J Mol Evol (2004) 59:806–814
DOI: 10.1007/s00239-004-0068-x

JOURNAL OF MOLECULAR EVOLUTION

Higher Gene Duplicabilities for Metabolic Proteins Than for Nonmetabolic Proteins in Yeast and *E. coli*

Elizabeth Marland,[1] Anuphap Prachumwat,[2] Natalia Maltsev,[1] Zhenglong Gu,[3] Wen-Hsiung Li[3]

[1] Mathematics & Computer Science Division, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439, USA
[2] Committee on Genetics, University of Chicago, Chicago, IL 60637, USA
[3] Ecology and Evolution, University of Chicago, 1101 East 57th Street, Chicago, IL 60637, USA

Received: 15 April 2004 / Accepted: 29 June 2004 [*Reviewing Editor:* Dr. Rüdiger Cerff]

Abstract. Although the evolutionary significance of gene duplication has long been appreciated, it remains unclear what factors determine gene duplicability. In this study we investigated whether metabolism is an important determinant of gene duplicability because cellular metabolism is crucial for the survival and reproduction of an organism. Using genomic data and metabolic pathway data from the yeast (*Saccharomyces cerevisiae*) and *Escherichia coli*, we found that metabolic proteins indeed tend to have higher gene duplicability than nonmetabolic proteins. Moreover, a detailed analysis of metabolic pathways in these two organisms revealed that genes in the central metabolic pathways and the catabolic pathways have, on average, higher gene duplicability than do other genes and that most genes in anabolic pathways are single-copy genes.

Key words: Gene duplicability — Metabolism — Functional requirement — Yeast genome

Introduction

In every genome sequenced to date, there are genes that are present in only a single copy and there are genes that are present in two or more copies. This observation suggests that different genes have different duplicabilities. However, it is far from clear what factors determine gene duplicability. Recently, Papp et al. (2003) proposed the dosage balance hypothesis, which postulates that genes coding for subunits of protein complexes (multimers) tend to have a lower duplicability than do genes coding for monomers because duplication of a single subunit may cause dosage imbalance among the subunits of the protein complex. Pursuing this issue further, Yang et al. (2003) hypothesized that dosage sensitivity increases while gene duplicability decreases with the number of subunits in a protein (i.e., protein complexity), and they indeed found support for this hypothesis from genomic and protein structure data of human and yeast.

Gene function is likely another important determinant of gene duplicability because it is well known that high dosages of some genes (e.g., histone genes) are required for a complex organism and that in many cases (e.g., MHC genes) multiple gene copies are required for functional diversities. In this study we investigated whether metabolic proteins tend to have higher gene duplicability than nonmetabolic proteins. It is well known that cellular metabolism is crucial for the survival and reproduction of cells. All cells in the three domains of life (Bacteria, Archaea, and Eukaryota) obtain energy and universal precursors during the biochemical assimilation and dissimilation of nutrients via metabolic pathways. The metabolic axis of a cell is represented by the pathways of central metabolism (e.g., glycolysis, pentose–phosphate shunt, and the

Correspondence to: Wen-Hsiung Li; *email:* whli@uchicago.edu

Krebs cycle). The crucial roles that metabolic pathways play in the survival of an organism may affect the duplicability of metabolic genes. Moreover, the patterns of gene duplication may depend on the metabolic role of the gene product (e.g., catabolic, anabolic).

Escherichia coli and *Saccharomyces cerevisiae* are good prokaryotic and eukaryotic model organisms, respectively, for studying gene duplication patterns in metabolic pathways, because their genomes have been completely sequenced and their metabolic pathways have been well characterized. In the present study, an analysis of metabolic pathways in these organisms revealed that genes in the central metabolic pathways and catabolic pathways have, on average, higher gene duplicabilities than do other genes. In contrast, single-copy genes (singletons) were predominant in anabolic pathways.

Materials and Methods

Identification of Duplicate and Singleton Genes

As described in Gu et al. (2002, 2003), the whole sets of *S. cerevisiae* and *E. coli* K-12 MG1655 protein sequences were downloaded from SGD (http://genome-www.stanford.edu/Saccharomyces/) and from *E. coli* Genome Project (http://www.genome.wisc.edu/sequencing/k12.htm), respectively. An all-against-all FASTA search was conducted on each protein dataset independently. A singleton was defined as a protein that did not hit any other proteins in the FASTA search with $E = 0.1$. Duplicate genes were identified as described in Gu et al. (2003) ($E < 10^{-10}$). We have also used less stringent criteria to detect duplicate genes and obtained essentially the same results.

Metabolic Pathways

Genes in *S. cerevisiae* and *E. coli* metabolic pathways are defined according to the KEGG (http://www.genome.ad.jp/kegg/; Ogata et al. 1999) and WIT (http://wit.mcs.anl.gov/WIT2/; Overbeek et al. 2000) databases. The *S. cerevisiae* and *E. coli* ORFs (denoted ALL) are categorized into metabolic (M) and nonmetabolic (non-M) genes. Metabolic genes are those that are involved in any metabolic pathways but not in signal transduction and transport. The metabolic genes are further classified into genes in central metabolic (denoted CM) and non-central metabolic pathway genes (denoted non-CM). The numbers of metabolic steps within CM and non-CM with singletons and duplicates are counted. A metabolic step represents a biochemical reaction catalyzed by an enzyme. When a step has both singleton and duplicate enzymes, we count it as one for singleton and one for duplicate. Although many reactions are reversible, the *glucose dissimilation* is the direction used to define the non-CM upstream (predominantly catabolic) and downstream of CM (predominantly anabolic) pathways (upstream- and downstream-CM, respectively). For example, galactose, starch, and sucrose catabolism are upstream-CM pathways, whereas amino acid biosynthesis is a downstream-CM pathway.

Proportion of Unduplicated Genes and Number of Duplications per Gene

For each category (i.e., a pathway) under study, the number of unique types of genes is defined as the number of singletons plus the number of duplicated gene types in that category. The number of duplications per gene (n) is the total number of genes divided by the total number of unique types of genes. The proportion of unduplicated genes (P) is the proportion of singletons in the total number of unique types of genes. While n roughly indicates how often a gene has been duplicated in the genome, $1 - P$ denotes the proportion of gene types that have been duplicated in the genome. Both n and $1 - P$ can be used as measures of gene duplicability (Yang et al. 2003). In addition, we also consider the proportion of duplicate genes in each category. The latter measure and n are less desirable than P because they can be strongly affected by the presence of large gene families.

Our statistical analyses were conducted in *R* (Version 1.7.1; http://www.r-project.org/). All statistical tests were Fisher's exact test.

Results

Duplication Patterns of Genes in Metabolic and Non-metabolic Pathways

The genes involved in 72 yeast metabolic pathways as defined by the KEGG and WIT databases were downloaded, but only 43 pathways (Table 1) were used in this study because the others showed small numbers of steps or overlapped with other pathways. These genes, which are called metabolic (M) genes, were further divided into two categories: central metabolic (CM) and non-central metabolic (non-CM) genes. The duplication patterns of genes in these 43 *S. cerevisiae* pathways are compared with those in nonmetabolic genes (non-M) and all genes (ALL). The proportions of duplicates in the ALL and non-M categories are similar (34–36%), but the proportion is significantly higher for metabolic genes (56%; $p < 10^{-40}$; Table 2 and Fig. 1A); all p values in this paper were obtained by Fisher's exact test. Furthermore, the proportion of duplicates in CM is about 1.5-fold higher than that in non-CM ($p < 10^{-8}$; Table 2 and Fig. 1A). A similar pattern of gene duplication is observed in *E. coli*, where CM also has the highest proportion of duplicates, being significantly higher than non-CM ($p < 0.003$). Moreover, the metabolic pathways as a whole (M) show a significantly higher proportion of duplicates than non-M ($p < 10^{-7}$) and ALL genes (Table 2 and Fig. 1B).

The proportion of unduplicated genes (P) in the central metabolic pathways (CM) show the lowest P (i.e., the highest duplicability) for both *S. cerevisiae* and *E. coli* (Table 2). In *S. cerevisiae*, non-CM has a P value similar to that for the whole metabolic category (M), which is, however, lower than those for ALL and non-M (Table 2). Similar conclusions hold for the *E. coli* data (Table 2).

With respect to the number of duplications per gene (n) for each category in *S. cerevisiae*, CM has the highest value (2.46; Table 2), non-CM has an intermediate value (1.63), and non-M has the lowest value

Table 1. Distributions of duplicates in 43 *S. cerevisiae* metabolic pathways (pathways are ordered by the proportion of duplicates; pathway numbers are assigned according to the KEGG database)

Pathway No.	Pathway name	Singletons	Duplicated gene types[a]	Unique types of genes[b]	Proportion of duplicates	Total number of genes	n[c]	Location relative to CM[d]
630	Glyoxylate metabolism	1	8	9	88.89	17	1.89	CM
460	Cyanoamino acid metabolism	1	2	3	66.67	5	1.67	D
620	Pyruvate metabolism	9	14	23	60.87	40	1.74	CM
52	Galactose metabolism	2	3	5	60.00	8	1.60	U
330	Arginine and proline metabolism	2	3	5	60.00	8	1.60	D
20	Citrate (TCA) cycle	7	10	17	58.82	33	1.94	CM
10	Glycolysis/gluconeogenesis	9	12	21	57.14	50	2.38	CM
530	Aminosugars metabolism	6	7	13	53.85	22	1.69	D
272	Cysteine metabolism	4	4	8	50.00	13	1.63	D
440	Aminophosphonate metabolism	1	1	2	50.00	4	2.00	D
500	Starch and sucrose metabolism	9	9	18	50.00	29	1.61	U
61	Fatty acid biosynthesis (path 1)	1	1	2	50.00	3	1.50	D
72	Synthesis and degradation of ketone bodies	1	1	2	50.00	3	1.50	D
280	Valine, leucine, and isoleucine degradation	2	2	4	50.00	6	1.50	U
30	Pentose–phosphate shunt	5	5	10	50.00	22	2.20	CM
910	Nitrogen metabolism	5	4	9	44.44	16	1.78	D
271	Methionine metabolism	8	4	12	33.33	17	1.42	D
480	Glutathione metabolism	4	2	6	33.33	10	1.67	D
730	Thiamine metabolism	2	1	3	33.33	6	2.00	D
920	Sulfur metabolism	4	2	6	33.33	8	1.33	D
580	Phospholipid degradation	2	1	3	33.33	6	2.00	U
260	Glycine, serine, and threonine metabolism	15	7	22	31.82	30	1.36	D
450	Selenoamino acid metabolism	9	4	13	30.77	18	1.38	D
790	Folate biosynthesis	7	3	10	30.00	13	1.30	D
252	Alanine and aspartate metabolism	12	5	17	29.41	25	1.47	D
561	Glycerolipid metabolism	15	6	21	28.57	34	1.62	U
251	Glutamate metabolism	19	7	26	26.92	41	1.58	D
410	β-Alanine metabolism	3	1	4	25.00	9	2.25	D
770	Pantothenate and CoA biosynthesis	7	2	9	22.22	11	1.22	D
220	Urea cycle and metabolism of amino groups	11	3	14	21.43	17	1.21	D
290	Valine, leucine, and isoleucine biosynthesis	11	3	14	21.43	19	1.36	D
300	Lysine biosynthesis	8	2	10	20.00	12	1.20	D
51	Fructose and mannose metabolism	15	3	18	16.67	24	1.33	U
740	Riboflavin metabolism	5	1	6	16.67	10	1.67	D
230	Purine metabolism	63	11	74	14.86	86	1.16	D
240	Pyrimidine metabolism	55	8	63	12.70	72	1.14	D
860	Porphyrin and chlorophyll metabolism	8	1	9	11.11	10	1.11	D
100	Sterol biosynthesis	8	1	9	11.11	10	1.11	D
400	Phenylalanine, tyrosine, and tryptophan biosynthesis	14	1	15	6.67	16	1.07	D
340	Histidine metabolism	12	0	12	0.00	12	1.00	D
430	Taurine and hypotaurine metabolism	6	0	6	0.00	6	1.00	D
780	Biotin metabolism	9	0	9	0.00	9	1.00	D
900	Terpenoid biosynthesis	4	0	4	0.00	4	1.00	D
All		401	165	566	29.15	814	1.44	

[a]"Duplicated gene types" (duplication groups) are defined as the number of gene families whose proteins participate in a pathway studied.
[b]Unique types of genes = singletons + duplicated gene types.
[c]Number of duplicates per gene (n) = (singletons + duplicates)/unique types of genes.
[d]Position of the pathways relative to the central metabolism pathways (CM), where U and D denote upstream- and downstream-CM, respectively. The *glucose dissimilation* is the direction used to define the upstream- and downstream-CM pathways (see Materials and Methods).

(1.31). A similar pattern holds for the *E. coli* data (Table 2). These data together with the P values suggest that genes in the central metabolic pathways have, on average, the highest gene duplicability.

Table 2. Distribution pattern of duplicates for the whole genome, nonmetabolic, metabolic, central metabolic, and non-central metabolic pathways for *S. cerevisiae* and *E. coli*

	Singletons	Duplicates	% of duplicates	Duplicated gene types[a]	Unique types of genes[b]	P[c]	n[d]
S. cerevisiae							
All genes	3920	2255	36.52	715	4635	84.57	1.33
Nonmetabolic genes	3468	1828	34.52[e]	566	4034	85.97	1.31
Metabolic genes	449	595	59.99[e]	188	637	70.49	1.63
Genes in noncentral metabolism	434	514	54.22[f]	164	598	72.58	1.59
Genes in central metabolism	15	81	84.38[f]	24	39	38.46	2.46
E. coli							
All genes	2677	1613	37.60	459	3136	85.36	1.37
Nonmetabolic genes	1994	1191	37.39[g]	318	2312	86.25	1.38
Metabolic genes	675	583	46.34[g]	195	870	77.59	1.45
Genes in noncentral metabolism	644	531	45.19[h]	176	820	78.54	1.43
Genes in central metabolism	31	52	62.65[h]	19	50	62.00	1.66

[a]"Duplicated gene types" (duplication groups) are defined as the number of gene type families whose genes are duplicated and whose proteins participate in such a category.
[b]Unique types of genes = singletons + duplicated gene types.
[c]% of unduplicated genes (P) = singletons/unique types of genes.
[d]Number of duplicates per gene (n) = (singletons + duplicates)/unique types of genes.
[e]Significant difference; $p < 10^{-40}$, Fisher's exact test.
[f]Significant difference; $p < 10^{-8}$, Fisher's exact test.
[g]Significant difference; $p < 10^{-7}$, Fisher's exact test.
[h]Significant difference; $p < 0.003$, Fisher's exact test.

The above comments still apply when the criteria used to detect duplicates are relaxed to $E < 10^{-5}$ in both the *S. cerevisiae* and the *E. coli* data.

Pattern of Duplicates in Each Step of the Metabolic Pathways in S. cerevisiae

In *S. cerevisiae* (Table 3) the proportion of singleton steps in non-CM (68.4%) is much higher than that in CM (42.85%; $p = 0.001$). Indeed, in non-CM there are more steps with a singleton than steps with duplicates (158 vs. 73), whereas in CM there are roughly equal proportions of steps with singletons and duplicates (21 vs. 28).

Interestingly, the non-CM pathways upstream of the CM pathways (upstream-CM) show a high proportion of duplicate genes and a high number of duplications per gene (Table 1 and Fig. 2), in comparison with the non-CM pathways downstream of CM pathways (downstream-CM; Fig. 3). Indeed, steps in downstream-CM pathways are dominant with singletons; for example, steps with singletons are overrepresented in the histidine, urea, glutamate, biotin, pyrimidine and purine metabolism pathways (Fig. 3). These results suggest that CM and upstream-CM pathways have a higher gene duplicability than do downstream-CM pathways.

Discussion

The gene duplication patterns in both *S. cerevisiae* and *E. coli* reveal a higher average duplicability for genes that are involved in metabolism, especially central metabolism, than for nonmetabolic genes. We note that both species studied are fast-growing organisms and this could be the reason for the higher duplicability for central metabolic enzymes. It will therefore be interesting to see whether our observation holds for other organisms in general.

It is also possible that certain protein families have been preferentially duplicated in the central metabolic pathways. For this possibility we consider the enzymes with a $(\beta\alpha)_8$ (TIM) barrel because Copley and Bork (2000) have noted the presence of many TIM barrel-containing enzymes in the pathways of central metabolism; from this observation they suggested that early on, enzyme recruitment was a driving force behind the evolution of metabolic pathways. In yeast the proportion of unduplicated genes is 42.9% for TIM barrel-containing enzymes and 37.5% for enzymes containing no TIM barrel. In *E. coli*, the corresponding proportions are 62.5 and 61.9%. In both cases, the difference between the two proportions is not significant, so TIM barrel-containing enzymes and non-TIM-barrel enzymes have approximately the same gene duplicability. It should be noted that while Copley and Bork (2000) were concerned with ancient duplications, we are concerned with more recent duplications, i.e., duplicate proteins whose homology can still be readily detected from sequence alignment. Therefore, TIM barrel-containing enzymes in the central metabolic pathways do not seem to have been preferentially duplicated during the evolution of yeast and *E. coli* at least in recent times.

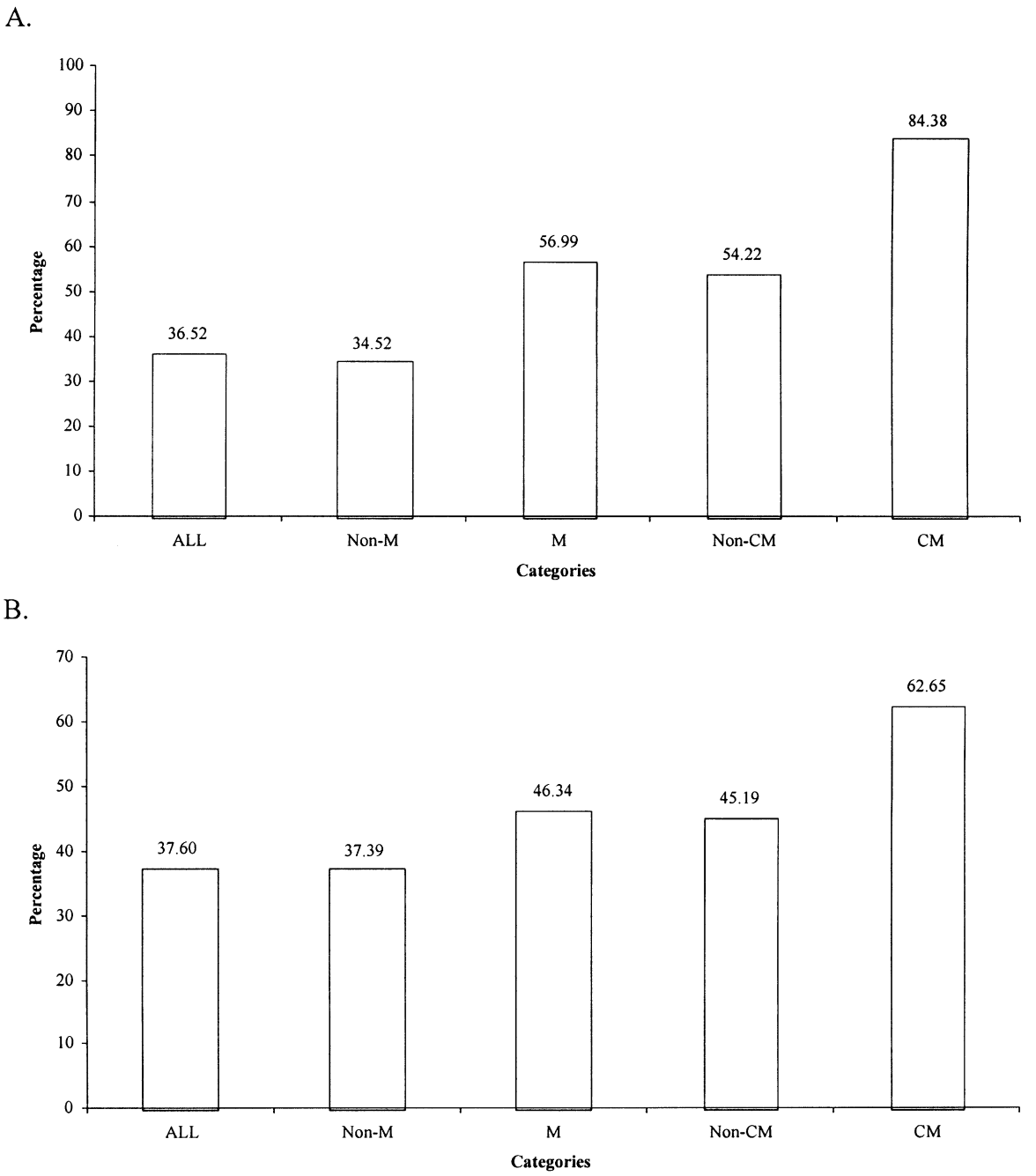

Fig. 1. Distributions of duplicates among all genes (denoted ALL), nonmetabolic genes (non-M), metabolic genes (M), non-central metabolic pathway genes (non-CM), and central metabolic pathway genes (CM) for *S. cerevisiae* (**A**) and *E. coli* (B). The number at the top of each bar represents the proportion of duplicate genes.

Table 3. Numbers of steps with singletons and duplicates for metabolic, non-central metabolic, and central metabolic pathways of *S. cerevisiae*

	Number of steps	
	With singletons (%)	With duplicates (%)
Metabolic pathways	179 (63.92)	101 (36.08)
Non-central metabolic pathways	158 (68.40)	73 (31.60)[a]
Central metabolic pathways	21 (42.85)	28 (57.15)[a]

[a]Significant difference; $p = 0.001$, Fisher's exact test.

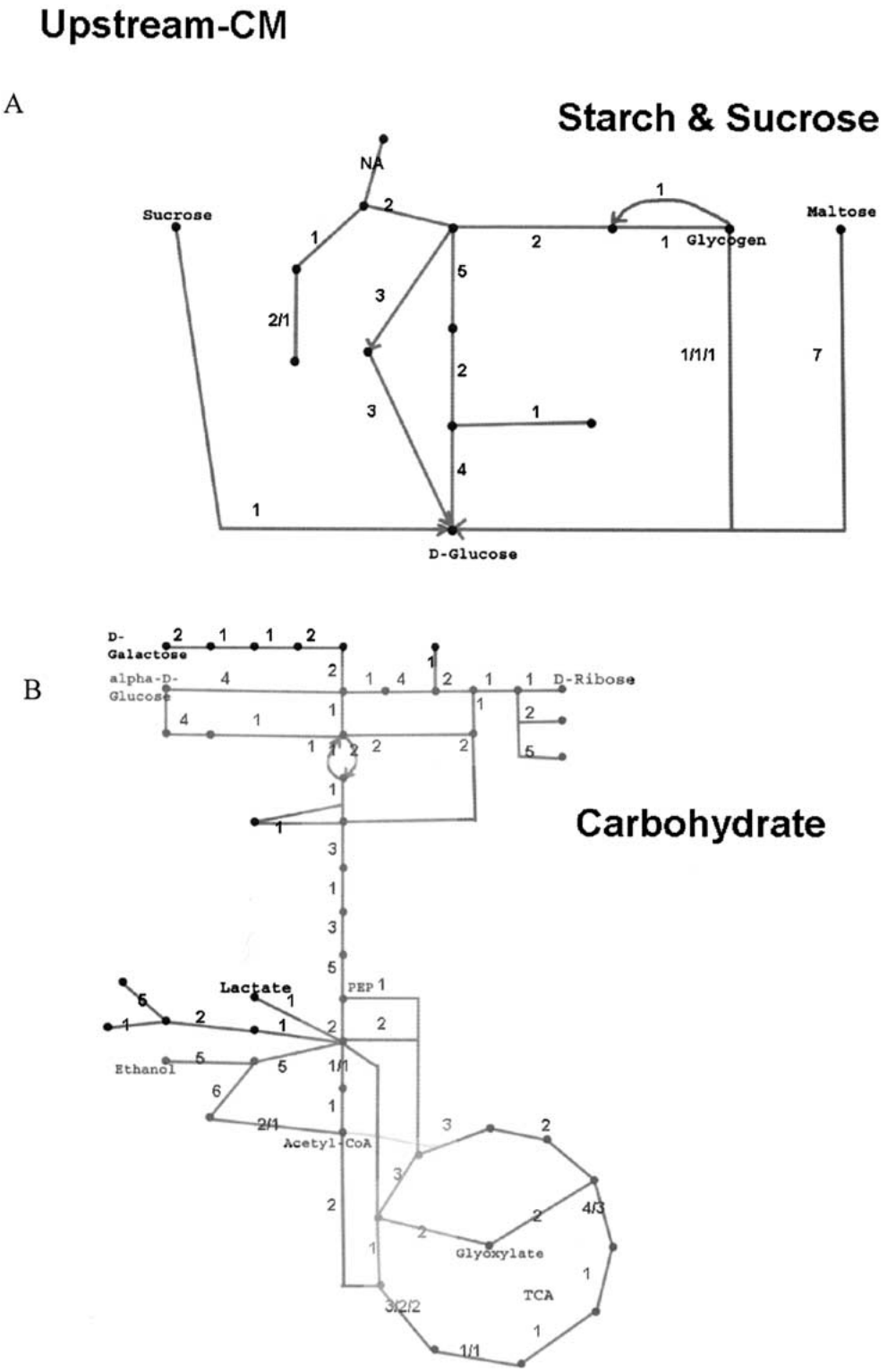

Fig. 2. Diagrams showing the locations of duplicates and singletons in both upstream-central metabolic (**A**) and central metabolic (**B**; highlighted in light gray) pathways in *S. cerevisiae*. An arrowed branch indicates the direction of the metabolic reaction, whereas an unarrowed branch indicates a reversible reaction. The numbers beside each branch indicate the number of duplicate genes on that branch; 1 represents a singleton. If more than one duplicate family or singleton is present for a step, the numbers of genes are separated by a slash.

Generally, a gene duplicate accumulates deleterious mutations more quickly than advantageous ones and has a high chance of becoming a pseudogene as long as the other copy maintains the original function. Thus, the persistence of both duplicates in a genome would require a selective advantage such as

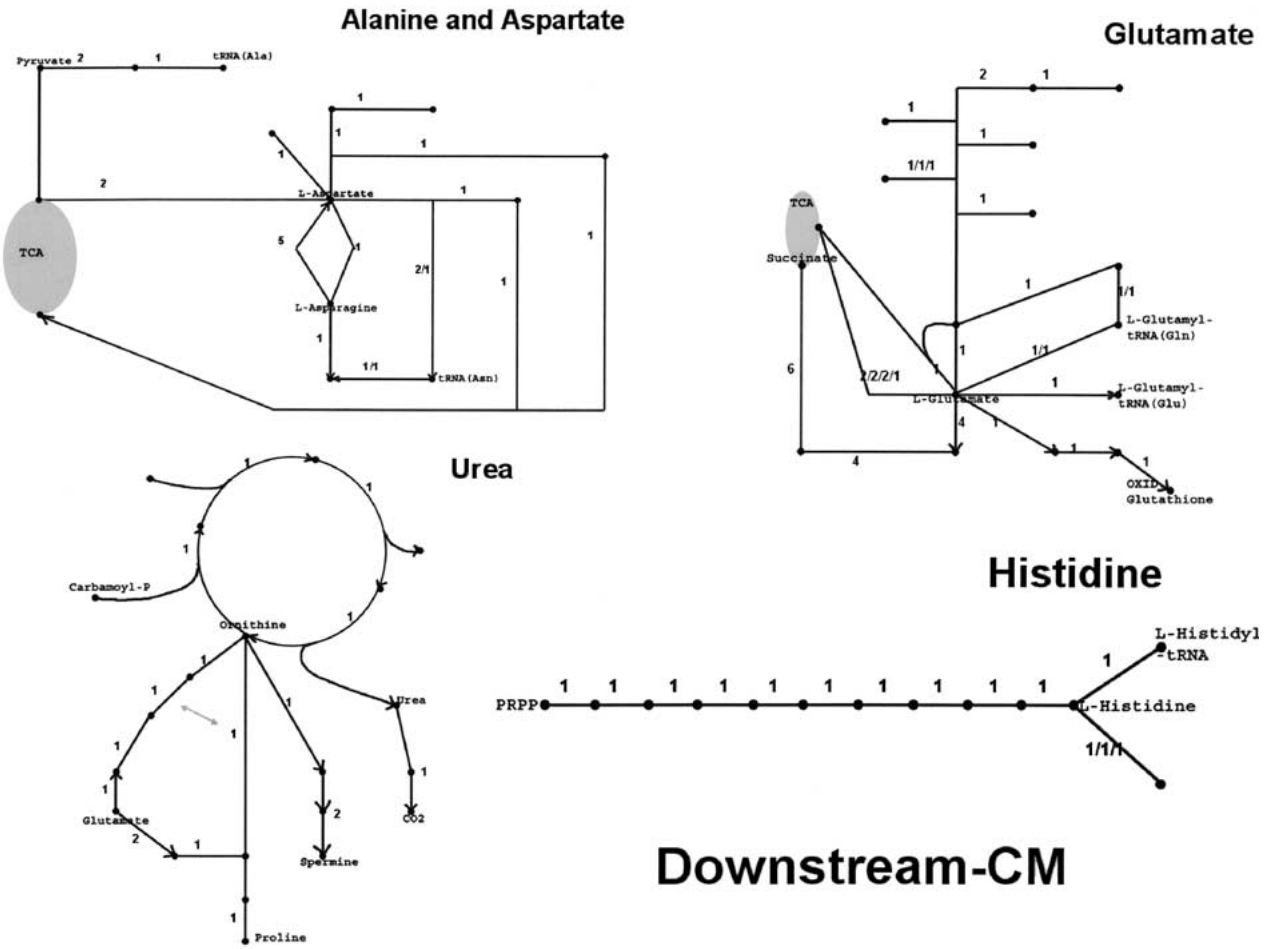

Fig. 3. Diagrams showing the locations of duplicates and singletons on downstream-central metabolic pathways in *S. cerevisiae*. An arrowed branch indicates the direction of the reaction, whereas an unarrowed branch indicates a reversible reaction. The number beside each branch indicates the number of duplicate genes on that branch; 1 represents a singleton. If more than one duplicate or singleton family is present on a step, the numbers of genes are separated by a slash. The two-headed arrow indicates that the enzymes at the tip of the arrow are duplicate members of a gene family.

functional diversification or a larger dosage requirement. Therefore, it seems that duplication of a metabolic gene tends to have a higher chance to become advantageous than duplication of a nonmetabolic gene.

Most universal precursors for biosynthesis are produced by the central metabolic pathways (e.g., glyceroldehyde 3-phosphate, fructose 6-phosphate, citrate, α-ketoglutarate [Neidhardt et al. 1990]). For this reason, duplication of a gene in a central metabolic or upstream-CM pathway might have been favored. As noted above, in *S. cerevisiae* and *E. coli*, genes in the central metabolic and upstream-CM pathways have the highest gene duplicability (Table 2, Figs. 1 and 2).

This argument may be strengthened by the following observation. In *S. cerevisiae* intracellular hexoses (mainly glucose) that enter the glycolytic pathway are converted to pyruvate and oxidized to ethanol via fermentation. After the fermentable hexoses are exhausted, ethanol is used as a carbon source for aerobic growth, which involves the TCA cycle. Alternatively, glucose can be oxidized in the pentose–phosphate shunt. This pathway provides the cell with pentose sugar and cytosolic NADPH. Ribose sugars generated are used further in the biosynthesis of nucleic acid precursors and nucleotide coenzymes. Therefore, in order to utilize the hexoses rapidly, duplication of an enzyme in an upstream-CM or CM pathway might have been an advantage during some period in evolution. Furthermore, the importance of glycolysis is obvious in view of the fact that glycolytic enzymes are present around 30–68% of soluble protein in the yeast cell (Banuelos and Fraenkel 1982).

The presence of gene duplicates may also increase genetic robustness against null mutations (Gu et al. 2003). Using the data on the fitness effects of single-gene deletions for the whole yeast genome (Steinmetz et al. 2002), we find that essential genes in the central metabolic pathways are all singletons (i.e., in CM 100% of genes with lethal single-gene deletions are singletons), but no deletion of a duplicate is lethal.

Enzyme duplication could provide an opportunity for an enzyme with a multiple substrate specificity to specialize in different functions. Recent biochemical studies provide evidence that many enzymes in central metabolic pathways have binding specificities to not-normally-known substrates (e.g., O'Brien and Herschlag 1999; for a review, see D'Ari and Casadesus 1998). For example, the glycolytic kinases such as 6-phosphofructokinases, phosphoglycerate kinases, pyruvate kinases, and acetate kinases of the small genome wall-less Mollicutes (*Mycoplasma* species) could use other nucleoside diphosphates besides their normally known reactants (Pollack et al. 2002). Such usages of unnatural reactants of these glycolytic kinases are reported in various organisms including *E. coli*, dog, and cat (Brenda Enzyme Database; http://www.brenda.uni-koeln.de [Schomburg et al. 2002]). Moreover, duplicates may be regulated and/or expressed in different environmental conditions. In yeast, *pyk1* (pyruvate kinase 1) mutants fail to grow on fermentable carbon sources but can grow normally on ethanol or other gluconeogenic carbon sources (a very low glycolytic flux). Under such conditions, pyruvate kinase 2 (*PYK2*), a *PYK1* paralog, is expressed (Boles et al. 1997). Such an "underground metabolism" could provide functional diversification, which in turn provides metabolic plasticity for organisms to survive in wider environmental habitats (D'Ari and Casadesus 1998).

Gene duplication has been the major process proposed for the evolution of enzymes and the metabolic pathways, but the issue has been under intense debate for more than 50 years. Possible models for describing its evolutionary mechanism have been proposed such as duplication of either enzymes or pathways, recruitment of enzymes from other pathways, or retro-evolution of the pathways (e.g., for a review, see Schmidt et al. 2003). As metabolic data from various organisms increased, it became clear that the lower part of glycolysis has been well conserved across eubacteria, archaea and eukaryotes, whereas major variations are found in the upper part from glucose to 3-phosphoglycerate (Ronimus and Morgan 2003; Verhees et al. 2003). Although archaeal enzymes in the upper part of the glycolysis have less sequence similarity than, and diverse functions from, eubacteria and eukaryote counterparts, their structures are homologous. In addition to this observation, many downstream-CM pathways (e.g., individual amino acid biosyntheses) in *E. coli* show high conservation in the number of orthologs in all three domains of life (Peregrin-Alvarez et al. 2003). Thus, in ancient times duplication in the central metabolic and upstream-CM pathways might have been a driving force for an organism to cope with changes in metabolites.

These data provide evidence for gene function as an important determinant of gene duplicability, especially genes functioning in metabolism in *S. cerevisiae* and *E. coli*. Given that these free-living unicellular organisms make a contact to the environment directly, their source of nutrients depends on the habitats. Often their inhabiting environments are short in nutrient supplies, so that they have to compete with each other in a species and/or with different species for the available metabolites. The ability to process these nutrients into metabolic precursors quickly directly increases the growth and survival rates. Therefore, duplication in upstream metabolic genes may increase the ability to compete for resources. In this study, we have indeed found that many gene duplicates have been retained in the upstream-CM and CM pathways.

Acknowledgments. We thank the two reviewers for valuable comments. This study was supported by the International Balzan Foundation and NIH grants to W.H.L.

References

Banuelos M, Fraenkel DG (1982) *Saccharomyces carlsbergensis* fdp mutant and futile cycling of fructose 6-phosphate. Mol Cell Biol 8:921–929

Boles E, Schulte F, Miosga T, Freidel K, Schluter E, Zimmermann FK, Hollenberg CP, Heinisch JJ (1997) Characterization of a glucose-repressed pyruvate kinase (Pyk2p) in *Saccharomyces cerevisiae* that is catalytically insensitive to fructose-1,6-bisphosphate. J Bacteriol 179:2987–2993

Copley RR, Bork P (2000) Homology among $(\beta\alpha)_8$ barrels: Implications for the evolution of metabolic pathways. J Mol Biol 303:627–640

D'Ari R, Casadesus J (1998) Underground metabolism. Bioessays 20:181–186

Gu Z, Nicolae D, Lu HH, Li WH (2002) Rapid divergence in expression between duplicate genes inferred from microarray data. Trends Genet 18:609–613

Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li WH (2003) Role of duplicate genes in genetic robustness against null mutations. Nature 421:63–66

Neidhardt FC, Ingraham J, Schaechter M (1990) Physiology of the bacterial cell: A molecular approach. Sinauer Associates, Sunderland, MA

O'Brien PJ, Herschlag D (1999) Catalytic promiscuity and the evolution of new enzymatic activities. Chem Biol 6:R91–R105

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 27:29–34

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

Papp B, Pal C, Hurst LD (2003) Dosage sensitivity and the evolution of gene families in yeast. Nature 424:194–197

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

Pollack JD, Myers MA, Dandekar T, Herrmann R (2002) Suspected utility of enzymes with multiple activities in the small genome Mycoplasma species: The replacement of the missing ''household'' nucleoside diphosphate kinase gene and activity by glycolytic kinases. OMICS 6:247–258

Ronimus R, Morgan H (2003) Distribution and phylogenies of enzymes of the Embden–Meyerhof–Parnas pathway from archaea and hyperthermophilic bacteria support a gluconeogenic origin of metabolism. Archaea 1:199–221

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

Schomburg I, Chang A, Schomburg D (2002) BRENDA, enzyme data and metabolic information. Nucleic Acids Res 30:47–49

Steinmetz LM, Scharfe C, Deutschbauer AM, Mokranjac D, Herman ZS, Jones T, Chu AM, Giaever G, Prokisch H, Oefner PJ, Davis RW (2002) Systematic screen for human disease genes in yeast. Nat Genet 31:400–404

Verhees CH, Kengen SW, Tuininga JE, Schut GJ, Adams MW, De Vos WM, Van Der OoJ (2003) The unique features of glycolytic pathways in Archaea. Biochem J 375:231–246

Yang J, Lusk R, Li WH (2003) Organismal complexity, protein complexity, and gene duplicability. Proc Natl Acad Sci USA 100:15661–15665