Nutrients **2015**, *7*, 1-16; doi:10.3390/nu7010001

OPEN ACCESS

nutrients

ISSN 2072-6643

www.mdpi.com/journal/nutrients

Article

Metazoan Remaining Genes for Essential Amino Acid Biosynthesis: Sequence Conservation and Evolutionary Analyses

Igor R. Costa [1], Julie D. Thompson [2], José Miguel Ortega [3] and Francisco Prosdocimi [1,*]

[1] Instituto de Bioquímica Médica Leopoldo de Meis, Universidade Federal do Rio de Janeiro, Rio de Janeiro 21941-902, RJ, Brazil; E-Mail: igorc@ufrj.br

[2] Department of Computer Science Research, ICube Laboratoire des sciences de l'ingénieur, de l'informatique et de l'imagerie, CNRS/Université de Strasbourg, 11 rue Humann, Strasbourg F-67000, France; E-Mail: thompson@unistra.fr

[3] Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte 31270-901, MG, Brazil; E-Mail: miguel@icb.ufmg.br

* Author to whom correspondence should be addressed; E-Mail: prosdocimi@bioqmed.ufrj.br; Tel.: +55-21-3938-6789; Fax: +55-21-3938-8647.

Received: 16 May 2014 / Accepted: 25 November 2014 / Published: 24 December 2014

Abstract: Essential amino acids (EAA) consist of a group of nine amino acids that animals are unable to synthesize via *de novo* pathways. Recently, it has been found that most metazoans lack the same set of enzymes responsible for the *de novo* EAA biosynthesis. Here we investigate the sequence conservation and evolution of all the metazoan remaining genes for EAA pathways. Initially, the set of all 49 enzymes responsible for the EAA *de novo* biosynthesis in yeast was retrieved. These enzymes were used as BLAST queries to search for similar sequences in a database containing 10 complete metazoan genomes. Eight enzymes typically attributed to EAA pathways were found to be ubiquitous in metazoan genomes, suggesting a conserved functional role. In this study, we address the question of how these genes evolved after losing their pathway partners. To do this, we compared metazoan genes with their fungal and plant orthologs. Using phylogenetic analysis with maximum likelihood, we found that acetolactate synthase (ALS) and betaine-homocysteine *S*-methyltransferase (BHMT) diverged from the expected Tree of Life (ToL) relationships. High sequence conservation in the paraphyletic group Plant-Fungi was identified for these two genes using a newly developed Python algorithm. Selective pressure analysis of ALS and BHMT protein sequences showed higher non-synonymous mutation ratios in comparisons between metazoans/fungi and metazoans/plants, supporting the hypothesis that these two genes have undergone non-ToL evolution in animals.

Keywords: comparative genomics; essential amino acids; molecular evolution

1. Introduction

It has been known for decades that most animal cells are incapable of growing in a medium lacking amino acid supplements [1]. This means that pathways for *de novo* amino acid biosynthesis are missing in their genomes, characterizing the Essential Amino Acid (EAA) phenotype. There is no consensus over the exact number of essential amino acids, but it is normally accepted that His, Ile, Leu, Lys, Met, Phe, Thr, Trp and Val belong to this group. Although this phenotype was discovered in 1932, it was only recently that a two research groups independently studied a large number of animal genomes in an attempt to identify precisely which genes encoding enzymes for the EAA pathways are absent in animal genomes [2,3].

The EAA phenotype was found to be apomorphic in the metazoan clade, indicating that a number of genes were lost in the genome of an ancestral heterotroph organism. Moreover, it has been reported that the loss of an entire pathway for amino acid biosynthesis could also be observed in a number of other groups of organisms, mainly in bacteria and protists [2] capable of obtaining that specific amino acid from their surroundings.

The metazoan ancestor must have had a considerable supply of amino acids in order for it to survive after the complete loss of many EAA biosynthesizing enzymes. Three non-exclusive explanations could justify the increased availability of dietary amino acids: (i) the development of heterotrophy, after the acquisition of a digestive cavity or unicellular predation; (ii) the association with symbiotic organisms that provided the amino acid supply [2]; or (iii) the acquisition of efficient transmembrane transporters. In the metazoan lineage, no organism has been found that produces an intermediate number of essential amino acids, with possible exceptions in *Cnidaria* [3,4], suggesting a unique deletion event.

Hypothetically, when organisms lose key enzymes in a pathway, genes encoding their upstream or downstream pathway partners should experience relaxation in the selection pressure. Such relaxation has been observed in whole genome duplication events, since many gene copies turn out to be unnecessary [5–7]. In other extreme cases, such as in the metazoan EAA pathways, we suggest that this relaxation may lead to a pseudogenization cascade that might result in the deletion of each and every gene participating in a given pathway. However, when performing the evolutionary analysis of EAA biosynthetic pathways [2,8,9] in animals, our group found that some genes from these pathways are still present in metazoan genomes. In this manuscript, we analyze the sequence conservation and evolutionary fate of those metazoan REmaining GENes (ReGens).

Two different hypotheses could explain the presence of these genes in metazoan genomes: (i) the first and most obvious reason would be that they continue to perform the same functions they once did, or a subset of those. Most of the biochemical reactions originally performed by enzymes in the EAA biosynthesis pathway might be performed by symbiotic bacteria. This might provide the animal cells with the intermediate metabolites for the missing steps and the enzymes could use these as substrates to complement their biosynthesis pathways. It is also known that some proteins are able to perform different reactions at the same catalytic site, using similar substrates and generating distinct, but related products.

The enzymes for the EAA synthesis might participate in such anaplerotic pathways and perform the same biochemical reaction. Moreover, it is well known that enzymes involved in biosynthetic pathways are often capable of working in the reverse reactions and at least some of the remaining enzymes might be used in the degradation steps for their respective amino acid. Thus, the selective pressure relaxation caused by the loss of pathway partners might not be enough to modify these proteins considerably or cause their complete pseudogenization.

An alternative hypothesis is that these remaining genes may have evolved in some particular way only in the metazoan clade, as a consequence of losing their partners. We can imagine a scenario in which entire regions or protein domains related to the lost function would accumulate neutral substitutions and/or undergo positive selection. It is also possible that the remaining genes could accumulate mutations and acquire a new function. In the evolution after gene duplication framework such events would be called subfunctionalization and neofunctionalization, respectively.

As we know from gene duplication studies, genes may change their function with only a small number of mutations, characterizing neofunctionalization. This makes neofunctionalization events especially difficult to infer based solely on sequence analyses of distant clades. Nevertheless, we believe that an EAA remaining gene showing strong signals of non-ToL evolution in metazoans is unlikely to have exactly the same functions as their autotrophic homologs.

Using a three clade comparison model (plants, fungi and metazoans), we investigated the effect of the loss of pathway partners on the conservation of the ReGens. We took advantage of the fact that (i) plants and fungi are known to be autotrophic organisms capable of producing all amino acids by *de novo* pathways and (ii) fungi and metazoans share a most recent common ancestor and group together in the *Opisthokonta* clade according to the Tree of Life (ToL) (Figure 1). Therefore, we used a set of molecular evolutionary metrics, including nucleotide and amino acid conservation, phylogeny using maximum likelihood, and non-synonymous mutation (Ka) ratio, as well as newly developed bioinformatics strategies to determine the fate of metazoan ReGens. Genes will be classified under ToL or non-ToL evolutionary models whether the observed results respect or not the *Opisthokonta* clade topology.

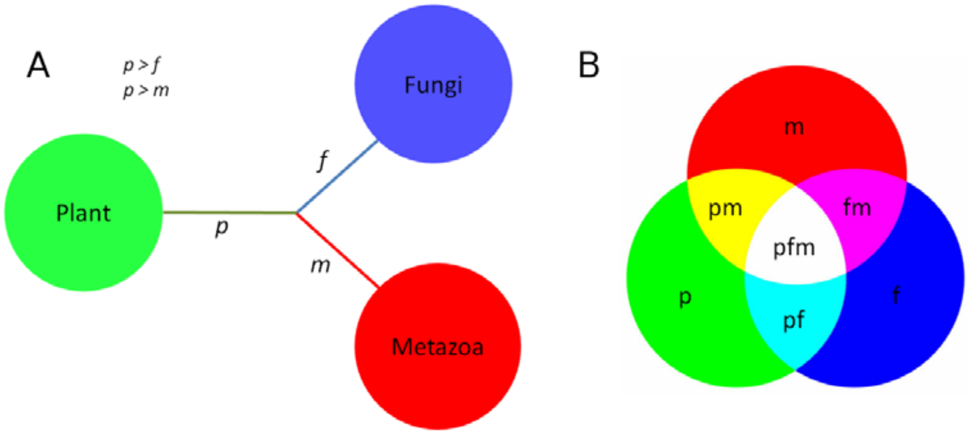


Figure 1. Clade comparison and color code definitions. (**A**) Tree of Life topology of the clades compared here: a more recent common ancestor is known between fungi and metazoans (*Opisthokonta* clade); (**B**) color code for clade comparison.

Here we study the sequence conservation of the whole set of ReGens, consisting of eight enzymes originally involved in the EAA biosynthetic pathways and shown to be present in the human and most animal genomes [2].

2. Experimental Section

2.1. Database Searches for Genes Involved in Essential Amino Acid Biosynthesis

Based on previous articles [2], we performed sequence searches to find all the enzymes responsible for the biosynthesis of essential amino acids in yeast (*Saccharomyces cerevisiae*). Each gene in this set was used as a query for a BLAST-based similarity search [10] in metazoan genomes (i) to retrieve putative homologs and (ii) to search for deleted genes. Genes were considered homologs based on annotation and on being able to retrieve the original query using BLAST in the yeast genome.

2.2. Curation of Metazoan Homologs of Yeast EAA Biosynthetic Enzymes

All of the eight metazoan homologous genes involved in EAA biosynthesis were manually investigated in databases such as UNIPROT [11], BRENDA [12] and KEGG [13], for their known molecular function.

2.3. Phylogenetic Analysis of ReGens

BLAST searches in the NCBI Refseq database were conducted, in order to find putative homologs from other species of plants, fungi and metazoans for each of the eight curated genes. These protein sequences were multiply aligned using ClustalW [14], Muscle [15] and MAFFT [16] using default parameters. Phylogenetic analyses were then performed using the PhyML software [17]. The evolutionary history was inferred using the Maximum Likelihood method based on the JTT matrix [18] of amino acid substitution. Five-hundred bootstrap replicates were used to evaluate the phylogenetic consistency.

2.4. Conservation Pattern

A Python script was developed to visualize the amino acid conservation profiles of ReGens and their putative homologs in fungi and plants. Amino acid sequences from all clades were multiply aligned using ClustalW and each residue was compared, using the BLOSUM 62 matrix, with all residues located in the same alignment column (representing the same molecular character) (Supplementary material 1). Amino acid conservation scores were divided into groups for pairwise clade comparisons: Metazoan X Fungi, Fungi X Plants and Plants X Metazoan. The character-to-character BLOSUM value is summed for each alignment column and finally averaged for each pairwise clade comparison. This average value is further filtered to remove the high frequency noise using a low pass filter (See formula below).

$$y_n = \alpha x_n + (1 - \alpha) y_{n-1} \qquad (1)$$

This filter is used to observe local tendencies of conservation, instead of the score of a single position. It filters the average comparison score for every position in the multiple alignment (x_n), multiplying this value by a constant alpha, $0 < \alpha < 1$ (we used $\alpha = 0.05$, higher means less noise and less resolution) to produce the final value shown on the y-axis at the n^{th} position (y_n). As this filter is directional, we plot the average results for both the original alignment and the reverse alignment. This program uses the graphical library MatPlotLib [19] to plot the results.

2.5. Back Translation

The ClustalW multiple alignments of proteins from all clades were back-translated to nucleotide sequences using a Python script that searches and replaces amino acids by their original codons.

2.6. Ka/Ks Estimation Using PAML

Pairwise and branch specific Ka/Ks estimation was done using the codeml program of the PAML package, using the PhyML maximum likelihood tree and the back translated multiple alignment as inputs. Pairwise comparisons were separated into three categories, Fungi X Metazoans, Fungi X Plants and Metazoans X Plants. Finally, the Ka/Ks values were averaged for each clade comparison.

3. Results

3.1. Finding and Describing Human ReGens

A complete dataset containing all the yeast (*Saccharomyces cerevisiae*) protein sequences for EAA biosynthesis, consisting of 49 proteins, was produced through manual curation in the UniProt database [11] (see Supplementary material 2). All these sequences were used as BLAST queries to search the RefSeq [20] database for putative homologs in metazoan genomes. Eight of the 49 enzymes were found to have clearly identifiable homologs in metazoans, as estimated by sequence identity and protein coverage (Table 1). They were carefully annotated and their molecular function was manually investigated in both the literature and databases such as UNIPROT [21], KEGG [13], GO [22], Brenda [12] and CDD [23].

We confirmed that five of the eight ReGens are currently annotated as members of other non-EAA metabolic pathways. For instance, besides participating in the biosynthetic pathway for the EAA methionine, cystathionine gamma-lyase also participates in the biosynthesis of the non-EAA cysteine. We also verified that at least one ReGen, the aromatic/aminoadipate aminotransferase 1 was not found in plants, since organisms in this clade use a different pathway to synthesize lysine (Table 1). This enzyme was therefore excluded from subsequent analyses.

Thus, saccharopine dehydrogenase (SD), betaine-homocysteine *S*-methyltransferase (BHMT) and acetolactate synthase (ALS) were the three remaining genes with no other known function assigned besides EAA biosynthesis.

Table 1. Essential amino acids (EAA) biosynthesis enzymes with homologs in metazoans.

Enzyme Name	Acronym	Yeast ID	Human ID	EC Number	EAA	Pathway Functions
Acetolactate synthase	ALS	P07342.1	NP_006835.2	2.2.1.6	Val, Leu, Ile	Second reaction of the branched-chain amino acid biosynthesis pathway
Betaine-homocysteine *S*-methyltransferase	BHMT	Q12525.1	NP_001704.2	2.1.1.5 2.1.1.10	Met	Last reaction of the Met biosynthesis pathway
Branched-chain-amino acid aminotransferase cytosolic	BCA	P47176.1	NP_005495.2	2.6.1.42	Val, Leu, Ile	Last reaction of the branched-chain amino acid biosynthesis pathway
Saccharopine dehydrogenase	SD	P38999.1	NP_005754.2	1.5.1.7	Lys	Last reaction of the Lys biosynthesis pathway
Cystathionine gamma-lyase	CTH	P31373.2	NP_001893.2	4.4.1.1	Met	Biosynthesis of Cys and Met
Aspartate aminotransferase, mitochondrial	AATm	Q01802.2	NP_002071.2	2.6.1.1	Phe	Ala, Asp, Glu, Cys, Met, Arg, Pro, Tyr, and Phe metabolism
Aspartate aminotransferase, cytoplasmic	AATc	P23542.3	NP_002070.1	2.6.1.1	Phe	Ala, Asp, Glu, Cys, Met, Arg, Pro, Tyr, and Phe metabolism
Aromatic/aminoadipate aminotransferase 1	AadAT *	P53090	NP_057312.1	2.6.1.39	Lys	Fifth reaction of the Lys biosynthesis pathway

* AadAT is absent in plants, since they use an alternative pathway to produce Lys. This last enzyme was not analyzed here.

3.2. Known Functions of the SD, BHMT and ALS Proteins

The human homolog of saccharopine dehydrogenase (SD) was first described in 2000 [24] and confirmed to function in the degradation pathway of lysine in metazoans and plants [25,26]. SD in these clades is bi-functional and catalyzes the first and second reactions of the lysine degradation pathway. In fungi, this enzyme is encoded as two independent genes, each of them being responsible for one reaction.

Betaine-homocysteine *S*-methyltransferase (BHMT) catalyzes the reversible methylation of homocysteine to methionine [27] and has been the subject of gained renewed interest after the discovery that the blood concentration of homocysteine is related to human health problems such as: thrombosis, vascular [28–32] and congenital diseases, such as spina bifida [33,34], as well as Alzheimer disease [35,36]. The inhibition or deletion of BHMT causes hyperhomocysteinemia in mice [37,38]. Functional assays of BHMT in metazoans have shown that it uses betaine and L-homocysteine as substrates, generating L-methionine and dimethylglycine [39]. However, in fungi and plants, it uses *S*-methyl-L-methionine and L-homocystheine as substrates producing 2 L-methionine; or uses *S*-adenosyl-L-methionine and L-homocysteine generating *S*-adenosyl-L-homocysteine and L-methionine [40,41]. Moreover, a paralogous gene of BHMT has been found in therians, called BHMT2 (see Supplementary material 3). This gene performs the same molecular reaction as the plant and fungi orthologs, suggesting that BHMT may have duplicated in a therian ancestor allowing one of the copies to neofunctionalize and stop using betaine as substrate [42].

Acetolactate synthase (ALS) has been extensively studied in plants, since it is the target of four classes of high spectrum herbicides: (i) sulfonylureas [43]; (ii) imidazolines [44]; (iii) triazolopyrimidines and (iv) pyrimidyl-oxy-benzoates [45]. These herbicides are considered relatively safe for humans, since it is assumed we lack the ALS enzyme due to our inability to synthetize branched-chain amino acids [46]. Joutel *et al.* [47] described a human ALS homolog for the first time, while looking for the genetic cause of a mental disorder called CADASIL (Cerebral Autosomal Dominant Arteriopathy with Subcortical Infarcts and Leucoencephalopathy). They subsequently found that this gene is conserved in every metazoan they tested and it is ubiquitously expressed in human tissues [47]. We were unable to find a definitive source of information about human ALS functions. One group published an attempt to express the human homolog of the ALS in *E. coli* [48], but they failed to measure significant acetolactase synthase activity of the recombinant enzyme and proposed that the human gene does not encode ALS. However, due to incorrect folding and lack of posttranslational modifications and other limitations of the bacterial expression model, this experiment did not determine whether or not the human ALS has the original acetolactate synthase function. To our knowledge, there have been no other studies published using knockout animal models, or any other attempts to characterize metazoan ALS homologs.

3.3. Conservation of Protein Sequences between Clades

The sequences encoding the homologs of ReGens in a number of selected organisms were retrieved from the RefSeq database (Table 2). Manually curated orthologous groups were created for each ReGen. In the case where a given ReGen presented multiple family members (paralogs) in a given genome, we chose the version that was the most similar to the fungi sequence used as query.

All multiple alignment tools provided efficient and congruent sequence alignments (data not shown) and we proceeded with ClustalW data (see Supplementary material 4).

Using the Python programming language, together with Biopython and MatProtLib libraries, we developed a visualization script to provide an easy and intuitive view of sequence conservation between tens of proteins derived from distant clades. Figure 2 and Supplementary material 5 show the amino acid conservation profiles of metazoan ReGens and their putative orthologs in fungi and plants and Supplementary material 6 presents the graphs containing standard deviation information. The profiles showed higher conservation in the *Opisthokonta* clade for the BCA, AATm and CTH (Figure 2C,D and Supplementary material 5C, magenta line), in line with the ToL-like model, although it did not identify any clear significant differences between clades for the SD and AATc (Supplementary material 5A,B).

Surprisingly, we found that some regions in ALS and BHMT protein sequences were more conserved between fungi and plants (Figure 2A,B; cyan line) than between metazoans and fungi (magenta). This result contradicts the known ToL relationships and might suggest that these two ReGens have diverged from their original function as EAA biosynthesis enzymes. Less conserved *C*-terminal and *N*-terminal protein regions produced a smaller number of comparisons due to a higher number of gaps and were removed from the plot (original alignments are provided in Supplementary material 4). In a more detailed analysis, we focused on the highly conserved regions between plants and fungi in the ALS and BHMT proteins (Supplementary material 7). The Weblogo plots [49] show that fungi and plant proteins share a larger number of differentially conserved amino acids than the other pairwise clade comparisons in the selected regions.

Table 2. Organisms studied.

Organism Name	**Clade**
Neurospora crassa OR74A	Fungi
Pyrenophora tritici-repentis	Fungi
Sclerotinia sclerotiorum 1980 UF-70	Fungi
Schizosaccharomyces pombe	Fungi
Gibberella zeae PH-1	Fungi
Aspergillus niger	Fungi
Fusarium oxysporum f. sp. *lycopersici*	Fungi
Puccinia graminis f. sp. *Tritici*	Fungi
Saccharomyces cerevisiae S288c	Fungi
Ustilago maydis 521	Fungi
Homo sapiens	Metazoan
Pan troglodytes	Metazoan
Mus musculus	Metazoan
Monodelphis domestica	Metazoan
Taeniopygia guttata	Metazoan
Anolis carolinensis	Metazoan
Xenopus tropicalis	Metazoan
Danio rerio	Metazoan
Drosophila melanogaster	Metazoan
Ciona intestinalis	Metazoan
Caenorhabditis elegans	Metazoan
Solanum tuberosum	Plant
Vitis vinifera	Plant
Populus trichocarpa	Plant
Arabidopsis thaliana	Plant
Physcomitrella patens	Plant
Chlamydomonas reinhardtii	Plant
Selaginella moellendorffii	Plant
Oryza sativa	Plant
Sorghum bicolor	Plant
Zea mays	Plant

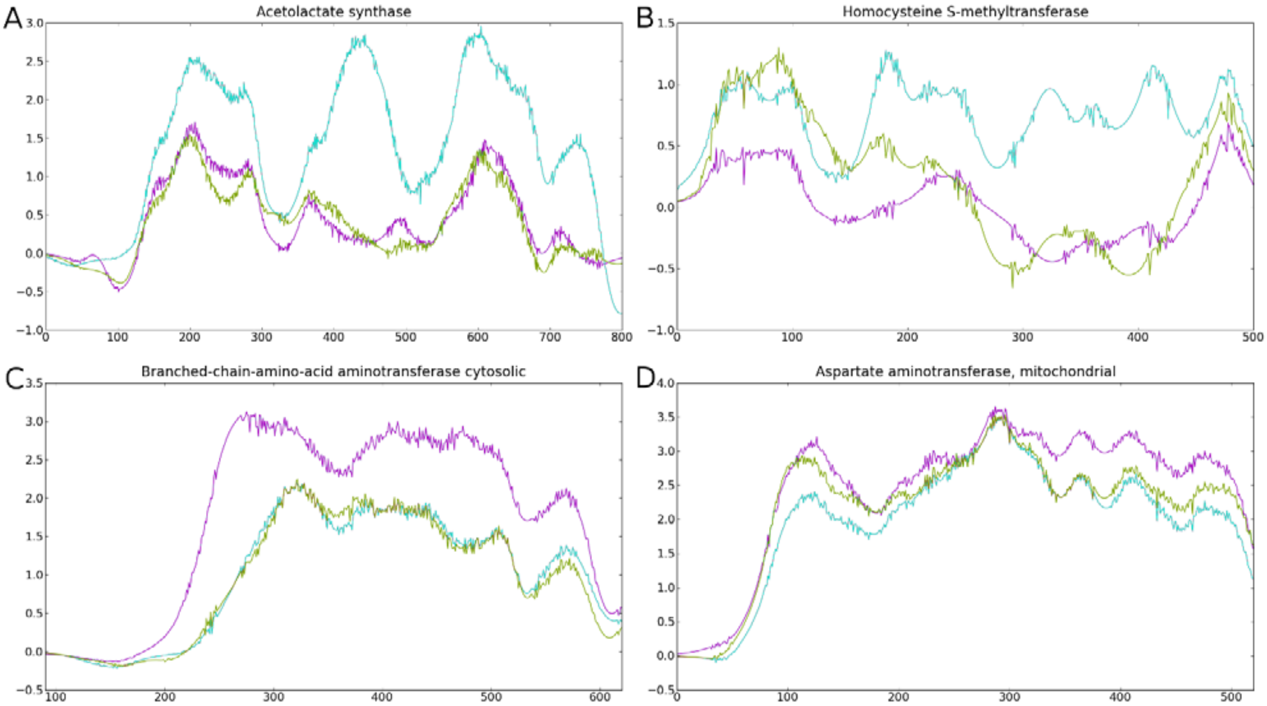


Figure 2. Amino acid conservation between clades for four ReGens. Lines are colored according to the clade comparison code defined in Figure 1B. Pairwise comparisons of proteins from each clade were plotted along their multiple alignment extension (*x*-axis). Higher values (*y*-axis) represent more similar regions in the interclade comparison. The *y*-axis represents the proportion of amino acid substitutions in a given position of the multiple alignments (*x*-axis); positions with gaps were removed. (**A**) ALS; (**B**) BHMT; (**C**) BCA; (**D**) AATm.

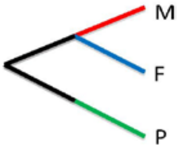
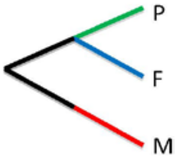
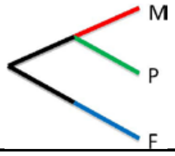
3.4. Phylogenies of ReGens Using Maximum Likelihood

Phylogenetic analysis was performed for all eight ReGens except AadAT (absent in plants), using maximum likelihood (ML) approaches based on the JTT method implemented in the PhyML package [17]. As most gene trees agree with the species tree, metazoan ReGens were expected to cluster with their fungi orthologs. If non-ToL evolution occurred in the metazoan genes, fungi and plant enzymes (still being used for amino acid biosynthesis and therefore subject to similar functional constraints) might retain a higher level of similarity. Thus, a phylogenetic analyses of these genes would result in the autotrophic (plant and fungi) genes being clustered as sister groups, while metazoan ones would be seen as an outgroup.

Corroborating the conservation analysis results, phylogeny data showed that most observed ReGens fit the ToL model, and metazoan proteins were found to cluster together with fungal proteins in the expected *Opisthokonta* clade. Table 3 resumes the tree topology information and Supplementary material 8 provides the ML trees with bootstrap values.

However, we found that the genes with considerable conservation between Fungi and Plants (Figure 2), such as ALS and BHMT, also showed tree topologies that did not support the *Opisthokonta* clade (Table 3), further supporting the hypothesis of non-ToL evolution. The BHMT and ALS trees showed the greatest ratio between the metazoan-fungi and fungi-plant relative branch lengths and represent clear examples of molecular anagenesis in the metazoan lineage. The ALS tree showed that most fungi proteins were found as a sister group of the plant clade, supporting an autotrophic paraphyletic group and suggesting non-ToL evolution of metazoan orthologs. Finally, the SD phylogenetic tree topology showed fungi as an outgroup, probably due to the existence of two distinct genes coding each of the functions.

Table 3. Summary of maximum likelihood (ML) tree topology.

Topology Name	Topology Schema	ReGen	mf/fp Branch Length Ratio
ToL-like topology		BCA	0.17
		CTH	0.31
		AATm	0.34
		AATc	1.30
Autotrophic paraphyly		ALS	3.80
		BHMT	3.59
Fungi as outgroup		SD	0.73

3.5. Synonymous and Non-Synonymous Mutation Rates

Another standard metric for analyzing the effects of natural selection on genes and proteins is the ratio between synonymous (Ka) and non-synonymous (Ks) substitutions [50]. In this evaluation, Ka/Ks scores less than one indicate that the sequences under analysis were subject to purifying selection, while values close to one indicate neutral variation and scores greater than one are evidence of positive selection.

Here, all-against-all Ka/Ks ratios for all ReGens were calculated for each pair of species using PAML pairwise algorithms and then averaged within each clade. We found that both ALS and BHMT had the highest Ka ratios for comparisons involving the metazoan clade, suggesting a faster rate of evolution (Table 4) [51]. Ks estimation was saturated mainly because of the early divergence between the clades (>1 billion years). Thus, Ka/Ks could not be precisely estimated for the complete sequences of ReGens in this timescale.

We also performed a branch test with PAML using the maximum likelihood phylogenetic trees of the previous section (Table 4), in which we tested whether the Ka/Ks ratio for the metazoan branch was significantly higher than for the rest of the tree. We found *p*-values < 0.05 for ALS, BHMT and for both the cytoplasmic and mitochondrial orthologs of AAT.

Table 4. Ka/Ks ratio as calculated by PAML.

ReGen	Clades Compared	Color Code	Ka	Ks
ALS	*fp*	🟪	0.58	57.29
	mf	🟨	1.15	37.93
	mp	🟦	1.26	28.65
BHMT	*fp*	🟪	0.89	54.54
	mf	🟨	1.74	31.84
	mp	🟦	1.45	36.84
BCA	*fp*	🟪	0.70	48.79
	mf	🟨	0.45	57.63
	mp	🟦	0.70	61.02
SD	*fp*	🟪	0.64	35.62
	mf	🟨	0.66	44.35
	mp	🟦	0.67	17.31
CTH	*fp*	🟪	0.74	48.33
	mf	🟨	0.53	53.39
	mp	🟦	0.92	34.77
AATm	*fp*	🟪	0.46	64.12
	mf	🟨	0.34	61.14
	mp	🟦	0.40	62.58
AATc	*fp*	🟪	0.47	65.63
	mf	🟨	0.43	56.60
	mp	🟦	0.46	56.57

4. Discussion

Here we have performed a conservation study of all genes involved in EAA biosynthesis that remained in metazoan genomes long after the deletion of their pathway partners. We performed manual curation of the enzymes involved in biosynthetic pathways for EAA in autotrophic organisms [2], selected the eight homologous genes present in metazoan genomes and studied the evolutionary fate of seven of them (excluding AadAT). Why are these genes retained in the metazoan genomes if they no longer participate in amino acid biosynthesis? We used standard molecular evolution metrics and developed new strategies to understand the evolutionary fate of these remaining genes.

We have shown that five of the analyzed ReGens show evidence of standard ToL-like evolution in metazoans. These genes most likely act in the biosynthesis of other amino acids or can hypothetically be used together with metabolites provided by symbionts or commensals to complement the biosynthetic pathway. Interactions between proteins from hosts and symbionts have been shown in a number of organisms, sometimes in a given tissue or during a brief development stage [52,53].

We observed favored permanence of genes at the start or end of EAA pathways, possibly because these are usually responsible for the regulation of the entire pathway. Moreover, the last genes of a biosynthetic pathway might also take part in the degradation of the final product, as is the case with BHMT and SD.

On the other hand, several of our experiments suggested non-ToL evolution of the ALS and BHMT genes in metazoan genomes (Table 5). These results are further supported by the observation that amongst the 13 amino acids annotated as belonging to the catalytic site of *Arabidopsis thaliana*'s ALS protein structure (deposited as 1YI1 [54] in the PDB database [55] and verified using the Catalytic Site Atlas [56]), 10 were found to be conserved in fungi and only 5 in metazoan proteins. Further experimental studies on ALS might indicate whether this enzyme has neofunctionalized or not.

Concerning BHMT, we found a complex evolutionary history. The eukaryotic ancestor used *S*-methyl-L-methionine as a methyl donor, while the metazoan ancestror used betaine instead. The gene seems to be duplicated as BHMT and BMHT2 in the Therian clade, allowing BHMT2 to change its substrate from betaine back to *S*-methyl-L-methionine [42].

Table 5. Summary of results.

ReGen	Tree Topology *	Conservation Diagram	Ka/Ks Branch Test ($p < 0.05$)	Ka/Ks Clade Average > 1 for Metazoans
ALS	Non-ToL	Non-ToL	+	Yes
BHMT	Non-ToL	Non-ToL	+	Yes
BCA	ToL	ToL	−	No
SD	FO	ToL	−	No
CTH	ToL	ToL	−	No
AATm	ToL	ToL	+	No
AATc	ToL	ToL	+	No

***** ToL, Tree of Life topology; FO, fungi outgroup

5. Conclusions

Eight of the 49 genes participating in the EAA biosynthetic pathways in autotrophs have been retained in metazoans (ReGens). Two of them (ALS and BHMT) show phylogenetic evidence, conservation profiles and non-synonymous mutation rates that suggest non-ToL evolution in metazoans.

Acknowledgment

The authors thank Krzysztof Siemion for beginning the manual curation and the catalog of the genes coding for essential amino acid biosynthesis pathways.

This work was funded by Brazilian funding agencies FAPERJ (E-26/111.806/2011) and CNPq (482839/2012-0).

Author Contributions

Francisco Prosdocimi and José Miguel Ortega defined the research program on the great genomic deletion of enzymes for the EAA biosynthesis. Francisco Prosdocimi and Julie D. Thompson defined the research program on the gene evolution in metazoans. Igor R. Costa developed most of the scripts, multiple alignments, performed most of the manual curation, phylogenetic analyses and Ka/Ks calculations. Igor R. Costa and Francisco Prosdocimi made the figures and drafted the first version of the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Rose, W.C. The amino acids in nutrition. *Yale J. Biol. Med.* **1932**, *4*, 519–536.
2. Guedes, R.L.; Prosdocimi, F.; Fernandes, G.R.; Moura, L.K.; Ribeiro, H.A.; Ortega, J.M. Amino acids biosynthesis and nitrogen assimilation pathways: A great genomic deletion during eukaryotes evolution. *BMC Genomics* **2011**, *12*, S2, doi:10.1186/1471-2164-12-S4-S2.
3. Hernandez-Montes, G.; Diaz-Mejia, J.J.; Perez-Rueda, E.; Segovia, L. The hidden universal distribution of amino acid biosynthetic networks: A genomic perspective on their origins and evolution. *Genome Biol.* **2008**, *9*, R95, doi:10.1186/gb-2008-9-6-r95.
4. Starcevic, A.; Akthar, S.; Dunlap, W.C.; Shick, J.M.; Hranueli, D.; Cullum, J.; Long, P.F. Enzymes of the shikimic acid pathway encoded in the genome of a basal metazoan, nematostella vectensis, have microbial origins. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 2533–2537.
5. Zhang, J. Evolution by gene duplication: An update. *Trends Ecol. Evol.* **2003**, *18*, 292–298.
6. Lynch, M.; Conery, J.S. The evolutionary fate and consequences of duplicate genes. *Science* **2000**, *290*, 1151–1155.
7. Kondrashov, F.A.; Rogozin, I.B.; Wolf, Y.I.; Koonin, E.V. Selection in the evolution of gene duplications. *Genome Biol.* **2002**, *3*, RESEARCH0008, doi:10.1186/gb-2002-3-2-research0008.
8. Prosdocimi, F.; Mudado, M.A.; Ortega, J.M. A set of amino acids found to occur more frequently in human and fly than in plant and yeast proteomes consists of non-essential amino acids. *Comput. Biol. Med.* **2007**, *37*, 159–165.
9. Santana-Santos, L.; Prosdocimi, F.; Ortega, J.M. Essential amino acid usage and evolutionary nutrigenomics of eukaryotes—Insights into the differential usage of amino acids in protein domains and extra-domains. *Genet. Mol. Res.* **2008**, *7*, 839–852.
10. Altschul, S.F.; Madden, T.L.; Schaffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402.
11. The UniProt Consortium. Reorganizing the protein space at the universal protein resource (UniProt). *Nucleic Acids Res.* **2012**, *40*, D71–D75.
12. Scheer, M.; Grote, A.; Chang, A.; Schomburg, I.; Munaretto, C.; Rother, M.; Sohngen, C.; Stelzer, M.; Thiele, J.; Schomburg, D. BRENDA, the enzyme information system in 2011. *Nucleic Acids Res.* **2011**, *39*, D670–D676.
13. Kanehisa, M. The KEGG database. *Novartis Found. Symp.* **2002**, *247*, 91–101.
14. Larkin, M.A.; Blackshields, G.; Brown, N.P.; Chenna, R.; McGettigan, P.A.; McWilliam, H.; Valentin, F.; Wallace, I.M.; Wilm, A.; Lopez, R.; *et al.* Clustal W and clustal X version 2.0. *Bioinformatics* **2007**, *23*, 2947–2948.
15. Edgar, R.C. Muscle: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **2004**, *5*, 113, doi:10.1186/1471-2105-5-113.

16. Katoh, K.; Misawa, K.; Kuma, K.; Miyata, T. MAFFT: A novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Res.* **2002**, *30*, 3059–3066.
17. Guindon, S.; Dufayard, J.F.; Lefort, V.; Anisimova, M.; Hordijk, W.; Gascuel, O. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst. Biol.* **2010**, *59*, 307–321.
18. Jones, D.T.; Taylor, W.R.; Thornton, J.M. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **1992**, *8*, 275–282.
19. Hunter, J.D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95.
20. Pruitt, K.D.; Tatusova, T.; Brown, G.R.; Maglott, D.R. NCBI reference sequences (RefSeq): Current status, new features and genome annotation policy. *Nucleic Acids Res.* **2012**, *40*, D130–D135.
21. Apweiler, R.; Bairoch, A.; Wu, C.H. Protein sequence databases. *Curr. Opin. Chem. Biol.* **2004**, *8*, 76–80.
22. Gene-Ontology-Consortium. The gene ontology: Enhancements for 2011. *Nucleic Acids Res.* **2012**, *40*, D559–D564.
23. Marchler-Bauer, A.; Lu, S.; Anderson, J.B.; Chitsaz, F.; Derbyshire, M.K.; DeWeese-Scott, C.; Fong, J.H.; Geer, L.Y.; Geer, R.C.; Gonzales, N.R.; *et al.* CDD: A conserved domain database for the functional annotation of proteins. *Nucleic Acids Res.* **2011**, *39*, D225–D229.
24. Sacksteder, K.A.; Biery, B.J.; Morrell, J.C.; Goodman, B.K.; Geisbrecht, B.V.; Cox, R.P.; Gould, S.J.; Geraghty, M.T. Identification of the alpha-aminoadipic semialdehyde synthase gene, which is defective in familial hyperlysinemia. *Am. J. Hum. Genet.* **2000**, *66*, 1736–1743.
25. Markovitz, P.J.; Chuang, D.T. The bifunctional aminoadipic semialdehyde synthase in lysine degradation. Separation of reductase and dehydrogenase domains by limited proteolysis and column chromatography. *J. Biol. Chem.* **1987**, *262*, 9353–9358.
26. Serrano, G.C.; Rezende e Silva Figueira, T.; Kiyota, E.; Zanata, N.; Arruda, P. Lysine degradation through the saccharopine pathway in bacteria: LKR and SDH in bacteria and its relationship to the plant and animal enzymes. *FEBS Lett.* **2012**, *586*, 905–911.
27. Finkelstein, J.D.; Martin, J.J. Homocysteine. *Int. J. Biochem. Cell Biol.* **2000**, *32*, 385–389.
28. Jacobsen, D.W. Homocysteine and vitamins in cardiovascular disease. *Clin. Chem.* **1998**, *44*, 1833–1843.
29. McCully, K.S. Vascular pathology of homocysteinemia: Implications for the pathogenesis of arteriosclerosis. *Am. J. Pathol.* **1969**, *56*, 111–128.
30. Blom, H.J.; Smulders, Y. Overview of homocysteine and folate metabolism. With special references to cardiovascular disease and neural tube defects. *J. Inherit. Metab. Dis.* **2011**, *34*, 75–81.
31. Weisberg, I.S.; Park, E.; Ballman, K.V.; Berger, P.; Nunn, M.; Suh, D.S.; Breksa, A.P., III; Garrow, T.A.; Rozen, R. Investigations of a common genetic variant in betaine-homocysteine methyltransferase (BHMT) in coronary artery disease. *Atherosclerosis* **2003**, *167*, 205–214.
32. Heil, S.G.; Lievers, K.J.; Boers, G.H.; Verhoef, P.; den Heijer, M.; Trijbels, F.J.; Blom, H.J. Betaine-homocysteine methyltransferase (BHMT): Genomic sequencing and relevance to hyperhomocysteinemia and vascular disease in humans. *Mol. Genet. Metab.* **2000**, *71*, 511–519.
33. Morin, I.; Platt, R.; Weisberg, I.; Sabbaghian, N.; Wu, Q.; Garrow, T.A.; Rozen, R. Common variant in betaine-homocysteine methyltransferase (BHMT) and risk for spina bifida. *Am. J. Med. Genet. A* **2003**, *119A*, 172–176.

34. Zhu, H.; Curry, S.; Wen, S.; Wicker, N.J.; Shaw, G.M.; Lammer, E.J.; Yang, W.; Jafarov, T.; Finnell, R.H. Are the betaine-homocysteine methyltransferase (BHMT and BHMT2) genes risk factors for spina bifida and orofacial clefts? *Am. J. Med. Genet. A* **2005**, *135*, 274–277.
35. Smach, M.A.; Jacob, N.; Golmard, J.L.; Charfeddine, B.; Lammouchi, T.; Ben Othman, L.; Dridi, H.; Bennamou, S.; Limem, K. Folate and homocysteine in the cerebrospinal fluid of patients with Alzheimer's disease or dementia: A case control study. *Eur. Neurol.* **2011**, *65*, 270–278.
36. Morris, M.S. Homocysteine and Azheimer's disease. *Lancet Neurol.* **2003**, *2*, 425–428.
37. Teng, Y.W.; Mehedint, M.G.; Garrow, T.A.; Zeisel, S.H. Deletion of betaine-homocysteine *S*-methyltransferase in mice perturbs choline and 1-carbon metabolism, resulting in fatty liver and hepatocellular carcinomas. *J. Biol. Chem.* **2011**, *286*, 36258–36267.
38. Collinsova, M.; Strakova, J.; Jiracek, J.; Garrow, T.A. Inhibition of betaine-homocysteine *S*-methyltransferase causes hyperhomocysteinemia in mice. *J. Nutr.* **2006**, *136*, 1493–1497.
39. Pajares, M.A.; Perez-Sala, D. Betaine homocysteine *S*-methyltransferase: Just a regulator of homocysteine metabolism? *Cell. Mol. Life Sci.* **2006**, *63*, 2792–2803.
40. Thomas, D.; Becker, A.; Surdin-Kerjan, Y. Reverse methionine biosynthesis from *S*-adenosylmethionine in eukaryotic cells. *J. Biol. Chem.* **2000**, *275*, 40718–40724.
41. Ranocha, P.; Bourgis, F.; Ziemak, M.J.; Rhodes, D.; Gage, D.A.; Hanson, A.D. Characterization and functional expression of cDNAs encoding methionine-sensitive and -insensitive homocysteine *S*-methyltransferases from arabidopsis. *J. Biol. Chem.* **2000**, *275,* 15962–15968.
42. Szegedi, S.S.; Castro, C.C.; Koutmos, M.; Garrow, T.A. Betaine-homocysteine *S*-methyltransferase-2 is an *S*-methylmethionine-homocysteine methyltransferase. *J. Biol. Chem.* **2008**, *283*, 8939–8945.
43. Chaleff, R.S.; Mauvais, C.J. Acetolactate synthase is the site of action of two sulfonylurea herbicides in higher plants. *Science* **1984**, *224*, 1443–1445.
44. Saxena, P.K.; King, J. Herbicide resistance in *Datura innoxia*: Cross-resistance of sulfonylurea-resistant cell lines to imidazolinones. *Plant Physiol.* **1988**, *86*, 863–867.
45. Subramanian, M.V.; Hung, H.Y.; Dias, J.M.; Miner, V.W.; Butler, J.H.; Jachetta, J.J. Properties of mutant acetolactate synthases resistant to triazolopyrimidine sulfonanilide. *Plant Physiol.* **1990**, *94*, 239–244.
46. Pang, R.G.D.A.S.S. Acetohydroxyacid synthase. *J. Biochem. Mol. Biol.* **2000**, *33*, 1–36.
47. Joutel, A.; Ducros, A.; Alamowitch, S.; Cruaud, C.; Domenga, V.; Marechal, E.; Vahedi, K.; Chabriat, H.; Bousser, M.G.; Tournier-Lasserve, E. A human homolog of bacterial acetolactate synthase genes maps within the cadasil critical region. *Genomics* **1996**, *38*, 192–198.
48. Duggleby, R.G.; Kartikasari, A.E.R.; Wunsch, R.M.; Lee, Y.T.; Kil, M.W.; Shin, J.Y.; Chang, S.I. Expression in *Escherichia coli* of a putative human acetohydroxyacid synthase. *J. Biochem. Mol. Biol.* **2000**, *33*, 195–201.
49. Crooks, G.E.; Hon, G.; Chandonia, J.M.; Brenner, S.E. Weblogo: A sequence logo generator. *Genome Res.* **2004**, *14*, 1188–1190.
50. Hurst, L.D. The Ka/Ks ratio: Diagnosing the form of sequence evolution. *Trends Genet.* **2002**, *18*, 486–487.
51. Wang, D.; Liu, F.; Wang, L.; Huang, S.; Yu, J. Nonsynonymous substitution rate (Ka) is a relatively consistent parameter for defining fast-evolving and slow-evolving protein-coding genes. *Biol. Direct.* **2011**, *6*, 13, doi:10.1186/1745-6150-6-13.

52. Arumugam, M.; Raes, J.; Pelletier, E.; le Paslier, D.; Yamada, T.; Mende, D.R.; Fernandes, G.R.; Tap, J.; Bruls, T.; Batto, J.M.; *et al.* Enterotypes of the human gut microbiome. *Nature* **2011**, *473*, 174–180.
53. Hansen, A.K.; Moran, N.A. Aphid genome expression reveals host-symbiont cooperation in the production of amino acids. *Proc. Natl. Acad Sci. USA* **2011**, *108*, 2849–2854.
54. McCourt, J.A.; Pang, S.S.; King-Scott, J.; Guddat, L.W.; Duggleby, R.G. Herbicide-binding sites revealed in the structure of plant acetohydroxyacid synthase. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 569–573.
55. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The protein data bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
56. Furnham, N.; Holliday, G.L.; de Beer, T.A.; Jacobsen, J.O.; Pearson, W.R.; Thornton, J.M. The catalytic site atlas 2.0: Cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Res.* **2014**, *42*, D485–D489.