERs were observed to have occurred in the vertebrate lineage. More interestingly 5 ERs out of 6 had emerged in the vertebrate lineage after the divergence between *D. melanogaster* and *H. sapiens* (Figure 2-4). The products derived from the gains of these ERs were all hormones that work in various physiological regulations such as the cell proliferation and constriction of sooth muscle (Table 2-5). These results suggest that the ER gains of the lipid and complex lipid metabolisms had played an important role in the evolution of the vertebrates whose system of physiological regulation is exquisitely complex.

2.4 Discussion

I examined the gain and loss of ERs in the metabolic networks for evaluating their evolutionary process. It is clear that the number of ERs examined is not enough to construct the complete metabolic networks in the 6 eukaryotes because the complete set of ERs for the metabolic networks has not been determined. There are so many ERs that have not been discovered yet. Therefore, it is hard to I conclude the correctness of the number of ERs in the ancestor and the absolute values of gain and loss of ERs. In other words, ascertain bias must exist in this study because the depth of the study for the metabolic networks is different between examined species. This means that the number of gain of ERs will increases in any lineages because of the progress of the study of the novel ERs. However, number of ER gains is not larger than that of lost ERs in any lineage. This result suggests that a total number of ER gains will not be larger than that of lost ERs in this study even if the novel ERs are found. In fact, when I conducted the comparative analysis of the metabolic networks by the use of the set of 1,046 ERs, the number of gain of ERs increased in any lineage. However, the loss of ERs have occurred more frequently than the gain of them (Figure 2-5).

The ascertain bias was affected by the dataset of sequences that I used. In this study, I used the protein sequences of only 6 eukaryotes whose genome sequences were completely determined and I estimated the gain of ERs based on the distribution of homologous sequences involved with the ERs. However, if there are homologous sequences in other species that I did not use in this study, the time of the gain of ER must be changed. I evaluated the possibility by conducting the homology search of the

58 protein sequences discovered in only one species examined to the non redundant protein database as of Jan. 7th 2004. When I made the 58 phylogenetic trees using the matched sequences, there were 6 (10%) trees that the timing of ER gains may be doubtful in my estimation (See the supplementary results2). This result suggests that the timing of ER gain may be changeable by the increase of the sequences examined.

Moreover, In this study, the estimation of the gain and loss of genes was conducted with special care of the following two points; horizontal gene transfer and parallel evolution. Although I understand that these are important in the understanding of evolutionary mechanisms, I consider that they should not affect seriously the estimation of ER gains and losses in this study because of the reasons that I mentioned in the section of the introduction in this chapter. My results may be affected by a topology of the phylogenetic tree used. When I used two different topologies particularly regarding a plant and yeasts, the number of ERs of the common ancestor of the 6 eukaryotes and the numbers of the gains and losses of ERs were changed in some lineages (Figure 2-5). Because those changes are not substantial, however, the essential points of the results obtained remain unchanged.

In this study, I found that a lot of losses and gains of ERs had occurred during evolution of the 6 eukaryotic species whose complete genomes have been sequenced. In particular, the number of ERs in the extant 6 species varied considerably with the species. I fully understand that the estimated number of ERs may not be accurate. In particular, because the ERs were identified by the gene sequence families that were classified by homology search, I may have underestimated the number of ERs. However, although I may have missed other ERs that were not identified in the present study, it should not affect much the comparisons of ERs between different species.

This is because the same criteria of identifying ERs were always adopted equally to all the 6 eukaryotic species. From the same reasons, I must take with caution that the common ancestor of the 6 eukaryotic species must have had 622 ERs. This number of the ancestral ERs can be underestimated. However, I am not much interested in the absolute number of ERs, but am interested in the numbers of gains and losses of ERs in the evolutionary lineages from the common ancestor to each of the 6 species, as long as the ERs identified here are concerned. Thus, the present comparative study of ER gains and losses in the metabolic network should be valid. It is interesting to note that a total of 804 events of gains and losses were observed in the evolution of the ancestral 622 ERs. These tremendous evolutionary changes were, for the first time, found in the eukaryotic species by the comparative genome approach.

It is reasonable to consider that losses of ERs might affect the metabolic networks more seriously than gains of ERs, because even a single event of loss of ER possibly causes destruction of the metabolic network. Thus, it is quite possible that such losses of ERs would be selected against by severe pressures of negative selection. It is natural that the previous studies of metabolic networks have concentrated mostly on gains of ERs. In the present study, I found that the losses of ERs had occurred much more frequently than the gains of ERs in the eukaryotic species examined: The former occurred about five times frequently than the latter. This may be explained by heterotrophy in which the essential compounds for sustaining the life can be acquired from the other organisms. It is well exemplified by the amino acid metabolism. In this study, I found that a total of 156 ER losses in the amino acid metabolism and that most of those occurred in the animal lineages. It has been well known that for animals, the essential amino acids are obtained from the foods or symbiotic/parasitic bacteria

(Reeds and Garlick 2003), though most plants are completely autotrophic. Moreover, the metabolisms of carbohydrates and cofactors/vitamins had undergone more numbers of ER losses than the other metabolisms except the amino acid metabolism; 74 and 64 losses of ERs are observed, respectively. More than half of losses for both metabolisms occurred particularly in the animal lineages. These observations are also consistent with the well known facts that the animals usually need these essential compounds of nourishment from the environments. It is also possible that the lost ER can be complemented by another ER in the metabolic pathway. Thus, the result in the study is consistent with Jeong and his colleagues' speculation that the metabolic networks are robust, error-tolerant and scale-free (Jeong at al. 2000).

The ER gain is often used for explaining the evolution of the metabolic networks (Horowitz 1945; Jensen 1976; Schmidt *et al.* 2003). It is easily conceivable that the species gaining the new ER may become able to use a new substrate, leading to formation of the new system. In this study, I found that over one third of total ER gains concentrated on the lipid and complex lipid metabolisms in the vertebrate lineages from the common ancestor. Moreover, after the divergence from *D. melanogaster*, the vertebrate lineage had a total of 17 ER gains, though only one loss of ER had occurred in these categories of metabolic networks. Because some of the fat-soluble hormonal compounds were produced through the lipid and complex lipid metabolisms (Liston and Roberts 1985; Wen *et al.* 2003; Soberman and Christmas 2003), the ER gains of the lipid and complex lipid metabolisms enabled the species to develop the advanced hormonal system. It eventually led the species to the completion of the complex system of physiological regulation during the evolution in the vertebrate lineage.

In conclusion, I successfully estimated the gains or the losses of ERs, which

should give us important insight into the understanding of the evolutionary process of the metabolic networks.

Table 2-1. List of species

Symbol	Name	ERs	Gain*	Loss*	Total
ATH	*Arabidopsis thaliana*	538	15	99	114
CEL	*Caenorhabditis elegans*	465	52	209	261
DME	*Drosophila melanogaster*	454	67	235	302
HSA	*Homo sapiens*	578	107	151	258
SCE	*Saccharomyces cerevisiae S288C*	424	23	221	244
SPO	*Schizosaccharomyces pombe*	417	20	225	245

*Number of the gain and loss was a total of counts from the common ancestor of the 6 eukaryotes to each extant species.

Table 2-2. Number of ERs in each metabolic networks in the 6 eukaryotes

Metabolic network	Number of ER						
	ANC	H SA	DME	CEL	SCE	SPO	ATH
Amino Acid Metabolism							
Urea cycle and metabolism of amino groups	23	21	14	13	19	20	22
Glutamate metabolism	26	22	22	23	21	19	20
Alanine and aspartate metabolism	19	19	17	17	16	16	16
Glycine, serine and threonine metabolism	37	32	24	22	25	28	24
Methionine metabolism	15	9	8	8	12	10	9
Cysteine metabolism	11	7	5	7	9	9	7
Valine, leucine and isoleucine degradation	17	18	15	17	6	5	14
Valine, leucine and isoleucine biosynthesis	11	6	6	6	11	10	11
Lysine biosynthesis	11	4	3	3	9	9	10
Lysine degradation	14	17	13	12	10	10	13
Arginine and proline metabolism	26	30	22	21	17	21	20
Histidine metabolism	17	11	5	7	11	12	12
Tyrosine metabolism	13	19	13	15	5	8	11
Phenylalanine metabolism	9	9	7	8	5	7	9
Tryptophan metabolism	18	25	15	20	13	11	14
Phenylalanine, tyrosine and tryptophan biosynthesis	22	6	6	6	20	21	21
Biodegradation of Xenobiotics							
gamma-Hexachlorocyclohexane degradation	5	5	4	4	4	4	3
Benzoate degradation via hydroxylation	2	1	1	1	2	1	1
Tetrachloroethene degradation	1	1	1	1	1	0	1
1,4-Dichlorobenzene degradation	1	0	0	0	1	0	0
1,2-Dichloroethane degradation	1	1	1	1	1	1	1
Benzoate degradation via CoA ligation	8	8	8	6	4	4	7
Styrene degradation	4	4	4	4	1	1	2
Atrazine degradation	2	1	1	1	1	2	2
Caprolactam degradation	1	1	0	1	1	1	1
Biosynthesis of Secondary Metabolites							
Streptomycin biosynthesis	3	3	3	3	3	1	3
Erythromycin biosynthesis	4	3	3	3	3	3	3
Terpenoid biosynthesis	7	7	5	5	6	6	7
Flavonoids, stilbene and lignin biosynthesis	8	4	4	6	3	4	13
Alkaloid biosynthesis I	4	4	4	4	1	2	3
Alkaloid biosynthesis II	2	2	1	1	1	2	3
Carbohydrate Metabolism							
Glycolysis / Gluconeogenesis	26	26	25	22	23	22	25
Citrate cycle (TCA cycle)	17	16	14	16	13	13	14
Pentose phosphate pathway	18	16	16	16	15	15	16
Inositol metabolism	2	2	2	2	1	1	2
Pentose and glucuronate interconversions	8	7	6	7	6	5	8
Fructose and mannose metabolism	18	16	14	15	12	12	17
Galactose metabolism	13	15	13	11	12	11	12
Ascorbate and aldarate metabolism	2	1	2	2	2	2	4
Pyruvate metabolism	31	21	18	20	21	19	28
Glyoxylate and dicarboxylate metabolism	14	9	5	7	9	8	14
Propanoate metabolism	16	15	10	14	6	7	12
Butanoate metabolism	17	18	15	16	9	9	12
C5-Branched dibasic acid metabolism	2	2	1	2	1	1	1

Table 2-2. (Continue)

Metabolic network	Number of ER						
	ANC	H SA	DME	CEL	SCE	SPO	ATH
Energy Metabolism							
Oxidative phosphorylation	7	6	5	6	6	6	6
ATP synthesis	1	1	1	1	1	1	1
Photosynthesis	2	2	2	2	1	2	2
Methane metabolism	5	4	3	4	5	4	5
Carbon fixation	19	12	12	12	12	11	20
Reductive carboxylate cycle (CO2 fixation)	9	7	5	8	5	4	7
Nitrogen metabolism	19	13	12	13	13	12	16
Sulfur metabolism	15	7	5	6	12	10	11
Lipid Metabolism							
Fatty acid biosynthesis (path 1)	7	8	5	9	9	9	5
Fatty acid biosynthesis (path 2)	5	5	4	5	4	3	4
Fatty acid metabolism	14	16	15	15	9	7	13
Synthesis and degradation of ketone bodies	5	5	4	5	2	3	3
Sterol biosynthesis	16	15	9	8	12	12	16
Bile acid biosynthesis	6	11	7	8	5	4	5
C21-Steroid hormone metabolism	4	9	5	5	2	3	3
Androgen and estrogen metabolism	9	13	7	8	4	5	7
Metabolism of Cofactors and Vitamins							
Ubiquinone biosynthesis	4	2	2	2	2	2	4
One carbon pool by folate	16	16	12	11	13	13	15
Thiamine metabolism	5	1	1	1	4	4	5
Riboflavin metabolism	9	5	4	7	7	8	8
Vitamin B6 metabolism	4	5	3	3	4	4	4
Nicotinate and nicotinamide metabolism	9	11	6	8	7	6	7
Pantothenate and CoA biosynthesis	13	8	6	8	10	10	12
Biotin metabolism	3	2	2	1	3	2	3
Folate biosynthesis	13	10	9	6	10	10	9
Retinol metabolism	2	2	1	2	0	0	2
Porphyrin and chlorophyll metabolism	16	15	14	5	11	13	14
Metabolism of Complex Carbohydrates							
Starch and sucrose metabolism	30	21	18	19	18	16	30
N-Glycans biosynthesis	11	16	15	12	8	7	11
N-Glycan degradation	8	8	7	7	1	1	6
O-Glycans biosynthesis	2	5	2	3	0	0	2
Aminosugars metabolism	14	14	13	12	9	7	8
Glycosaminoglycan degradation	4	9	8	6	0	1	4
Chondroitin / Heparan sulfate biosynthesis	4	6	3	6	0	0	4
Keratan sulfate biosynthesis	0	4	3	2	0	0	0
Lipopolysaccharide biosynthesis	3	0	0	0	0	0	3
Peptidoglycan biosynthesis	2	1	1	1	1	1	2
Metabolism of Complex Lipids							
Glycerolipid metabolism	33	37	34	33	29	27	30
Inositol phosphate metabolism	8	11	10	10	7	5	8
Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	0	1	0	0	0	0	0
Sphingophospholipid biosynthesis	1	2	2	2	2	1	0
Phospholipid degradation	5	5	5	5	5	5	5
Prostaglandin and leukotriene metabolism	9	15	6	6	5	4	10
Sphingoglycolipid metabolism	8	12	7	9	3	2	5
Blood group glycolipid biosynthesis - lact series	1	5	1	2	0	0	1
Blood group glycolipid biosynthesis - neolact series	2	9	4	5	0	0	2
Globoside metabolism	5	9	5	4	1	1	5
Ganglioside biosynthesis	1	5	1	1	0	0	1

Table 2-2. (Continue)

Metabolic network	Number of ER						
	ANC	H SA	DME	CEL	SCE	SPO	ATH
Metabolism of Other Amino Acids							
beta-Alanine metabolism	13	13	9	11	7	7	11
Taurine and hypotaurine metabolism	2	5	4	4	2	1	2
Aminophosphonate metabolism	3	3	3	3	4	3	3
Selenoamino acid metabolism	14	9	8	10	13	10	10
Cyanoamino acid metabolism	6	5	5	6	5	6	5
D-Glutamine and D-glutamate metabolism	2	2	2	2	0	0	1
D-Arginine and D-ornithine metabolism	1	2	2	2	1	2	1
D-Alanine metabolism	1	0	0	0	0	1	0
Glutathione metabolism	10	9	10	10	8	8	10
Nucleotide Metabolism							
Purine metabolism	56	51	43	46	42	43	46
Pyrimidine metabolism	36	36	30	33	26	27	32
Nucleotide sugars metabolism	6	7	6	6	3	4	5

In the column of ANC, the number represented a number of ERs found in the common ancestor among all the 6 species. In the column of each species, for a given metabolic network, I estimated how many ERs exist for each of all the 6 species.

Table 2-3. Number of ERs in each category of metabolic networks*

Categories of metabolic networks	No. of ERs in ancestral ER set*	No. of lost ER**	No. of gained ER*
Amino Acid Metabolism	214	156	32
Biodegradation of Xenobiotics	24	16	0
Biosynthesis of Secondary Metabolites	26	17	7
Carbohydrate Metabolism	129	74	11
Energy Metabolism	73	41	4
Lipid Metabolism	51	34	18
Metabolism of Cofactors and Vitamins	88	63	8
Metabolism of Complex Carbohydrates	71	58	27
Metabolism of Complex Lipids	62	35	32
Metabolism of Other Amino Acids	48	31	6
Nucleotide Metabolism	86	43	3

*The numbers of ERs were counted at every category of metabolic networks in KEGG.

** These numbers were total numbers of gains or losses of ERs in each categories, respectively.

Table 2-4. Numbers of gains of ERs in lipid and complex lipid metabolism in the vertebrate lineage after the divergence with *D. melanogaster*

Metabolic networks	The number of the ER gain*
Prostaglandin and leukotriene metabolism	5(6)
Blood group glycolipid biosynthesis - neolact se	3(7)
Ganglioside biosynthesis	3(4)
C21-Steroid hormone metabolism	2(5)
Androgen and estrogen metabolism	2(4)
Globoside metabolism	2(4)
Glycerolipid metabolism	1(6)
Bile acid biosynthesis	1(5)
Sphingoglycolipid metabolism	1(4)
Blood group glycolipid biosynthesis - lact series	1(4)
Fatty acid metabolism	1(2)
Glycosylphosphatidylinositol(GPI)-anchor biosy	1(1)
Fatty acid biosynthesis (path 1)	0(3)
Inositol phosphate metabolism	0(3)
Sphingophospholipid biosynthesis	0(1)
Fatty acid biosynthesis (path 2)	0(0)
Synthesis and degradation of ketone bodies	0(0)
Sterol biosynthesis	0(0)
Phospholipid degradation	0(0)

*The numbers of gain in the parentheses represented total numbers of ER gains from the common ancestor among the 6 eukaryotic species to *H. sapiens*.

Table 2-5. Function of products in prostaglandin and leukotriene metabolism

ER number	Enzymatic reaction	Products	Function	Ref.
14	prostaglandin-F synthase	11-epi PGF 2α	Blood pressure	(Liston and Roberts 1985)
82	arachidonate 12-lipoxygenase	12-HPETE	Cell proliferation	(Wen et al. 2003)
414	LTC4 synthase	LTC4	Constriction of smooth muscle	(Soberman and Christmas 2003)
971	prostaglandin-D synthase	PGD2	Chemotaxis, Allergic asthma	(Funk 2001)
972	PGI2 synthase	PGI2	Declumping, Vasodilation	(Funk 2001)
973	thromboxane-A synthase	TXA2	Vasoconstriction, Aggregation	(Funk 2001)

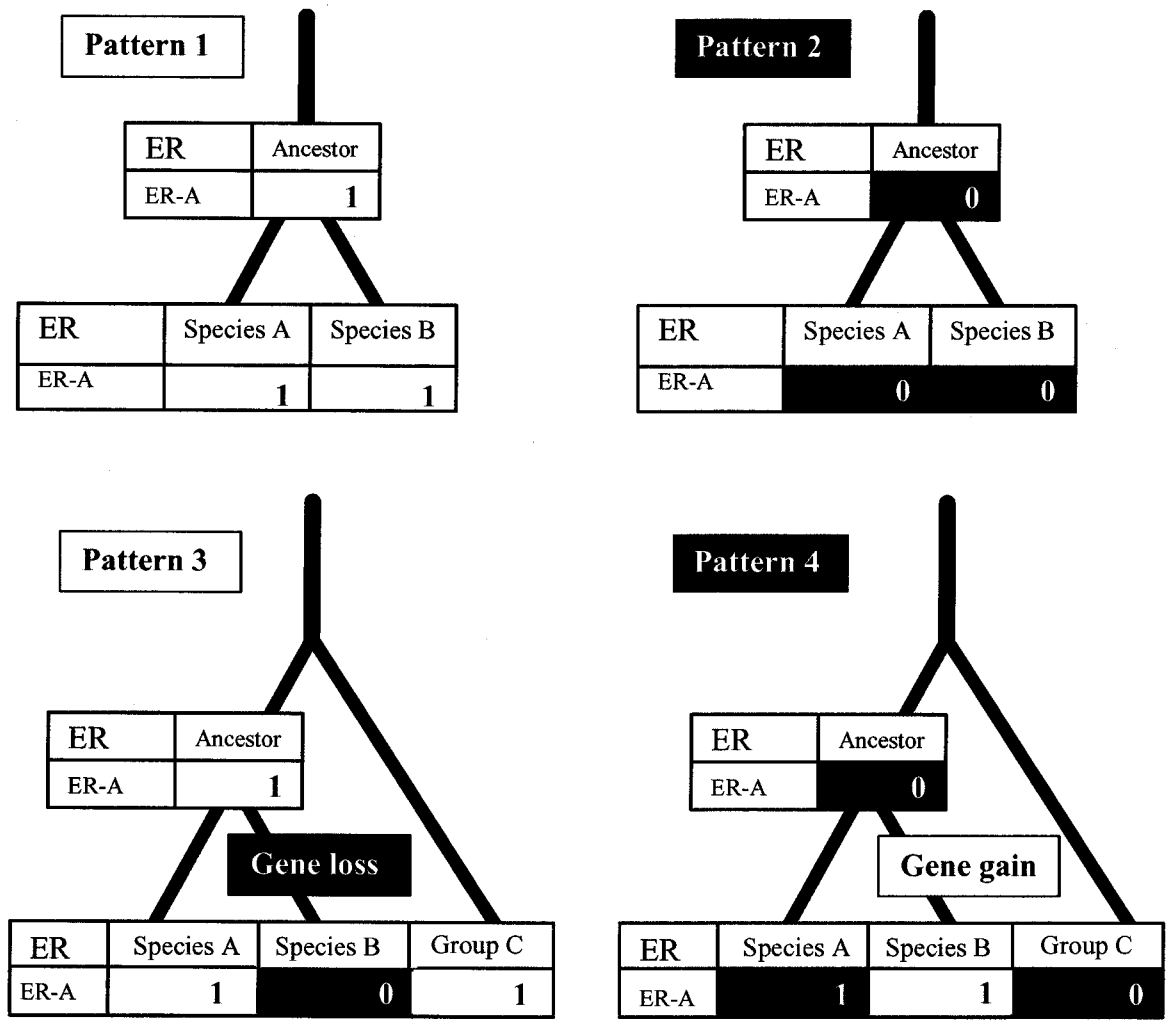


Figure 2-1. Estimation of the ancestral gene set and the timing of the gain and loss of ER

ER-A represents one of the enzymatic reactions. In this figure, the state "1" is defined when the species had at least one gene with the ER-A, and "0" is defined when the species had no gene with the ER-A.

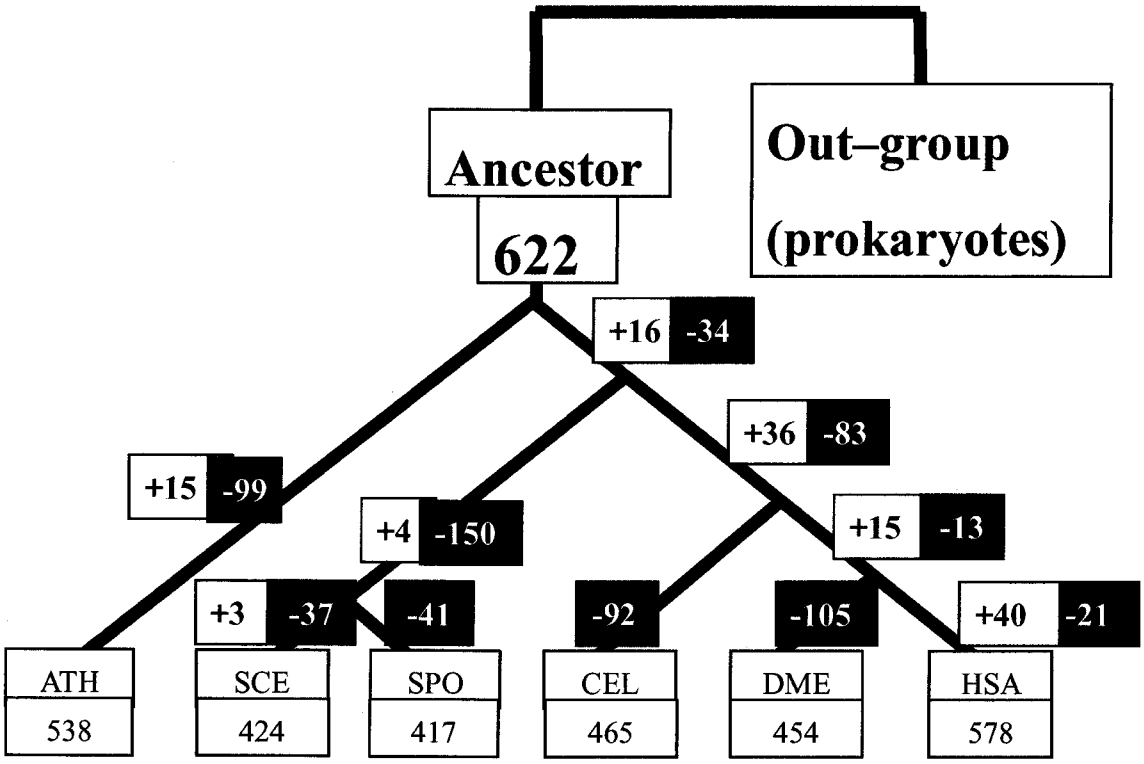


Figure 2-2. Phylogenetic tree of 6 eukaryotes

The numbers below the 6 species and the ancestor were total numbers of ERs existing in each organism. The red number indicated the total number of gains of ERs in the branch and the white number indicated the total number of losses of ERs in the branch.

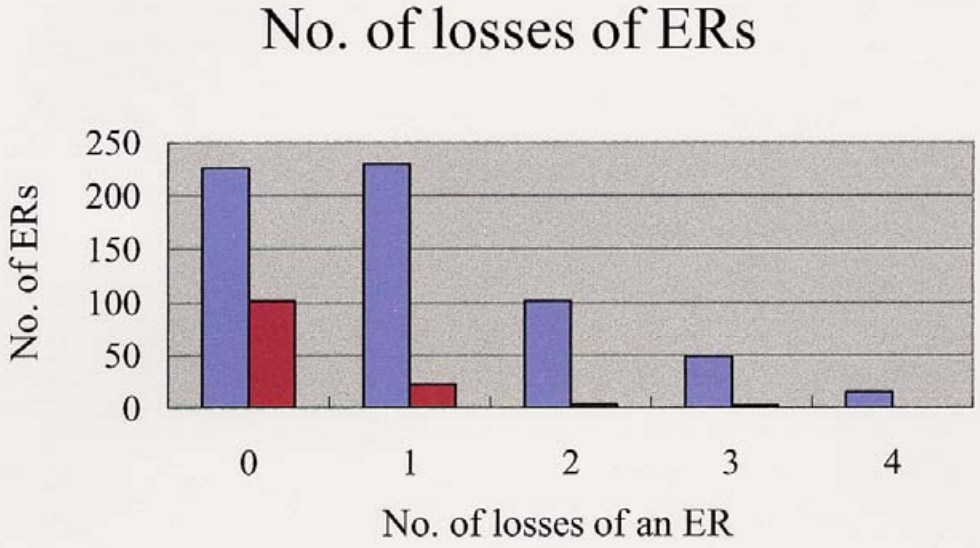


Figure 2-3. Distribution of total numbers of the gain and loss in ER

Blue bar represents the number of ERs that were only lost in this study. Red bar represents the number of ERs that were gained once and then were lost in this study.

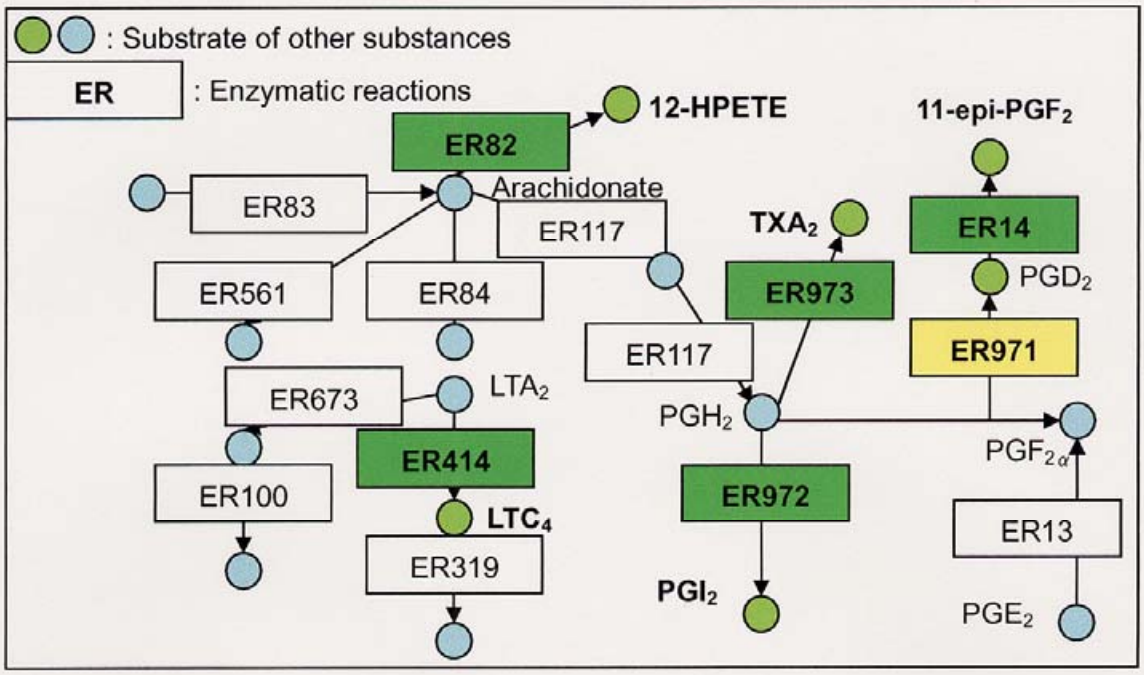


Figure 2-4. Prostaglandin and leukotriene metabolism in *H. sapiens*

ER colored in white represents that the ER existed in the common ancestor of the 6 eukaryotes examined. ERs colored in yellow or green were observed after the divergence between yeast and animal or invertebrate and vertebrate, respectively. Substrate colored in green represents that it was produced by the ER in the vertebrate lineage after the divergence from *D. melanogaster*. See the Table 2-5 for more detail about the gained ERs and the compounds produced newly by these ERs. ER414 is functional in the metabolic networks of the taurine and hypotaurine metabolism, the selenoamino acid metabolism, the cyanoamino acid metabolism and the glutathione metabolism.

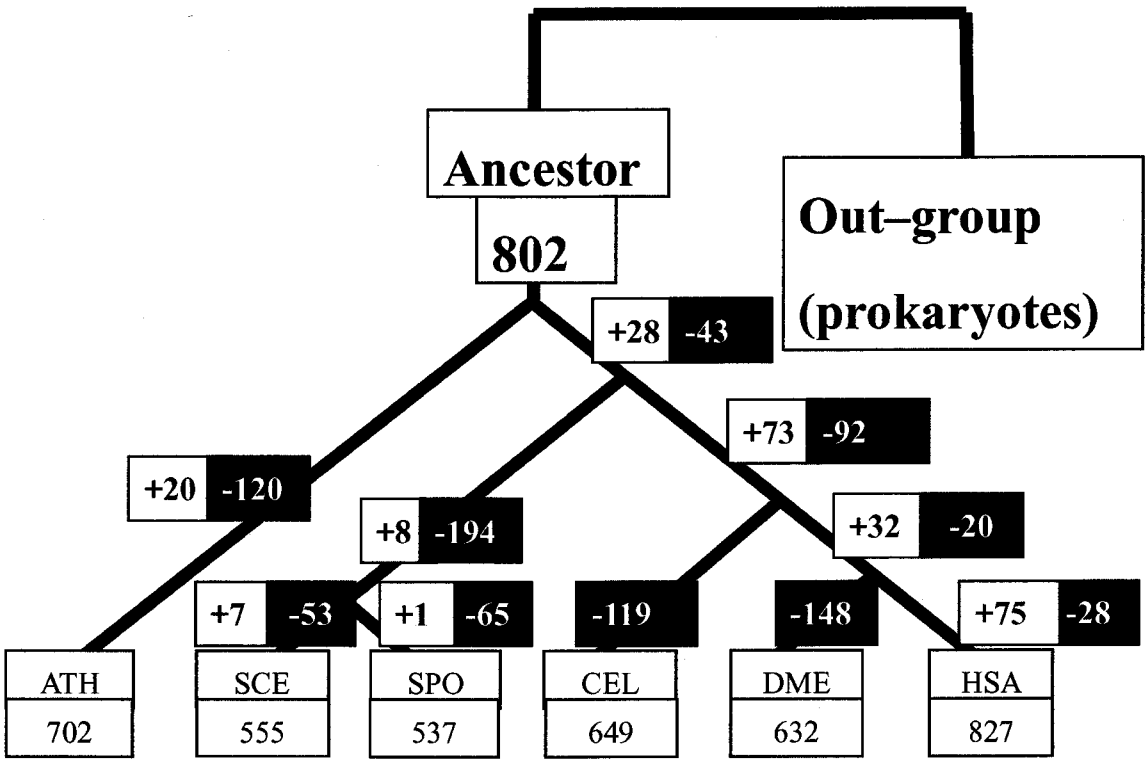


Figure 2-5. Phylogenetic tree of 6 eukaryotes

The numbers below the 6 species and the ancestor were total numbers of ERs existing in each organism. The red number indicated the total number of gains of ERs in the branch and the white number indicated the total number of losses of ERs in the branch.

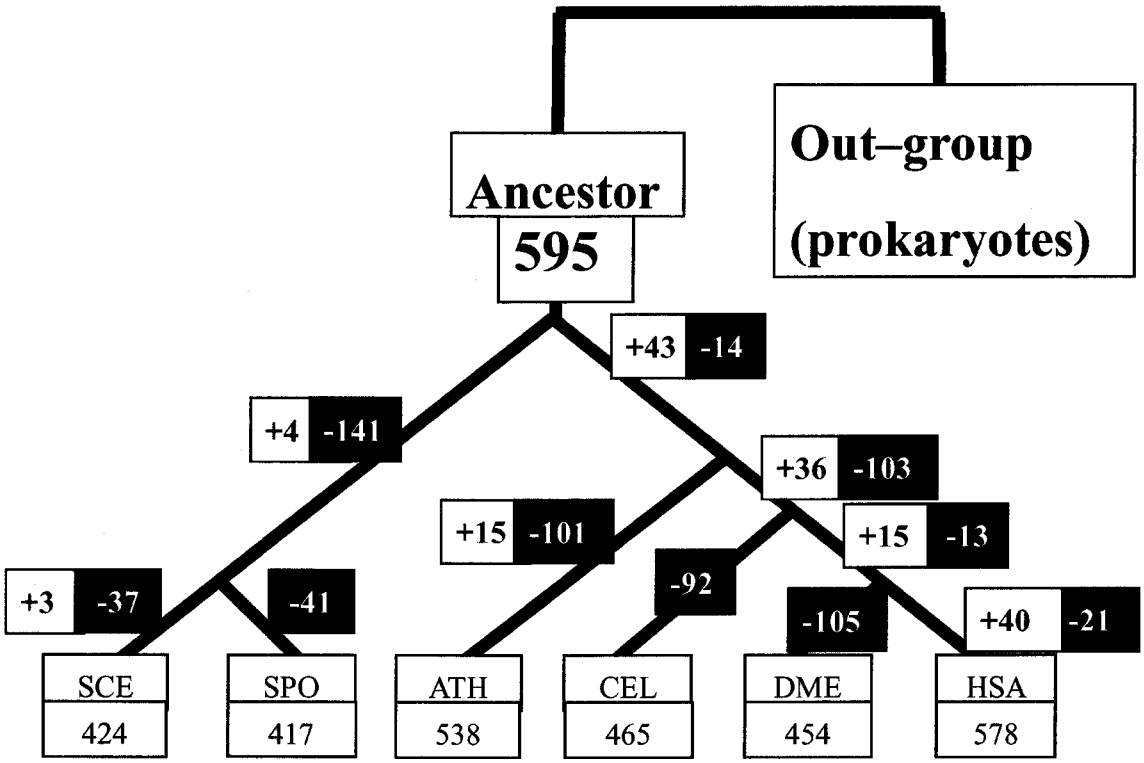


Figure 2-6. Effect to the numbers of ERs in the common ancestor or gain and loss of ERs in each branch by the topology of the phylogenetic tree

In this figure, the topology between a plant and yeasts were different from Figure 4-1. The numbers below the 6 species and the ancestor were total numbers of ERs existing in each organism. The red number indicated the total number of gains of ERs in the branch and the white number indicated the total number of losses of ERs in the branch.

Chapter 3

Evolution of vitamin B_6 (Pyridoxine) metabolism by gain and loss of genes

3.1 Introduction

In the previous chapter, I examined the gain and loss of ERs in the metabolic networks of the 6 eukaryotes for evaluating their evolutionary process. For this examination, I dealt with only ERs in which the gene products inferred from the complete genome sequence data can possibly work as an enzyme. This is because more than one enzyme is often involved with an ER. However, if I deal with the metabolic network in which I can easily identify the enzymes encoded by particular genes that are involved with the particular ER, I can discuss directly the gain and loss of genes instead of ERs.

In this analysis, I take the Vitamin B_6 (VB6) metabolism as a concrete example because of the basis of the following backgrounds. VB6 functions as a cofactor of many enzymes. In particular, pyridoxal 5'-phosphate (PLP), which is the active form of VB6, has multiple roles as a versatile cofactor of enzymes mainly in the metabolism of amino acid compounds (Grogan 1988; Rottmann *et al.* 1991; Helmreich 1992; Mihara *et al.* 1997; Kack *et al.* 1999). Moreover, VB6 appears to play an important role against photosensitization in fungi (Bilski *et al.* 2000). Most unicellular organisms and plants can biosynthesize PLP by themselves.

Studies of the VB6 metabolism have been conducted in particular in *Escherichia*

coli (White and Dempsey 1970; Lam and Winkler 1990; Zhao *et al.* 1995; Yang *et al.* 1996; Man *et al.* 1996) and fungi such as *Cercospora nicotianae, Neurospora crassa, Aspergillus nidulans*, and *Saccharomyces cerevisiae* (Ehrenshaft *et al.* 1999; Osmani *et al.* 1999; Ehrenshaft and Daub 2001; Bean *et al.* 2001). For the PLP biosynthetic pathways, some of these studies have characterized three pathways and a total of 12 genes involved in the pathways in the bacteria and fungi (Figure 3-1). In the case of *E. coli*, both *de novo* and salvage pathways were discovered. The two pathways include the enzymes encoded by 10 genes in total (Mittenhuber 2001). The corresponding *de novo* pathway has not been discovered in fungi or plants. Instead, fungi were found to have another biosynthesis pathway. This pathway, the fungi-type pathway, has two genes, *SNZ* and *SNO* whose functions are unknown. These reports indicate that the species synthesizing PLP have at least one of the three PLP biosynthetic pathways.

Animals such as insects and mammals have to take VB6 compounds as nourishment. In particular, humans take VB6 from meats and vegetables (Manore 2000). This fact suggests that there is a problem for VB6 metabolism in animal lineages. It is conceivable that their PLP biosynthetic pathways became dysfunctional due to the loss of some of the 12 genes in their evolutionary lineages, on the condition that the ancestor of the animals had PLP biosynthetic pathways. Actually, *pdxA* and *pdxJ*, which are the members of the 10 genes mentioned above, have not been found in animals (Ehrenshaft *et al.* 1999). Moreover, *SNZ* and *SNO* have been reported to be lost in animals except for the marine sponge *Suberites domuncula* (Seack *et al.* 2001). Even though the functions of *SNZ* and *SNO* are unknown, they should have a role indispensable in the biosynthetic pathway in fungi. The loss of these two genes has been considered as the cause of the inability for PLP biosynthesis in animal lineages, in

particular in the Eumetazoa lineage.

To study the evolution of the VB6 metabolism, I paid my attention to the three pathways for PLP biosynthesis and the 12 genes involved in these pathways. I was particularly interested to know when individual genes of them were gained or lost in evolution.

In this study, to elucidate the evolutionary processes of VB6 metabolism, I focused on the gain and loss of the 12 genes in 122 species in the three domains of life, eubacteria, archaebacteria and eukaryotes (Woese *et al.* 1990). I estimated a particular gene set of the common ancestor of the 122 species on the basis of their genealogical relationships. Then, I identified the gain and loss events for the genes by comparing the gene sets between the ancestral species and the extant one. On the basis of the results obtained, I will report evolutionary features of the formation and dysfunction processes of VB6 metabolism in the view of gains and losses of the genes.

3.2 Materials and methods

3.2.1 Genes related to the VB6 metabolism

There are 12 genes involved in the PLP biosynthetic pathways of VB6 metabolism; *gapA*, *epd*, *dxs*, *pdxA*, *pdxB*, *pdxF*, *pdxH*, *pdxJ*, *pdxK*, *pdxY*, *SNZ*, and *SNO* (Mittenhuber 2001). *GapA* is paralogous to *epd*, and so is *pdxK* and *pdxY*, and each pair's functions was also similar in the PLP biosynthetic pathways (Yang *et al.* 1998a; Yang *et al.* 1998b). From this reason, I regarded *epd* as *gapA*, and *pdxY* as *pdxK*. Thus, I studied the total of 10 genes instead of 12 genes for the VB6 metabolism.

In order to identify the genes in the VB6 metabolism in the complete genome sequences of 122 species (9 eukaryotes, 16 archaebacteria, and 97 eubacteria) (Table 3-1), BLAST homology searches for the 10 genes were performed against them (Altschul *et al.* 1990). As query sequences, I used the protein sequences of *gapA* (the accession number of b1779), *dxs* (b0420), *pdxA* (b0052), *pdxB* (b2320), *pdxF* (b0907), *pdxH* (b1638), *pdxJ* (b2564), and *pdxK* (b2418) that were derived from *E. coli* K-12 MG1655, and *SNZ* (NP_013814) and *SNO* (NP_013813) that were derived from *S. cerevisiae*. If I could not find any of the 10 genes by the homology search against the complete sequences of a species, I assumed that the species did not have it. Thereafter, I compared my results with the KEGG Orthology (KO) dataset in KEGG database (Bono *et al.* 1998; Kanehisa *et al.* 2002), in order to make my results more reliable. The KO dataset contains orthologous gene families that are categorized on the basis of experimental information, sequence homology, and gene order in the genome. If a gene was not related to the VB6 metabolism, I removed it from my dataset.

3.2.2 Phylogenetic tree

I used phylogenetic trees that were revised based on those of Baldauf *et al.* (2000) and Nelson *et al.* (2000) to estimate the gene sets which were involved with VB6 metabolism of the ancestor of the 122 species (See the next section). Because some of eubacterial and archaebacterial species examined were not included in those trees, I estimated the genealogical relationships by constructing a phylogenetic tree in the subdivision of eubacteria and archaebacteria. To do so I applied their 16s rRNA sequences to the CLUSTALW program with 1,000 bootstrap trials (Thompson *et al.* 1994). I excluded the positions with gaps and corrected for multiple substitutions.

3.2.3 Estimation of the gene set of the VB6 metabolism in the ancestor

I assumed that a single gene was acquired only once during evolution of the 122 species examined in this model, ignoring horizontal gene transfer and parallel evolution. I show the method for estimating the set of genes in the ancestor. This method is exactly the same as the method used in the previous chapter except that a gene instead of an ER is discussed in this study. In Figure 3-2, I used the following two states for representing whether a species had a particular gene (Gene-A): the state "1" is used when the species had at least one homologous gene with Gene-A, and "0" is used when the species had no homologous gene with Gene-A.

If the states of two species are the same, namely (0,0) or (1,1), I assume that the ancestor has the same state (Patterns 1 and 2 in Figure 3-2). If the state is different between the two closest species compared, namely (0,1) or (1,0), I use the states for all of the species that were located in the outside of the two comparing species in the tree and out-group species. Let me denote a group of those species by species group C. If

at least one species in Group C has the state of "1", I regard the ancestral state as "1" (Pattern 3 in Figure 3-2). If all species of group C have the state of "0", I regard the ancestral state as "0" (Pattern 4 in Figure 3-2). In this way, I estimated the states for a given gene at all internal-nodes.

3.2.4 The order of the losses among the genes

Using the result of the estimation of gene loss, I investigated the order of the losses among the 10 genes. For any pair of the 10 genes, I examined which was lost earlier during evolution of the 122 species. I statistically tested the frequency of the order of gene loss by the binominal distribution. If the losses of two genes occurred on the same branch in the phylogenetic tree, I did not count them because I did not know which gene lost earlier than the other.

3.3 Results

3.3.1 Comparison of the gene sets among species

For the VB6 metabolism, the three PLP biosynthetic pathways have been reported as mentioned earlier; the *de novo* pathway and the salvage pathway in *E. coli* (the bacterial-type) and the fungi-type pathway. When I examined whether the 10 genes existed in the 122 species whose genome sequences were completely or almost completely determined, I found that no species had all genes (Table 3-1). Moreover, 7 genes, *gapA*, *dxs*, *pdxA*, *pdxB*, *pdxF*, *pdxH*, and *pdxJ*, for the *de novo* pathway were found only in 17 eubacteria. These bacterial species were all gamma-proteobacteria, indicating the the *de novo* pathway may be functioning only in gamma-proteobacteria.

3.3.2 Distribution of the genes in the three domains of life

Based on the gene sets of the 122 species, I compared the gene sets among the three domains of life (Figure 3-3). Since *SNZ*, *SNO* and *pdxF* were observed in all domains, these genes were considered to exist before the divergence into the three domains. This result indicates that the fungi-type pathway composed by *SNZ* and *SNO* was formed before the divergence into the three domains. Other 4 genes, *gapA, dxs, pdxH* and *pdxK,* in the bacterial-type pathways were discovered in eukaryotes and eubacteria. This result means that part of the bacterial-type pathways was found in eukaryotes and eubacteria. Based on my assumption that a single gene was acquired only once during evolution of the 122 species examined, I interpret this result as indicating that the salvage pathway composed by *pdxH* and *pdxK* was formed before the divergence into the three domains. I also note that *PdxA*, *pdxB*, and *pdxJ* were

observed only in eubacteria. When I focused on the eubacteria group, I found that both *pdxA* and *pdxJ* were observed not only in Proteobacteria but also Firmicutes, Cyanobacteria, Chlorobi and Aquificae. On the other hand, *pdxB* existed only in gamma-proteobacteria. Therefore, I consider that both *pdxA* and *pdxJ* were generated in eubacterial lineage after the divergence from the other two domains and that *pdxB* was generated in the gamma-proteobacterial lineage after the divergence from the other lineages. These three genes are thus considered to contribute to the formation of the *de novo* pathway in the eubacterial lineage. In particular, *pdxB* may be the most important gene for the completion of this pathway because it is considered that this gene was generated only in gamma-proteobacteria.

3.3.3 Estimation of the losses of genes

Because the gene sets of the VB6 metabolism were different among the 122 species examined, it was considered that the loss as well as gain of genes had occurred in the 10 genes. I thus examined how many losses of the 10 genes had occurred in the evolution of the 122 species. To estimate the occurrence of the gain or loss of a particular gene in each evolutionary lineage from the common ancestor to the extant species, I needed to estimate whether the gene had existed in the common ancestor of the 122 species. Using the phylogenetic tree in Figure 3-4, I estimated the ancestral gene set of each node. As a result, I found that the total of 133 losses of genes had occurred during the evolution from the common ancestor of the 122 species (Table 3-2). The numbers of losses of *gapA*, *pdxB*, and *pdxJ* were 3, 3 and 8, respectively. These numbers were all smaller than those of losses of the other seven genes. In the case of *gapA*, a less number of gene losses may be explained by the following functional

constraint: *GapA* functions not only in VB6 metabolism but also in glycolysis (Seta *et al.* 1997), and therefore, the functional constraint of *gapA* is not expected to be lost. Since I estimated that *pdxB* had emerged in the gamma-proteobacterial lineage, I did not need to count the number of losses of the genes before the divergence of this lineage from the other lineages. Therefore, it was considered that the total numbers of losses of this gene were smaller than those of the other genes. In the case of *pdxJ*, I estimated that this gene emerged in eubacterial lineage and was lost in an early period in the particular lineages of the 97 eubacterial species. The total number of loss of *pdxJ* was smaller than those of the other genes, because this gene did not exist for a long time in the particular lineages.

3.3.4 Losses of *SNO* and *SNZ* in Animal lineages

The genes of *SNO* and *SNZ* were reported only in the sponge *S. domuncula* among Metazoa. By homology search against the genome sequences of six animals (*Homo sapiens, Mus musculs, Rattus nurvegicus, Ciona intestinalis, Caenorhabditis elegans* and *Drosophila melanogaster*), I found that these two genes exist in *C. intestinalis*. In addition, the positions of the two genes on the genome of *C. intestinalis*, in which they aligned head to head, were also the same as those in *S. domuncula*, *S. cerevisiae* and *S. pombe* (Table 3-3). In the case of eubacteria and archaebacteria, *SNO* was next to *SNZ* on the genomes of the 33 species studied. However, the two genes were aligned head to tail on all the 33 genomes. Therefore, it is reasonable that *SNO* and *SNZ* in *C. intestinalis* had not been transferred from bacterial species. From these results, I conclude that *SNO* and *SNZ* had existed in the Eumetazoan lineage, and the losses of *SNO* and *SNZ* occurred at least three times

independently in *C. elegans*, *D. melanogaster*, and vertebrate lineages (Figure 3-4).

3.3.5 Correlation of losses of genes among PLP biosynthetic pathways

I have found that the loss of *SNZ* occurred together with that of *SNO* in the same branch more often than other pairs of genes (Table 3-4). As I mentioned above, the two genes were often present side by side on the genome. Moreover, the two genes function only in the fungi-type pathway. These observations support that the losses of the two genes occurred at the same time.

Next, I examined the order of losses for the 10 genes. When I compared the order of losses between two genes, a significant bias was observed in particular in 9 combinations of genes; *pdxH* amd *dxs*, *pdxJ* and *pdxH*, *pdxH* and *pdxF*, *pdxA* and *dxs*, *pdxJ* and *pdxA*, *pdxJ* and *dxs*, *pdxJ* and *pdxK*, *pdxJ* and *SNO*, *pdxK* and *pdxF* (Table 3-4). In every pair of the 9 combinations, I observed that the loss of the latter gene occurred more frequently after the loss of the former gene. These biases were statistically significant against the binominal distribution ($P < 0.05$). From this result, I depicted the patterns of loss of the 5 genes in Figure 3-5. The loss of *pdxJ* caused the loss of at least one of *pdxA*, *pdxH* and *pdxK* as shown in the figure. These four genes encode the enzymes whose reactions are connected through the substance called pyridoxine 5'-phosphate (PNP). Moreover, the losses of these four genes caused the loss of at least one of *pdxF* and *dxs*. These two genes encode the enzymes whose reactions are connected to *pdxA* and *pdxJ* through 4-phosphohydroxy-L-threonine (4PHT) or 1-deoxyxyluiose 5-phosphate (DXP), respectively. If I call two genes encoding the enzymes that catalyze two consecutive reactions the neighbor genes, I would thus conclude that the loss of a gene accelerates the loss of the neighbor gene.

3.4 Discussion

To my knowledge, I have made the first attempt to elucidate the evolution of VB6 metabolism focusing on the gain and loss of the 10 genes involved in the PLP biosynthetic pathways. On the assumption that one of the genes was acquired only once during the evolution of the whole 122 species, I found that every gene in the VB6 metabolism had been lost more than once in the evolutionary lineages of the 122 species. This result suggests that the breakdown of the VB6 metabolism by gene loss might have occurred in many lineages, which should be examined by experiments.

I have also revealed three aspects in the evolution of VB6 metabolism by estimating the gain and loss of the 10 genes. One is related to the evolutionary order of the generations of three PLP biosynthetic pathways. From the distribution of the 10 genes over the 122 species examined, I found that the fungi-type and the salvage pathway were possibly older than the *de novo* pathway on the basis of the following two results. First, *pdxK* and *pdxH* in the salvage pathway exist only in eubacteria and eukaryotes, while *SNO* and *SNZ* in the fungi-type pathway are found in the three domains, eubacteria, archeabacteria, and eukaryotes. Therefore, I consider that both the fungi-type and the salvage pathways have existed before the separation of three domains of life. Second, I have found that *pdxB* in the *de novo* pathway exists only in gamma-proteobacteria indicating that this gene was generated in gamma-proteobacteria after the divergences of the three domains of life.

Applying the second result mentioned above to the existing model for explaining the evolution of the metabolic networks, the patchwork model and the *de novo* invention model (Jensen 1976; Schmidt *et al.* 2003), I have proposed the process of the

formation of the *de novo* pathway in gamma-proteobacteria as follows (Figure 3-6). Originally, the common ancestor of the 97 bacterial species studied had a part of the *de novo* pathway which had 5 genes, *dxs, pdxA, pdxF, pdxH,* and *pdxJ.* Because *gapA* functioned not only in pyridoxine biosynthesis but also in glycolysis (Seta *et al.* 1997), I think that this gene had also existed in the common ancestor. As shown Figure 3-6, when *pdxB* was generated in the lineage of gamma-proteobacteria, the reaction catalyzed by the product of *pdxB* was connected to the two enzymatic reactions that were respectively catalyzed by the products of *gapA* and *pdxF.* As a result, the *de novo* pathway was completed by the 7 genes, *dxs, gapA, pdxA,pdxB, pdxF, pdxH,* and *pdxJ* in gamma-proteobacteria. I have reached the same conclusion as Mittenhuber's (Mittenhuber 2001) but in a different way. He has explained the reason why the *de novo* pathway was largely restricted to gamma-proteobacteria on the basis of the functions of the two genes, *pdxA* and *pdxJ*, and the requirement of VB6 in the *de novo* pathway.

However, I could not answer which of the fungi-type and salvage pathways was established earlier as the PLP biosynthetic pathway, because I could not estimate the times when *pdxK*, *pdxH*, *SNO* and *SNZ* were originated in this study. In other words, I could not determine which gene sets of the two pathways have been generated earlier.

The second aspect is related to the losses of *SNO* and *SNZ* in animal lineages. In animals, the two genes have been discovered only in the marine sponge, *S. domuncula* (Seack *et al.* 2001). Therefore, it is plausible that the losses of *SNZ* and *SNO* occurred only once in the Eumetazoa lineage after its divergence from the Porifera lineage. I have found that there were the two genes in the complete genome sequence of *C. intestinalis* in the present study. This species is more closely related with mammals

than *D. melanogaster* and *C. elegans* both of which do not have the two genes (Figure 3-4). Therefore, I consider that *SNZ* and *SNO* had existed in the animals after the divergence between invertebrate and vertebrate, and the losses of them occurred independently at least three times in the animal lineage, as shown in Figure 3-4. I reject the possibility of horizontal gene transfer from the bacterial lineage to *C. intestinalis* not only by the homology of the two genes but also by the order and the direction of the two genes on the genome (Table 3-3).

The third aspect is related to the evolutionary order of the gene loss. As shown in Figure 3-5, the losses of the 5 genes occurred in that order in the figure during the evolution of the VB6 metabolism. Historically, five models have been proposed to explain the formation of metabolic pathways; the retrograde model, the patchwork model, *de novo* invention, specialization of a multifunctional enzyme and pathway duplication (Horowitz 1945; Jensen 1976; Schmidt *et al.* 2003). However, these models consider gene gain only. Since there are other reports that gene loss also have often occurred in the metabolic pathways in bacterial lineages (Tatusov *et al.* 1996; Shigenobu *et al.* 2000), it is not sufficient to consider the gain of genes only for the evolution of metabolic pathways in the prokaryotic lineages.

Therefore, I propose a new model based on my results that explain the evolution of metabolic pathways by gene loss. Once the loss of a gene occurred in a metabolic pathway, the neighboring gene is more easily lost than the other genes in the pathway. This could be explained by functional constraint. The breakdown of a metabolic pathway by gene loss will decrease the functional constraints of the other genes of the pathway. My model tells that the functional constraint decreases more extensively in the proximal genes to the lost gene than in the distant ones. Of course, it is possible

that the functional constraint is affected by others. For example, if a gene is involved also in another metabolic pathway such as *gapA*, its functional constraint can not be changed.

My approach for the estimation of the gain and loss of genes is affected at least by two points. First, in this study, I did not take into account horizontal gene transfer or parallel evolution. In other words, I have considered that gene gain occurred only once, while gene loss could have occurred more than once in the evolution of the 122 species examined. Therefore, if there is a difference in a gene set among closely related species, the number of gene losses is expected to be larger than that of gene gains. If horizontal gene transfer and parallel evolution occurred, gene loss would decrease whereas gene gain would increase. When I have the evidence for horizontal gene transfer and parallel evolution of the 10 genes in this study, it is possible to estimate more accurately the time of gain and loss of the genes. Second, my results are affected by the topology of a phylogenetic tree. If the topology is changed, the estimation of the evolutionary times of the gain and loss of genes are accordingly changed. As a result, it is possible to miscount the total number of gains or losses of genes. However, my results showed that the sets of genes are different between the 122 species examined (Table 3-1). Since the gain and loss of genes cause of the differentiation of the set of genes in the species, my conclusion that the evolutionary process of VB6 metabolism had been quite dynamic in the events of gain and loss of genes, under some constraints, is not fluctuated, even though the topology of the phylogenetic tree changes.

Reported studies using the comparative analysis have often shown the difference in a gene set involved in the metabolic pathways among species (Huynen *et al.* 1999). By the estimation of the gain and loss of genes, I would be able to know not only the

difference in a metabolic pathway among species examined but also the answer to the following question as to in which lineage the change in a metabolic pathway occurred in evolution. This means that I will understand the evolutionary processes of the metabolic networks by gains and losses of genes. In some of the metabolic pathways, dysfunctions in particular lineages are reported (Smirnoff 2001; Meganathan 2001). By applying my approach to these metabolic patnways, I will be able to elucidate the dysfunctions of the pathways by the gain and loss of genes. It is also possible that I further extend my approach to the other metabolic networks in the KEGG (Kanehisa *et al.* 2002) and EcoCyc (Karp *et al.* 2002) database, to understand more clearly the evolutionary process of the metabolic pathways included in them.

Table 3-1. 10 Gene families related to the vitamin B6 metabolism in the 122 species

Symbol	Species name	gapA	pdxB	pdxF	dxs	pdxA	pdxJ	pdxK	pdxH	SNZ	SNO
ECJ	Escherichia coli K-12 W3110	JW1413 JW1414 JW1768 *JW2894	JW2317	JW0890	JW0410	JW0051	JW2548	JW1628 JW2411	JW1630		
ECO	Escherichia coli K-12 MG1655	b1416 b1417 b1779 *b2927	b2320	b0907	b0420	b0052	b2564	b1636 b2418	b1638		
SFL	Shigella flexneri 301 (serotype 2a)	SF1444 SF1795 SF1796 *SF2912	SF2396	SF0902	SF0357	SF0049	SF2626	SF1661 SF2473	SF1663		
ECE	Escherichia coli O157:H7 EDL933Bacteria	Z2304 Z2818 *Z4266	Z3582	Z1253	Z0523	Z0061	Z3845	Z2648 Z3684	Z2652		
ECS	Escherichia coli O157:H7 Sakai	ECs2022 ECs2488 *ECs3798	ECs3204	ECs0990	ECs0474	ECs0057	ECs3430	ECs2345 ECs3290	ECs2347		
ECC	Escherichia coli CFT073	c1843 c2184 *c3505	c2865	c1045	c0531	c0065 c0764	c3088	c2028 c2953	c2030		
STM	Salmonella typhimurium LT2	STM1290	STM2370	STM0977	STM0422	STM0091 STM0163	STM2578	STM1450 STM2435	STM1448		
STY	Salmonella typhi	STY1825	STY2601	STY0977	STY0461	STY0106 STY0185	STY2824	STY1672 STY2672	STY1674		
STT	Salmonella typhi Ty2	t1169	t0494	t1957	t2441	t0094 t0168	t0279	t1318	t1316		
YPE	Yersinia pestis CO92	YPO2157	YPO2763	YPO1389	YPO3177	YPO0493	YPO2930	YPO2368	YPO2370		
YPK	Yersinia pestis KIM	y2165	y1597	y2784	y1008	y3682		y1967	y1965		
BUC	Buchnera sp. APS	BU298		BU312	BU464						
BAS	Buchnera aphidicola Sg	BUsg287		BUsg302	BUsg448						
BAB	Buchnera aphidicola Bp	bbp276		bbp289							
WBR	Wigglesworthia brevipalpis	Wbr0018	Wbr0458	Wbr0271	Wbr0633	Wbr0118	Wbr0578		Wbr0428		
VPA	Vibrio parahaemolyticus	VP2157 VP2970 *VP2601	VP2193	VP1247	VP0686	VP0337	VP2569	VPA1632	VPA1730		
VVU	Vibrio vulnificus CMCP6	VV11141 VV13140 *VV11539	VV11988	VV12813	VV10315	VV10662	VV11568	VV21237	VV21122		

Table 3-1. (Continue)

Symbol	Species name	gapA	pdxB	pdxF	dxs	pdxA	pdxJ	pdxK	pdxH	SNZ	SNO
VCH	Vibrio cholerae El Tor N16961	VC1069 VC2000 *VC0476	VC2108	VC1159	VC0889	VC0444	VC2458		VCA1079		
SON	Shewanella oneidensis MR-1	SO0538 SO2345 *SO0931	SO3071	SO2410	SO1525	SO3638	SO1351		SO2895		
HIN	Haemophilus influenzae Rd KW20Bacteria	HI0001		HI1167	HI1439			HI0405	HI0863	HI1647	HI1648
PMU	Pasteurella multocida PM70	PM0924		PM0837	PM0532	PM1650		PM0290		PM1232	PM1233
PAE	Pseudomonas aeruginosa PA01	PA3195 *PA0551	PA1375	PA3167	PA4044	PA0593 PA2212	PA0773	PA5516	PA1049		
PPU	Pseudomonas putida KT2440	PP1009 *PP4964	PP2117	PP1768	PP0527	PP0402	PP1436	PP5357	PP1129		
XAC	Xanthomonas axonopodis	XAC3352		XAC1648	XAC2565	XAC0864	XAC0012	XAC1524	XAC3009		
XCC	Xanthomonas campestris pv.	XCC3192		XCC1589	XCC2434	XCC0792	XCC0012	XCC1476	XCC2840		
XFA	Xylella fastidiosa 9a5c	XF0457		XF2326	XF2249	XF0839	XF0060		XF1337		
XFT	Xylella fastidiosa Temecula1	PD1626		PD1358	PD1293	PD1834	PD0040		PD0583		
NME	Neisseria meningitidis MC58	NMB0207 NMB2159		NMB1640	NMB1867	NMB0195	NMB0448		NMB1360		
NMA	Neisseria meningitidis Z2491	NMA0062 NMA0246		NMA1894	NMA0589	NMA0072	NMA2037		NMA1572		
RSO	Ralstonia solanacearum GMI1000	RS00105		RS04512	RS01378	RS04975	RS04138		RS01771 RS05086		
ATU	Agrobacterium tumefaciens C58 (Uwash/Dupont)	Atu3737		Atu3707	Atu0745	Atu1104 Atu5073	Atu2024	Atu2487	Atu0760		
ATC	Agrobacterium tumefaciens C58 (Cereon)	AGR_L_2195		AGR_L_2260	AGR_C_1351	AGR_C_2043 AGR_pAT_104	AGR_C_3668	AGR_C_4518	AGR_C_1381		
SME	Sinorhizobium meliloti 1021	SMc03979		SMc00640	SMc00972	SMb20146 SMb20772 SMc00580	SMc01407	SMc04084	SMc00069		
BME	Brucella melitensis 16M	BMEI0310		BMEI0347	BMEI1498	BMEI1266	BMEI0621	BMEI0221	BMEI1517		
BMS	Brucella suis 1330	BR1728		BR1687	BR0436	BR0683	BR1385	BR1830	BR0416		
MLO	Mesorhizobium loti MAFF303099	mlr3750		mll3876	mlr7474	mll7861	mll1418	mlr4132	mll7454		
BJA	Bradyrhizobium japonicum	bll1523		bll7402	bll2651	bll4103 blr3887 blr4374	bll5064	blr4233	bll2624		
CCR	Caulobacter crescentus CB15	CC3248		CC3216	CC2068	CC1686	CC1557	CC0256	CC0918		
RCO	Rickettsia conorii Malish 7										
RPR	Rickettsia prowazekii Madrid E										

Table 3-1. (Continue)

Symbol	Species name	gapA	pdxB	pdxF	dxs	pdxA	pdxJ	pdxK	pdxH	SNZ	SNO
HPJ	Helicobacter pylori J99	jhp0855			jhp0328	jhp1490	jhp1489				
		jhp1265									
HPY	Helicobacter pylori 26695	HP0921			HP0354	HP1583	HP1582				
		HP1346									
CJE	Campylobacter jejuni NCTC11168	Cj1403c		Cj0326	Cj0321	Cj1239	Cj1238				
CPJ	Chlamydophila pneumoniae J138	gapA			tktB_2						
CPA	Chlamydophila pneumoniae AR39	CP0123			CP0790						
CPN	Chlamydophila pneumoniae	CPn0624			CPn1060						
CMU	Chlamydia muridarum	TC0792			TC0608						
CTR	Chlamydia trachomatis serovar D	CT505			CT331						
ANA	Anabaena sp. PCC 7120	all2566		all1683	alr0599	alr4755	all0218		all1248		
		all5062									
		alr1095									
TEL	Thermosynechococcus elongatus	tll0043			tll0623	tll2045	tlr0838		tll0331		
		tll1466									
SYN	Synechocystis sp. PCC 6803	slr0884			sll1945	sll0660	slr1779		sll1440		
CTE	Chlorobium tepidum TLS	CT1480			CT0337	CT0591	CT0331				
BBU	Borrelia burgdorferi B31	BB0057						BB0768			
TPA	Treponema pallidum Nichols	TP0844			TP0824						
LIL	Leptospira interrogans 56601	LA1704		LA0366	LA3285	LA0800	LA1514		LA1166		
SPM	Streptococcus pyogenes MGAS8232	spyM18_0261									
SPY	Streptococcus pyogenes SF370	SPy0274									
SPG	Streptococcus pyogenes SGAS315	SpyM3_0201									
SPS	Streptococcus pyogenes SSI-1	SPs0207									
SAG	Streptococcus agalactiae 2603	SAG1768								SAV6830	SAV6831
SAN	Streptococcus agalactiae NEM316	gbs1811		gbs1621							
SPR	Streptococcus pneumoniae R6	spr1825								spr1322	spr1321
SPN	Streptococcus pneumoniae TIGR4	SP2012								SP1468	SP1467
SMU	Streptococcus mutans UA159	SMU.360		SMU.1656							
LLA	Lactococcus lactis subsp. lactis	L0004		L0083	L108911						
		L0005			L123365						
LPL	Lactobacillus plantarum WCFS1	lp_0789		lp_0204	lp_2610			lp_0863			
SAM	Staphylococcus aureus MW2	MW0734								MW0474	MW0475
		MW1630									
SAU	Staphylococcus aureus N315,	SA0727								SA0477	SA0478
		SA1510									
SAV	Staphylococcus aureus Mu50,	SAV0772								SAV0519	SAV0520
		SAV1687									
SEP	Staphylococcus epidermidis	SE0557								SE2262	SE2261
		SE1361									

Table 3-1. (Continue)

Symbol	Species name	gapA	pdxB	pdxF	dxs	pdxA	pdxJ	pdxK	pdxH	SNZ	SNO
BSU	Bacillus subtilis 168	BG10827 BG12592		BG12673	BG11715					BG10075	BG10076
BHA	Bacillus halodurans C-125	BH3149 BH3560		BH1188	BH2779	BH0804				BH0022	BH0023
LIN	Listeria innocua CLIP 11262	lin2553		lin2957	lin1402					lin2205	lin2206
LMO	Listeria monocytogenes EGD-e	lmo2459		lmo2825	lmo1365					lmo2101	lmo2102
OIH	Oceanobacillus iheyensis HTE831	OB2160 OB2438				OB1013				OB2687	OB2686
CAC	Clostridium acetobutylicum	CAC0709			CAC2077 CAC0106			CAC1622		CAC0594	CAC0595
CTC	Clostridium tetani E88	CTC00378			CTC01575			CTC00497			
CPE	Clostridium perfringens 13	CPE1304			CPE1819			CPE2318			
MPN	Mycoplasma pneumoniae M129	A05_orf337									
MGE	Mycoplasma genitalium G-37	MG301									
UUR	Ureaplasma urealyticum serovar 3										
MPE	Mycoplasma penetrans HF-2	MYPE8170			MYPE730						
MPU	Mycoplasma pulmonis UAB CTIP	MYPU_0460									
TTE	Thermoanaerobacter	TTE1762			TTE1298					TTE0823	TTE0822
MTC	Mycobacterium tuberculosis CDC1551	MT1480		MT0907	MT2756				MT2682	MT2681	MT2679
MTU	Mycobacterium tuberculosis H37Rv	Rv1436		Rv0884c	Rv2682c Rv3379c				Rv2607	Rv2606c	rV2604c
MLE	Mycobacterium leprae TN	ML0570		ML2136	ML1038				ML2131	ML0450	ML0474
CEF	Corynebacterium efficiens YS-314	CE1706		CE0903	CE1796			CE0975		CE1779	
CGL	Corynebacterium glutamicum	NCgl0900 NCgl1526		NCgl0794	NCgl1827			NCgl0876		Cgl0788	Cgl0789
SCO	Streptomyces coelicolor A3(2)	SCO1947 SCO7511		SCO4366	SCO6768				SCO4387	SCO1523	SCO1522
TWH	Tropheryma whipplei Twist	TW300			TW484					TW264	TW265
TWS	Tropheryma whipplei TW08/27	TW472			TW280					TW506	TW505
BLO	Bifidobacterium longum NCC2705Bacteria	BL1363		BL1660	BL1132			BL0934		BL1146	BL1145
FNU	Fusobacterium nucleatum	FN0652			FN1208 FN1464	FN0226				FN1463	
DRA	Deinococcus radiodurans R1	DR1343			DR1475			DRA0184	DR0495	DR1367	DR1366
TMA	Thermotoga maritima MSB8	TM0688			TM1770					TM0473	TM0472
AAE	Aquifex aeolicus VF5	aq_1065			aq_881	aq_852	aq_1423				
SSO	Sulfolobus solfataricus									SSO0570	SSO0571
STO	Sulfolobus tokodaii strain7									ST1441	ST1442
APE	Aeropyrum pernix K1									APE0246	APE0244
PAI	Pyrobaculum aerophilum IM2									PAE2819	PAE2820

Table 3-1. (Continue)

Symbol	Species name	gapA	pdxB	pdxF	dxs	pdxA	pdxJ	pdxK	pdxH	SNZ	SNO
MMA	Methanosarcina mazei Goe1			MM2911						MM2432	MM2433
MAC	Methanosarcina acetivorans C2A			MA2304						MA1567	MA1566
HAL	Halobacterium sp. NRC-1									VNG1793C	VNG2598G
TVO	Thermoplasma volcanium GSS1									TVG1050816	TVG0059350
TAC	Thermoplasma acidophilum									Ta0522	Ta0009
MTH	Methanobacterium									MTH666	MTH190
AFU	Archaeoglobus fulgidus VC-16									AF0508	AF0509
MJA	Methanococcus jannaschii									MJ0677	MJ1661
PHO	Pyrococcus horikoshii OT3									PH1355	PH1354
PAB	Pyrococcus abyssi GE5									PAB0537	PAB0538
PFU	Pyrococcus furiosus DSM 3638									PF1529	PF1528
MKA	Methanopyrus kandleri AV19									MK1371	MK1062
HSA	Homo sapiens	G3P2_HUMAN G3PT_HUMAN		SERC_HUMAN				PDXK_HUMAN	NP_060599		
MMU	Mus musuculus	G3P_MOUSE G3PT_MOUSE		NP_932136				NP_742146	NP_598782		
RNO	Rattus novegicus	G3P_RAT NP_076454		XP_227251 XP_215179				NP_113957	NP_072123		
CIN	Ciona intesitinalis	ci0100132109		ci0100142284				ci0100152209	ci0100131648	ci0100145397	ci0100145416
DME	Drosophila melanogaster	CG12055-PA CG8893-PA CG9010-PA		CG11899-PA				CG4446-PA			
CEL	Caenorhabditis elegans	F33H1.2 K10B3.7 K10B3.8 T09F3.3		F26H9.5				F57C9.1a	F57B9.1		
SCE	Saccharomyces cerevisiae	YGR192C YJL052W YJR009C		YOR184W				NP_010885 NP_014424	YBR035C	NP_013814 NP_014066 NP_116596	NP_013813 NP_014065 NP_116595
SPO	Schizosaccharomyces pombe	SPBC354.12		SPAC1F12.07				NP_593904 NP_588389	SPAC1093.02	SPAC29B12.04	SPAC222.08c
ATH	Arabidopsis thaliana	At1g12900 At1g13440 At1g16300 At1g42970 At1g79530 At3g04120 At3g26650		At2g17630 At4g35630	At3g21500 At4g15560 At5g11380			At5g37850	At5g49970	At2g38230 At3g16050 At5g01410	At5g60540

* Genes were named epd in KEGG database.

Table 3-2. Distribution and number of losses of the 10 genes

Gene name	No. of species	No. of gene loss
dxs	76	10
gapA	103	3
pdxA	46	13
pdxB	17	3
pdxF	67	24
pdxH	52	14
pdxJ	42	8
pdxK	45	20
SNO	46	20
SNZ	48	18

Table 3-3. Order between *SNZ* and *SNO* on the genome

Species	SNZ	+-	start	end	SNO	+-	start	end	
BHA	BH0022	+	32914	33810	BH0023	+	33818	34408	*
BLO	BL1146	-	1390876	1389905	BL1145	-	1389819	1389181	*
BSU	BG10075	+	19060	19944	BG10076	+	19966	20556	*
CAC	CAC0594	+	686139	687014	CAC0595	+	687014	687574	*
CEF	CE1779	-	1877157	1876264					
CGL	Cgl0788	+	831021	831974	Cgl0789	+	831971	832573	*
DRA	DR1367	-	1373812	1372889	DR1366	-	1372892	1372302	*
FNU	FN1463	+	2145142	2145984					
HIN	HI1647	+	1712590	1713465	HI1648	+	1713517	1714044	*
LIN	lin2205	+	2226138	2227025	lin2206	+	2227027	2227593	*
LMO	lmo2101	+	2181329	2182216	lmo2102	+	2182218	2182784	*
MLE	ML0450	-	551116	550193	ML0474	+	575469	576140	
MTC	MT2681	-	2930234	2929314	MT2679	-	2928432	2927836	
MTU	Rv2606c	-	2934068	2933169	Rv2604c	-	2932287	3921691	
OIH	OB2687	-	2757867	2756980	OB2686	-	2756929	2756354	*
PMU	PM1232	+	1424316	1425203	PM1233	+	1425206	1425787	*
SAG	SAV6830	+	8154087	8155001	SAV6831	+	8155015	8155620	*
SAM	MW0474	+	548163	549050	MW0475	+	549054	549614	*
SAU	SA0477	+	557598	558485	SA0478	+	558489	559049	*
SAV	SAV0519	+	581847	582734	SAV0520	+	582738	583298	*
SCO	SCO1523	-	1628367	1627456	SCO1522	-	1627443	1626835	*
SEP	SE2262	-	2330773	2329886	SE2261	-	2329883	2329326	*
SPN	SP1468	-	1381373	1380498	SP1467	-	1380497	1379916	*
SPR	spr1322	-	1309261	1308368	spr1321	-	1308367	1307786	*
TMA	TM0473	-	500776	499895	TM0472	-	499891	499891	*
TTE	TTE0823	-	830528	829650	TTE0822	-	829648	829079	*
TWH	TW264	+	348455	349318	TW265	+	349318	349884	*
TWS	TW506	-	544896	544033	TW505	-	544033	543467	*
AFU	AF0508	+	466652	467662	AF0509	+	467659	468255	*
APE	APE0246	-	177978	176965	APE0244	-	176917	176357	*
HAL	VNG1793C	+	1326538	1327446	VNG2598G	-	1943348	1942737	
MAC	MA1567	-	1856725	1855673	MA1566	-	1855406	1854807	*
MJA	MJ0677	-	604189	603197	MJ1661	-	1644079	1643519	
MKA	MK1371	-	1400016	1399129	MK1062	+	1030519	1031127	
MMA	MM2432	+	2901968	2902873	MM2433	+	2903175	2903786	*
MTH	MTH666	-	593940	593059	MTH190	+	134383	134961	
PAB	PAB0537	+	747283	748290	PAB0538	+	748321	748911	*
PAI	PAE2819	+	1666365	1667375	PAE2820	+	1667372	1667986	*
PFU	PF1529	-	1426935	1425928	PF1528	-	1425896	1525303	*
PHO	PH1355	-	1222832	1221825	PH1354	-	1221796	1221206	*
SSO	SSO0570	+	502607	503623	SSO0571	+	503620	504222	*
STO	ST1441	+	1446232	1447242	ST1442	+	1447239	1447841	*
TAC	Ta0522	-	551548	550538	Ta0009	+	8957	9568	
TVO	TVG1050816	-	1051826	1050826	TVG0059350	+	58751	59350	
ATH	At5g01410	-	173504	172575	At5g60540	-	23632473	23630700	
	At2g38230	+	15960019	15960948					
	At3g16050	-	5445065	5444121					
SCE	NP_013814	+	458407	459300	NP_013813	-	457958	457284	*
	NP_14066	+	13267	14163	NP_014065	-	12876	12208	*
	NP_116596	+	11363	12259	NP_116595	-	10301	10969	*
SPO	SPAC29B12.0	+	5376387	5377277	SPAC222.08c	-	968574	967870	
SDO	AJ27952				AJ277953				*
CIN	ci0100145397		40522	37684	ci0100145416		41897	44025	*

+- represented that the gene was on the sense strand or the antisense strand, respectively.

**SNZ* and *SNO* are neighbors on the genome.

Table 3-4. Patterns of gene losses between two genes

1	2	Same branch*	1 only**	2 only**	1 to 2***	2 to 1***
pdxH	pdxA	5	9	8	5	3
pdxH	dxs	1	13	9	8	0
pdxH	pdxJ	5	9	3	0	6
pdxH	pdxB	1	13	2	0	1
pdxH	gapA	0	14	3	1	1
pdxH	pdxK	6	8	14	4	3
pdxH	serC	7	7	17	7	1
pdxH	SNZ	2	12	16	8	4
pdxH	SNO	3	11	17	9	4
pdxA	dxs	1	12	9	7	0
pdxA	pdxJ	5	8	3	0	9
pdxA	pdxB	1	12	2	0	1
pdxA	gapA	0	13	3	1	1
pdxA	pdxK	3	10	17	5	4
pdxA	serC	6	7	18	4	1
pdxA	SNZ	0	13	18	8	3
pdxA	SNO	0	13	20	9	3
dxs	pdxJ	1	9	7	0	8
dxs	pdxB	0	10	3	0	1
dxs	gapA	1	9	2	0	1
dxs	pdxK	1	9	19	0	3
dxs	serC	3	7	21	2	5
dxs	SNZ	0	10	18	6	5
dxs	SNO	0	10	20	6	5
pdxJ	pdxB	2	6	1	0	0
pdxJ	gapA	0	8	3	2	1
pdxJ	pdxK	2	6	18	8	2
pdxJ	serC	5	3	19	6	0
pdxJ	SNZ	0	8	18	8	3
pdxJ	SNO	0	8	20	10	3
pdxB	gapA	0	3	3	0	0
pdxB	pdxK	0	3	20	0	1
pdxB	serC	0	3	24	0	0
pdxB	SNZ	0	3	18	0	1
pdxB	SNO	0	3	20	0	1
gapA	pdxK	0	3	20	3	0
gapA	serC	0	3	24	1	1
gapA	SNZ	2	1	16	0	1
gapA	SNO	2	1	18	0	1
pdxK	serC	5	15	19	9	2
pdxK	SNZ	3	17	15	4	7
pdxK	SNO	4	16	16	4	7
serC	SNZ	2	22	16	3	7
serC	SNO	2	22	18	4	7
SNZ	SNO	18	0	2	0	0

* Same branch indicated that the loss of the two genes had occurred in the same node.

** 1 only indicated that the loss of the gene in the first column had occurred and 2 only indicated that the loss of the gene in the second column

*** 1 to 2 indicated that the loss of gene in the first column had occurred before the loss of gene in the second column, and 2 to 1 indicated that the loss of gene in the second column had occurred before the loss of gene in the first column.

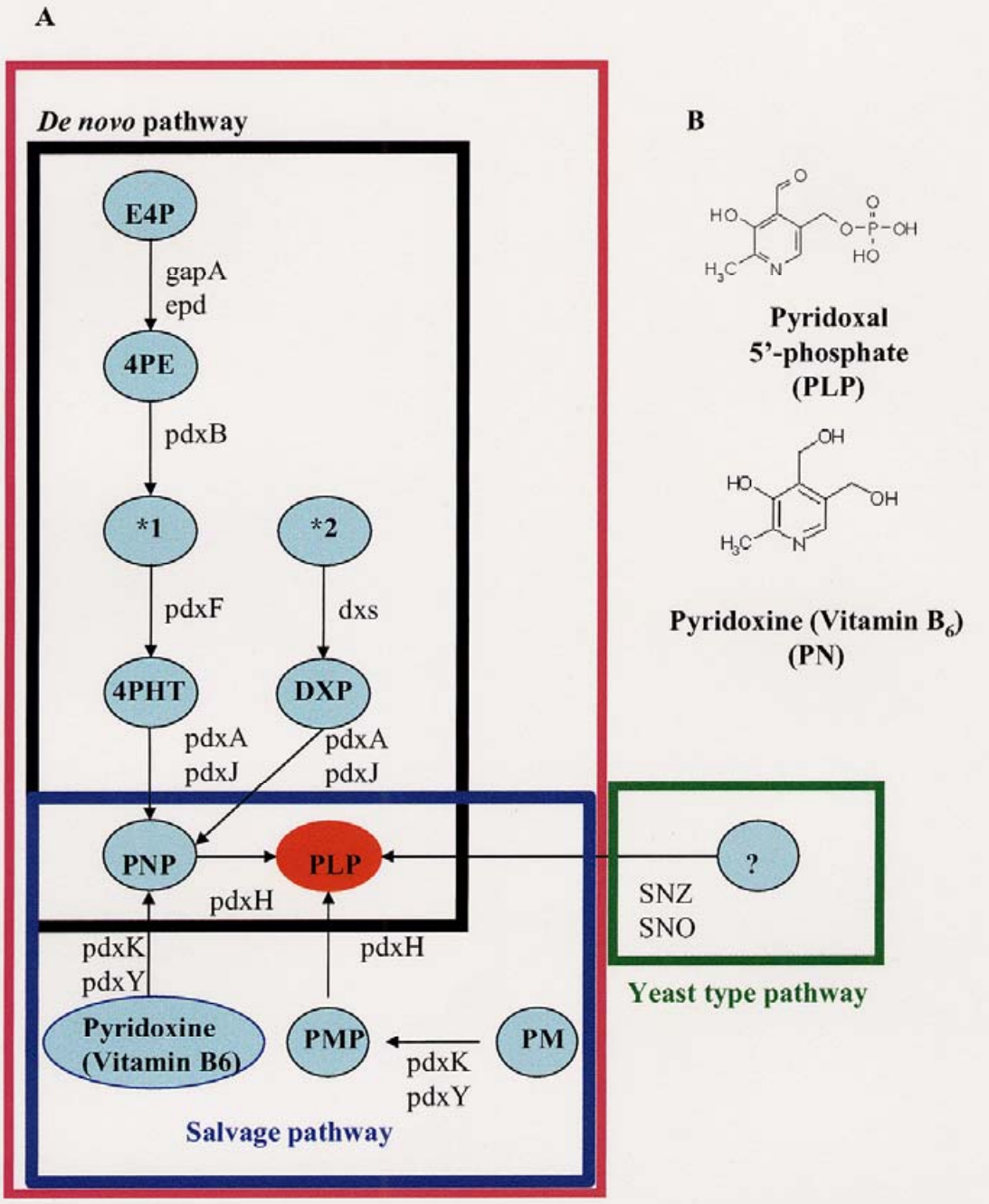


Figure 3-1. VB6 metabolism

A: PNP biosynthetic pathways. Circle indicated the substrates. PLP was an active form of VB6. *1: 3-Hydroxy-4-phophohydroxy-alpha-ketobutyrate *2: Glyceraldehyde-3-phosphate

B: Chemical structure of PLP and vitamin B_6

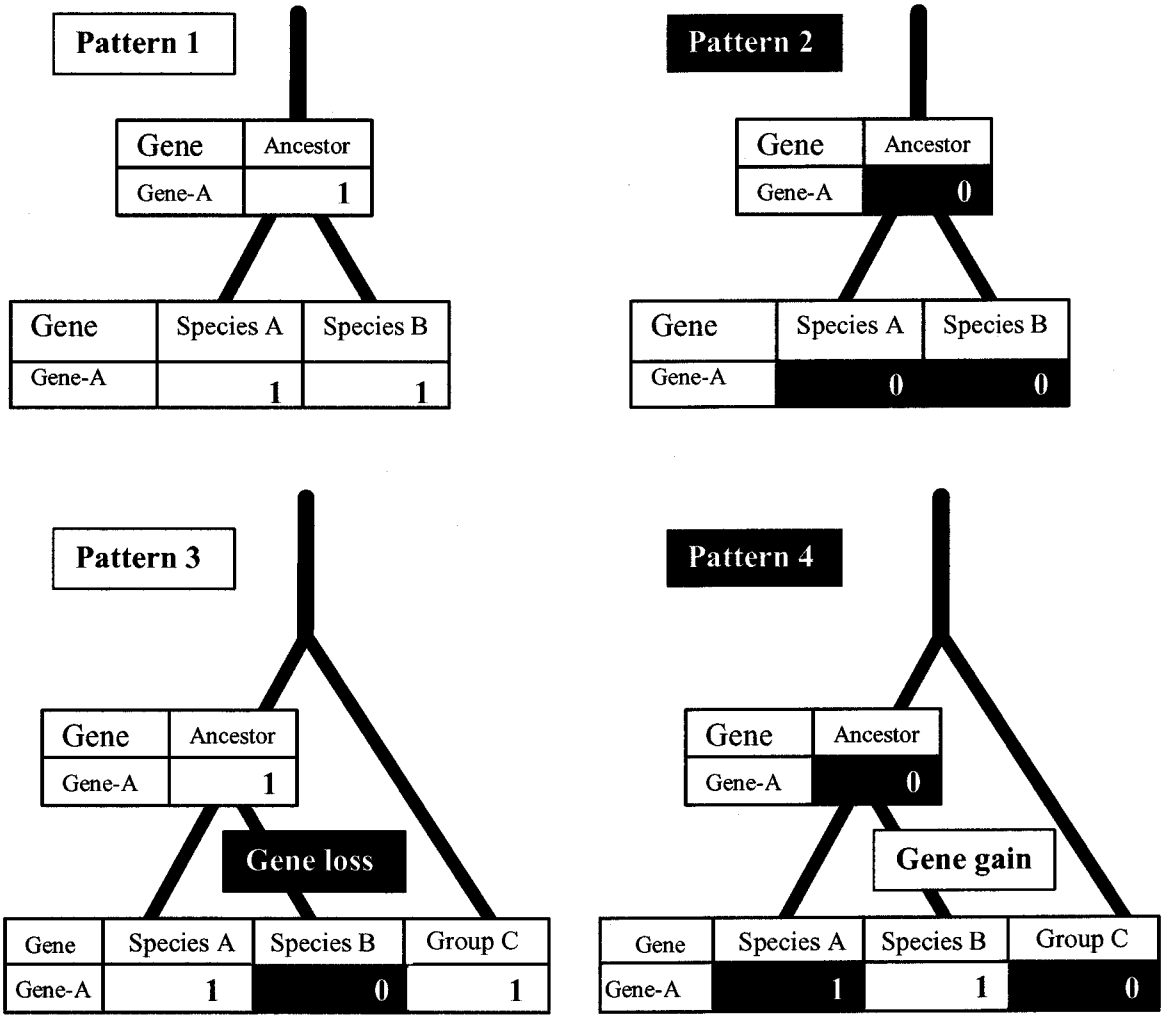


Figure 3-2. Estimation of the ancestral gene set by comparing the gene set in two species

In this figure, I used the two following states for representing whether the species had the Gene-A: the state "1" when the species had at least one gene with the Gene-A, and "0" when the species had no gene with the Gene-A.

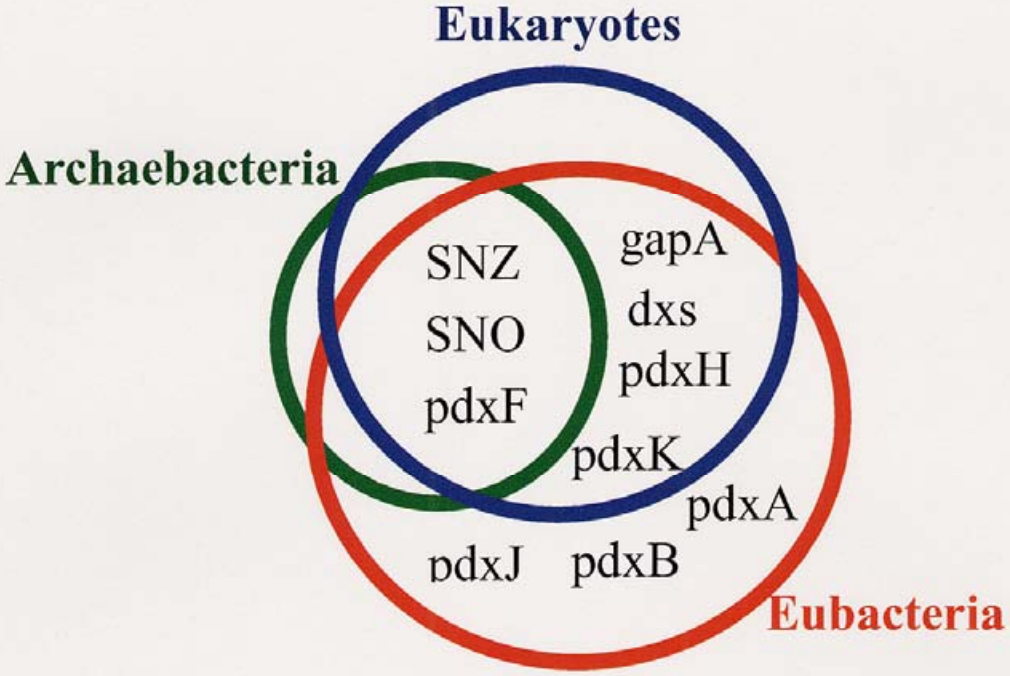


Figure 3-3. Distribution of 10 genes related to VB6 metabolism

A

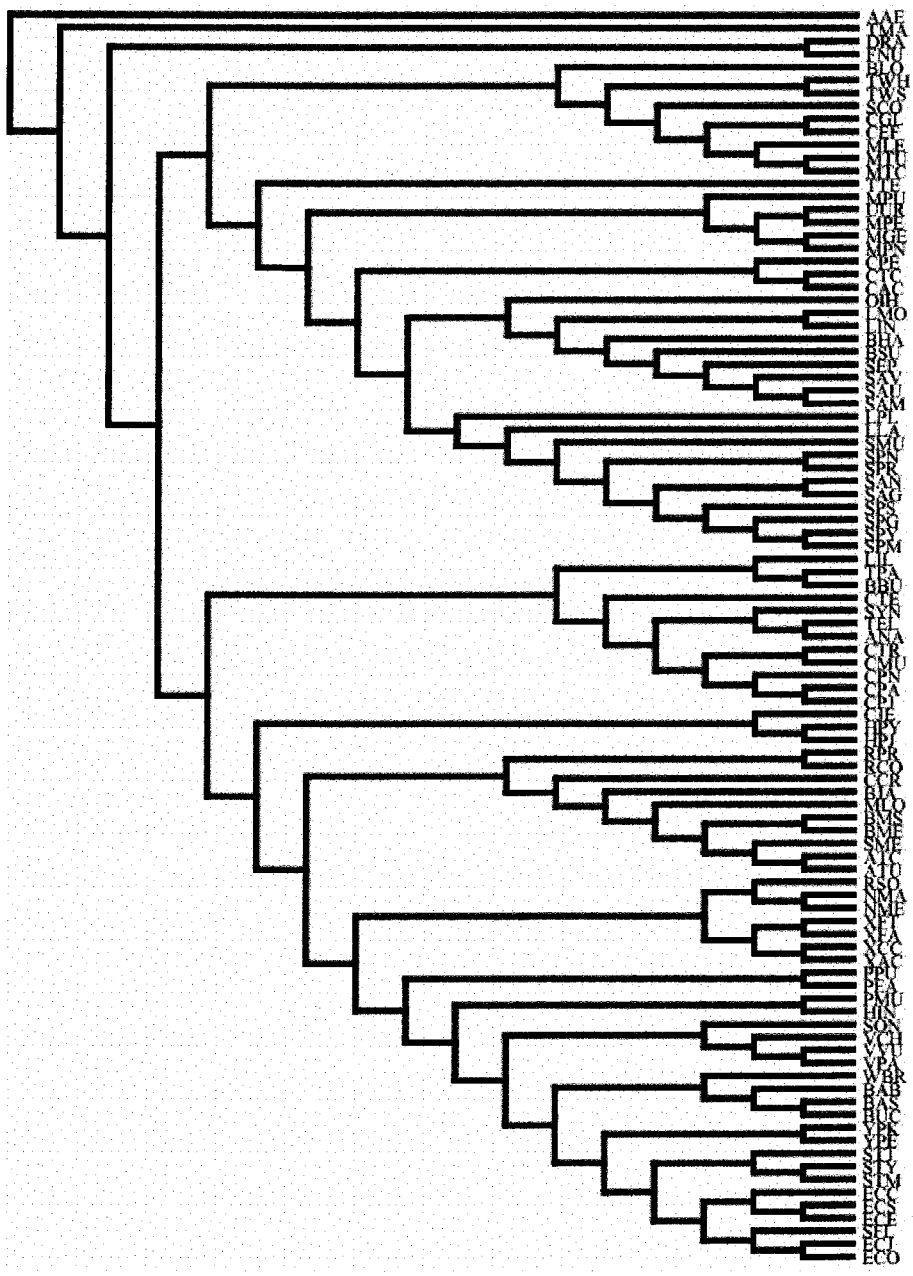


Figure 3-4. (Continue)

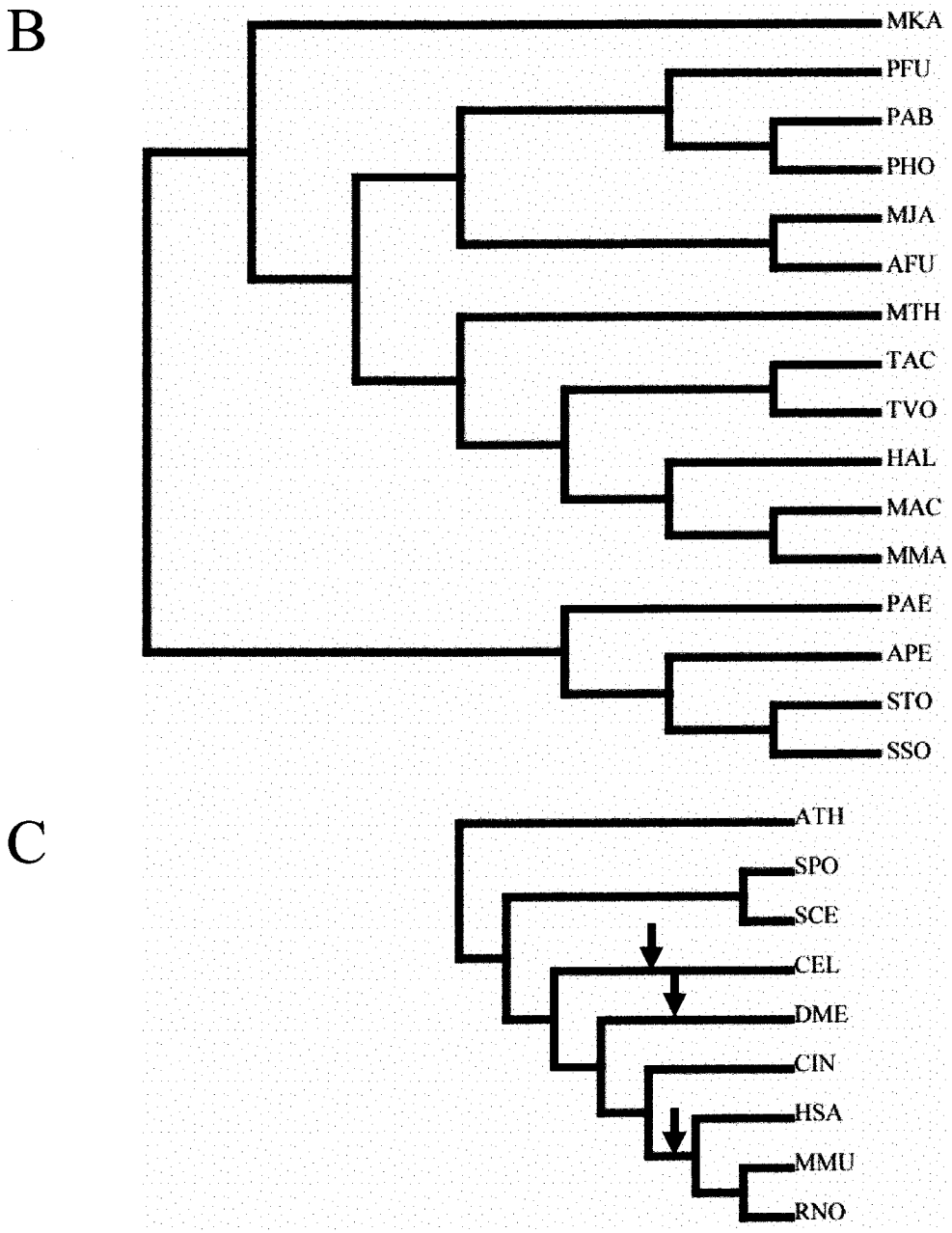


Figure 3-4. Genealogical relationships of 122 species

A: a genealogical relationship of 97 eubacteria. B: a genealogical relationship of 16 archaebacteria. C: a genealogical relationship of 9 eukaryotes. Red arrows on the branches represented the losses of *SNZ* and *SNO*.

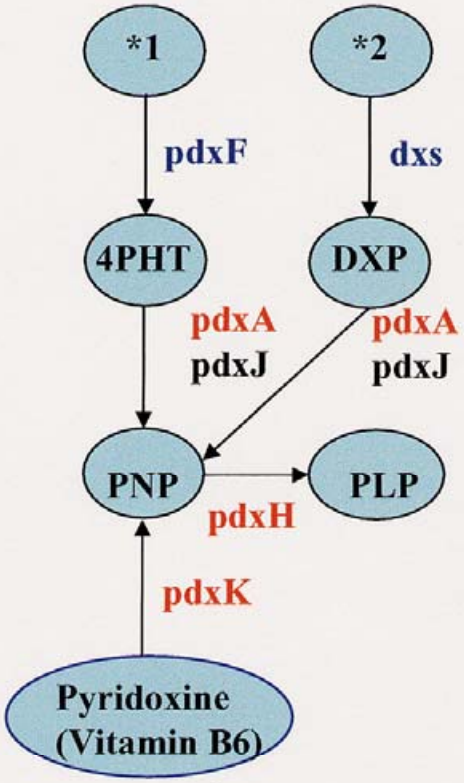


Figure 3-5. Loss of the 5 genes in VB6 metabolism

The pathway is a part of the *de novo* and the salvage pathways. The gene colored in black (*pdxJ*) was lost at first. The genes colored in red (*pdxA*, *pdxH*, and *pdxK*) were lost secondary. The genes colored in blue (*dxs* and *pdxF*) were lost thirdly. Circle indicated the substrates. *1: 3-Hydroxy-4-phophohydroxy-alpha-ketobutyrate *2: Glyceraldehyde-3-phosphate

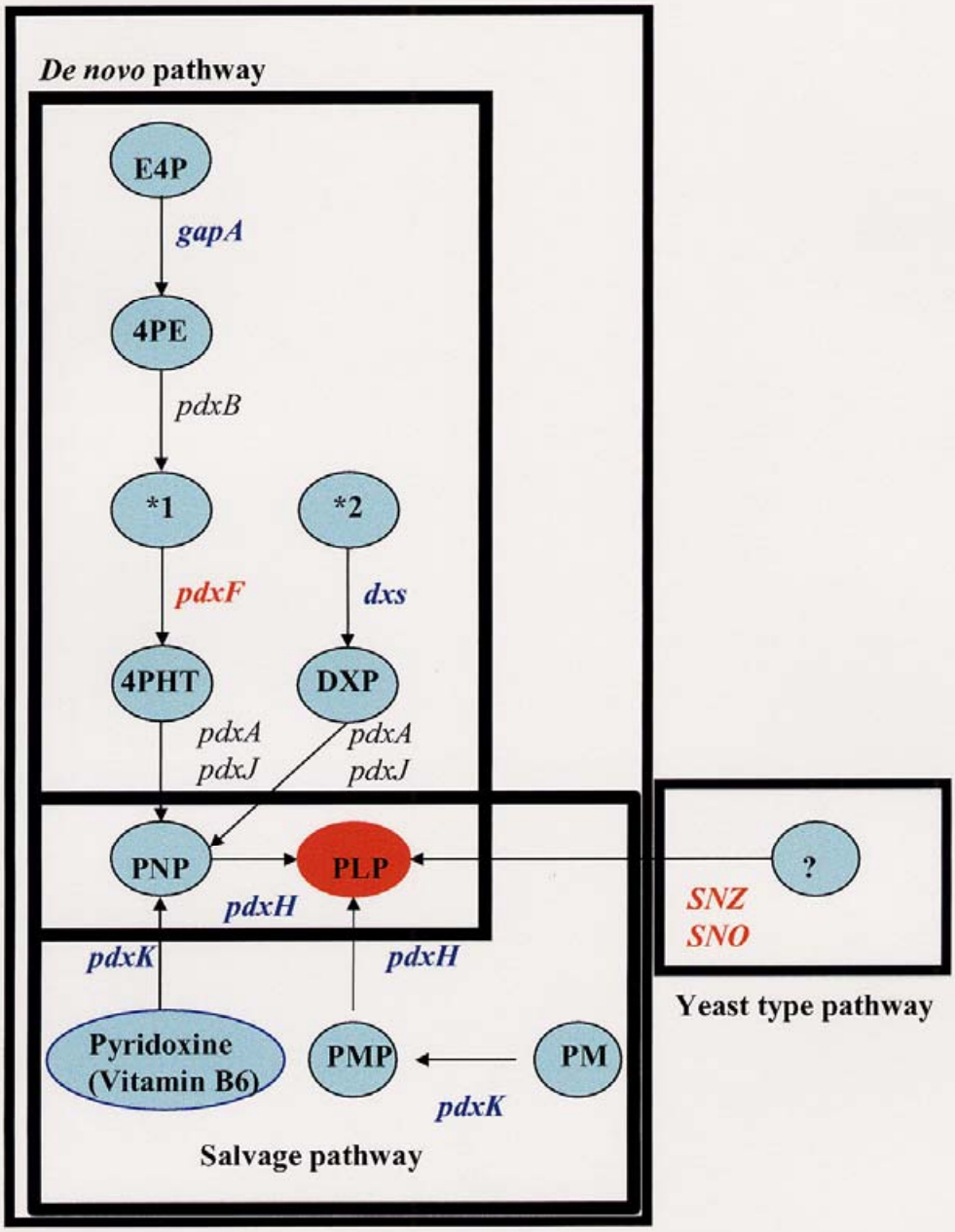


Figure 3-6. Formation of VB6 metabolism

The genes colored in red and blue were the primary genes for VB6 biosynthesis. The genes colored in black were generated in eubacteria. Circle indicated the substrates. PLP was an active form of VB6. *1: 3-Hydroxy-4-phophohydroxy-alpha-ketobutyrate *2: Glyceraldehyde-3-phosphate

Chapter 4

Conclusion

In this thesis, I focused upon the evolutionary process of the metabolic networks on the basis of gains and losses of the genes or the enzymatic reactions (ERs). From the studies in the chapters 2 and 3, I made an attempt to answer the following questions, as I mentioned in the chapter 1: How have the biological networks evolved and how were they diversified by genomic changes during evolution?

First, in the chapter 2, I studied the evolution of the metabolic networks for the 6 eukaryotic species whose complete genomes have been sequenced. In particular, I examined the evolutionary process of the metabolic networks from the viewpoint of the gain and loss of ERs, using the genome sequence data. My results have clearly shown that both gains and losses of ERs have occurred many times during evolution of the 6 eukaryotic species. Moreover, the loss of ERs was found to have occurred almost five times as frequently as the gain of ERs, so that the loss of ERs affected greatly evolutionary diversification of the metabolic networks in the eukaryotic species. This may give an explanation about why the heterotrophy is ubiquitously observed in a variety of the species, particularly animals. For the vertebrate lineages after their divergence from the insects such as *D. melanogaster*, however, I found a drastic increase in the occurrence frequency of the gain of ERs in the evolution of metabolic networks. In particular, 41% of the ER gains were deeply involved with the lipid and complex lipid metabolisms. Because some products of these two metabolisms function as hormones, I concluded that the ER gains of the two metabolisms accelerated the

development of hormonal signal transduction for the elaborated regulation of a physiological system during the evolution of vertebrates.

Second, I focused on the evolution of vitamin B_6 metabolism (the PLP biosynthetic pathways), as an example, among the 122 species in the three domains, eubacteria, archaebacteria and eukaryotes, of life. Here, I used the complete genome sequence data of the 122 species for identifying the genes encoding the relevant enzymes in this metabolic pathway. As shown in the chapter 3, I found that all the 10 genes related to this metabolism had been lost more than once during evolution of the 122 species examined. This result suggests that the vitamin B_6 metabolism has changed dynamically the set of genes by the events of gene loss. Taking into account the results of the gain and loss events, I speculated the mechanisms of formation and dysfunction of the PLP biosynthetic pathways. For formation of the pathway, I explained the formation process of the pathway by the known models, the patch work model and the *de novo* invention model. Moreover, I also explained the dysfunctional process of the pathway by a particular order of gene losses. Of course, the validity and versatility of these models to the other networks have to be evaluated in the future study.

As shown in this study, taking the advantage of utilizing the complete genome sequences, I successfully estimated the existence and the absence of particular genes in a species by conducting the comparative studies of different species of organisms.

Finally, I would like to make emphasis of that this line of studies will give us an important insight into the understanding of the evolutionary process of the metabolic networks.

References

Akman L, Yamashita A, Watanabe H, Oshima K, Shiba T, Hattori M, Aksoy S. (2002) Genome sequence of the endocellular obligate symbiont of tsetse flies, Wigglesworthia glossinidia. *Nat. Genet.* **32**:402-407.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**:403-10.

Baldauf SL, Roger AJ, Wenk-Siefert I, Doolittle WF. (2000) A kingdom-level phylogeny of Eukaryotas based on combined protein data. *Science* **290**:972-927.

Bean LE, Dvorachek WH Jr, Braun EL, Errett A, Saenz GS, Giles MD, Werner-Washburne M, Nelson MA, Natvig DO. (2001) Analysis of the pdx-1 (snz-1/sno-1) region of the Neurospora crassa genome: correlation of pyridoxine-requiring phenotypes with mutations in two structural genes. *Genetics.* **157**:1067-1075.

Bilski P, Li MY, Ehrenshaft M, Daub ME, Chignell CF. (2000) VB6 (pyridoxine) and its derivatives are efficient singlet oxygen quenchers and potential fungal antioxidants. *Photochem. Photobiol.* **71**:129-134.

Bono H, Ogata H, Goto S, Kanehisa M. (1998) Reconstruction of amino acid biosynthesis pathways from the complete genome sequence. *Genome Res.* **8**:203-210.

Bono H, Goto S, Fujibuchi W, Ogata H, Kanehisa M. (1998) Systematic Prediction of Orthologous Units of Genes in the Complete Genomes. *Genome Inform. Ser. Workshop Genome Inform.* **9**:32-40.

Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, Wheeler PR, Honore N, Garnier T, Churcher C, Harris D, Mungall K, Basham D, Brown D, Chillingworth T, Connor R, Davies RM, Devlin K, Duthoy S, Feltwell T, Fraser A, Hamlin N, Holroyd S, Hornsby T, Jagels K, Lacroix C, Maclean J, Moule S, Murphy L, Oliver K, Quail MA, Rajandream MA, Rutherford KM, Rutter S, Seeger K, Simon S, Simmonds M, Skelton J, Squares R, Squares S, Stevens K, Taylor K, Whitehead S, Woodward JR, Barrell BG. (2001) Massive gene decay in the leprosy bacillus. *Nature* **409**:1007-1111.

Covert MW, Schilling CH, Famili I, Edwards JS, Goryanin II, Selkov E, Palsson BO. (2001) Metabolic modeling of microbial strains in silico. *Trends Biochem. Sci.* **26**:179-186.

Dubey RK, Oparil S, Imthurn B, Jackson EK. (2002) Sex hormones and hypertension. *Cardiovasc. Res.* **53**:688-708.

Ehrenshaft M, Bilski P, Li MY, Chignell CF, Daub ME. (1999) A highly conserved sequence is a novel gene involved in de novo VB6 biosynthesis. *Proc. Natl. Acad. Sci. USA.* **96**:9374-938.

Ehrenshaft M, Daub ME. (2001) Isolation of PDX2, a second novel gene in the pyridoxine biosynthesis pathway of eukaryotes, archaebacteria, and a subset of eubacteria. *J. Bacteriol.* **183**:3383-3390.

Ellis LB, Hou BK, Kang W, Wackett LP. (2003) The University of Minnesota Biocatalysis/Biodegradation Database: post-genomic data mining. *Nucleic. Acids Res.* **31**:262-265.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, *et al.* (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**:496-512.

Forst CV, Schulten K. (2001) Phylogenetic analysis of metabolic pathways. *J. Mol. Evol.* **6**:471-489.

Funk CD. (2001) Prostaglandins and leukotrienes: advances in eicosanoid biology. *Science* **294**:1871-1875.

Gaasterland T, Selkov E. (1995) Reconstruction of metabolic networks using incomplete information. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **3**:127-135.

Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, FitzHugh W, Ma LJ, Smirnov S, Purcell S, Rehman B, Elkins T, Engels R, Wang S, Nielsen CB, Butler J, Endrizzi M, Qui D, Ianakiev P, Bell-Pedersen D, Nelson MA, Werner-Washburne M, Selitrennikoff CP, Kinsey JA, Braun EL, Zelter A, Schulte U, Kothe GO, Jedd G, Mewes W, Staben C, Marcotte E, Greenberg D, Roy A, Foley K, Naylor J, Stange-Thomann N, Barrett R, Gnerre S, Kamal M, Kamvysselis M, Mauceli E, Bielke C, Rudd S, Frishman D, Krystofova S, Rasmussen C, Metzenberg RL, Perkins DD, Kroken S, Cogoni C, Macino G, Catcheside D, Li W, Pratt RJ, Osmani SA, DeSouza CP, Glass L, Orbach MJ, Berglund JA, Voelker R, Yarden O, Plamann M, Seiler S, Dunlap J, Radford A, Aramayo R, Natvig DO, Alex LA, Mannhaupt G, Ebbole DJ, Freitag M, Paulsen I, Sachs MS, Lander ES, Nusbaum C, Birren B. (2003) The genome sequence of the filamentous fungus Neurospora crassa. *Nature* **422**:859-868.

Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. (2003) ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic. Acids Res.* **31**:3784-3788.

Fumoto M, Miyazaki S, Sugawara H. (2002) Genome Information Broker (GIB): data retrieval and comparative analysis system for completed microbial genomes and more. *Nucleic. Acids Res.* **30**:66-68.

Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. (2003) ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic. Acids Res.* **31**:3784-3788.

Grogan DW. (1988) Temperature-sensitive murein synthesis in an Escherichia coli pdx mutant and the role of alanine racemase. *Arch. Microbiol.* **150**:363-367.

Helmreich EJ. (1992) How pyridoxal 5'-phosphate could function in glycogen phosphorylase catalysis. *Biofactors.* **3**:159-172.

Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC, Herrmann R. (1996) Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae. *Nucleic. Acids Res.* **24**:4420-4449.

Horowitz N. H. (1945) On the evolution of biochemical synthesis. *Proc. Natl. Acad. Sci. USA.* **31**:152-157.

Huber J, Gruber C. (2001) Immunological and dermatological impact of progesterone. *Gynecol. Endocrinol.* **15**:18-21.

Huynen MA, Dandekar T, Bork P. (1999) Variation and evolution of the citric-acid cycle: a genomic perspective. *Trends. Microbiol.* **7**:281-291.

Jain R, Rivera MC, Moore JE, Lake JA. (2002) Horizontal gene transfer in microbial genome evolution. *Theor. Popul. Biol.* **61**:489-495.

Jensen R. A. (1976) Enzyme recruitment on evolution of new function. *Annu. Rev. Microbiol.* **30**:409-425.

Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL. (2000) The large-scale organization of metabolic networks. *Nature* **407**:651-654.

Kack H, Sandmark J, Gibson K, Schneider G, Lindqvist Y. (1999) Crystal structure of diaminopelargonic acid synthase: evolutionary relationships between pyridoxal-5'-phosphate-dependent enzymes. *J. Mol. Biol.* **291**:857-876.

Kanehisa M, Goto S, Kawashima S, Nakaya A. (2002) The KEGG databases at GenomeNet. *Nucleic. Acids Res.* **30**:42-46.

Karp PD, Riley M, Paley SM, Pellegrini-Toole A. (2002) The MetaCyc Database. *Nucleic. Acids Res.* **30**:59-61.

Karp PD, Riley M, Saier M, Paulsen IT, Collado-Vides J, Paley SM, Pellegrini-Toole A, Bonavides C, Gama-Castro S. (2002) The EcoCyc Database. *Nucleic. Acids Res.* **30**:56-58.

Kolesnick R, Fuks Z. (2003) Radiation and ceramide-induced apoptosis. *Oncogene.* **22**:5897-5906.

Lam HM, Winkler ME. (1990) Metabolic relationships between pyridoxine (VB6) and serine biosynthesis in Escherichia coli K-12. *J. Bacteriol.* **172**:6518-6528.

Lawrence JG. (2002) Gene transfer in bacteria: speciation without species? *Theor. Popul. Biol.* **61**:449-460.

Lazcano A, Miller SL. (1999) On the origin of metabolic pathways. *J. Mol. Evol.* **49**:424-431.

Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA. (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* **298**:799-804.

Liston TE, Roberts LJ 2nd. (1985) Transformation of prostaglandin D2 to 9 alpha, 11 beta-(15S)-trihydroxyprosta-(5Z,13E)-dien-1-oic acid (9 alpha, 11 beta-prostaglandin F2): a unique biologically active prostaglandin produced enzymatically in vivo in humans. *Proc. Natl. Acad. Sci. USA.* **82**:6030-6034.

Man TK, Zhao G, Winkler ME. (1996) Isolation of a pdxJ point mutation that bypasses the requirement for the PdxH oxidase in pyridoxal 5' -phosphate coenzyme biosynthesis in Escherichia coli K-12. *J. Bacteriol.* **178**:2445-2449.

Manore MM. (2000) Effect of physical activity on thiamine, riboflavin, and VB6 requirements. *Am. J. Clin. Nutr.* **72**:598S-606S.

Meganathan R. (2001) Biosynthesis of menaquinone (vitamin K2) and ubiquinone (coenzyme Q): a perspective on enzymatic mechanisms. *Vitam. Horm.* **61**:173-218

Mihara H, Kurihara T, Yoshimura T, Soda K, Esaki N. (1997) Cysteine sulfinate desulfinase, a NIFS-like protein of Escherichia coli with selenocysteine lyase and cysteine desulfurase activities. Gene cloning, purification, and characterization of a novel pyridoxal enzyme. *J. Biol. Chem.* **272**:22417-22424.

Mittenhuber G. (2001) Phylogenetic analyses and comparative genomics of VB6 (pyridoxine) and pyridoxal phosphate biosynthesis pathways. *J. Mol. Microbiol. Biotechnol.* **3**:1-20

Nelson KE, Paulsen IT, Heidelberg JF, Fraser CM. (2000) Status of genome projects for nonpathogenic bacteria and archaea. *Nat. Biotechnol.* **18**:1049-1054.

Osmani AH, May GS, Osmani SA. (1999) The extremely conserved pyroA gene of Aspergillus nidulans is required for pyridoxine synthesis and is required indirectly for resistance to photosensitizers. *J. Biol. Chem.* **274**:23565-23569.

Oshima K, Kakizawa S, Nishigawa H, Jung HY, Wei W, Suzuki S, Arashida R, Nakata D, Miyata SI, Ugaki M, Namba S. (2003) Reductive evolution suggested from the complete genome sequence of a plant-pathogenic phytoplasma. *Nat. Genet.* Epub.

Overbeek R, Larsen N, Pusch GD, D'Souza M, Selkov E Jr, 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.

Peregrin-Alvarez JM, Tsoka S, Ouzounis CA. (2003) The phylogenetic extent of metabolic enzymes and pathways. *Genome Res.* **13**:422-427.

Read TD, Peterson SN, Tourasse N, Baillie LW, Paulsen IT, Nelson KE, Tettelin H, Fouts DE, Eisen JA, Gill SR, Holtzapple EK, Okstad OA, Helgason E, Rilstone J, Wu M, Kolonay JF, Beanan MJ, Dodson RJ, Brinkac LM, Gwinn M, DeBoy RT, Madpu R, Daugherty SC, Durkin AS, Haft DH, Nelson WC, Peterson JD, Pop M, Khouri HM, Radune D, Benton JL, Mahamoud Y, Jiang L, Hance IR, Weidman JF, Berry KJ, Plaut RD, Wolf AM, Watkins KL, Nierman WC, Hazen A, Cline R, Redmond C, Thwaite JE, White O, Salzberg SL, Thomason B, Friedlander AM, Koehler TM, Hanna PC, Kolsto AB, Fraser CM. (2003) The genome sequence of Bacillus anthracis Ames and comparison to closely related bacteria. *Nature* **423**:81-86.

Reeds PJ, Garlick PJ. (2003) Protein and amino acid requirements and the composition of complementary foods. *J. Nutr.* **133**:2953S-2961S.

Rottmann WH, Peter GF, Oeller PW, Keller JA, Shen NF, Nagy BP, Taylor LP, Campbell AD, Theologis A. (1991) 1-aminocyclopropane-1-carboxylate synthase in tomato is encoded by a multigene family whose transcription is induced during fruit and floral senescence. *J. Mol. Biol.* **222**:937-961.

Salzberg SL, White O, Peterson J, Eisen JA. (2001) Microbial genes in the human genome: lateral transfer or gene loss? *Science* **292**:1903-1906.

Schuster S, Fell DA, Dandekar T. (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat. Biotechnol.* **18**:326-332.

Schmidt S, Sunyaev S, Bork P, Dandekar T. (2003) Metabolites: a helping hand for pathway evolution? *Trends Biochem. Sci.* **28**:336-341

Seack J, Perovic S, Gamulin V, Schroder HC, Beutelmann P, Muller IM, Muller WE. (2001) Identification of highly conserved genes: SNZ and SNO in the marine sponge Suberites domuncula: their gene structure and promoter activity in mammalian cells(1). *Biochim. Biophys. Acta.* **1520**:21-34.

Seta FD, Boschi-Muller S, Vignais ML, Branlant G. (1997) Characterization of Escherichia coli strains with gapA and gapB genes deleted. *J. Bacteriol.* **179**:5218-5221.

Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H. (2000) Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp. APS. *Nature* **407**:81-86.

Smirnoff N. (2001) L-ascorbic acid biosynthesis. *Vitam. Horm.* **61**:241-266.

Soberman RJ, Christmas P. (2003) The organization and consequences of eicosanoid signaling. *J. Clin. Invest.* **111**:1107-1113.

Tatusov RL, Mushegian AR, Bork P, Brown NP, Hayes WS, Borodovsky M, Rudd KE, Koonin EV. (1996) Metabolism and evolution of Haemophilus influenzae deduced from a whole-genome comparison with Escherichia coli. *Curr. Biol.* **6**:279-291.

Thompson JD, Higgins DG, Gibson TJ. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic. Acids Res.* **11**:4673-4680.

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

Unger RH, Orci L. (2002) Lipoapoptosis: its mechanism and its diseases. *Biochim. Biophys. Acta.* **1585**:202-212.

Wen Y, Gu J, Peng X, Zhang G, Nadler J. (2003) Overexpression of 12-lipoxygenase and cardiac fibroblast hypertrophy. *Trends Cardiovasc. Med.* **13**:129-136.

White RS, Dempsey WB. (1970) Purification and properties of VB6 kinase from Escherichia coli B. *Biochemistry* **9**:4057-4064.

Woese CR, Kandler O, Wheelis ML. (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. USA.* **87**:4576-4579.

Winkel-Shirley B. (2002) Biosynthesis of flavonoids and effects of stress. *Curr. Opin. Plant Biol.* **5**:218-223.

Yamashita T, Wada R, Sasaki T, Deng C, Bierfreund U, Sandhoff K, Proia RL. (1999) A vital role for glycosphingolipid synthesis during development and differentiation. *Proc. Natl. Acad. Sci. USA.* **96**:9142-9147.

Yang Y, Zhao G, Winkler ME. (1996) Identification of the pdxK gene that encodes pyridoxine (VB6) kinase in Escherichia coli K-12. *FEMS Microbiol. Lett.* **141**:89-95.

Yang Y, Zhao G, Man TK, Winkler ME. (1998a) Involvement of the gapA- and epd (gapB)-encoded dehydrogenases in pyridoxal 5'-phosphate coenzyme biosynthesis in Escherichia coli K-12. *J. Bacteriol.* **180**:4294-9.

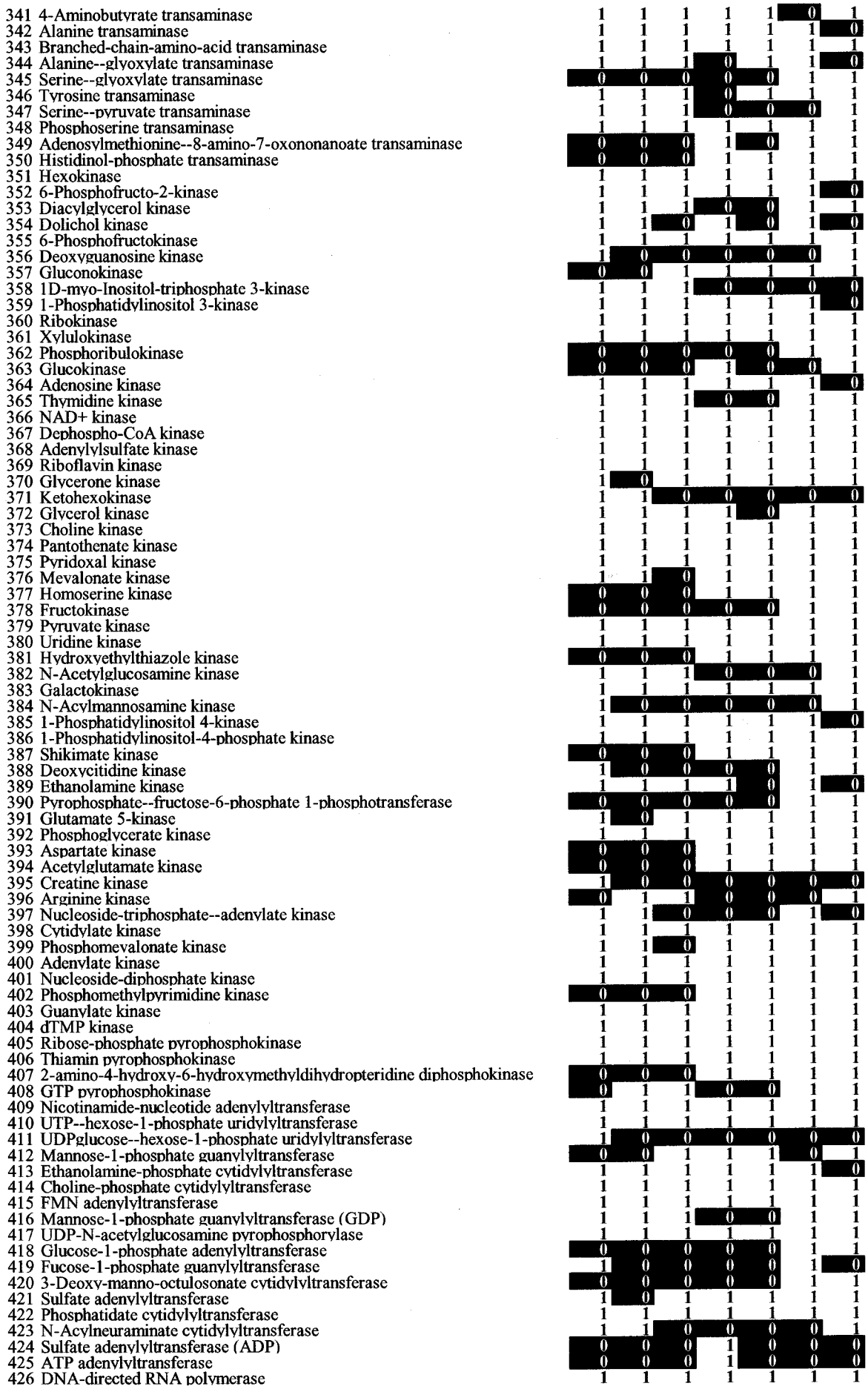
Yang Y, Tsui HC, Man TK, Winkler ME. (1998b) Identification and function of the pdxY gene, which encodes a novel pyridoxal kinase involved in the salvage pathway of pyridoxal 5'-phosphate biosynthesis in Escherichia coli K-12. *J. Bacteriol.* **180**:1814-21.

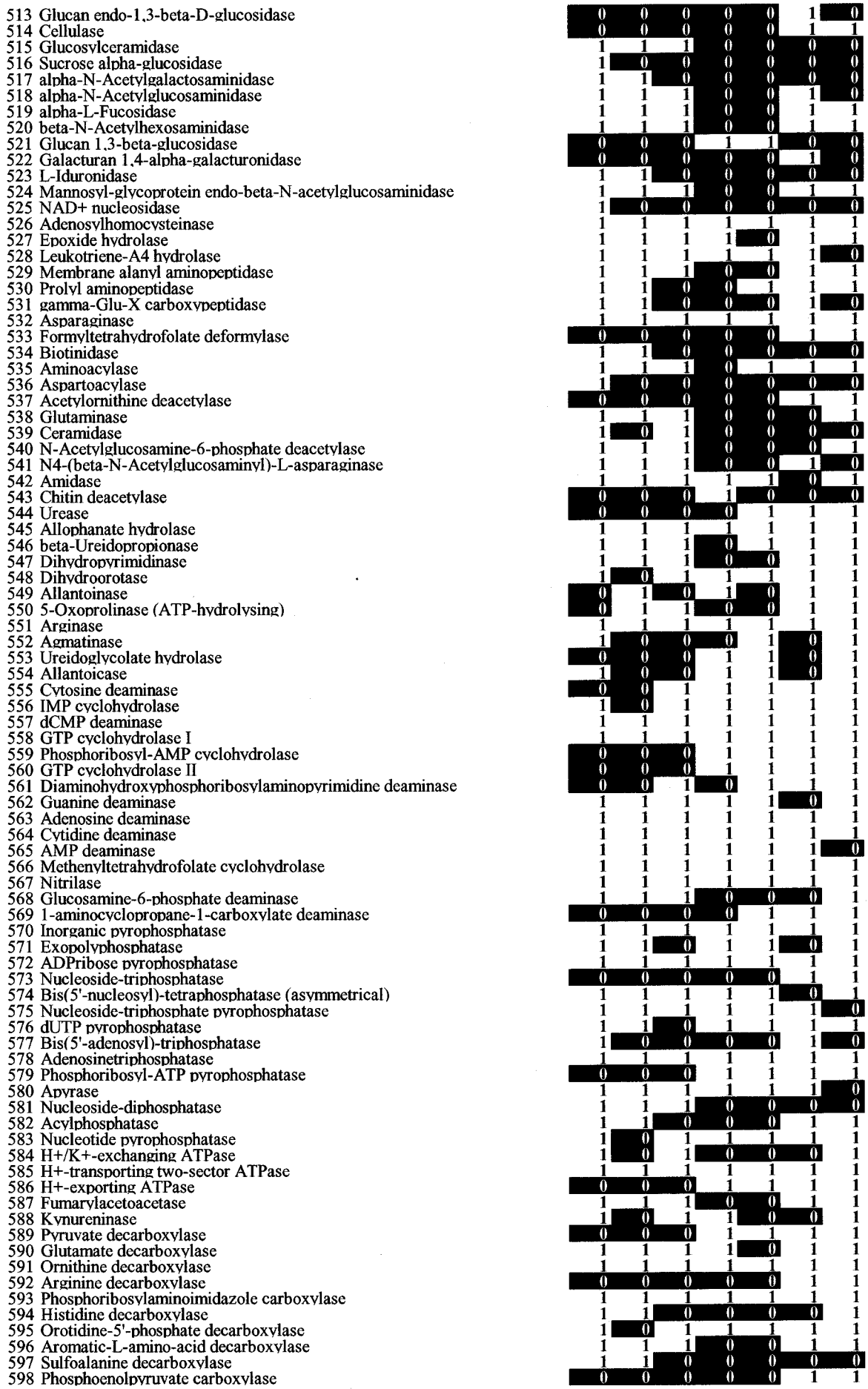
Zhao G, Pease AJ, Bharani N, Winkler ME. (1995) Biochemical characterization of gapB-encoded erythrose 4-phosphate dehydrogenase of Escherichia coli K-12 and its possible role in pyridoxal 5'-phosphate biosynthesis. *J. Bacteriol.* **177**:2804-2812.

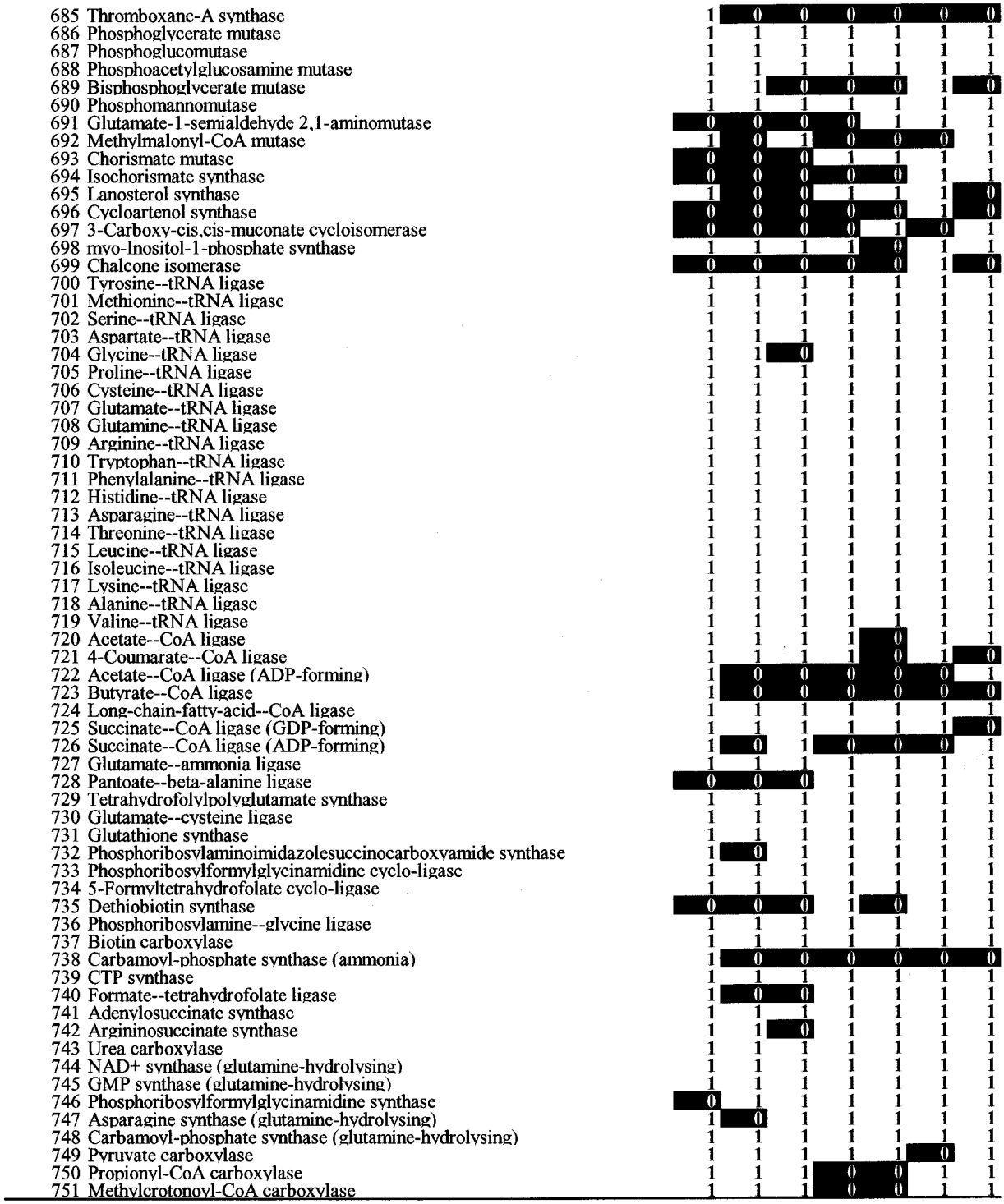
Supplementary table 1. Species-ER matrix

ER group	Enzyme involved with the ER	H	SA	DME	CEL	SCE	SPO	ATH	OUT
1	Alcohol dehydrogenase	1	1	1	1	1	1	1	1
2	3-Oxoacyl-[acyl-carrier-protein] reductase	1	1	1	1	1	1	1	1
3	Retinol dehydrogenase	1	0	1	0	0	1	0	
4	L-Iditol 2-dehydrogenase	1	1	1	1	1	1	1	
5	3beta-Hydroxy-delta5-steroid dehydrogenase	1	1	1	1	1	1	0	
6	11beta-Hydroxysteroid dehydrogenase	1	1	1	0	1	1	0	
7	Sepiapterin reductase	1	1	0	0	0	0	1	
8	3-Hydroxybutyryl-CoA dehydrogenase	1	1	0	0	0	1	1	
9	2-Dehydropantoate 2-reductase	0	0	0	1	1	0	1	
10	Carbonyl reductase (NADPH)	1	1	1	1	1	1	1	
11	Prostaglandin-F synthase	1	0	0	0	0	0	0	
12	Prostaglandin-E2 9-reductase	1	1	0	1	0	1	0	
13	5-Amino-6-(5-phosphoribosylamino)uracil reductase	0	0	1	1	1	1	1	
14	Cinnamyl-alcohol dehydrogenase	0	0	0	0	0	1	0	
15	Alcohol dehydrogenase(NADP+)	1	0	1	1	1	1	1	
16	Xanthine dehydrogenase	1	1	1	0	0	1	1	
17	IMP dehydrogenase	1	1	1	1	1	1	1	
18	Aldehyde reductase	1	1	1	1	1	1	1	
19	Dihydrokaempferol 4-reductase	0	0	1	1	1	1	1	
20	UDPglucose 6-dehydrogenase	1	1	1	0	0	1	1	
21	Histidinol dehydrogenase	0	0	0	1	1	1	1	
22	Shikimate 5-dehydrogenase	0	0	0	1	1	1	1	
23	Glyoxylate reductase	1	0	0	0	0	1	0	
24	L-Lactate dehydrogenase	1	1	1	0	1	1	1	
25	GDP-L-fucose synthase	1	1	1	0	0	1	1	
26	Homoserine dehydrogenase	0	0	0	1	1	1	1	
27	3-Hydroxybutyrate dehydrogenase	1	1	1	0	1	0	1	
28	3-Hydroxyisobutyrate dehydrogenase	1	1	1	0	0	1	1	
29	Hydroxymethylglutaryl-CoA reductase (NADPH)	1	1	1	1	1	1	1	
30	3-Hydroxyacyl-CoA dehydrogenase	1	1	1	0	0	1	1	
31	Malate dehydrogenase	1	1	1	1	1	1	1	
32	Malate dehydrogenase (oxaloacetate-decarboxylating)	1	0	0	1	1	1	1	
33	Malate dehydrogenase (decarboxylating)	0	0	0	0	0	1	1	
34	(R,R)-Butanediol dehydrogenase	1	1	1	0	0	1	1	
35	Malate dehydrogenase (oxaloacetate-decarboxylating)(NADP+)	1	1	1	0	0	1	1	
36	Isocitrate dehydrogenase (NAD+)	1	1	1	1	1	1	1	
37	Isocitrate dehydrogenase (NADP+)	1	1	1	1	1	1	1	
38	Phosphogluconate dehydrogenase (decarboxylating)	1	1	1	1	1	1	1	
39	Glucose 1-dehydrogenase	1	0	0	0	0	0	1	
40	Glucose-6-phosphate 1-dehydrogenase	1	1	1	1	1	1	1	
41	3alpha-Hydroxysteroid dehydrogenase (B-specific)	1	0	0	0	0	0	1	
42	Glycerol dehydrogenase	0	0	0	0	1	0	1	
43	Estradiol 17beta-dehydrogenase	1	1	1	1	1	1	0	
44	Glyoxylate reductase (NADP+)	1	0	0	0	0	1	0	
45	Glycerol-3-phosphate dehydrogenase(NAD+)	1	1	1	1	1	1	0	
46	Malate dehydrogenase (NADP+)	0	0	0	0	0	1	1	
47	3-Isopropylmalate dehydrogenase	0	0	0	1	1	1	1	
48	Ketol-acid reductoisomerase	0	0	0	1	0	1	1	
49	Phosphoglycerate dehydrogenase	1	1	1	1	1	1	1	
50	L-Lactate dehydrogenase (cytochrome)	1	1	0	1	1	1	1	
51	D-Lactate dehydrogenase (cytochrome)	0	0	1	1	1	1	1	
52	(S)-2-Hydroxy-acid oxidase	1	0	1	0	1	1	1	
53	Xanthine oxidase	1	1	1	0	0	1	0	
54	Choline dehydrogenase	1	0	1	0	0	0	1	
55	Glucose dehydrogenase (acceptor)	0	1	0	0	0	1	1	
56	Glycerol-3-phosphate dehydrogenase	1	1	1	1	1	1	1	
57	Ubiquinol--cytochrome-c reductase	1	1	1	1	1	1	1	
58	L-Ascorbate oxidase	0	1	1	1	1	1	1	
59	L-Ascorbate peroxidase	0	0	0	0	0	1	0	
60	Catalase	1	1	1	1	1	1	1	
61	Peroxidase	1	1	1	1	1	1	0	
62	Iodide peroxidase	1	1	1	0	0	0	0	
63	Glutathione peroxidase	1	1	1	1	1	1	1	
64	Tryptophan 2,3-dioxygenase	1	1	1	0	0	0	1	
65	Cysteine dioxygenase	1	1	1	0	0	0	0	
66	4-Hydroxyphenylpyruvate dioxygenase	1	1	1	0	0	1	1	
67	Arachidonate 12-lipoxygenase	1	0	0	0	0	0	0	
68	Arachidonate 15-lipoxygenase	1	0	0	0	0	1	0	
69	Arachidonate 5-lipoxygenase	1	0	0	0	0	1	0	
70	Indoleamine-pyrrole 2,3-dioxygenase	1	0	0	1	0	0	1	
71	Homogentisate 1,2-dioxygenase	1	1	1	0	0	1	1	
72	3-Hydroxyanthranilate 3,4-dioxygenase	1	0	1	1	0	0	0	
73	myo-Inositol oxygenase	1	0	1	0	0	1	0	
74	gamma-Butyrobetaine dioxygenase	1	1	1	1	0	0	1	
75	Procollagen-proline dioxygenase	1	1	1	0	0	1	0	
76	Procollagen-lysine 5-dioxygenase	1	1	1	0	0	1	0	
77	Trimethyllysine dioxygenase	1	0	0	0	0	0	0	
78	Naringenin 3-dioxygenase	0	1	0	0	1	1	0	
79	trans-Cinnamate 4-monooxygenase	0	0	0	0	0	1	0	
80	Cholestanetriol 26-monooxygenase	1	1	0	0	0	0	0	
81	Cholesterol 7alpha-monooxygenase	1	0	0	0	0	0	0	
82	Flavonoid 3'-monooxygenase	0	0	0	0	0	1	0	


169 L-Pipecolate oxidase
170 Sarcosine dehydrogenase
171 Dimethylglycine dehydrogenase
172 NAD(P)+ transhydrogenase (AB-specific)
173 NADH dehydrogenase (ubiquinone)
174 NAD(P)H dehydrogenase (quinone)
175 Dihydropteridine reductase
176 Nitrate reductase (NADH)
177 GMP reductase
178 Urate oxidase
179 Ferredoxin--nitrite reductase
180 Sulfite reductase (NADPH)
181 Dihydrolipoamide dehydrogenase
182 Glutathione-disulfide reductase
183 Thioredoxin-disulfide reductase
184 Sulfite oxidase
185 Phosphoadenylyl-sulfate reductase (thioredoxin)
186 Sulfite reductase (ferredoxin)
187 Cytochrome-c oxidase
188 Nicotinamide N-methyltransferase
189 Homocysteine S-methyltransferase
190 Caffeoyl-CoA O-methyltransferase
191 Uroporphyrin-III C-methyltransferase
192 Magnesium-protoporphyrin O-methyltransferase
193 5-Methyltetrahydrofolate--homocysteine S-methyltransferase
194 5-Methyltetrahydropteroyltriglutamate--homocysteine
195 Phosphatidylethanolamine N-methyltransferase
196 Guanidinoacetate N-methyltransferase
197 Glycine N-methyltransferase
198 Phenylethanolamine N-methyltransferase
199 DNA (cytosine-5-)-methyltransferase
200 Acetylserotonin N-methyltransferase
201 Histone-lysine N-methyltransferase
202 Thymidylate synthase
203 Amine N-methyltransferase
204 Betaine--homocysteine S-methyltransferase
205 Catechol O-methyltransferase
206 Quercetin 3-O-methyltransferase
207 Histamine N-methyltransferase
208 Glycine hydroxymethyltransferase
209 Aminomethyltransferase
210 3-Methyl-2-oxobutanoate hydroxymethyltransferase
211 Phosphoribosylglycinamide formyltransferase
212 Phosphoribosylaminoimidazolecarboxamide formyltransferase
213 Glutamate formiminotransferase
214 Methionyl-tRNA formyltransferase
215 Aspartate carbamoyltransferase
216 Ornithine carbamoyltransferase
217 Glycine amidinotransferase
218 Transketolase
219 Transaldolase
220 Acetolactate synthase
221 Amino-acid N-acetyltransferase
222 Dihydrolipoamide S-acetyltransferase
223 Acyl-[acyl-carrier-protein]-UDP-N-acetylglucosamine O-acyltransferase
224 Glycerol-3-phosphate O-acyltransferase
225 Acetyl-CoA C-acyltransferase
226 Diacylglycerol O-acyltransferase
227 Carnitine O-palmitoyltransferase
228 Sterol O-acyltransferase
229 Glycine C-acetyltransferase
230 Serine O-acetyltransferase
231 Homoserine O-acetyltransferase
232 Glutamate N-acetyltransferase
233 5-Aminolevulinate synthase
234 [Acyl-carrier-protein] S-acetyltransferase
235 [Acyl-carrier-protein] S-malonyltransferase
236 Glucosamine-phosphate N-acetyltransferase
237 3-Oxoacyl-[acyl-carrier-protein] synthase
238 Serine C-palmitoyltransferase
239 1-Acylglycerol-3-phosphate O-acyltransferase
240 Diamine N-acetyltransferase
241 Choline O-acetyltransferase
242 Dihydrolipoamide S-succinyltransferase
243 Glycine N-choloyltransferase
244 Carnitine O-acetyltransferase
245 Naringenin-chalcone synthase
246 Fatty-acid synthase
247 Fatty-acyl-CoA synthase
248 Aralkylamine N-acetyltransferase
249 Acetyl-CoA C-acetyltransferase
250 gamma-Glutamyltransferase
251 Citrate (Si)-synthase
252 Hydroxymethylglutaryl-CoA synthase
253 2-isopropylmalate synthase
254 Homocitrate synthase


341 4-Aminobutyrate transaminase
342 Alanine transaminase
343 Branched-chain-amino-acid transaminase
344 Alanine--glyoxylate transaminase
345 Serine--glyoxylate transaminase
346 Tyrosine transaminase
347 Serine--pyruvate transaminase
348 Phosphoserine transaminase
349 Adenosylmethionine--8-amino-7-oxononanoate transaminase
350 Histidinol-phosphate transaminase
351 Hexokinase
352 6-Phosphofructo-2-kinase
353 Diacylglycerol kinase
354 Dolichol kinase
355 6-Phosphofructokinase
356 Deoxyguanosine kinase
357 Gluconokinase
358 1D-myo-Inositol-triphosphate 3-kinase
359 1-Phosphatidylinositol 3-kinase
360 Ribokinase
361 Xylulokinase
362 Phosphoribulokinase
363 Glucokinase
364 Adenosine kinase
365 Thymidine kinase
366 NAD+ kinase
367 Dephospho-CoA kinase
368 Adenylylsulfate kinase
369 Riboflavin kinase
370 Glycerone kinase
371 Ketohexokinase
372 Glycerol kinase
373 Choline kinase
374 Pantothenate kinase
375 Pyridoxal kinase
376 Mevalonate kinase
377 Homoserine kinase
378 Fructokinase
379 Pyruvate kinase
380 Uridine kinase
381 Hydroxyethylthiazole kinase
382 N-Acetylglucosamine kinase
383 Galactokinase
384 N-Acylmannosamine kinase
385 1-Phosphatidylinositol 4-kinase
386 1-Phosphatidylinositol-4-phosphate kinase
387 Shikimate kinase
388 Deoxycitidine kinase
389 Ethanolamine kinase
390 Pyrophosphate--fructose-6-phosphate 1-phosphotransferase
391 Glutamate 5-kinase
392 Phosphoglycerate kinase
393 Aspartate kinase
394 Acetylglutamate kinase
395 Creatine kinase
396 Arginine kinase
397 Nucleoside-triphosphate--adenylate kinase
398 Cytidylate kinase
399 Phosphomevalonate kinase
400 Adenylate kinase
401 Nucleoside-diphosphate kinase
402 Phosphomethylpyrimidine kinase
403 Guanylate kinase
404 dTMP kinase
405 Ribose-phosphate pyrophosphokinase
406 Thiamin pyrophosphokinase
407 2-amino-4-hydroxy-6-hydroxymethyldihydropteridine diphosphokinase
408 GTP pyrophosphokinase
409 Nicotinamide-nucleotide adenylyltransferase
410 UTP--hexose-1-phosphate uridylyltransferase
411 UDPglucose--hexose-1-phosphate uridylyltransferase
412 Mannose-1-phosphate guanylyltransferase
413 Ethanolamine-phosphate cytidylyltransferase
414 Choline-phosphate cytidylyltransferase
415 FMN adenylyltransferase
416 Mannose-1-phosphate guanylyltransferase (GDP)
417 UDP-N-acetylglucosamine pyrophosphorylase
418 Glucose-1-phosphate adenylyltransferase
419 Fucose-1-phosphate guanylyltransferase
420 3-Deoxy-manno-octulosonate cytidylyltransferase
421 Sulfate adenylyltransferase
422 Phosphatidate cytidylyltransferase
423 N-Acylneuraminate cytidylyltransferase
424 Sulfate adenylyltransferase (ADP)
425 ATP adenylyltransferase
426 DNA-directed RNA polymerase


513 Glucan endo-1,3-beta-D-glucosidase
514 Cellulase
515 Glucosylceramidase
516 Sucrose alpha-glucosidase
517 alpha-N-Acetylgalactosaminidase
518 alpha-N-Acetylglucosaminidase
519 alpha-L-Fucosidase
520 beta-N-Acetylhexosaminidase
521 Glucan 1,3-beta-glucosidase
522 Galacturan 1,4-alpha-galacturonidase
523 L-Iduronidase
524 Mannosyl-glycoprotein endo-beta-N-acetylglucosaminidase
525 NAD+ nucleosidase
526 Adenosylhomocysteinase
527 Epoxide hydrolase
528 Leukotriene-A4 hydrolase
529 Membrane alanyl aminopeptidase
530 Prolyl aminopeptidase
531 gamma-Glu-X carboxypeptidase
532 Asparaginase
533 Formyltetrahydrofolate deformylase
534 Biotinidase
535 Aminoacylase
536 Aspartoacylase
537 Acetylornithine deacetylase
538 Glutaminase
539 Ceramidase
540 N-Acetylglucosamine-6-phosphate deacetylase
541 N4-(beta-N-Acetylglucosaminyl)-L-asparaginase
542 Amidase
543 Chitin deacetylase
544 Urease
545 Allophanate hydrolase
546 beta-Ureidopropionase
547 Dihydropyrimidinase
548 Dihydroorotase
549 Allantoinase
550 5-Oxoprolinase (ATP-hydrolysing)
551 Arginase
552 Agmatinase
553 Ureidoglycolate hydrolase
554 Allantoicase
555 Cytosine deaminase
556 IMP cyclohydrolase
557 dCMP deaminase
558 GTP cyclohydrolase I
559 Phosphoribosyl-AMP cyclohydrolase
560 GTP cyclohydrolase II
561 Diaminohydroxyphosphoribosylaminopyrimidine deaminase
562 Guanine deaminase
563 Adenosine deaminase
564 Cytidine deaminase
565 AMP deaminase
566 Methenyltetrahydrofolate cyclohydrolase
567 Nitrilase
568 Glucosamine-6-phosphate deaminase
569 1-aminocyclopropane-1-carboxylate deaminase
570 Inorganic pyrophosphatase
571 Exopolyphosphatase
572 ADPribose pyrophosphatase
573 Nucleoside-triphosphatase
574 Bis(5'-nucleosyl)-tetraphosphatase (asymmetrical)
575 Nucleoside-triphosphate pyrophosphatase
576 dUTP pyrophosphatase
577 Bis(5'-adenosyl)-triphosphatase
578 Adenosinetriphosphatase
579 Phosphoribosyl-ATP pyrophosphatase
580 Apyrase
581 Nucleoside-diphosphatase
582 Acylphosphatase
583 Nucleotide pyrophosphatase
584 H+/K+-exchanging ATPase
585 H+-transporting two-sector ATPase
586 H+-exporting ATPase
587 Fumarylacetoacetase
588 Kynureninase
589 Pyruvate decarboxylase
590 Glutamate decarboxylase
591 Ornithine decarboxylase
592 Arginine decarboxylase
593 Phosphoribosylaminoimidazole carboxylase
594 Histidine decarboxylase
595 Orotidine-5'-phosphate decarboxylase
596 Aromatic-L-amino-acid decarboxylase
597 Sulfoalanine decarboxylase
598 Phosphoenolpyruvate carboxylase


685 Thromboxane-A synthase
686 Phosphoglycerate mutase
687 Phosphoglucomutase
688 Phosphoacetylglucosamine mutase
689 Bisphosphoglycerate mutase
690 Phosphomannomutase
691 Glutamate-1-semialdehyde 2,1-aminomutase
692 Methylmalonyl-CoA mutase
693 Chorismate mutase
694 Isochorismate synthase
695 Lanosterol synthase
696 Cycloartenol synthase
697 3-Carboxy-cis,cis-muconate cycloisomerase
698 myo-Inositol-1-phosphate synthase
699 Chalcone isomerase
700 Tyrosine--tRNA ligase
701 Methionine--tRNA ligase
702 Serine--tRNA ligase
703 Aspartate--tRNA ligase
704 Glycine--tRNA ligase
705 Proline--tRNA ligase
706 Cysteine--tRNA ligase
707 Glutamate--tRNA ligase
708 Glutamine--tRNA ligase
709 Arginine--tRNA ligase
710 Tryptophan--tRNA ligase
711 Phenylalanine--tRNA ligase
712 Histidine--tRNA ligase
713 Asparagine--tRNA ligase
714 Threonine--tRNA ligase
715 Leucine--tRNA ligase
716 Isoleucine--tRNA ligase
717 Lysine--tRNA ligase
718 Alanine--tRNA ligase
719 Valine--tRNA ligase
720 Acetate--CoA ligase
721 4-Coumarate--CoA ligase
722 Acetate--CoA ligase (ADP-forming)
723 Butyrate--CoA ligase
724 Long-chain-fatty-acid--CoA ligase
725 Succinate--CoA ligase (GDP-forming)
726 Succinate--CoA ligase (ADP-forming)
727 Glutamate--ammonia ligase
728 Pantoate--beta-alanine ligase
729 Tetrahydrofolylpolyglutamate synthase
730 Glutamate--cysteine ligase
731 Glutathione synthase
732 Phosphoribosylaminoimidazolesuccinocarboxyamide synthase
733 Phosphoribosylformylglycinamidine cyclo-ligase
734 5-Formyltetrahydrofolate cyclo-ligase
735 Dethiobiotin synthase
736 Phosphoribosylamine--glycine ligase
737 Biotin carboxylase
738 Carbamoyl-phosphate synthase (ammonia)
739 CTP synthase
740 Formate--tetrahydrofolate ligase
741 Adenylosuccinate synthase
742 Argininosuccinate synthase
743 Urea carboxylase
744 NAD+ synthase (glutamine-hydrolysing)
745 GMP synthase (glutamine-hydrolysing)
746 Phosphoribosylformylglycinamidine synthase
747 Asparagine synthase (glutamine-hydrolysing)
748 Carbamoyl-phosphate synthase (glutamine-hydrolysing)
749 Pyruvate carboxylase
750 Propionyl-CoA carboxylase
751 Methylcrotonoyl-CoA carboxylase

Supplementary table 2. List of enzymatic reactions included in the 104 metabolic networks

Metabolic network	ER	Enzyme involved with the ER
Urea cycle and metabolism of amino groups		
	115	N-Acetyl-gamma-glutamyl-phosphate reductase
	116	Glutamate-5-semialdehyde dehydrogenase
	148	Glutamate dehydrogenase (NAD(P)+)
	161	Pyrroline-5-carboxylate reductase
	170	Sarcosine dehydrogenase
	196	Guanidinoacetate N-methyltransferase
	216	Ornithine carbamoyltransferase
	217	Glycine amidinotransferase
	221	Amino-acid N-acetyltransferase
	232	Glutamate N-acetyltransferase
	321	Spermidine synthase
	325	Spermine synthase
	338	Acetylornithine transaminase
	339	Ornithine-oxo-acid transaminase
	394	Acetylglutamate kinase
	391	Glutamate 5-kinase
	395	Creatine kinase
	544	Urease
	535	Aminoacylase
	537	Acetylornithine deacetylase
	545	Allophanate hydrolase
	551	Arginase
	591	Ornithine decarboxylase
	653	Argininosuccinate lyase
	742	Argininosuccinate synthase
	743	Urea carboxylase
	738	Carbamoyl-phosphate synthase (ammonia)
Glutamate metabolism		
	108	Succinate-semialdehyde dehydrogenase (NAD(P)+)
	111	Succinate-semialdehyde dehydrogenase
	147	Glutamate dehydrogenase
	148	Glutamate dehydrogenase (NAD(P)+)
	149	Glutamate dehydrogenase (NADP+)
	146	Glutamate synthase (NADPH)
	159	1-Pyrroline-5-carboxylate dehydrogenase
	182	Glutathione-disulfide reductase
	236	Glucosamine-phosphate N-acetyltransferase
	302	Amidophosphoribosyltransferase
	337	Aspartate transaminase
	342	Alanine transaminase
	340	Glutamine-fructose-6-phosphate transaminase (isomerizing)
	341	4-Aminobutyrate transaminase
	382	N-Acetylglucosamine kinase
	538	Glutaminase
	590	Glutamate decarboxylase
	592	Arginine decarboxylase
	707	Glutamate-tRNA ligase
	708	Glutamine-tRNA ligase
	727	Glutamate-ammonia ligase
	730	Glutamate-cysteine ligase
	731	Glutathione synthase
	738	Carbamoyl-phosphate synthase (ammonia)
	744	NAD+ synthase (glutamine-hydrolysing)
	745	GMP synthase (glutamine-hydrolysing)
	748	Carbamoyl-phosphate synthase (glutamine-hydrolysing)
Alanine and aspartate metabolism		
	150	D-Aspartate oxidase
	151	L-Aspartate oxidase
	215	Aspartate carbamoyltransferase

194 5-Methyltetrahydropteroyltriglutamate-homocysteine
199 DNA (cytosine-5-)-methyltransferase
214 Methionyl-tRNA formyltransferase
231 Homoserine O-acetyltransferase
335 Methionine adenosyltransferase
331 Cystathionine gamma-synthase
332 O-acetylhomoserine aminocarboxypropyltransferase
526 Adenosylhomocysteinase
628 Cystathionine beta-synthase
656 Cystathionine gamma-lyase
660 Cystathionine beta-lyase
701 Methionine-tRNA ligase

Cysteine metabolism

24 L-Lactate dehydrogenase
65 Cysteine dioxygenase
230 Serine O-acetyltransferase
330 Cysteine synthase
331 Cystathionine gamma-synthase
332 O-acetylhomoserine aminocarboxypropyltransferase
337 Aspartate transaminase
439 3-Mercaptopyruvate sulfurtransferase
647 L-serine ammonia-lyase
656 Cystathionine gamma-lyase
660 Cystathionine beta-lyase
706 Cysteine-tRNA ligase

Valine, leucine and isoleucine degradation

28 3-Hydroxyisobutyrate dehydrogenase
30 3-Hydroxyacyl-CoA dehydrogenase
113 Aldehyde dehydrogenase (NAD+)
112 Methylmalonate-semialdehyde dehydrogenase (acylating)
120 Aldehyde oxidase
123 3-Methyl-2-oxobutanoate dehydrogenase (lipoamide)
141 Butyryl-CoA dehydrogenase
142 Acyl-CoA dehydrogenase
139 Isovaleryl-CoA dehydrogenase
225 Acetyl-CoA C-acyltransferase
252 Hydroxymethylglutaryl-CoA synthase
343 Branched-chain-amino-acid transaminase
446 3-Oxoacid CoA-transferase
618 Hydroxymethylglutaryl-CoA lyase
624 Enoyl-CoA hydratase
692 Methylmalonyl-CoA mutase
750 Propionyl-CoA carboxylase
751 Methylcrotonoyl-CoA carboxylase

Valine, leucine and isoleucine biosynthesis

47 3-Isopropylmalate dehydrogenase
48 Ketol-acid reductoisomerase
121 Pyruvate dehydrogenase (lipoamide)
220 Acetolactate synthase
253 2-isopropylmalate synthase
343 Branched-chain-amino-acid transaminase
641 Dihydroxy-acid dehydratase
631 3-Isopropylmalate dehydratase
715 Leucine-tRNA ligase
716 Isoleucine-tRNA ligase
719 Valine-tRNA ligase

Lysine biosynthesis

26 Homoserine dehydrogenase
106 Aspartate-semialdehyde dehydrogenase
114 L-Aminoadipate-semialdehyde dehydrogenase
129 Dihydrodipicolinate reductase
165 Saccharopine dehydrogenase (NAD+, L-lysine forming)

153 Amine oxidase (flavin-containing)
155 Amine oxidase (copper-containing)
207 Histamine N-methyltransferase
213 Glutamate formiminotransferase
303 ATP phosphoribosyltransferase
350 Histidinol-phosphate transaminase
464 Histidinol-phosphatase
536 Aspartoacylase
559 Phosphoribosyl-AMP cyclohydrolase
579 Phosphoribosyl-ATP pyrophosphatase
594 Histidine decarboxylase
596 Aromatic-L-amino-acid decarboxylase
625 Imidazoleglycerol-phosphate dehydratase
635 Urocanate hydratase
649 Histidine ammonia-lyase
674 phosphoribosylformiminoaminophosphoribosylimidazolecarboxamide isomerase
712 Histidine-tRNA ligase

Tyrosine metabolism

1 Alcohol dehydrogenase
118 Aldehyde dehydrogenase (NAD(P)+)
108 Succinate-semialdehyde dehydrogenase (NAD(P)+)
120 Aldehyde oxidase
153 Amine oxidase (flavin-containing)
155 Amine oxidase (copper-containing)
62 Iodide peroxidase
71 Homogentisate 1,2-dioxygenase
66 4-Hydroxyphenylpyruvate dioxygenase
91 Tyrosine 3-monooxygenase
93 Dopamine beta-monooxygenase
94 Monophenol monooxygenase
205 Catechol O-methyltransferase
198 Phenylethanolamine N-methyltransferase
337 Aspartate transaminase
346 Tyrosine transaminase
350 Histidinol-phosphate transaminase
587 Fumarylacetoacetase
596 Aromatic-L-amino-acid decarboxylase
651 Phenylalanine ammonia-lyase
672 Maleylacetoacetate isomerase

Phenylalanine metabolism

118 Aldehyde dehydrogenase (NAD(P)+)
153 Amine oxidase (flavin-containing)
155 Amine oxidase (copper-containing)
61 Peroxidase
66 4-Hydroxyphenylpyruvate dioxygenase
79 trans-Cinnamate 4-monooxygenase
337 Aspartate transaminase
346 Tyrosine transaminase
350 Histidinol-phosphate transaminase
542 Amidase
596 Aromatic-L-amino-acid decarboxylase
651 Phenylalanine ammonia-lyase

Tryptophan metabolism

30 3-Hydroxyacyl-CoA dehydrogenase
113 Aldehyde dehydrogenase (NAD+)
120 Aldehyde oxidase
122 Oxoglutarate dehydrogenase (lipoamide)
145 Glutaryl-CoA dehydrogenase
153 Amine oxidase (flavin-containing)
155 Amine oxidase (copper-containing)
60 Catalase
72 3-Hydroxyanthranilate 3,4-dioxygenase

145 Glutaryl-CoA dehydrogenase
249 Acetyl-CoA C-acetyltransferase
542 Amidase
567 Nitrilase
582 Acylphosphatase
624 Enoyl-CoA hydratase

Styrene degradation

71 Homogentisate 1,2-dioxygenase
542 Amidase
587 Fumarylacetoacetase
672 Maleylacetoacetate isomerase

Atrazine degradation

544 Urease
545 Allophanate hydrolase

Caprolactam degradation

15 Alcohol dehydrogenase(NADP+)

Streptomycin biosynthesis

466 myo-Inositol-1(or 4)-monophosphatase
687 Phosphoglucomutase
698 myo-Inositol-1-phosphate synthase

Erythromycin biosynthesis

351 Hexokinase
363 Glucokinase
633 dTDPglucose 4,6-dehydratase
687 Phosphoglucomutase

Terpenoid biosynthesis

100 Squalene monooxygenase
318 Dimethylallyltranstransferase
319 Geranyltranstransferase
324 Farnesyl-diphosphate farnesyltransferase
327 Farnesyltranstransferase
655 Strictosidine synthase
680 Isopentenyl-diphosphate delta-isomerase

Flavonoids, stilbene and lignin biosynthesis

14 Cinnamyl-alcohol dehydrogenase
19 Dihydrokaempferol 4-reductase
61 Peroxidase
78 Naringenin 3-dioxygenase
79 trans-Cinnamate 4-monooxygenase
82 Flavonoid 3'-monooxygenase
94 Monophenol monooxygenase
206 Quercetin 3-O-methyltransferase
190 Caffeoyl-CoA O-methyltransferase
245 Naringenin-chalcone synthase
297 Flavonol 3-O-glucosyltransferase
503 beta-Glucosidase
699 Chalcone isomerase
721 4-Coumarate-CoA ligase

Alkaloid biosynthesis I

94 Monophenol monooxygenase
337 Aspartate transaminase
346 Tyrosine transaminase
596 Aromatic-L-amino-acid decarboxylase

Alkaloid biosynthesis II

155 Amine oxidase (copper-containing)
591 Ornithine decarboxylase

610 Fructose-bisphosphate aldolase
666 Ribulose-phosphate 3-epimerase
676 Ribose-5-phosphate epimerase
678 Glucose-6-phosphate isomerase
687 Phosphoglucomutase

Inositol metabolism

112 Methylmalonate-semialdehyde dehydrogenase (acylating)
673 Triose-phosphate isomerase

Pentose and glucuronate interconversions

15 Alcohol dehydrogenase(NADP+)
18 Aldehyde reductase
20 UDPglucose 6-dehydrogenase
277 Glucuronosyltransferase
361 Xylulokinase
429 UTP-glucose-1-phosphate uridylyltransferase
447 Pectinesterase
499 Polygalacturonase
666 Ribulose-phosphate 3-epimerase

Fructose and mannose metabolism

4 L-Iditol 2-dehydrogenase
18 Aldehyde reductase
25 GDP-L-fucose synthase
351 Hexokinase
371 Ketohexokinase
378 Fructokinase
355 6-Phosphofructokinase
390 Pyrophosphate-fructose-6-phosphate 1-phosphotransferase
352 6-Phosphofructo-2-kinase
412 Mannose-1-phosphate guanylyltransferase
416 Mannose-1-phosphate guanylyltransferase (GDP)
419 Fucose-1-phosphate guanylyltransferase
461 Fructose-bisphosphatase
471 Fructose-2,6-bisphosphate 2-phosphatase
610 Fructose-bisphosphate aldolase
634 GDPmannose 4,6-dehydratase
673 Triose-phosphate isomerase
677 Mannose-6-phosphate isomerase
690 Phosphomannomutase

Galactose metabolism

18 Aldehyde reductase
280 Lactose synthase
351 Hexokinase
363 Glucokinase
383 Galactokinase
355 6-Phosphofructokinase
429 UTP-glucose-1-phosphate uridylyltransferase
410 UTP-hexose-1-phosphate uridylyltransferase
411 UDPglucose-hexose-1-phosphate uridylyltransferase
477 Glucose-6-phosphatase
502 alpha-Glucosidase
504 alpha-Galactosidase
505 beta-Galactosidase
508 beta-Fructofuranosidase
494 Lactase
669 UDPglucose 4-epimerase
687 Phosphoglucomutase

Ascorbate and aldarate metabolism

113 Aldehyde dehydrogenase (NAD+)
132 Galactonolactone dehydrogenase
58 L-Ascorbate oxidase

725 Succinate-CoA ligase (GDP-forming)
726 Succinate-CoA ligase (ADP-forming)
722 Acetate-CoA ligase (ADP-forming)
750 Propionyl-CoA carboxylase

Butanoate metabolism

34 (R,R)-Butanediol dehydrogenase
27 3-Hydroxybutyrate dehydrogenase
30 3-Hydroxyacyl-CoA dehydrogenase
8 3-Hydroxybutyryl-CoA dehydrogenase
113 Aldehyde dehydrogenase (NAD+)
108 Succinate-semialdehyde dehydrogenase (NAD(P)+)
111 Succinate-semialdehyde dehydrogenase
121 Pyruvate dehydrogenase (lipoamide)
141 Butyryl-CoA dehydrogenase
220 Acetolactate synthase
249 Acetyl-CoA C-acetyltransferase
252 Hydroxymethylglutaryl-CoA synthase
341 4-Aminobutyrate transaminase
446 3-Oxoacid CoA-transferase
590 Glutamate decarboxylase
618 Hydroxymethylglutaryl-CoA lyase
624 Enoyl-CoA hydratase
723 Butyrate-CoA ligase

C5-Branched dibasic acid metabolism

220 Acetolactate synthase
726 Succinate-CoA ligase (ADP-forming)

Oxidative phosphorylation

138 Succinate dehydrogenase (ubiquinone)
173 NADH dehydrogenase (ubiquinone)
187 Cytochrome-c oxidase
57 Ubiquinol-cytochrome-c reductase
570 Inorganic pyrophosphatase
586 H+-exporting ATPase
584 H+/K+-exchanging ATPase

ATP synthesis

585 H+-transporting two-sector ATPase

Photosynthesis

104 Ferredoxin-NADP+ reductase
585 H+-transporting two-sector ATPase

Methane metabolism

105 Formaldehyde dehydrogenase (glutathione)
110 Formate dehydrogenase
60 Catalase
61 Peroxidase
208 Glycine hydroxymethyltransferase

Carbon fixation

31 Malate dehydrogenase
33 Malate dehydrogenase (decarboxylating)
35 Malate dehydrogenase (oxaloacetate-decarboxylating)(NADP+)
46 Malate dehydrogenase (NADP+)
218 Transketolase
337 Aspartate transaminase
342 Alanine transaminase
362 Phosphoribulokinase
379 Pyruvate kinase
392 Phosphoglycerate kinase
436 Pyruvate,orthophosphate dikinase
461 Fructose-bisphosphatase

234 [Acyl-carrier-protein] S-acetyltransferase
235 [Acyl-carrier-protein] S-malonyltransferase
237 3-Oxoacyl-[acyl-carrier-protein] synthase
246 Fatty-acid synthase
247 Fatty-acyl-CoA synthase
458 Oleoyl-[acyl-carrier-protein] hydrolase
638 3-Hydroxypalmitoyl-[acyl-carrier-protein] dehydratase
737 Biotin carboxylase

Fatty acid biosynthesis (path 2)

30 3-Hydroxyacyl-CoA dehydrogenase
249 Acetyl-CoA C-acetyltransferase
225 Acetyl-CoA C-acyltransferase
458 Oleoyl-[acyl-carrier-protein] hydrolase
624 Enoyl-CoA hydratase

Fatty acid metabolism

1 Alcohol dehydrogenase
30 3-Hydroxyacyl-CoA dehydrogenase
113 Aldehyde dehydrogenase (NAD+)
137 Acyl-CoA oxidase
141 Butyryl-CoA dehydrogenase
142 Acyl-CoA dehydrogenase
145 Glutaryl-CoA dehydrogenase
140 Long-chain-acyl-CoA dehydrogenase
86 Unspecific monooxygenase
87 Alkane 1-monooxygenase
249 Acetyl-CoA C-acetyltransferase
225 Acetyl-CoA C-acyltransferase
227 Carnitine O-palmitoyltransferase
624 Enoyl-CoA hydratase
682 Dodecenoyl-CoA delta-isomerase
724 Long-chain-fatty-acid-CoA ligase

Synthesis and degradation of ketone bodies

27 3-Hydroxybutyrate dehydrogenase
249 Acetyl-CoA C-acetyltransferase
252 Hydroxymethylglutaryl-CoA synthase
446 3-Oxoacid CoA-transferase
618 Hydroxymethylglutaryl-CoA lyase

Sterol biosynthesis

29 Hydroxymethylglutaryl-CoA reductase (NADPH)
127 7-Dehydrocholesterol reductase
134 Lathosterol oxidase
174 NAD(P)H dehydrogenase (quinone)
100 Squalene monooxygenase
98 Carotene 7,8-desaturase
318 Dimethylallyltranstransferase
319 Geranyltranstransferase
324 Farnesyl-diphosphate farnesyltransferase
327 Farnesyltranstransferase
376 Mevalonate kinase
399 Phosphomevalonate kinase
600 Diphosphomevalonate decarboxylase
680 Isopentenyl-diphosphate delta-isomerase
681 Cholestenol delta-isomerase
695 Lanosterol synthase
696 Cycloartenol synthase

Bile acid biosynthesis

1 Alcohol dehydrogenase
41 3alpha-Hydroxysteroid dehydrogenase (B-specific)
113 Aldehyde dehydrogenase (NAD+)
143 3-Oxo-5alpha-steroid 4-dehydrogenase

573 Nucleoside-triphosphatase

Riboflavin metabolism

13 5-Amino-6-(5-phosphoribosylamino)uracil reductase
94 Monophenol monooxygenase
336 Riboflavin synthase
369 Riboflavin kinase
415 FMN adenylyltransferase
465 Acid phosphatase
560 GTP cyclohydrolase II
561 Diaminohydroxyphosphoribosylaminopyrimidine deaminase
583 Nucleotide pyrophosphatase

Vitamin B6 metabolism

120 Aldehyde oxidase
154 Pyridoxamine-phosphate oxidase
348 Phosphoserine transaminase
375 Pyridoxal kinase
643 Threonine synthase

Nicotinate and nicotinamide metabolism

120 Aldehyde oxidase
172 NAD(P)+ transhydrogenase (AB-specific)
188 Nicotinamide N-methyltransferase
299 Purine-nucleoside phosphorylase
301 Nicotinate phosphoribosyltransferase
305 Nicotinate-nucleotide pyrophosphorylase (carboxylating)
366 NAD+ kinase
409 Nicotinamide-nucleotide adenylyltransferase
472 5'-Nucleotidase
525 NAD+ nucleosidase
583 Nucleotide pyrophosphatase
744 NAD+ synthase (glutamine-hydrolysing)

Pantothenate and CoA biosynthesis

48 Ketol-acid reductoisomerase
9 2-Dehydropantoate 2-reductase
126 Dihydropyrimidine dehydrogenase (NADP+)
210 3-Methyl-2-oxobutanoate hydroxymethyltransferase
220 Acetolactate synthase
343 Branched-chain-amino-acid transaminase
367 Dephospho-CoA kinase
374 Pantothenate kinase
546 beta-Ureidopropionase
547 Dihydropyrimidinase
583 Nucleotide pyrophosphatase
641 Dihydroxy-acid dehydratase
728 Pantoate-beta-alanine ligase

Biotin metabolism

349 Adenosylmethionine-8-amino-7-oxononanoate transaminase
440 Biotin synthase
534 Biotinidase
735 Dethiobiotin synthase

Folate biosynthesis

7 Sepiapterin reductase
162 Dihydrofolate reductase
175 Dihydropteridine reductase
202 Thymidylate synthase
320 Dihydropteroate synthase
407 2-amino-4-hydroxy-6-hydroxymethyldihydropteridine diphosphokinase
460 Alkaline phosphatase
531 gamma-Glu-X carboxypeptidase
558 GTP cyclohydrolase I

678 Glucose-6-phosphate isomerase
687 Phosphoglucomutase

N-Glycans biosynthesis

287 Beta-N-acetylglucosaminylglycopeptide beta-1,4-galactosyltransferase
292 Glycoprotein 6-alpha-L-fucosyltransferase
295 Dolichyl-phosphate beta-D-mannosyltransferase
258 alpha-1,3-mannosyl-glycoprotein 2-beta-N-acetylglucosaminyltransferase
261 Dolichyl-phosphate beta-glucosyltransferase
262 Dolichyl-diphosphooligosaccharide-protein glycosyltransferase
265 Glycolipid 2-alpha-mannosyltransferase
269 alpha-1,6-mannosyl-glycoprotein 2-beta-N-acetylglucosaminyltransferase
270 beta-1,4-mannosyl-glycoprotein 4-beta-N-acetylglucosaminyltransferase
271 alpha-1,3-mannosyl-glycoprotein 4-beta-N-acetylglucosaminyltransferase
275 alpha-1,6-mannosyl-glycoprotein 6-beta-N-acetylglucosaminyltransferase
312 beta-Galactoside alpha-2,6-sialyltransferase
354 Dolichol kinase
432 chitobiosylpyrophosphoryldolichol synthase
493 Mannosyl-oligosaccharide glucosidase
495 Mannosyl-oligosaccharide 1,2-alpha-mannosidase
496 Mannosyl-oligosaccharide 1,3-1,6-alpha-mannosidase

N-Glycan degradation

500 Exo-alpha-sialidase
505 beta-Galactosidase
506 alpha-Mannosidase
507 beta-Mannosidase
519 alpha-L-Fucosidase
520 beta-N-Acetylhexosaminidase
524 Mannosyl-glycoprotein endo-beta-N-acetylglucosaminidase
541 N4-(beta-N-Acetylglucosaminyl)-L-asparaginase

O-Glycans biosynthesis

289 Polypeptide N-acetylgalactosaminyltransferase
259 beta6-N-acetylglucosaminyltransferase
263 Glycoprotein-N-acetylgalactosamine 3-beta-galactosyltransferase
313 beta-Galactoside alpha-2,3-sialyltransferase
315 sialyltransferase

Aminosugars metabolism

236 Glucosamine-phosphate N-acetyltransferase
276 Chitin synthase
340 Glutamine-fructose-6-phosphate transaminase (isomerizing)
351 Hexokinase
382 N-Acetylglucosamine kinase
384 N-Acylmannosamine kinase
417 UDP-N-acetylglucosamine pyrophosphorylase
423 N-Acylneuraminate cytidylyltransferase
497 Chitinase
520 beta-N-Acetylhexosaminidase
540 N-Acetylglucosamine-6-phosphate deacetylase
543 Chitin deacetylase
568 Glucosamine-6-phosphate deaminase
671 N-Acylglucosamine 2-epimerase
668 UDP-N-acetylglucosamine 2-epimerase
687 Phosphoglucomutase
688 Phosphoacetylglucosamine mutase

Glycosaminoglycan degradation

488 N-Acetylgalactosamine-6-sulfatase
484 N-Acetylgalactosamine-4-sulfatase
485 Iduronate-2-sulfatase
486 N-Acetylglucosamine-6-sulfatase
505 beta-Galactosidase
511 beta-Glucuronidase

607 Phosphatidylserine decarboxylase
673 Triose-phosphate isomerase

Inositol phosphate metabolism

73 myo-Inositol oxygenase
385 1-Phosphatidylinositol 4-kinase
386 1-Phosphatidylinositol-4-phosphate kinase
358 1D-myo-Inositol-triphosphate 3-kinase
359 1-Phosphatidylinositol 3-kinase
466 myo-Inositol-1(or 4)-monophosphatase
473 Inositol-1,4,5-trisphosphate 5-phosphatase
474 Inositol-1,4-bisphosphate 1-phosphatase
475 Inositol-3,4-bisphosphate 4-phosphatase
478 1-Phosphatidylinositol-4,5-bisphosphate phosphodiesterase
698 myo-Inositol-1-phosphate synthase

Glycosylphosphatidylinositol(GPI)-anchor biosynthesis 03-4-10

279 Phosphatidylinositol N-acetylglucosaminyltransferase

Sphingophospholipid biosynthesis

430 Ethanolaminephosphotransferase
479 Sphingomyelin phosphodiesterase

Phospholipid degradation

452 Prophospholipase A2
455 Lisophospholipase
469 Phosphatidate phosphatase
481 Phospholipase D
482 Glycerophosphodiester phosphodiesterase

Prostaglandin and leukotriene metabolism

10 Carbonyl reductase (NADPH)
11 Prostaglandin-F synthase
12 Prostaglandin-E2 9-reductase
67 Arachidonate 12-lipoxygenase
68 Arachidonate 15-lipoxygenase
69 Arachidonate 5-lipoxygenase
83 Leukotriene-B4 20-monooxygenase
95 Prostaglandin-endoperoxide synthase
250 gamma-Glutamyltransferase
329 Leukotriene-C4 synthase
452 Prophospholipase A2
528 Leukotriene-A4 hydrolase
642 Hydroperoxide dehydratase
683 Prostaglandin-D synthase
684 Prostaglandin-I synthase
685 Thromboxane-A synthase

Sphingoglycolipid metabolism

238 Serine C-palmitoyltransferase
290 Ganglioside galactosyltransferase
294 Ceramide glucosyltransferase
442 Galactosylceramide sulfotransferase
483 Arylsulfatase
489 Cerebroside-sulfatase
500 Exo-alpha-sialidase
504 alpha-Galactosidase
505 beta-Galactosidase
515 Glucosylceramidase
539 Ceramidase
612 Sphinganine-1-phosphate aldolase

Blood group glycolipid biosynthesis - lact series

286 Fucosylglycoprotein 3-alpha-galactosyltransferase
288 Fucosylglycoprotein alpha-N-acetylgalactosaminyltransferase

250 gamma-Glutamyltransferase
335 Methionine adenosyltransferase
330 Cysteine synthase
331 Cystathionine gamma-synthase
368 Adenylylsulfate kinase
421 Sulfate adenylyltransferase
438 Selenide,water dikinase
526 Adenosylhomocysteinase
628 Cystathionine beta-synthase
656 Cystathionine gamma-lyase
660 Cystathionine beta-lyase
701 Methionine-tRNA ligase

Cyanoamino acid metabolism

208 Glycine hydroxymethyltransferase
250 gamma-Glutamyltransferase
503 beta-Glucosidase
532 Asparaginase
542 Amidase
567 Nitrilase

D-Glutamine and D-glutamate metabolism

148 Glutamate dehydrogenase (NAD(P)+)
538 Glutaminase

D-Arginine and D-ornithine metabolism

107 Glyceraldehyde-3-phosphate dehydrogenase (phosphorylating)
152 D-Amino-acid oxidase

D-Alanine metabolism

664 Alanine racemase

Glutathione metabolism

37 Isocitrate dehydrogenase (NADP+)
40 Glucose-6-phosphate 1-dehydrogenase
182 Glutathione-disulfide reductase
63 Glutathione peroxidase
250 gamma-Glutamyltransferase
322 Glutathione transferase
529 Membrane alanyl aminopeptidase
550 5-Oxoprolinase (ATP-hydrolysing)
730 Glutamate-cysteine ligase
731 Glutathione synthase

Purine metabolism

16 Xanthine dehydrogenase
17 IMP dehydrogenase
53 Xanthine oxidase
177 GMP reductase
178 Urate oxidase
103 Ribonucleoside-diphosphate reductase
211 Phosphoribosylglycinamide formyltransferase
212 Phosphoribosylaminoimidazolecarboxamide formyltransferase
299 Purine-nucleoside phosphorylase
308 Thymidine phosphorylase
309 Adenine phosphoribosyltransferase
310 Hypoxanthine phosphoribosyltransferase
302 Amidophosphoribosyltransferase
364 Adenosine kinase
368 Adenylylsulfate kinase
379 Pyruvate kinase
388 Deoxycitidine kinase
356 Deoxyguanosine kinase
400 Adenylate kinase
401 Nucleoside-diphosphate kinase

555 Cytosine deaminase
564 Cytidine deaminase
557 dCMP deaminase
580 Apyrase
581 Nucleoside-diphosphatase
574 Bis(5'-nucleosyl)-tetraphosphatase (asymmetrical)
575 Nucleoside-triphosphate pyrophosphatase
576 dUTP pyrophosphatase
595 Orotidine-5'-phosphate decarboxylase
639 Pseudouridylate synthase
739 CTP synthase
748 Carbamoyl-phosphate synthase (glutamine-hydrolysing)

Nucleotide sugars metabolism

20 UDPglucose 6-dehydrogenase
429 UTP-glucose-1-phosphate uridylyltransferase
410 UTP-hexose-1-phosphate uridylyltransferase
411 UDPglucose-hexose-1-phosphate uridylyltransferase
633 dTDPglucose 4,6-dehydratase
669 UDPglucose 4-epimerase
667 dTDP-4-dehydrorhamnose 3,5-epimerase

Supplementary results 1

I analyzed the difference of the metabolic networks between the 6 eukaryotes based on the existence of ERs in detail. In this study, I separated 751 ERs into 104 metabolic networks. As results, I discovered the following four patterns for explaining the difference of the metabolic networks between species. Therefore, I show the examples of these patterns in supplementary results 1.

Pattern 1: Conservation

In this pattern, the sets of ERs constructing the metabolic network are conserved between all the 6 eukaryotes examined. There are only three out of 104 metabolic networks, ATP synthesis, 1,2-Dichloroethane degradation, and phospholipid degradation. As an example, I show the metabolic network of the phospholipids degradation. In this network, a total of five ERs are found in all the 6 eukaryotes.

Pattern 2: Alternative conservation

In this pattern, the metabolic network is different between the 6 eukaryotes examined by the existence of the alternative pathway. Therefore, the difference of the sets of the ER affects the difference of the metabolic networks but not affects the capability of the metabolic networks between the 6 eukaryotes examined. In other words, the losses of ERs in either lineage were complemented by the alternative pathways by the different ERs. As examples, I show the metabolic networks of the glycolysis and the citrate cycle (TCA cycle).

Glycolysis

The sets of ERs for the glycolysis are different between the 6 eukaryotes, however, the metabolic pathways from D-glucose to pyruvate or acetyl-CoA are functional among them.

Citrate cycle (TCA cycle)

Even if the sets of ERs for the citrate cycle are different between the 6 eukaryotes, the degrading pathway of pyruvate is functional among them.

Pattern 3: ER losses

In this pattern, the extremely losses of ERs are found in the metabolic networks in the particular lineages. The deficiency of the metabolic networks by the losses of ERs correlates well with the acquirement of the products in these networks as nourishment. As examples, I show the metabolic networks of the valine, leucine and isoleucine biosynthesis and the phenylalanine, tyrosine and tryptophan biosynthesis.

Valine, leucine and isoleucine biosynthesis

The five losses of ERs in this network were observed in the animal lineage. This result suggests that this network may not be functional in the animal lineages and, in fact, is consistent with the acquirements of valine, leucine and isoleucine as the essential amino acids for the animals.

Phenylalanine, tyrosine and tryptophan biosynthesis

In the case of plant and yeasts lineages, loss of ER is complemented by the different

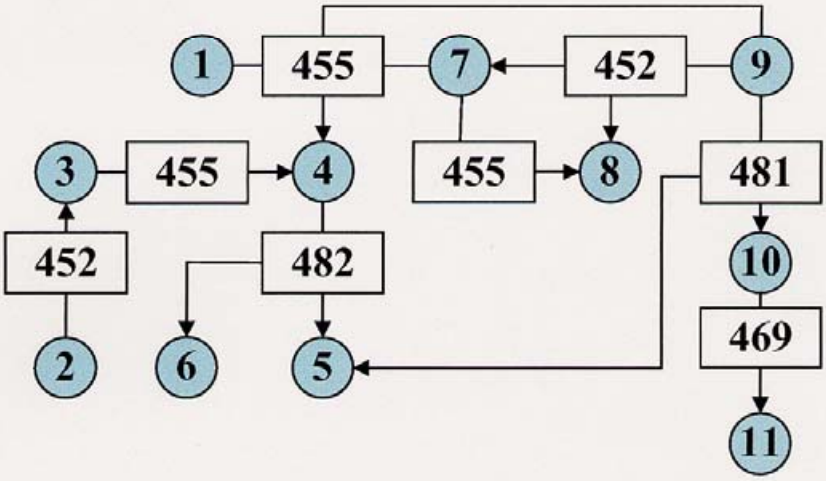
ERs constructing the alternative pathways. On the other hand, a total of 16 losses of ERs are found in the animal lineages. This result suggests that this network may not be functional in the animal lineages. In fact, phenylalanine and tryptophan are the essential amino acid for the animals and tyrosine is not an essential amino acid by the existence of ER90, phenylalanine 4-monooxygenase, in the animal lineages.

Pattern 4: ER gains

In this pattern, the extremely gains of ERs were found in the metabolic networks in the particular lineages. It is considered that the gains of ERs give the species the capability to make the novel products. In my thesis, I have already explained about the metabolic network of prostaglandin and leukotriene metabolism as the example of this pattern. As another example, I show the metabolic network of flavonoid, stilbene and lignin biosynthesis.

Flavonoid, stilbene and lignin biosynthesis

Six ERs are found only in *A. thaliana* in this study. This means that gains of these ERs occurred only in the plant lineage. Since flavonoids are the plant pigments and essential compounds for plant reproduction, the gains of ERs could have contributed to the establishment of the plant-specific system of the reproduction.



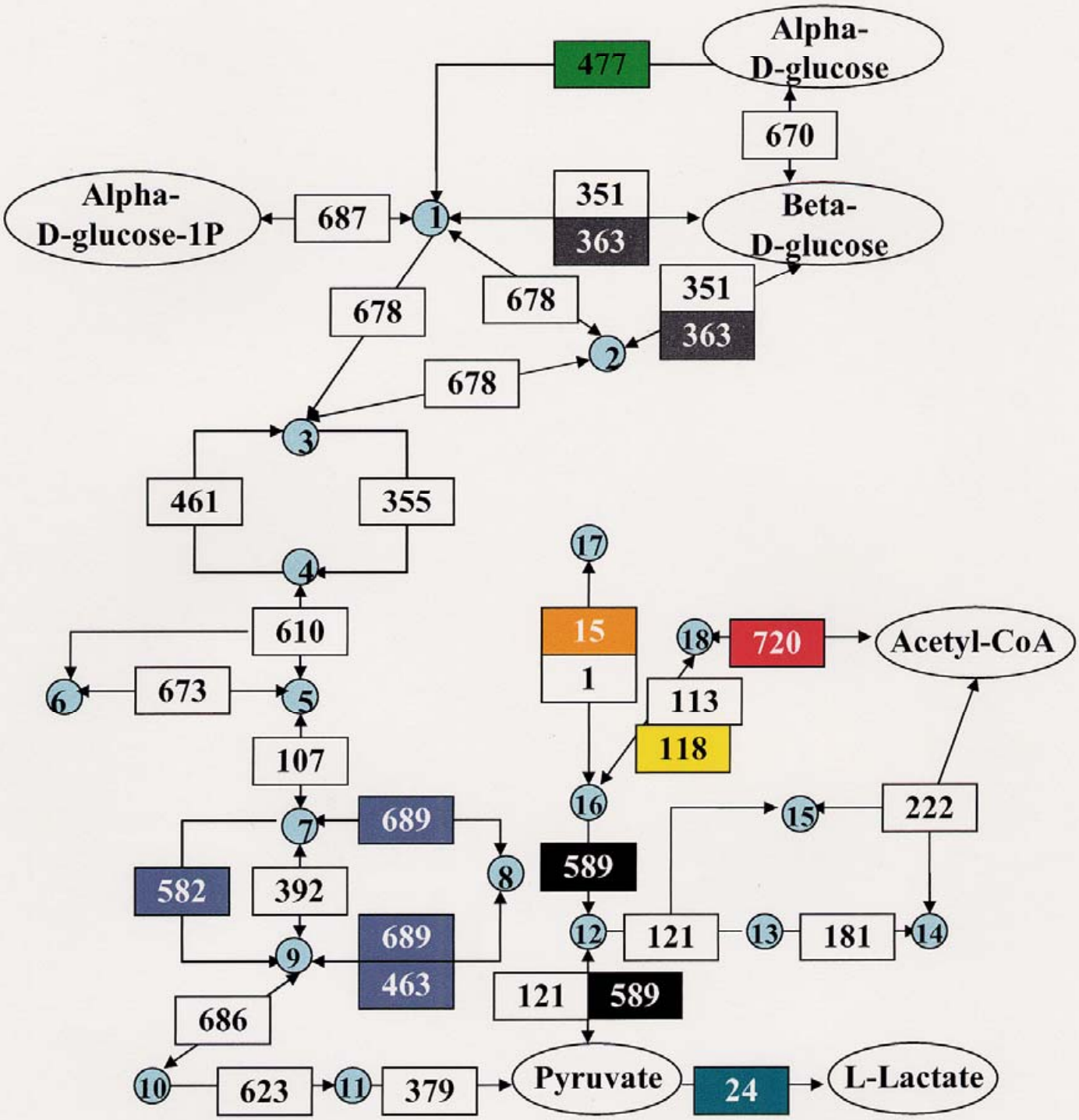
Phospholipid degradation

Square indicates the enzymatic reaction existing in the 6 eukaryotes examined and circle indicates the substance.

Species-ER matrix for the Phospholipid degradation

		Species						
ER	Enzymes involved with the ER	H SA	DME	CEL	SCE	SPO	ATH	OUT
452	Prophospholipase A2	1	1	1	1	1	1	0
455	Lisophospholipase	1	1	1	1	1	1	1
469	Phosphatidate phosphatase	1	1	1	1	1	1	0
481	Phospholipase D	1	1	1	1	1	1	0
482	Glycerophosphodiester phosphodiesterase	1	1	1	1	1	1	1

List of substrates in the phospholipid degradation
1: 2-Acylglycero-3-phosphocholine
2: 1-Alkenyl-2-acyl-glycero-3-phosphocholine
3: 1-Alkenyl-glycero-3-phosphocholine
4: Glycero-3-phosphocholine
5: Choline
6: Clycero-3P
7: 1-Acylglycero-3-phosphocholine
8: Fatty acid
9: Lecithin
10: 1,2-Diacylglycero-3P
11: 1,2-Diacylglycerol



Glycolysis

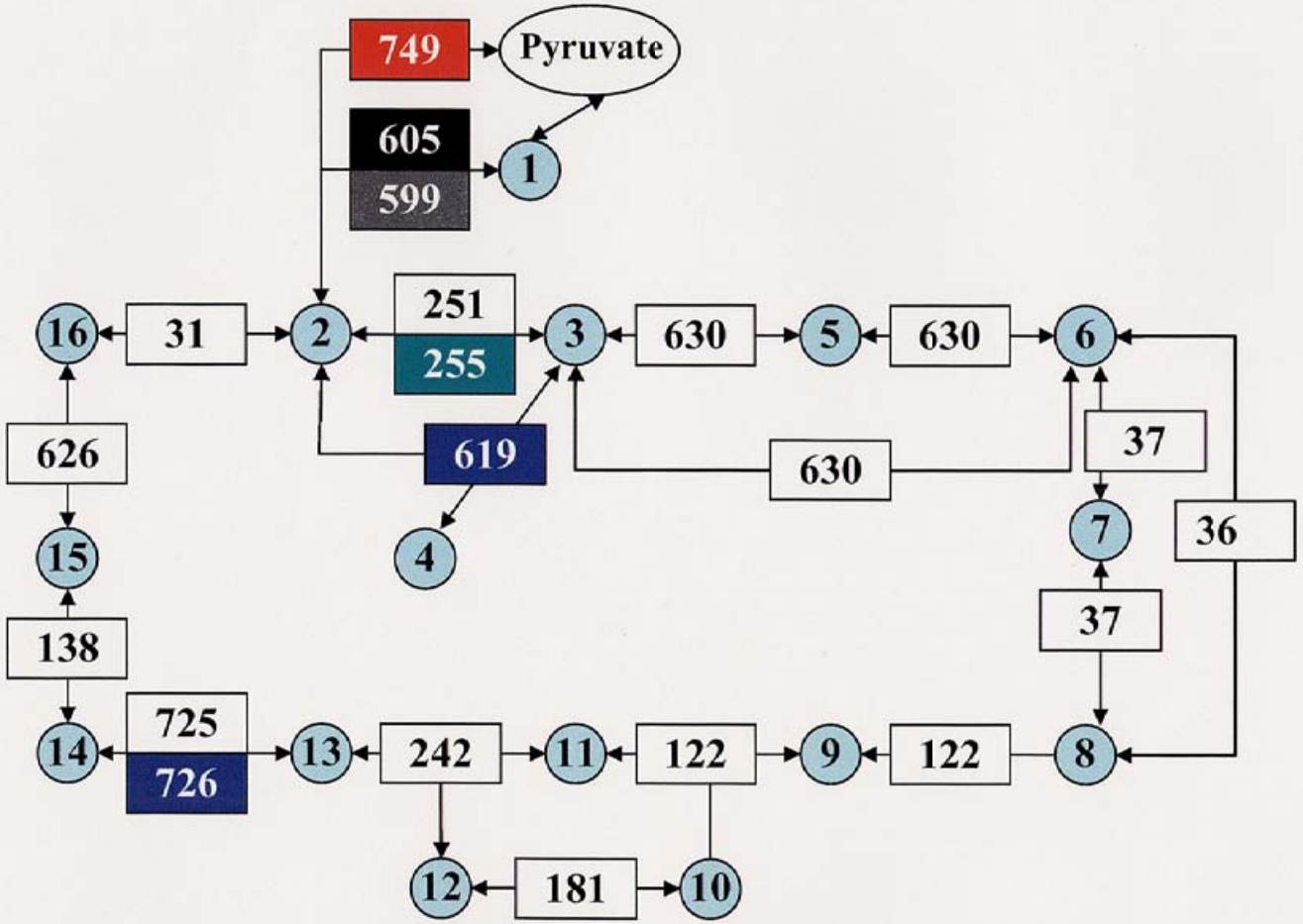
Square indicates the enzymatic reaction and circle indicates the substance. The color of each squares correspond to the status changes of genes encoding the protein that is involved with the ER. White : Exist in all the 6 eukaryotes. Black: Lost in the animal lineages. Grey: Lost in the all lineages except for the *S. cerevisiae*. Grey blue: Lost in the yeasts and *C. elegans* lineages. Orange: Lost in the *D. melanogaster* lineage. Dark green: Lost in the *S. cerevisiae* lineage. Green: Gain in the vertebrate and insect lineages. Yellow: Gain in the yeast and animal lineage. Number in the circles indicate the name of the substrates (See the next page).

Species-ER matrix for the glycolysis

		Species						
ER	Enzymes involved with the ER	H SA	DME	CEL	SCE	SPO	ATH	OUT
1	Alcohol dehydrogenase	1	1	1	1	1	1	1
15	Alcohol dehydrogenase(NADP+)	1	0	1	1	1	1	1
24	L-Lactate dehydrogenase	1	1	1	0	1	1	1
107	Glyceraldehyde-3-phosphate dehydrogenase (phosphorylating)	1	1	1	1	1	1	1
113	Aldehyde dehydrogenase (NAD+)	1	1	1	1	1	1	1
118	Aldehyde dehydrogenase (NAD(P)+)	1	1	1	1	1	0	0
121	Pyruvate dehydrogenase (lipoamide)	1	1	1	1	1	1	1
181	Dihydrolipoamide dehydrogenase	1	1	1	1	1	1	1
222	Dihydrolipoamide S-acetyltransferase	1	1	1	1	1	1	1
351	Hexokinase	1	1	1	1	1	1	1
355	6-Phosphofructokinase	1	1	1	1	1	1	1
363	Glucokinase	0	0	0	1	0	0	1
379	Pyruvate kinase	1	1	1	1	1	1	1
392	Phosphoglycerate kinase	1	1	1	1	1	1	1
461	Fructose-bisphosphatase	1	1	1	1	1	1	1
463	Bisphosphoglycerate phosphatase	1	1	0	0	0	1	0
477	Glucose-6-phosphatase	1	1	0	0	0	0	0
582	Acylphosphatase	1	1	0	0	0	1	1
589	Pyruvate decarboxylase	0	0	0	1	1	1	1
610	Fructose-bisphosphate aldolase	1	1	1	1	1	1	1
623	Phosphopyruvate hydratase	1	1	1	1	1	1	1
670	Aldose 1-epimerase	1	1	1	1	1	1	1
673	Triose-phosphate isomerase	1	1	1	1	1	1	1
678	Glucose-6-phosphate isomerase	1	1	1	1	1	1	1
686	Phosphoglycerate mutase	1	1	1	1	1	1	1
687	Phosphoglucomutase	1	1	1	1	1	1	1
689	Bisphosphoglycerate mutase	1	1	0	0	0	1	0
720	Acetate-CoA ligase	1	1	1	1	0	1	1

Substrates in Glycolysis

1: alpha-D-Glucose-6P 2: beta-D-Glucose-6P 3: beta-D-Fructose-6P
4: beta-D-Fructose-1,6P2 5: Glycraldehyde-3P 6: Glycerone-P
7: Glycerate-1,3P2 8: Glycerate-2,3P2 9: Glycerate-3P
10: Glycerate-2P 11: Phosphoenolpyruvate 12: 2-Hydroxy-ethyl-ThPP
13: Lipoamide 14: Dihydrolipoamide 15: 6-S-Acetyl-dihydrolipoamide
16: Acetaldehyde 17: Ethanol 18: Acetate



Citrate cycle (TCA cycle)

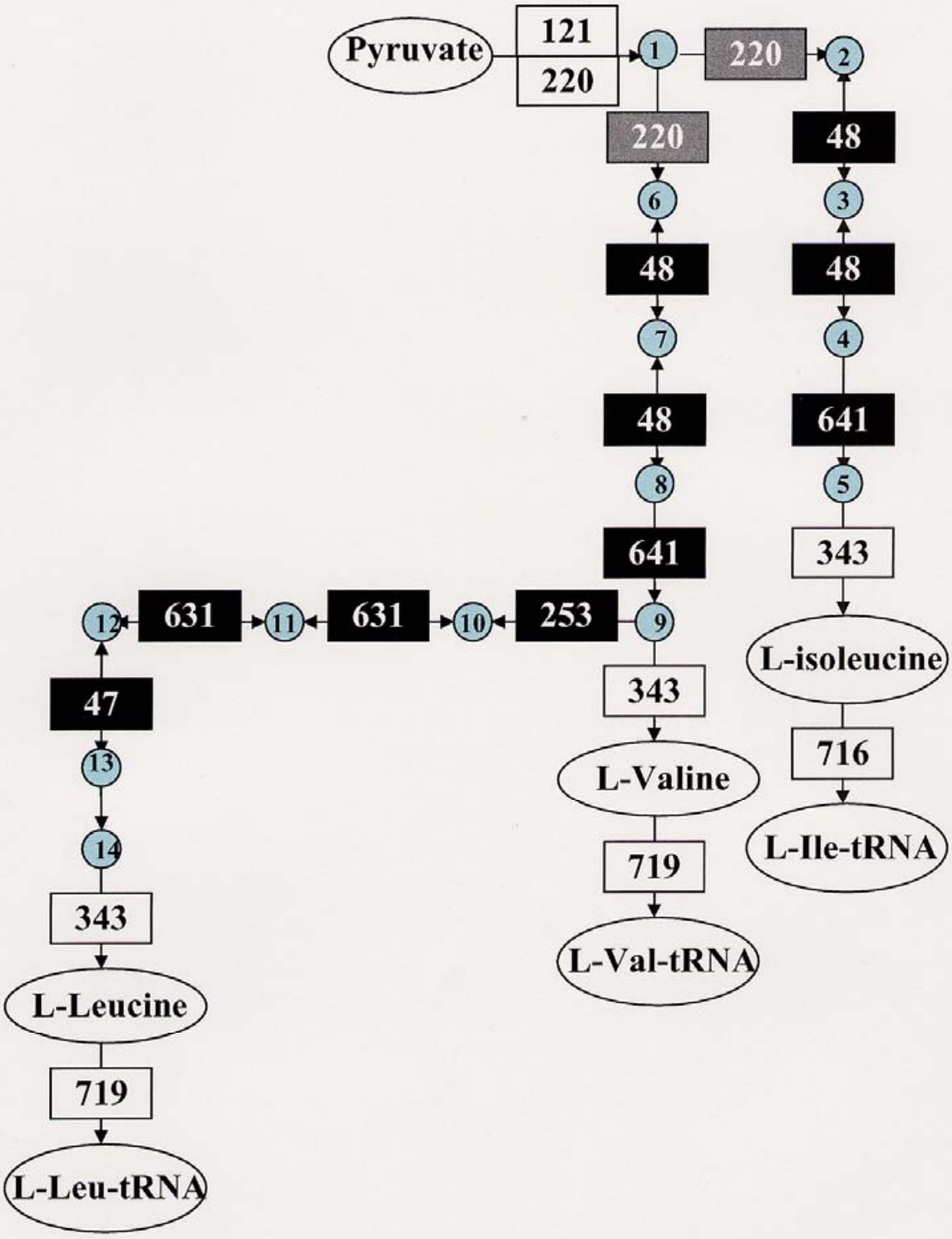
Square indicates the enzymatic reaction and circle indicates the substance. White : Exist in all the 6 eukaryotes. Black: Lost in the animals and *S. pombe* lineages. Grey: Lost in the plant and yeasts lineages. Dark Blue: Lost in the plant, yeasts and *D. melanogaster* lineages. Orange: Lost in the *A. thaliana*. Green: Lost in the *S.* cerevisiae lineage. Number in the circles indicate the name of the substrates (See the next page).

Species-ER matrix for the Citrate cycle (TCA cycle)

		Species						
ER	Enzymes involved with the ER	H SA	DME	CEL	SCE	SPO	ATH	OUT
31	Malate dehydrogenase	1	1	1	1	1	1	1
36	Isocitrate dehydrogenase (NAD+)	1	1	1	1	1	1	1
37	Isocitrate dehydrogenase (NADP+)	1	1	1	1	1	1	1
122	Oxoglutarate dehydrogenase (lipoamide)	1	1	1	1	1	1	1
138	Succinate dehydrogenase (ubiquinone)	1	1	1	1	1	1	1
181	Dihydrolipoamide dehydrogenase	1	1	1	1	1	1	1
242	Dihydrolipoamide S-succinyltransferase	1	1	1	1	1	1	1
251	Citrate (Si)-synthase	1	1	1	1	1	1	1
255	ATP citrate synthase	1	1	1	0	1	1	0
599	Phosphoenolpyruvate carboxykinase (GTP)	1	1	1	0	0	0	1
605	Phosphoenolpyruvate carboxykinase (ATP)	0	0	0	1	0	1	1
619	Citrate lyase	1	0	1	0	0	0	1
626	Fumarate hydratase	1	1	1	1	1	1	1
630	Aconitate hydratase	1	1	1	1	1	1	1
725	Succinate-CoA ligase (GDP-forming)	1	1	1	1	1	1	0
726	Succinate-CoA ligase (ADP-forming)	1	0	1	0	0	0	1
749	Pyruvate carboxylase	1	1	1	1	1	0	1

List of the substrates in the citrate cycle (TCA cycle)

1: Phosphoenol-pyruvate
2: Oxaloacetate
3: Citrate
4: Acetate
5: cis-Acetate
6: Isocitrate
7: Oxalosuccinate
8: 2-Oxoglutarate
9: 3-Carbosy-1-hyddroxypropyl-ThPP
10: Lipoamide
11: S-Succinyl-dihydrolipoamide
12: Dihydrolipoamide
13: Succinyl-CoA
14: Succinate
15: Fumarate
16: (S)-Malate



Valine, leucine and isoleucine biosynthesis

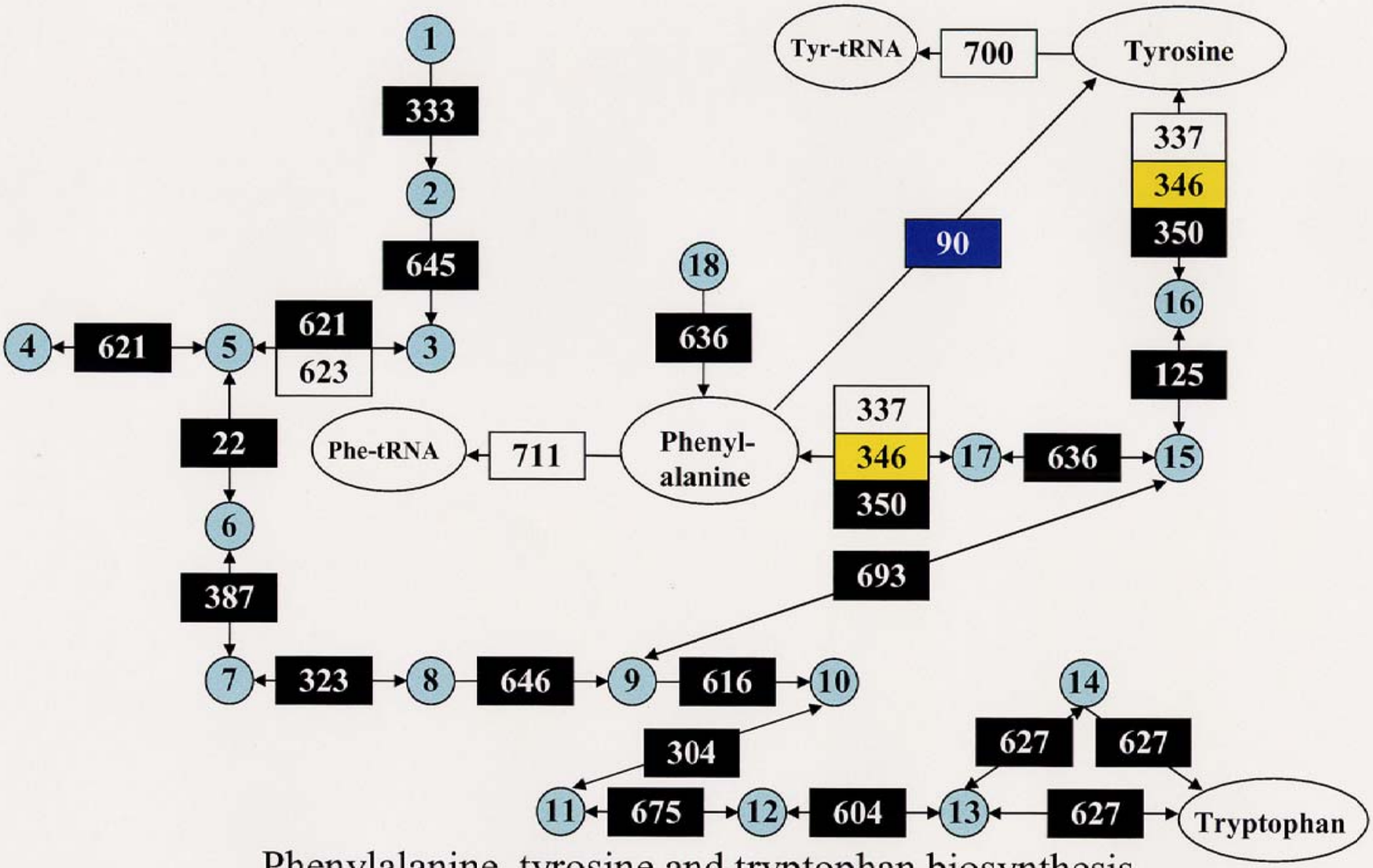
Square indicates the enzymatic reaction and circle indicates the substance. The color of each squares correspond to the status changes of genes encoding the protein that is involved with the ER. White : Exist in all the 6 eukaryotes. Black: Lost in the animal lineage. Grey: Lost in the animals and *S. pombe* lineage. Number in the circles indicate the name of the substrates (See the next page).

Species-ER matrix for the valine, leucine and isoleucine biosynthesis

		Species						
ER	Enzymes involved with the ER	H SA	DME	CEL	SCE	SPO	ATH	OUT
47	3-Isopropylmalate dehydrogenase	0	0	0	1	1	1	1
48	Ketol-acid reductoisomerase	0	0	0	1	0	1	1
121	Pyruvate dehydrogenase (lipoamide)	1	1	1	1	1	1	1
220	Acetolactate synthase	1	1	1	1	1	1	1
253	2-isopropylmalate synthase	0	0	0	1	1	1	1
343	Branched-chain-amino-acid transaminase	1	1	1	1	1	1	1
631	3-Isopropylmalate dehydratase	0	0	0	1	1	1	1
641	Dihydroxy-acid dehydratase	0	0	0	1	1	1	1
715	Leucine-tRNA ligase	1	1	1	1	1	1	1
716	Isoleucine-tRNA ligase	1	1	1	1	1	1	1
719	Valine-tRNA ligase	1	1	1	1	1	1	1

List of the substrates in the valine, leucine and isoleucine biosynthesis

1: 2-Hydroxyethyl-ThPP
2: (S)-2-Aceto-2-hydroxybutanoate
3: (S)-2-Hydroxy-3-methyl-3-oxopentanoate
4: (R)-2,3-Dihydroxy-3-methylpentanoate
5: (R)-2-Oxo-3-methyl-pentanoate
6: (S)-2-Acetolactate
7: (R)-3-Hydroxy-3-methyl-2-oxobutanoate
8: (R)-2,3-Dihydrosy-3-methylbutanoate
9: 2-Oxoisovalerate
10: 2-Isopropylmalate
11: 2-Isopropylmaleate
12: 3-Isopropylmalate
13: 2-Oxo-4-methyl-3-carboxypentanoate
14: e-Methyl-w-oxopentanoate



Phenylalanine, tyrosine and tryptophan biosynthesis

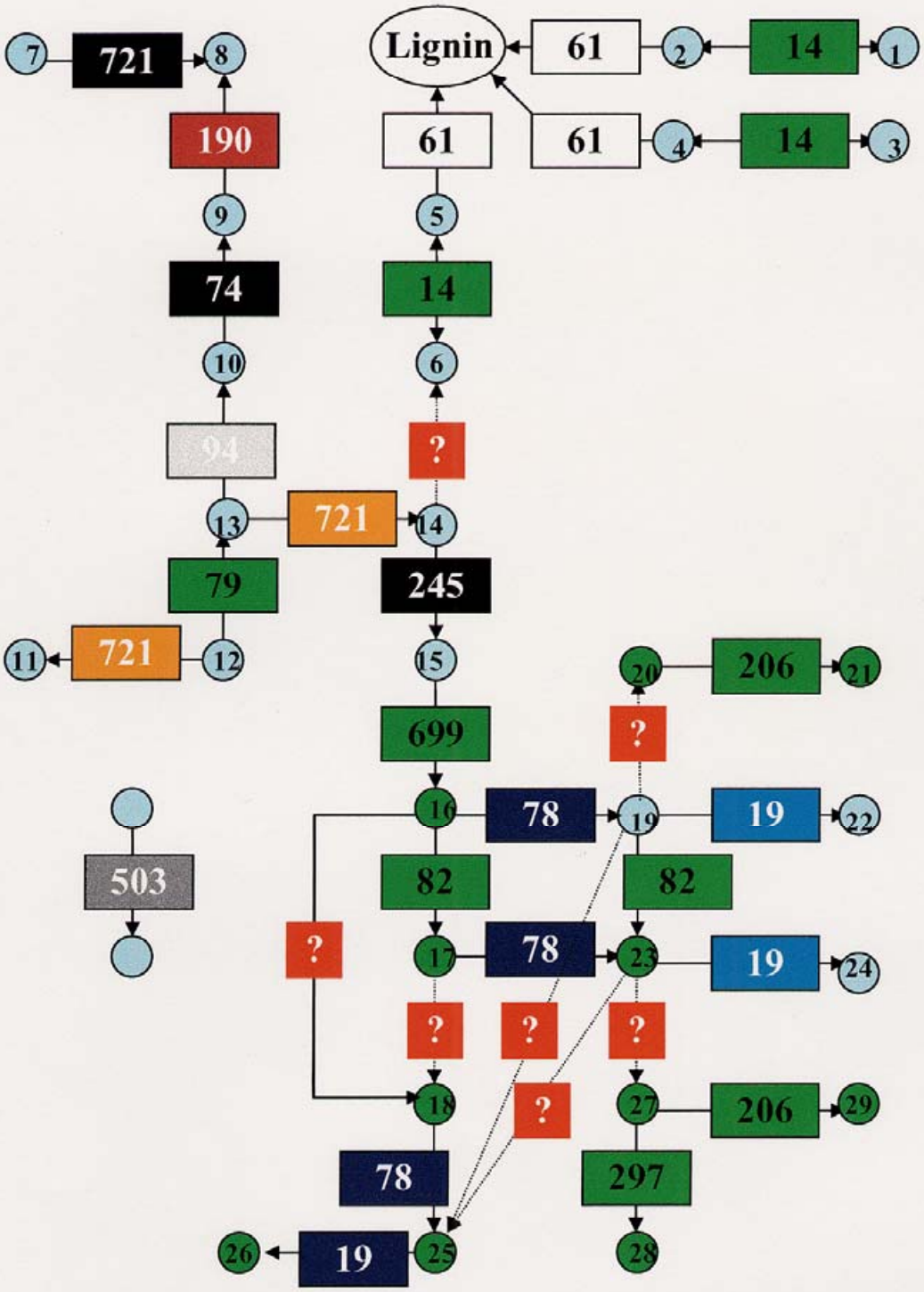
Square indicates the enzymatic reaction and circle indicates the substance. Black : Lost in the animal lineage. White : Exist in all the 6 eukaryotes. Yellow : Lost in the *S. pombe* lineage. Blue : Lost in the plant and yeast lineages. Number in the circles indicate the name of the substrates (See the next page).

Species-ER matrix for the phenylalanine, tyrosine and tryptophan biosynthesis

		Species						
ER	Enzymes involved with the ER	H SA	DME	CEL	SCE	SPO	ATH	OUT
22	Shikimate 5-dehydrogenase	0	0	0	1	1	1	1
90	Phenylalanine 4-monooxygenase	1	1	1	0	0	0	1
125	Prephenate dehydrogenase (NADP+)	0	0	0	1	1	1	0
304	Anthranilate phosphoribosyltransferase	0	0	0	1	1	1	1
323	3-Phosphoshikimate 1-carboxyvinyltransferase	0	0	0	1	1	1	1
333	3-deoxy-7-phosphoheptulonate synthase	0	0	0	1	1	1	1
337	Aspartate transaminase	1	1	1	1	1	1	1
	Tyrosine transaminase	1	1	1	0	1	1	1
350	Histidinol-phosphate transaminase	0	0	0	1	1	1	1
387	Shikimate kinase	0	0	0	1	1	1	1
604	Indole-3-glycerol-phosphate synthase	0	0	0	1	1	1	1
616	Anthranilate synthase	0	0	0	1	1	1	1
621	3-Dehydroquinate dehydratase	0	0	0	1	1	1	1
623	Phosphopyruvate hydratase	1	1	1	1	1	1	1
627	Tryptophan synthase	0	0	0	1	1	1	1
636	Prephenate dehydratase	0	0	0	1	1	1	1
645	3-dehydroquinate synthase	0	0	0	1	1	1	1
646	Chorismate synthase	0	0	0	1	1	1	1
675	Phosphoribosylanthranilate isomerase	0	0	0	1	1	1	1
693	Chorismate mutase	0	0	0	1	1	1	1
700	Tyrosine-tRNA ligase	1	1	1	1	1	1	1
711	Phenylalanine-tRNA ligase	1	1	1	1	1	1	1

List of the substrates in the phenylalanine, tyrosine and tryptophan biosynthesis

1: Phosphoenol-pyruvate and D-Erythrose-4-phosphate
2: 7P-2-Dehydro-3-deoxy-D-arabino-heptonate 3: 3-Dehydro-quinate
4: Protocatechuate 5: 3-Dehydro-shikimate 6: Shikimate
7: Shikimate-3-phosphate 8: 5-O-(1-Carboxyvinyl)-3-phosphoshikimate
9: Chorismate 10: Anthranilate
11: N-(5-Phospho-beta-D-ribosyl)-anthranilate
12: 1-(2-Carbosy-phenylamino)-1-deoxy-D-ribulose-5-phosphate
13: (3-Indolyl)-glycerol-phosphate 14: Indole 15: Prephenate
16: 4-Hydrosy-phenylpyruvate 17: Phenylpyruvate 18: Pretyrosine



Flavonoid, stilbene and lignin biosynthesis

Square indicates the enzymatic reaction and circle indicates the substance. The color of each squares correspond to the status changes of genes encoding the protein that is involved with the ER. White : Exist in all the 6 eukaryotes. Black: Lost in the animals and yeasts lineage. Dark grey: Lost in the *H. sapiens, D. melanogaster* and S. cerevisiae lineages. Pale grey: Lost in plant and yeasts lineages. Dark blue: Lost in the *H. sapiens, C. elegans* and *S. cerevisiae* lineages. Pale blue: Lost in the *H. sapiens* and *D. melanogaster* lineage. Purple: Lost in the *D. melanogaster* and yeasts lineages. Orange: Lost in the *S. pombe* lineage. Green: Gain in the *A. thaliana* lineage. "?" indicates that the enzyme corresponding to the ER is unknown in this study. Circle colored in the green indicates the Flavonoids. Number in the circles indicate the name of the substrates (See the next page).

Species-ER matrix for the flavonoids, stilbene and lignin biosynthesis

		Species						
ER	Enzymes involved with the ER	H SA	DME	CEL	SCE	SPO	ATH	OUT
14	Cinnamyl-alcohol dehydrogenase	0	0	0	0	0	1	0
19	Dihydrokaempferol 4-reductase	0	0	1	1	1	1	1
61	Peroxidase	1	1	1	1	1	1	0
78	Naringenin 3-dioxygenase	0	1	0	0	1	1	0
79	trans-Cinnamate 4-monooxygenase	0	0	0	0	0	1	0
82	Flavonoid 3'-monooxygenase	0	0	0	0	0	1	0
94	Monophenol monooxygenase	1	1	1	0	0	0	1
190	Caffeoyl-CoA O-methyltransferase	1	0	1	0	0	1	1
206	Quercetin 3-O-methyltransferase	0	0	0	0	0	1	0
245	Naringenin-chalcone synthase	0	0	0	0	0	1	1
297	Flavonol 3-O-glucosyltransferase	0	0	0	0	0	1	0
503	beta-Glucosidase	0	0	1	0	1	1	1
699	Chalcone isomerase	0	0	0	0	0	1	0
721	4-Coumarate-CoA ligase	1	1	1	1	0	1	0

List of the substrates in the flavonoid, stilbene and lignin biosynthesis

1: Sinapoyl aldehyde
2: Sinapoyl alcohol
3: Coniferyl aldehyde
4: Coniferyl alcohol
5: 4-Hydroxy-cinnamyl alcohol
6: 4-Hydroxy-cinnamyl aldehyde
7: Ferulate
8: Feruloyl-CoA
9: Caffeoyl-CoA
10: trans-Caffeate
11: Cinnamoyl-CoA
12: trans-Cinnamate
13: trans-4-Hydroxy-cinnamate
14: 4-Hydroxy-cinnamoyl-CoA
15: Naringenin-chalcone
16: Naringenin
17: Eriodictyol
18: Pentahydroxy-flavanone
19: Dihydrokaempferol
20: Kaempferol
21: 3-Methosyapigenin
22: cis-3,4-Leucopelargondin
23: Dihydroquercetin
24: Leucocyanidin
25: Dihydromyricetin
26: Leucodelphinidin
27: Quercetin
28: Quercetin-3-O-glucoside
29: 3-Methoxy-luteolin

The name of substrates in bold indicate the flavonoids.

Supplementary results 2

I evaluated the possibility of the emergence of ERs in particular lineage by conducting the homology search of the 58 protein sequences discovered in only one species examined to the non redundant protein database as of Jan. 7th 2004 in NCBI. After making the 58 phylogenetic trees using the matched sequences, I categorized each tree into the following four patterns named origination, duplication, diversification and false-positive.

Origination means that homologous sequence of the protein involved with the ER is found only in the ER-gained lineage. For example, ER207, histamine N-methyltransferase, was found in *H. sapiens* between the 6 eukaryotes examined and I concluded that the ER had gained in the vertebrate lineage after the separation from the *D. melanogaster*. As a result of the homology search, homologous sequences of the proteins involved in ER207 were found only in the vertebrate lineages. Therefore, I conclude that ER207 was gained in the vertebrate lineage.

Duplication means that the gained ER is caused by the gene duplication only in the ER-gained lineage. For example, ER96, steroid 21alpha-monooxygenase, was found only in *H. sapiens* in this study and I concluded that the ER had been gained in the vertebrate lineage after the separation from the *D. melanogaster*. From the topology of the phylogenetic tree, it was found that the gene involved in the ER generated by the gene duplication from the other ER, steroid 17alpha-monooxygenase. Therefore, I conclude that ER96 was gained in the vertebrate lineage by the gene duplication and neo-functionalization of the duplicated gene.

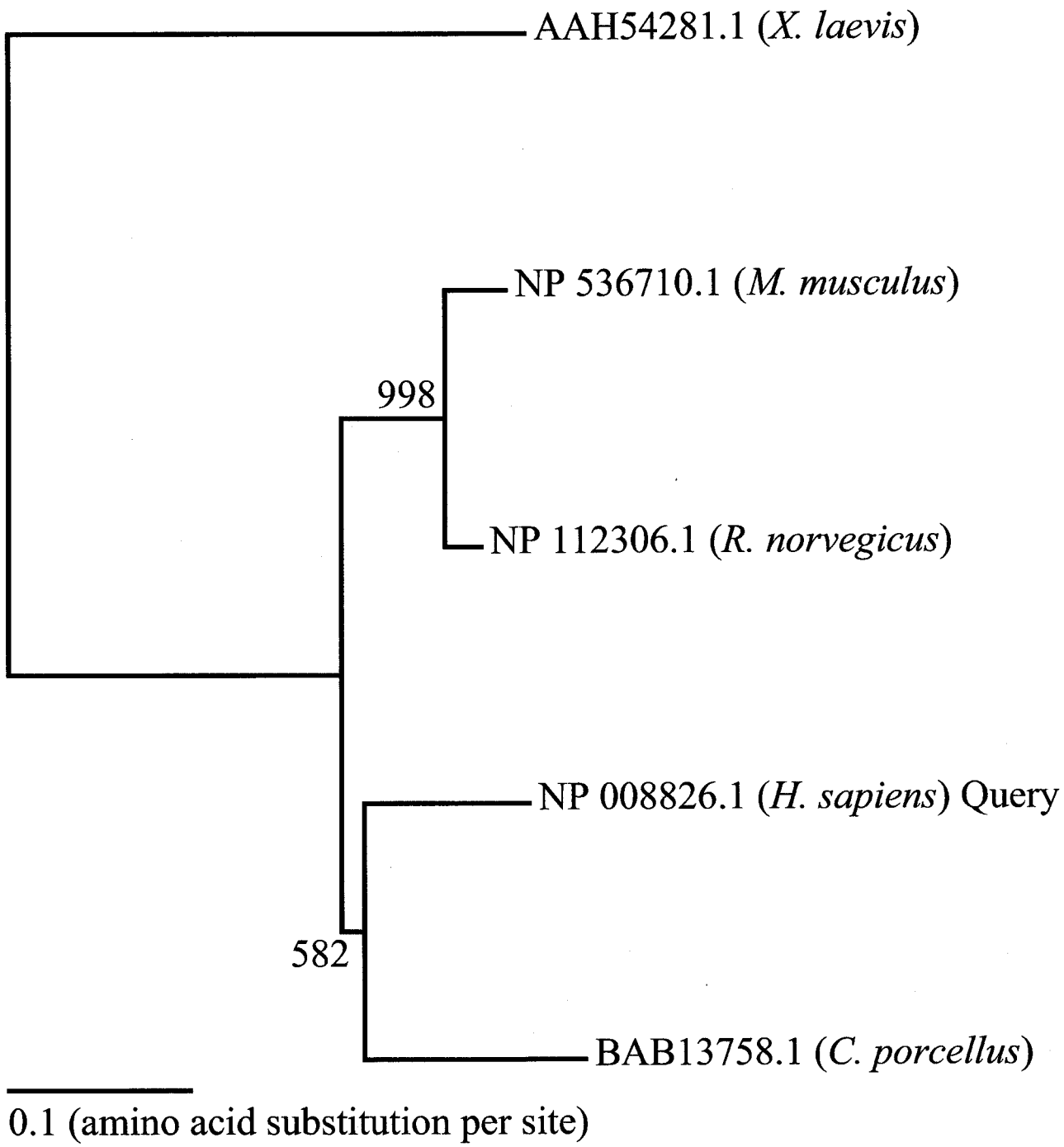
Diversifiation means that the new ER generates by the change of the

substrate-specificity of enzyme. For example, ER217, Glycine amidonotransferase, was found in *H. sapiens* in this study and I concluded that the ER had been gained in the vertebrate lineage after the separation from the *D. melanogaster*. As a result of the homology search, the sequences homologous to the enzyme involved in the ER were also found in bacterial lineages. However, the substrate-specificities of the enzymes were different between the vertebrate and the bacteria, glycine and inosamine-phosphate, respectively. Therefore, I conclude that ER217 was gained in the vertebrate lineage by the change of substrate-specificity of enzyme.

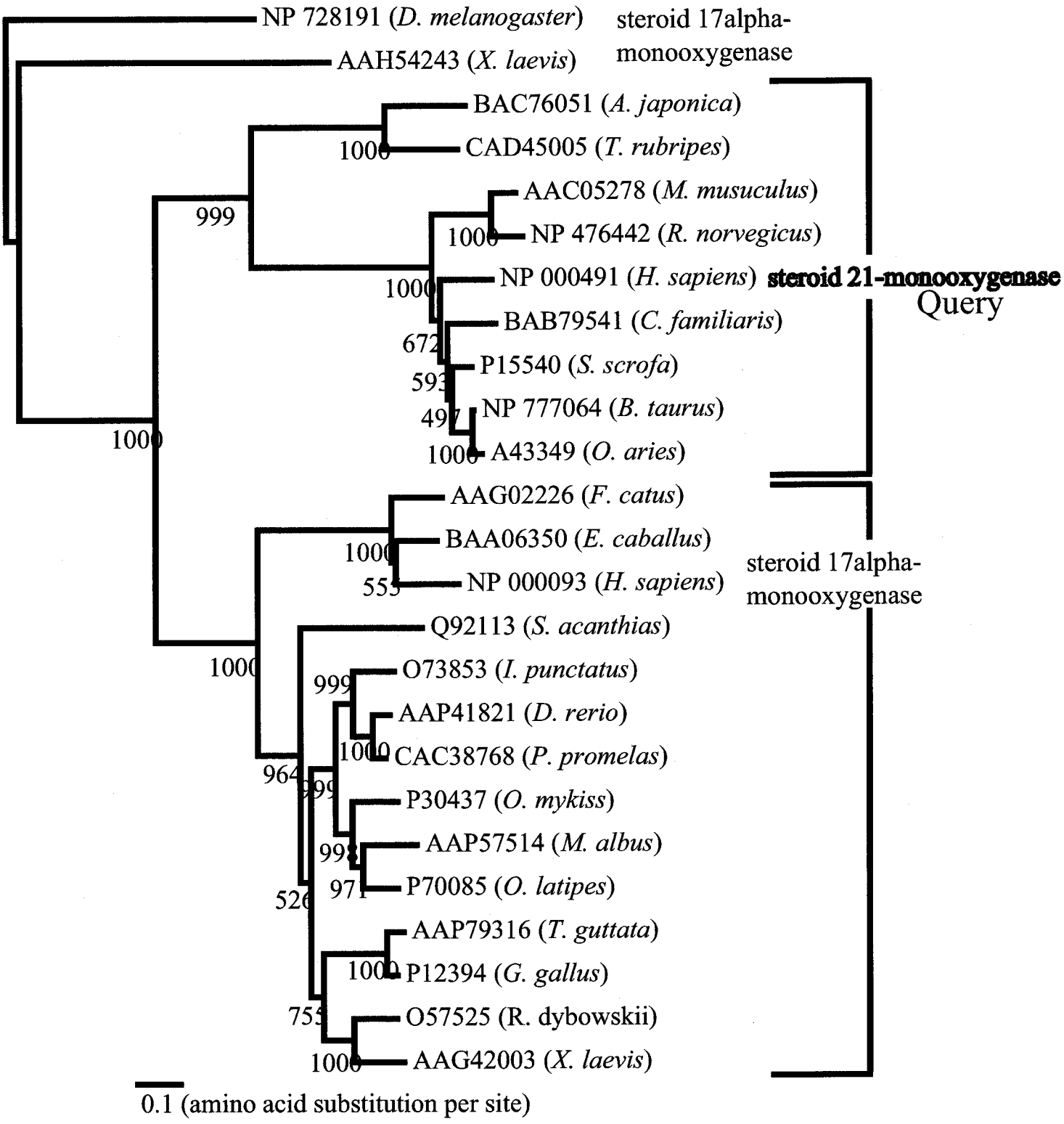
False-positive means that the timing of the ER gain may date back to the backward of my estimation. For example, ER671, N-acylglucosamine 2-epimearase, was found in *H. sapiens* in this study and I concluded that the ER had been gained in the vertebrate lineage after the separation from the *D. melanogaster*. However, as a result of the homology search, the homologous sequences of the enzyme involved in the ER were found in bacteria. Because the function of the genes discovered in bacteria were unknown, I could not decide when the ER generated. Therefore, it is conceivable that the generation of ER671 was older than my estimation. The false-positive result is caused by the limitation of the dataset that I used in this study. By the increase of the sequences in the database, the possibility to discover the homologous sequence of the targeted gene will also increase. This result suggests that the timing of ER gain may be changeable by the increase of the sequences examined.

Out of 58 ERs, 18 ERs (31%) were in the origination, 26 ERs (45%) were in the duplication, 8 ERs (14%) were in the diversification and 6 ERs (10%) were in the false-positive. These results suggest that the 10% of ER gains that I estimated was affected by the dataset of the sequences and may date back to the backward.

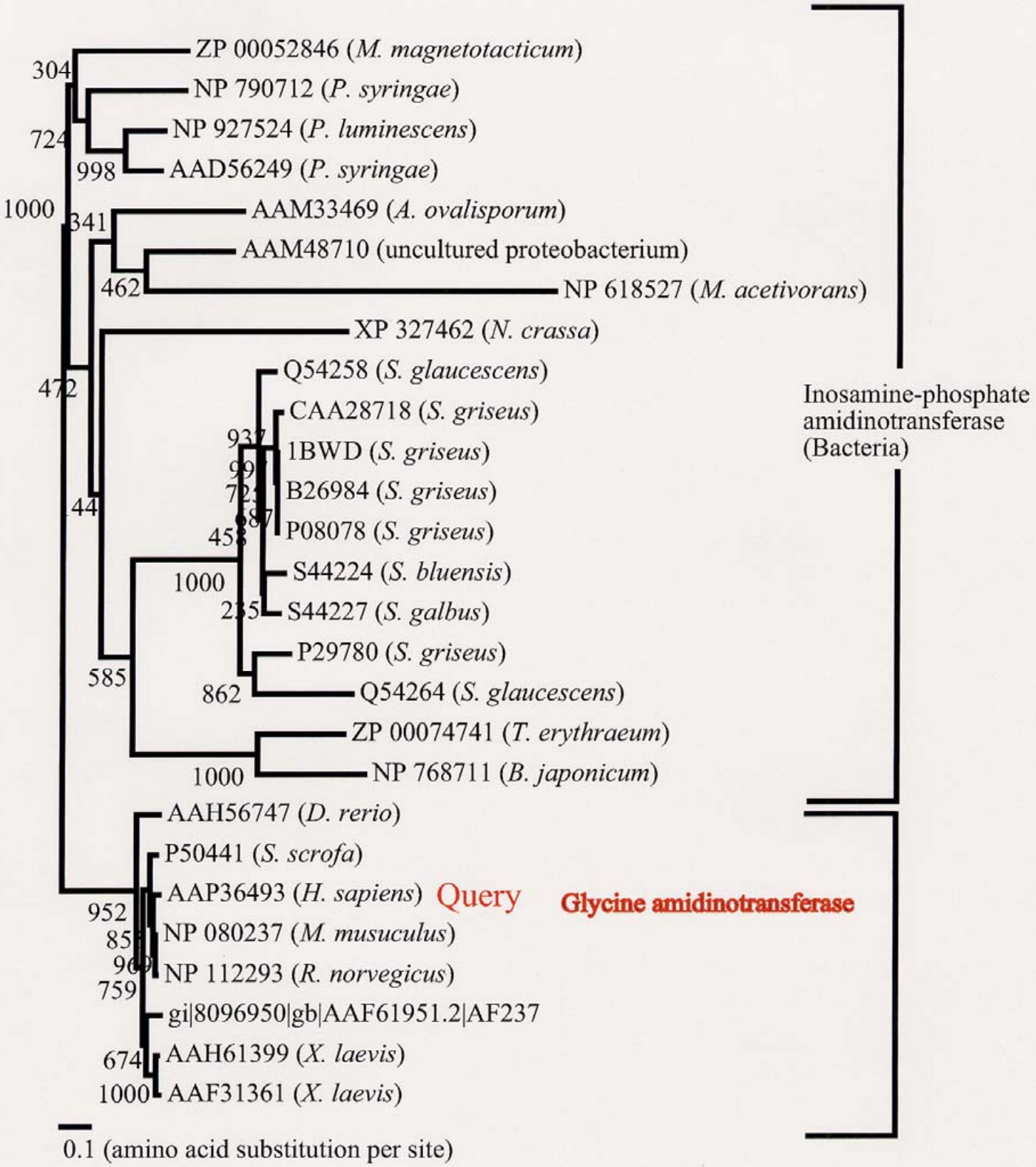
Pattern 1Histamine N-methyltransferase (ER207)



Pattern 2 Steroid 21-monooxygenase (ER96)



Pattern 3 Glycine amidinotransferase (ER217)



Pattern 4 N-acylglucosamine 2-epimerase (ER671)