J Mol Evol (2002) 54:258–266
DOI: 10.1007/s0023901-0007-z

JOURNAL OF MOLECULAR EVOLUTION

On the Occurrence of Horizontal Gene Transfer Among an Arbitrarily Chosen Group of 26 Genes

Michael Syvanen

Department of Medical Microbiology and Immunology, School of Medicine, University of California, Davis, CA 95616-8645, USA

Received: 12 July 2001 / Accepted: 27 July 2001

Abstract. The deduced amino acid sequences from 1200 *Haemophilus influenzae* genes was compared to a data set that contained the orfs from yeast, two different Archaea and the Gram+ and Gram− bacteria, *Bacillus subtilis* and *Escherichia coli.* The results of the comparison yielded a 26 orthologous gene set that had at least one representative from each of the four groups. A four taxa phylogenetic relationship for these 26 genes was determined. The statistical significance of each minimal tree was tested against the two alternative four taxa trees. The result was that four genes significantly supported the (Archaea, Eukaryota) (Gram+, Gram−) topology, two genes supported the one where Gram− and Eukaryota form a clade, and one gene supported the tree where Gram+ and Eukaryota define one clade. The remaining genes do not uniquely support any phylogeny, thereby collapsing the two central nodes into a single node. These are referred to as star phylogenies.

I offer a new suggestion for the mechanism that gave rise to the star phylogenies. Namely, these are genes that are younger than the underlying lineages that currently harbor them. This hypothesis is examined with two proteins that display the star phylogeny; namely onithine transcarbamylase and tryptophan synthetase. It is shown, using the distance matrix rate test, that the rate of evolution of these two proteins is comparable to a control gene when rates are determined by comparing closely related species. This implies that the genes under comparison experience comparable functional constraint. However, when the genes from remotely related species are compared, a plateau is encountered. Since we see no unusual levels of functional constraint this plateau cannot be attributed to the divergence of the protein having reached saturation. The simplest explanation is that the genes displaying the star phylogenies were introduced after Archaea, Eukaryota, and Bacteria had diverged from one another. They presumably spread through life by horizontal gene transfer.

Key words: Genetic code — Arginine biosynthesis — Tryptophan biosynthesis

Introduction

The recent accumulation of whole genome sequences has revealed numerous cases where gene trees appear highly incongruent with what we normally consider the underlying species trees i.e. the phylogeny that is based on small subunit RNA. Recent references include: Koonin and Galperin 1997; Saito and Tomita 1999; Makarova et al. 1999; Philippe and Forterre 1999; Worning et al. 2000; Kitabatake et al. 2000; Brinkman and Philippe 1999; Nelson et al. 1999; Wolf et al. 1999; Ponting et al. 1999. Though there remains uncertainty in assigning orthology or parology to many of the genes in question, these idiosyncratic gene trees are increasingly interpreted as being the result of horizontal gene transfer (Syvanen 1987; Smith et al. 1992; Woese 1998; Jain et al. 1999; Doolittle 1999).

The current unifying model provides us with the trifurcation of life into the three domains—Bacteria, Archaea, and Eukaryota. This organization is supported by many other character traits besides those from 18S RNA,

Email: syvanen@ucdavis.edu

including the sequences from many proteins especially those involved in protein, RNA, and DNA synthesis; that is, in those involved in central processing of the genome. However, phylogenetic analysis of the sequences for so many other proteins fail to support this major trifurcation and questions are beginning to be raised about the generality of this unifying trifurcation. In many of the idiosynchratic gene trees, Gram negative Bacteria appear to be more closely related to Eukaryota than they are to Gram positive Bacteria and Archaea; these are often considered as possible genes from the endosymbiotes that gave rise to mitochondria and chloroplasts. Because so many other possible horizontal gene transfers have been reported based on major incongruencies between ribosomal gene trees and given protein gene trees, other endosymbiotic events have now been suggested. In particular, Gupta and Golding (1993) saw that one of the heat shock proteins (hsp70) from the Gram positive bacteria was more closely related to the homologs from Archaea compared to those from Gram negative bacteria and this pattern was seen for a number of other proteins (Gupta and Golding 1994; Golding and Gupta 1995). [However, not all Archaea seem to contain a Gram+ hsp70 (Gribaldo et al. 1999).] In a survey of other protein sequence comparisons, they noted that only two major topologies were supported—i.e. on the one hand, the classical trifurcation or, on the other, a topology congruent with that seen by *hsp70.* This has led these authors to suggest that there was a major endosymbiosis event between a Gram positive bacteria and an ancestral Archaea (Golding and Gupta 1995).

The complete genome sequences available for multiple Archaea, Gram negative bacteria, and the Gram positive *Bacillus subtilis,* and Eukaryota has opened up the possibility of investigating the three domain hypothesis from a whole genomic perspective. In the current work, I have conducted a comparative survey of the genomes of representative Archaea, Bacteria, and Eukaryota with the goal of uncovering how frequently specific gene trees are incongruent with the ribosomal RNA trees. Twenty-six genes were arbitrarily selected that are present in the same Archaea, Eukaryota, a Gram positive bacteria, and a Gram negative bacteria, and two simple questions were posed: 1) When four taxa from each group are compared, what is the most parsimonious tree, and 2) Is the parsimonious tree significantly shorter than the two alternative trees? This effort was undertaken to gain some insight into the prevalence of horizontal gene transfer events.

Materials and Methods

Computations were performed on a vax using the VMS operating system and GCG (Genetics Computer Group at Wisconsin) software. All sequences were obtained directly from the National Center for Biological Information using Query (query@ncbi.nlm.nih.gov). This included the deduced protein sequence from the genomes of *H. influenzae, E. coli, B. subtilis, Methanococcus jannaschii, Archaeglobus fulgidus,* and *Saccharomyces cerevisiae.* The compiled orfs from each sequence were assembled into a local data file (using GCGTOBLAST) that could be queried with a defined sequence using the BLAST program. The 1200 BLAST searches, using *H. influenzae* queries, were executed through shell script instructions. The output file was manually analyzed to find those that had significantly high "expect values" (Altshul et al. 1997) ($<10^{-20}$ for an average sized orf) against a sequence from *B. subtilis* and at least one of the two *Archaea* and Yeast orthologs. At this point those sets that had multiple homologues from a single genome were excluded; this screen removed most paralogous gene sets. From this survey, about 300 genes were identified that had reasonably good homology with representatives from each of the four clades. A sequence from each putative orthologous set was then used to query the complete NCBI data files using BLAST. More genes that belonged to ambiguous paralogous groups were detected and removed. This resulted in the removal of most genes involved in DNA processing. The list of genes to be compared was further refined by selecting those for which a multiple sequence alignment was easily performed. After placing a few ribosomal proteins into a separate group, 26 orthologs were chosen and the multi-sequence alignments were edited by hand to improve the alignment, C and N terminal indels trimmed, and large internal indels trimmed and coded as a single substitution.

Statistical test of tree congruence. A simple statistical test has been suggested to test the significance of a four taxa tree. Multiple aligned multi-sequences are edited to remove all indels larger than 2. Shorter indel are coded to be the equivalent of amino acid difference. The length of each of the three four-taxa trees is determined by maximum parsimony using Swofford's PAUP. Each amino acid difference in the aligned sequences is weighed equally. For convenience the four taxa phylogenies will be presented using the parentheses convention, i.e. (ab)(cd) means that a and b define one clade and is linked to the second clade, c and d. Let us assume that the (Gram+, Gram−) (Arc, Euk) is the minimal tree and has, say, a length of 300 with 25 homoplastic substitutions. Let us further assume that homoplastic substitutions arose purely by chance. This latter assumption is implied, either implicitly or explicitly, in all tree building algorithms. We can then define the number of homoplastic substitutions in the minimal tree as the expected number. There are two alternative trees—namely, (Gram+, Arc) (Euk, Gram−) and (Gram+, Euk) (Arc, Gram−)—that will have a number of homoplastic changes, $N \geq 25$. If we assume homoplastic replacements occur by chance, then we can use chi square to test whether or not N is significantly greater than 25.

Statistical test of the star phylogeny. As will be apparent in Results, none of the three trees is given support for many of the genes, which leads to the possibility of a star phylogeny; i.e. (ab) and (cd) are linked with an internal branch of length zero. This conclusion would be supported if the parsimonious informative characters are randomly distributed among the three trees. A separate test is designed to assess this possibility. In this test, the number of informative characters, I, that support each of the three trees is determined and a chi square test, with two degrees of freedom, addresses whether or not they are equally distributed. This latter test is only subtly different from the one above because it is based on the same data in that the sum of informative characters supporting the three trees, I, is directly related to the amount of homoplasy in the three trees, H, simply by $I = H/2$. These statistical tests are closely related to tests of informative characters suggested by Willson and discussed in Felsenstein (1988).

Distance matrix rate test. Kimura protein distances (Kimura 1983) are computed from the multiply aligned sequences—these result in estimated Dij values which are the number of likely replacements per hundred amino acid substitutions. These values are used directly in the distance matrix rate test as described in Results. Computation of the 95% confidence intervals requires consideration of the total number of replacements and is equal to Dij times the number of sites compared divided by 100.

Table 1. Genes examined in this study. The abbreviation used is mostly the genetic symbol for *E. coli* genes

Aprt	AMIDOPHOSPHORIBOSYLTRANSFERASE
*OTC	ORNITHINE TRANSCARBAMYLASE
*ArgH	ARGININOSUCCINATE LYASE
*ArgT	ARGINYL-TRNA SYNTHETASE
AroA	3-PHOSPHOSHIKIMATE 1-CARBOXYVINYLTRANSFERASE
*AroE	SHIKIMATE 5-DEHYDROGENASE
CtpS	CTP SYNTHASE
*Enol	ENOLASE
*Fmu	SUN PROTEIN
*Gpd	GLYCEROL-3-PHOSPHATE DEHYDROGENASE
*Hyp1	HYPOTHETICAL 20KD PROTEIN
Hyp2	HYPOTHETICAL 19KD PROTEIN
*IlvH	ACETOLACTATE SYNTHASE III SMALL SUBUNIT
*IlvI	ACETOLACTATE SYNTHASE III LARGE SUBUNIT
Map	METHIONINE AMINOPEPTIDASE
*MoaA	MOLYBDENUM COFACTOR BIOSYNTHESIS PROTEIN A
*MoaC	MOLYBDENUM COFACTOR BIOSYNTHESIS PROTEIN C
*Ndk	NUCLEOSIDE DIPHOSPHATE KINASE
*PurA	ADENYLOSUCCINATE SYNTHETASE
*PurB	ADENYLOSUCCINATE LYASE
*Pur5	PHOSPHORIBOSYLFORMYLGLYCINAMIDINE CYCLO-LIGASE
Pur6	PHOSPHORIBOSYLAMINOIMIDAZOLE CARBOXYLASE CATALYTIC SUBUNIT
*RisB	RIBOFLAVIN SYNTHASE BETA CHAIN
*TrpA	TRYPTOPHAN SYNTHASE ALPHA CHAIN
*TrpB	TRYPTOPHAN SYNTHASE BETA CHAIN
*TrpG	ANTHRANILATE SYNTHASE COMPONENT II

* Genes displaying star phylogeny.

Results

In the present study, a group of genes were arbitrarily chosen that have at least one orthologue in each of the following groups 1) *E. coli,* 2) *B. subtilis,* 3) One of two Archaeae and 4) Yeast or another Eukaryota. A gene for each taxa was chosen, the sequences aligned and the distances of all minimal replacement trees was determined. With four taxa, there are three different trees. We can then ask two questions: Which of the three trees is shorter and is that tree significantly shorter than the other two?

Selection of orthologues. Table 1 lists the genes that were selected in this study. These genes were selected in the following manner. We constructed two data sets. Once contained 1200 orfs from *H. influenzae.* The second data set contains the entire set of orfs from each of the following known genome sequences: 1) The Proteobacter *E. coli,* a Gram negative, 2) the Gram positive *Bacillus subtilis,* 3) the Archaea, *Methanococcus jannaschii* and *Archaeologlobis fulgidis,* and 4) the Eukarya, *Saccharomyces cerevisiae.* The sequences in each of the 1200 orfs from the first group were compared to sequences in the second data set using the BLAST program, which resulted in a list of about 300 orfs who had at least one representative from each of the four groups. If there were good representative orthologues in all the taxa, a second Blast search was performed against gene bank; if possible, *A. thaliana* or *C. elegans* genes were chosen for the eukaryotic representative. This list was reduced by eliminating paralogous gene families and resulted in the elimination of transport, regulatory, and most DNA interacting proteins such as helicases, gyrases, and polymerases. This effort resulted in a group of 63 orthologs. Efforts were made to align these sequences relying entirely on multi-sequence alignment programs. We found about 35 orfs that had at least one representative in each of the four groups that would conveniently align; the other 28 were too highly diverged to align automatically. There were a number of ribosomal proteins in this group that were separated into another group.

In summary, we have selected a group of 26 orthologues on the basis of two criteria. 1) they have the highest degree of sequence homology among all four groups, and 2) the sample is biased against translational proteins.

Phylogenetic analysis. For convenience the four taxa phylogenies will be presented using the parentheses convention, i.e. (ab)(cd) means that a and b define one clade and is linked to the second clade, c and d. For a minimal replacement tree the number of informative characters in its support will be reflected by the length of the internal branch. The remaining informative characters will give the number of homoplasies for that tree. Where possible the four orthologs are from *E. coli* for the Gram (−), *B. subtilis* for the Gram (+), *M. jannaschii* for the Arc and Yeast for the Euk. The results of this analysis is shown in Table 2. This table shows the seven cases where one of

Table 2. Four taxa topology test. Of the 26 genes listed in Table 1, these six contained phylogenetically useful information i.e. would support one of the three topologies to the exclusion of the other two

Gene	Favored topology	χ^2
Aprt	(Eco, Yea)(Bsu, Mja)	54, 59
AroA	(Eco, Bsu)(Mja, Yea)	9.6, 12
	(Hin, Arf)(Bsu, Art)	27, 33
CtpS	(Eco, Bsu)(Yea, Mja)	4.8, 5.5
	(Hin, Bsu)(Art, Arf)	14, 27
Hyp2	(Eco, Bsu)(Mja, Yea)	<1, 3.6
	(Hin, Bsu)(Cae, Mth)	6, 10
Map	(Eco, Bsu)(Mja, Yea)	<1, 2.5
	(Hin, Art)(Bsu, Arf)	8, 8
Pur6	(Eco, Bsu)(Yea, Mja)	2, 4.5
	(Hin, Bsu)(Cae, Arf)	8, 12.5

The χ^2 values result from the number of homoplasies in the two alternative trees, when compared to those from the favored topology. Abbreviations: Eco, *Escherichia coil;* yea, *Saccharomyces cerevisiae;* BSU, *Bacillus subtilis;* Mja, *Methanococcus jannaschii;* Hin, *Haemophilus influenzae;* Arf, *Archaeoglobus fulgidis;* Art, *Arabidopsis thaliana;* Cae, *Caenorhabditis elegans;* Mth, *Methanobacterium thermoautotrophicum.*

Table 3. Numbers of phylogenetically informative characters supporting each tree

	L	T_1	T_2	T_3	P
Aprt	648	4	32	5	>.999
ArgH	654	10	15	5	.92
ArgT	810	2	1	1	.05
AroA	726	8	19	5	.99
AroE	492	3	7	4	.6
CtpS	627	14	4	6	.96
Enol	520	12	7	9	.4
Fmu	673	9	2	0	.99
Gpd	598	4	5	5	.25
Hyp1	327	0	2	5	.85
Hyp2	389	6	0	5	.93
IlvH	259	7	5	2	.75
IlvI	749	10	18	13	.7
Map	404	6	4	2	.65
MoaA	540	2	6	5	.65
MoaC	184	3	4	3	.05
Ndk	174	3	3	6	.4
OTC	342	9	10	4	.6
PurA	533	2	5	4	.3
PurB	810	2	6	7	.75
Pur5	454	5	6	2	.65
Pur6	174	3	0	1	.8
RisB	233	3	3	1	.3
TrpA	425	7	7	6	.1
TrpB	391	8	5	8	.2
TrpG	527	*7*	*6*	*7*	*.1*
Sum of 22 Taxa[1]		105	115	101	.4
Informational Genes					
ssuRNA	452	50	6	19	>.99
Rpl14	148	18	0	0	>.99
Rps11	138	19	3	0	>.99

L = length of minimal tree, T_n = number of phylogenetic useful characters where tree 1, 2, and 3 refers to (−, +)(Arc, Euk), (−, Euk)(Arc, +), and (−, Arc)(Euk, +), respectively.

[1] The 22 taxa are those for which $p < .95$, i.e., those whose parsimonious informative characters are uniformly distributed over the three trees.

the three trees is significantly supported over the other two. The (Arc, Euk), (Gram+, Gram−) relationship is encountered four times, while the (Arc, Gram+) (Euk, Gram−) is encountered twice. However, even in some of these cases where one of the three trees is significantly supported, the use of different taxa to represent the four major groups may yield a different topology (as, for example, AroA, Hyp2, and Map). This supports observations made by earlier investigations. That is, the eukaryotic cell has a major contribution of genes from two sources—namely, the Archaea on the one hand and the Gram (−) on the other. Possible movement of genes from Gram (+) Bacteria to Eukaryota is less common though we can see one possible example in Table 2.

In summary, of the 26 genes listed in Table 1, there are surprisingly few that yield trees that support any unique phylogeny.

Star phylogenies. In the other 20 orfs submitted to this study, we found that no unique phylogeny was supported (noted by asterisk in Table 1). Indeed, in most of these samples, the length of the three difference trees was not significantly different from each other. This would suggest that the gene tree could be represented by a star phylogeny where (a,b) is linked to (c,d) with a zero branch length. (Of course, forcing the internal branch of a four taxa tree to have a length of zero results in an overall tree that is longer than the most parsimonious tree. This is just a consequence of calling homoplastic characters phylogenetically informative and has no biological significance.) A second test was used to determine whether or not the number of phylogenetically informative characters are randomly distributed among the three different topologies. This test is described in Methods. The taxa analyzed were *E. coli, B. subtilis, M. janashi,* and yeast. The data is shown in Table 3. This table lists the same genes given in Table 1. The first column gives the length of the most parsimonious tree and the next three columns give the number of phylogenetic informative characters which support each of the three respective trees. Simple examination of the numbers in this table shows that for most of the genes informative characters support equally each of the three trees. Indeed, there are only four proteins—Aprt, AroA, Ctps, and Fmu—whose informative characters are not randomly distributed among the three alternative trees (i.e. $p > .95$).

At the bottom of the table for comparison, are listed the results for ribosomal genes that have traditionally supported the Tree 1 [(Euk, Arc), (Gram+, Gram−)] topology. As is clear, Tree 1 is highly supported by each of the three genes.

In general, the genes shown in Table 1 give very little support for the hypothesis that Archaea and Eukaryota define one clade and Gram+ and Gram− Bacteria define another. The suggestion has been made on a number of occasions that single genes may possess an insufficient number of informative characters to uniquely support any specific phylogeny and that a possible means of overcoming this obstacle is to unite the multiple genes in a single analysis—a phylogenetic meta analysis, as it were. Thus, for the 22 genes from Table 3 that yielded ambiguous topologies, the total number of informative characters are summed (see line near bottom of Table 3). As is abundantly clear, no single tree is favored by such an approach.

Possible explanation. If an underlying species phylogeny supports a bifurcated four-taxa tree but the protein sequence tree supports a star phylogeny, then the simplest explanation for this difference is that these genes have reached saturation in their divergence curves. The high degree of homology between these genes could argue against this possibility because we could imagine that each is separated into two groups; one subset of sites are highly constrained while a second has become so highly diverged that phylogenetic information is lost. An example of this problem would be to determine the phylogeny of, say, four taxa from yeast, angiosperm, mammals and porifera using, for example, the silent sites in the otherwise highly conserved actin gene. Such an analysis would yield star phylogenies since the silent site differences will have reached saturation. This explanation leads to a number of predictions as to the number of functionally constrained sites.

Ornithine transcarbamylase and its functional constraint. Let us consider the case of one gene, that for onrnithine transcarbamylase (OTCase), and address the issue if this protein's high degree of conservation among remotely related taxa could be due to unusually high levels of functional constraint. In order to make this argument, it is desirable to quantitatively compare the divergence of OTCase to an enzyme that is as closely related to OTCase as possible, and catalyses a comparable reaction. But also a protein that does not display a star phylogeny. The enzyme is aspartate transcarbamylase (ATCase), which is a paralogue of OTCase (Labedan et al. 1999). ATCase catalyses the conjugation of the carbamyl group of carbamyl phosphate to aspartate and OTCase catalyses the same conjugation to ornithine. I will first show that the rate of evolution of OTCase is nearly the same as ATCase when we compare recently diverged organisms. This is done by performing the distance matrix rate test—a graphical test that allows the rate of evolution of two different genes to be compared. In order to carry out this test, it is necessary to obtain both the ATCase and OTCase sequences from a large group of organisms and, furthermore, that both sequences are available from the same organisms. In this test, we calculate a distance matrix for ATCase from 34 different species and calculate a similar distance matrix for OTCase obtained from the same 34 species. Each matrix will have 748 entries. The entry comparing, for example, *E. coli* and *B. subtilis* in the ATCase matrix is plotted against the entry for the same two organisms from the OTCase matrix. Thus, 748 points in one matrix are plotted against the comparable 748 from the second. Figure 1 shows this plot.

For the closely related taxa (e.g., comparisons between two Eukaryotas or between two enterics), the curve in Fig. 1 shows that the rate of evolution of OTCase and ATCase are nearly the same (slope of the line through the origin is nearly one). Traditionally, in interpretations of molecular evolution data, this result suggests that the two proteins experience similar functional constraint. In this portion of the curve, I will argue that the common ancestor, that gave rise to the two organisms being compared, contained both an ATCase and and OTCase and that subsequently they evolved at nearly normal rates. In a control run using the ribosomal proteins rps11 and rpl14 (see Table 3) the distance matrix rate test was linear even to the most remotely related taxa (data not shown). However, for the more distantly related taxa, a plateau is seen in Fig. 1; even though we see that ATCase continued to diverge, OTCase has stopped. If we assume that the common ancestor that gave rise to these comparisons contained an ATCase and an OTCase that subsequently evolved to give today's taxa, we would further have to assume that, after the OTCase diverged to approximately 40% replacements, further divergence stopped. The explanation that OTCase has a subset of sites that are much more functionally constrained than comparable sites in ATCase is in apparent conflict with our earlier conclusion that ATCase and OTCase are similarly functionally constrained when we compare sequences from species that are more closely related.

There is a second way to compare the number of functionally constrained sites in these two proteins from a different type of comparison of the more closely related sequences. We can directly answer the question: Does ATCase have a larger number of variable sites than does OTCase? The results are shown in Table 4. In this comparison, the ATCase sequence from *E. coli* was aligned separately to the ATCase sequence from six different species and OTCase was aligned to OTCases from six different species. These species in each case were chosen so that the range of similarity scores was comparable (e.g., .72, .70, .69, .67, .64, and .63 for ATCase and .73, .72, .70, .67, .63, and .60 for OTCase). For each aligned sequence, I then determined whether a site in *E. coli* matched, or didn't match. The number of sites that did not vary or varied conservatively was then computed. Basically, the two proteins have the same number of variant sites. What we see is that 41% of the sites in ATCase are perfectly conserved and 56% functionally conserved while, with OTCase, the values are 37% and 54%, respectively. This shows that the number of sites

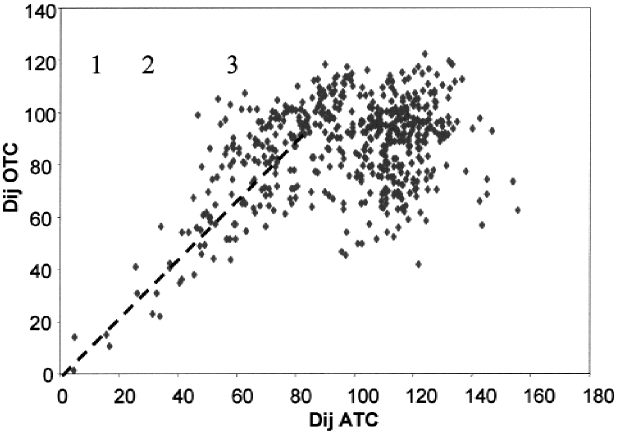

Fig. 1. Whole Matrix Rate test comparing aspartate transcarbamylase to ornithine transcarbamylase. Comparison between (1) *E. coli* and *S. typhimurium,* (2) *E. coli* and *H. influenzae,* and (3) Plants and Fungi, Fungi and Metazoa, or plants and metazoa. Representative species pairs directly below the three numbers shown in the figure are: 1) pea/ arabidopsis, *Mycobacterium tuberculosis/leprae; Thermoplasma acidophilim/volcanium,* 2) *Emericella nidulans/Schizosaccharomyces pombe, E. coli/Vibrio cholerae/Neisseria meningitidis, Lactococcus latis/Enterococcus faecalis/Streptococcus pyogenes,* 3) Human/two fungi/two plants. The taxa out on plateau include comparisons between *Archaea/Bacteria*/Eukaryotes as well as most comparison between different *Bacteria* and different *Archaea.*

Table 4. Funtionally constrained sites

	Identical	Similar
ATCase	37%	54%
OTCase	41%	56%

The ATCase and OTCase from *E. coli* were used to query the protein sequences stored in GeneBank using BLAST (Altschul et al. 1997). The outputs from each blast search were compared so that six comparisons from each could be matched on the basis of their similarity scores as noted in the text. The percentage of sites that were completely conserved (% identical) or had only conservative differences (% similar) over all six comparisons are shown.

free to vary in OTCase and ATCase are nearly the same; and the small difference cannot account for the plateau shown in Fig. 1. We repeated the above analysis except that the ATCase from *Arabidopsis* and the OTCase from pea was used. This resulted in comparing completely different sequences than were compared in Table 1—the results were nevertheless the same (data not shown).

The young gene hypothesis. Functional constraint doesn't seem to explain the high conservation of OTCase from remotely related species. I would like to suggest an alternative explanation that does not conflict with the facts. For purposes of explanation let us consider the remotely related pairs in Fig. 1 that define the plateau. I will posit that the last common ancestor for any of those pairs of taxa in the plateau did not have the modern OTCase gene. Rather, the same OTCase gene was introduced into these lineages at a time well after lineage splitting. This was not an isolated gene transfer event, but rather introductions of OTCase into multiple lineages. Indeed, this explains the star phylogeny, shown originally in Fig. 1. That is, OTCase was not present in the ancestral branch points that gave rise to modern Eukaryotas, Gram (−) and (+) Bacteria and the Archaea. Rather, I would suggest that the OTCase gene is younger than the common ancestors for these groups, and that the gene entered those already mature lineages more or less at the same time.

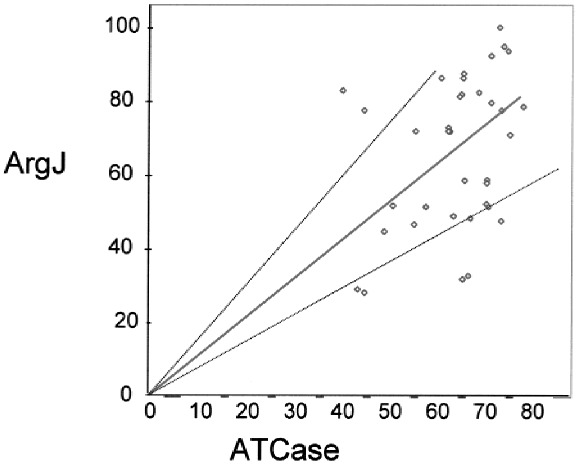

Fig. 2. Whole Matrix Rate test comparing aspartate transcarbamylase and glutamate N-acetyl transferase (argJ).

This means that horizontal gene transfer events occurred during the same period. I would suggest that OTCase likely evolved relatively recently (clearly from a gene duplication event from an ATCase gene) and that it must have offered some selective advantage for all of life's major kingdoms to take it on by horizontal gene transfer.

Examination of the genes displaying the star phylogeny in Table 3 shows that there are genes for other enzymes in the arginine biosynthetic pathway. These are enzymes that are involved in the conversion of ornithine to arginine. Enzymes responsible for the biosynthesis of ornithine, on the other hand, do not display this signature of "youthfulness." As an example, Fig. 2 shows the whole matrix rate test between argJ (the gene for acetylglutamate transferase) and ATCase. The pattern of divergence displayed here is consistent with the last common ancestors having orthologues for both arjJ and ATCase.

This would indicate that the pathway for the biosynthesis of ornithine is older than that part of the pathway that converts ornithine to arginine.

Tryptophan biosynthesis. Three genes for tryptophan biosynthesis display a star phylogeny (Table 3). (The two other tryptophan biosynthetic enzymes were eliminated from consideration because they are encoded by parologous gene families.) The sequence for the tryptophan synthetase α subunit is submitted to the whole matrix

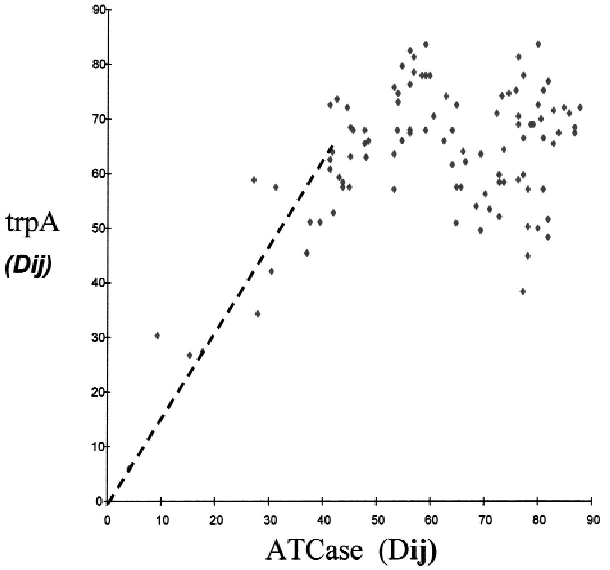

Fig. 3. Whole Matrix Rate test comparing aspartate transcarbamylase (ATCase) to tryptophan synthetase α subunit (TrpA). The dotted lines are estimated 95% confidence levels calculated assuming a random molecular clock.

rate test against ATCase (Fig. 3). As can be seen, tryptophan synthetase plotted against ATCase displays a pattern similar to that of OTCase. For the distances from the closely related taxa we see that tryptophan synthetase is evolving at approximately twice as fast as ATCase (i.e. the initial slope is close to two), indicating that TrpA is under less functional constraint than is ATCase. Though I have not established, as was done in the ATCase/OTCase comparison in Table 4, that tryptophan synthetase doesn't have a subset of invariant sites, it would appear that TrpA is even less functionally constrained than is OTCase. Therefore, I will suggest that tryptophan synthetase, like OTCase, was not present in the last common ancestor of the distantly related taxa.

Comparison of TrpA and OTC. Based on where the plateaus begin in Fig. 1 and Fig. 3, we can infer that OTCase and TrpA entered life sometime before the diversification of plants, fungi, and metazoa (more than ca. 1.1 bya) and after the divergence of Gram+ and Gram− bacteria (less than ca. 2.2 bya). If they both spread among the ancient lineages at about the same time, we would predict that when these two are compared with one another, using the distance matrix rate test, then the result should be linear. As can be seen in Fig. 4, this test is linear with most of the data falling in the 95% confidence range for a stochastic process. This result says that the last common ancestors giving rise to extant life either had both TrpA and OTC or had neither.

Further Discussion

The results in this paper describe the phylogenetic relationships of 26 genes that code for highly conserved

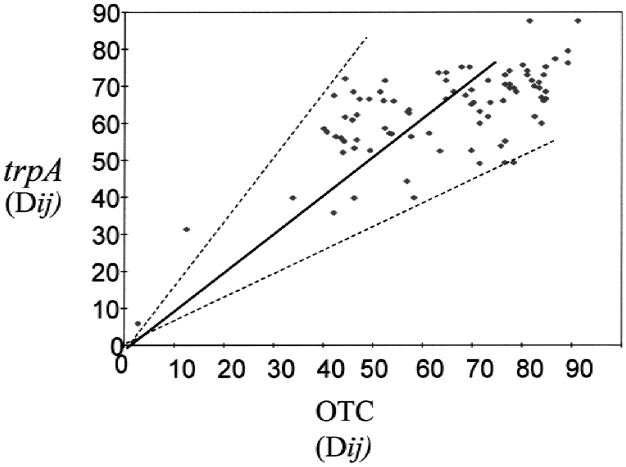

Fig. 4. Whole Matrix Rate test between ornithine transcarbamylase (OTCase) and tryptophan synthetase (TrpA). The dotted lines are estimated 95% confidence levels calculated assuming a random molecular clock.

enzymes that are found in each of the four major clades—Eukaryota, Archaea, Gram(+) and Gram(−) Bacteria. The genes were selected using the following criteria—they are orthologous and they are sufficiently highly conserved that multi-sequence alignment is convenient. Probably because of these restrictive criteria, most of the identified genes code for enzymes involved in biosynthesis of common and often essential metabolites. This was not a precondition, but it is the result.

The goal of this analysis was to gain some insight into how often horizontal gene transfer occurred during the evolution of life's major kingdoms. An effort was made to select genes arbitrarily from whole genome sequences independently of the hypothesis that I am exploring. Nevertheless, there are two biases that were introduced. First, and this was deliberate, ribosomal proteins, RNA polymerases, DNA polymerases, and their associated cofactors were excluded. The reason for this is that repeated analyses of these proteins have shown that the gene trees for these proteins support the 18S rRNA tree. The second bias was that the most highly conserved proteins were selected. A number of proteins that display high functional constraint have been noted—these include actins, tubulins, histones, and calmodulins. However, these are proteins that do not have any apparent homologs among the Archaea and Bacteria and thus would not have been selected.

It, therefore, appears that a group of proteins were selected in the current study that would not have attracted notice as a group from earlier analyses. Of those that have informative phylogenies, we see that four are consistent with the three kingdom hypothesis of Woese, two place Gram(+) and Archaea in a clade, and one places Gram(−) and Archaea in a clade. This result is consistent with earlier analyses. The trees for most genes seem to be consistent with the three-kingdom hypothesis but there are enough exceptions to show that horizontal gene transfer was common between the kingdoms, and, for some organisms, quite common (Nelson et al. 1999).

The unexpected finding was the large number of gene trees that displayed the star phylogeny. Of course, encountering such a pattern would lead one to assume that such genes are so highly diverged that phylogenetic information, especially for deep branches, has been lost. However, as I have shown in the current analysis, this does not appear to be the explanation for the star phylogenies exhibited by the genes for tryptophan synthase and OTCase. Neither enzyme appears to be particularly highly functionally constrained. Rather, it appears these enzymes are highly similar among remotely related taxa as if they were younger than the extant lineages that harbor them. Though, in the current study, all of the genes in Table 1 have not been analyzed as rigorously as was done for OTCase and tryptophan synthetase, I would like to suggest that the pathways for the biosynthesis of tryptophan and for arginine (from ornithine) also behave in the same manner. Besides OTC, we see two other arginine genes in Table 1. Independently of this work, and using molecular distances and rate estimations, Atkins and Li (1998) have estimated the age of divergence for the enzyme Argininosuccinate lyase from Archaea, Bacteria, and Eukaryota at 1.48 bya. If we date the metazoan, fungi, and plant radiation at 1.0 bya (see cluster 3 in Fig. 1), we estimate that the plateau and ascending line meet between 1.5–1.9 bya. That is, this would be the best estimate for the age of OTCase, which in consistent with the age of argininosuccinate lysate measured by Adkins and Li (1998). The different enzymes for the entire tryptophan biosynthetic pathway also looks as if it may display the star phylogeny. Table 1 shows that TrpA, B, and G are listed. Trp C, D, and E could also have been included except that they are members of paralogous gene families and they were eliminated from further study when paralogues were removed. Efforts to go back and study those genes is underway; the presence of paralogues complicates interpretation of these data.

This raises the question, why do these pathways, though found in all of life's kingdoms, appear to be younger than the lineages that carry them? The argument that I will make is that these two pathways did not exist in the last common ancestor for life's major kingdoms, rather, these pathways are more recent evolutionary inventions and moved into the multiple lineages by horizontal gene transfer. A number of scenarios by which this may have occurred can be imagined ranging from improved efficiencies of biosynthetic pathways to continued evolution of the genetic code after the diversification of life's major kingdoms (Syvanen 1985).

A main suggestion of this paper is that the presence of the same biochemistry in two different lineages does not mean that their last common ancestor necessarily showed this trait. In fact, considerable energy has been expended in examining the biological unities found in all life as a means of gaining insight into the last universal cellular ancestor (LUCA) (to cite just some of the most recent papers, DiRuggiero et al. 1999; Philippe and Fortrerre 1999; Castresana and Moreisa 1999; Leipe et al. 1999; Doolittle 2000; Labedan et al. 1999). The implications from the present work is that many of these determinations will reveal little about early ancestors. There has been recent discussion that horizontal gene transfer is so frequent that it may never be possible to reconstruct the last common ancestor (Kandler 1994; Doolittle 1999, 2000). However, if biochemical unities could be achieved after speciation events by horizontal gene transfer, then there is no reason to even postulate that a LUCA ever existed. If horizontal gene transfer is as common as I am implying, the modern cell could have evolved in multiple parallel lineages. Earliest life could have been truly polyphyletic.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Adkins RM, Li W-H (1998) Dating the age of last common ancestor of all living organisms with a protein clock. In: Horizontal gene transfer Syvanen M, C Kado (eds) Chapman and Hall, London p. 463

Brinkmann H, Philippe H (1999) Archaea sister group of *Bacteria?* Indications from tree reconstruction artifacts in ancient phylogenies. Mol Biol Evol 16:817–825

Castresana J, Moreira D (1999) Respiratory chains in the last common ancestor of living organisms. J Mol Evol 49:453–460

DiRuggiero J, Brown JR, Bogert AP, Robb FT (1999) DNA repair systems in Archaea: mementos from the last universal common ancestor? J Mol Evol 49:474–484

Doolittle WF (1999) Phylogenetic classification and the universal tree. Science 284:2124–2129

Doolittle RF (2000) Searching for the common ancestor. Res Microbiol 151:85–89

Felsenstein J (1988) Phylogenies from molecular sequences: inference and reliability. Ann Rev Gen 22:521–565

Golding GB, Gupta RS (1995) Protein-based phylogenies support a chimeric origin for the eukaryotic genome. Mol Biol Evol 12:1–6

Gribaldo S, Lumia V, Creti R, de Macario EC, Sanangelantoni A, Cammarano P (1999) Discontinuous occurrence of the *hsp70* (*dnaK*) gene among Archaea and sequence features of HSP70 suggest a novel outlook on phylogenies inferred from this protein. J Bact 181:434–443

Gupta RS, Golding GB (1993) Evolution of HSP70 gene and its implications regarding relationships between archae *Bacteria,* eu *Bacteria,* and Eukaryotes. J Mol Evol 37:573–582

Gupta RS, Golding GB, Singh B (1994) HSP70 phylogeny and the relationship between archae *Bacteria,* eu *Bacteria,* and Eukaryotes. J Mol Evol 39:537–540

Jain R, Rivera MC, Lake JA (1999) Horizontal gene transfer among genomes: the complexity hypothesis. Proc Nat Acad Sci USA 96: 3801–3806

Kandler O (1994) Cell wall biochemistry and three-domain concept of life. Syst Appl Microbiol 16:501–509

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge p. 75

Kitabatake M, So MW, Tumbula DL, Soll D (2000) Cysteine biosynthesis pathway in the archaeon Methanosarcina barkeri encoded by acquired *Bacteria* 1 genes? J Bact 182:143–145

Koonin EV, Galperin MY (1997) Prokaryotic genomes: the emerging paradigm of genome-based microbiology. Curr Op Gen Devel 7: 757–763

Labedan B, Boyen A, Baetens M, Charlier D, Chen P, Cunin R, Durbeco V, Glansdorff N, Herve G, Legrain C, et al. (1999) The evolutionary history of carbamoyltransferases: a complex set of paralogous genes was already present in the last universal common ancestor. J Mol Evol 49:461

Leipe DD, Aravind L, Koonin EV (1999) Did DNA replication evolve twice independently? Nucl Acids Res 27:3389–3401

Makarova KS, Aravind L, Galperin MY, Grishin NV, Tatusov RL, Wolf YI (1999) Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. Gen Res 9:608–628

Nelson KE, Clayton RA, Gill SR, Gwinn MI, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA, et al. (1999) Evidence for lateral gene transfer between Archaea and *Bacteria* from genome sequence of Thermotoga maritima. Nature 399:323–329

Philippe H, Forterre P (1999) The rooting of the universal tree of life is not reliable. J Mol Evol 49:509–523

Ponting CP, Aravind L, Schultz J, Bork P, Koonin EV (1999) Eukaryotic signaling domain homologues in Archaea and *Bacteria.* Ancient ancestry and horizontal gene transfer. J Mol Biol 289:729–745

Saito R, Tomita M (1999) Computer analyses of complete genomes suggest that some archae *Bacteria* employ both eukaryotic and eu *Bacteria* 1 mechanisms in translation initiation. Gene 238:79–83

Smith MW, Feng DF, Doolittle RF (1992) Evolution by acquisition: the case for horizontal gene transfers. Trends Biochem Sci 17:489–493

Syvanen M (1985) Cross-species gene transfer; implications for a new theory of evolution. J Theo Biol 112:333–343

Syvanen M (1987) Molecular clocks and evolutionary relationships: possible distortions due to horizontal gene flow. J Mol Evol 26:16–23

Woese C (1998) The universal ancestor. Proc Natl Acad Sci USA 95: 6854–6859

Wolf YI, Aravind L, Grishin NV, Koonin EV (1999) Evolution of aminoacyl-tRNA synthetases—analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. Gen Res 9:689–710

Worning P, Jensen LJ, Nelson KE, Brunak S, Ussery DW (2000) Structural analysis of DNA sequence: evidence for lateral gene transfer in Thermotoga maritima. Nuc Acids Res 28:706–709