Evo Edu Outreach (2012) 5:367–381
DOI 10.1007/s12052-012-0439-5

ORIGIN OF LIFE

The Origin and Evolution of Metabolic Pathways: *Why* and *How* did Primordial Cells Construct Metabolic Routes?

Renato Fani

Published online: 15 September 2012

Abstract The emergence and evolution of metabolic pathways represented a crucial step in molecular and cellular evolution. In fact, the exhaustion of the prebiotic supply of amino acids and other compounds that were likely present on the primordial Earth imposed an important selective pressure, favoring those primordial heterotrophic cells that became able to synthesize those molecules. Thus, the emergence of metabolic pathways allowed primitive organisms to become increasingly less dependent on exogenous sources of organic compounds. Comparative analyses of genes and genomes from organisms belonging to Archaea, Bacteria, and Eukarya reveal that, during evolution, different forces and molecular mechanisms might have driven the shaping of genomes and the emergence of new metabolic abilities. Among these gene elongations, gene and operon duplications played a crucial role since they can lead to the (immediate) appearance of new genetic material that, in turn, might undergo evolutionary divergence, giving rise to new genes coding for new metabolic abilities. Concerning the mechanisms of pathway assembly, both the analysis of completely sequenced genomes and directed evolution experiments strongly support the patchwork hypothesis, according to which metabolic pathways have been assembled through the recruitment of primitive enzymes that could react with a wide range of chemically related substrates. However, the analysis of the structure and organization of genes belonging to ancient metabolic pathways, such as histidine biosynthesis, suggests that other different hypothesis, i.e., the retrograde hypothesis, may account for the evolution of some steps within metabolic pathways.

R. Fani (✉)
Laboratory of Microbial and Molecular Evolution,
Department of Evolutionary Biology, University of Florence,
Via Romana 17-19,
50125 Florence, Italy
e-mail: renato.fani@unifi.it

Keywords Gene duplication · Gene elongation · Histidine biosynthesis · Patchwork hypothesis

The Primordial Cells and Metabolism

Fascinating. Exciting. The reconstruction of the history of life on Earth represents one of the most intriguing issues of science. And even more intriguing is trying to understand the (very) first molecular steps leading to the primordial cells and their early evolution. The extant cells are quite complex entities constituted from a myriad of different molecules that, however, have to act and interact in a concerted manner in order to assure the survival and reproduction of cells (and multicellular organisms). In each moment of cell life, billions of molecules are transformed into different ones through reactions that are accelerated (catalyzed) by the so-called enzymes, most of which are represented by proteins. Even though these proteins might interact with a plethora of different molecules during their chaotic trip within the cell, they bind only to specific molecules representing their *substrate*, and transform it into another and different molecules called *product* (of the reaction). Overall, this is not true for all enzymes; each enzyme interacts with one substrate giving rise to a specific product. Hence, in each moment of cell life *billions* of substrates are transformed into *billions* of products by *billions* of enzyme molecules. These reactions are extremely fast, and we can imagine the cell as a viscous environment where these reactions occur in an ordered (and only apparently chaotic) fashion. The whole body of these reactions is called *metabolism*, a circular "entity" in the sense that molecules can be destroyed (catabolism) to obtain energy and "bricks" that are required to construct other different molecules (anabolism) (Fig. 1). It is thus clear that within a cell an "equilibrium" between catabolic and anabolic reactions

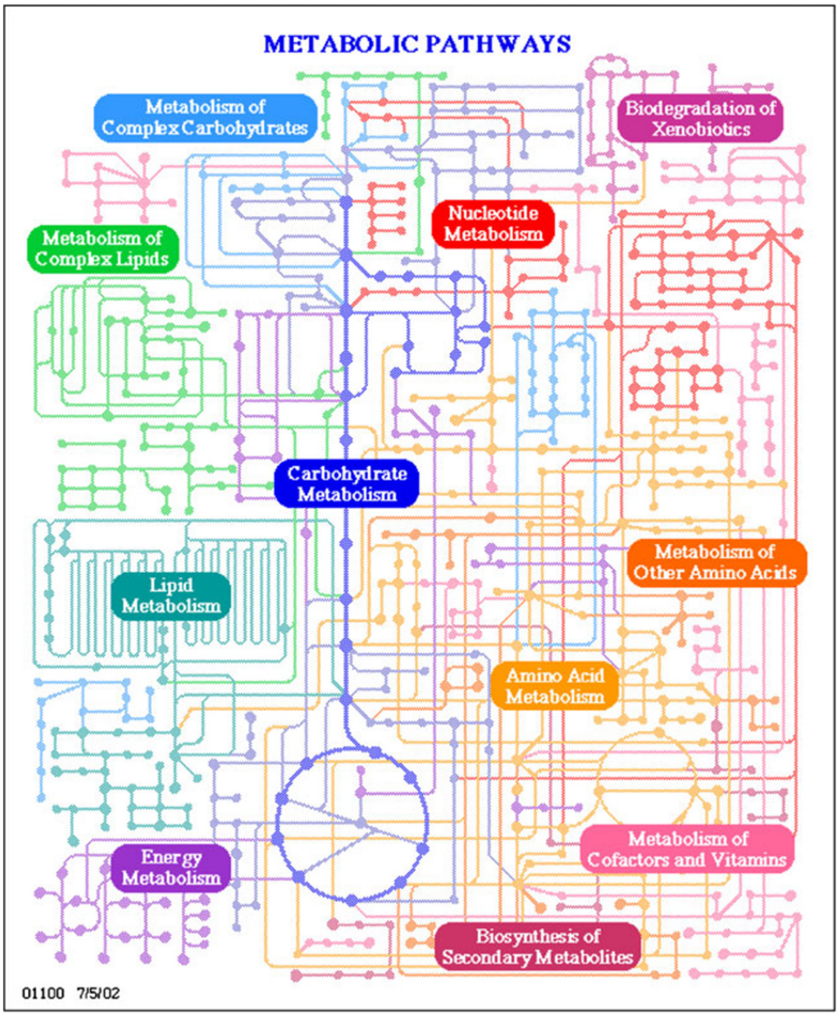
METABOLIC PATHWAYS
Metabolism of Complex Carbohydrates
Biodegradation of Xenobiotics
Nucleotide Metabolism
Metabolism of Complex Lipids
Carbohydrate Metabolism
Metabolism of Other Amino Acids
Lipid Metabolism
Amino Acid Metabolism
Metabolism of Cofactors and Vitamins
Energy Metabolism
Biosynthesis of Secondary Metabolites
01100 7/5/02

Fig. 1 Schematic representation of metabolic networks existing in the extant cells (from http://manet.illinois.edu/pathways.php)

exists. Thus, metabolism of the extant cells is quite complex, but we can also consider it extremely ordered. Figure 2 charts an example of catabolic (the degradation of glucose during glycolysis) and anabolic (the biosynthesis of the amino acid histidine) systems. As we can see from Fig. 2, both glycolysis and histidine biosynthesis proceed through a sort of "cascade" of reactions where the destruction of glucose and the construction of histidine requires the sequential action of different enzymes, each of which is able to catalyze a single step of this cascade. The set of reactions starting from the substrate and leading to the final product of the reaction is called the *metabolic pathway*. In most cases, each step of a metabolic pathway is catalyzed by a single enzyme, which (in a third of the cases) is a single protein that is encoded by a single gene (Holliday et al. 2011).

If we assume that the extant and very complex cells originated from much simpler ancestral cells, it is also plausible to imagine that the latter had a simpler metabolism in respect to the extant one. This, in turn, implies that they should possess much simpler genomes, constituted very likely by a few hundreds of genes. If this is so, the question is: *why* and *how* did primordial cells assemble and evolve their metabolic pathways? The question can be rephrased as follows: why and how did the early cells increase the number of their genes and the complexity of their genomes? The answer(s) that we can try to give to these questions clearly depend on the conditions of primitive Earth and what primordial living beings looked like. However, this is one of the foggiest issues; in fact, although considerable efforts have been made to understand the emergence of the first living beings, we still do not know when and how life originated (Peretò et al. 1998). Still, it is commonly assumed that early organisms arose and inhabited aquatic environments (oceans, rivers, ponds, etc.) rich in organic compounds spontaneously formed in the prebiotic world. This heterotrophic origin of life is generally assumed and is frequently referred to as the Oparin–Haldane theory (Oparin 1924; Lazcano and Miller 1996). If this idea is correct, life evolved from a

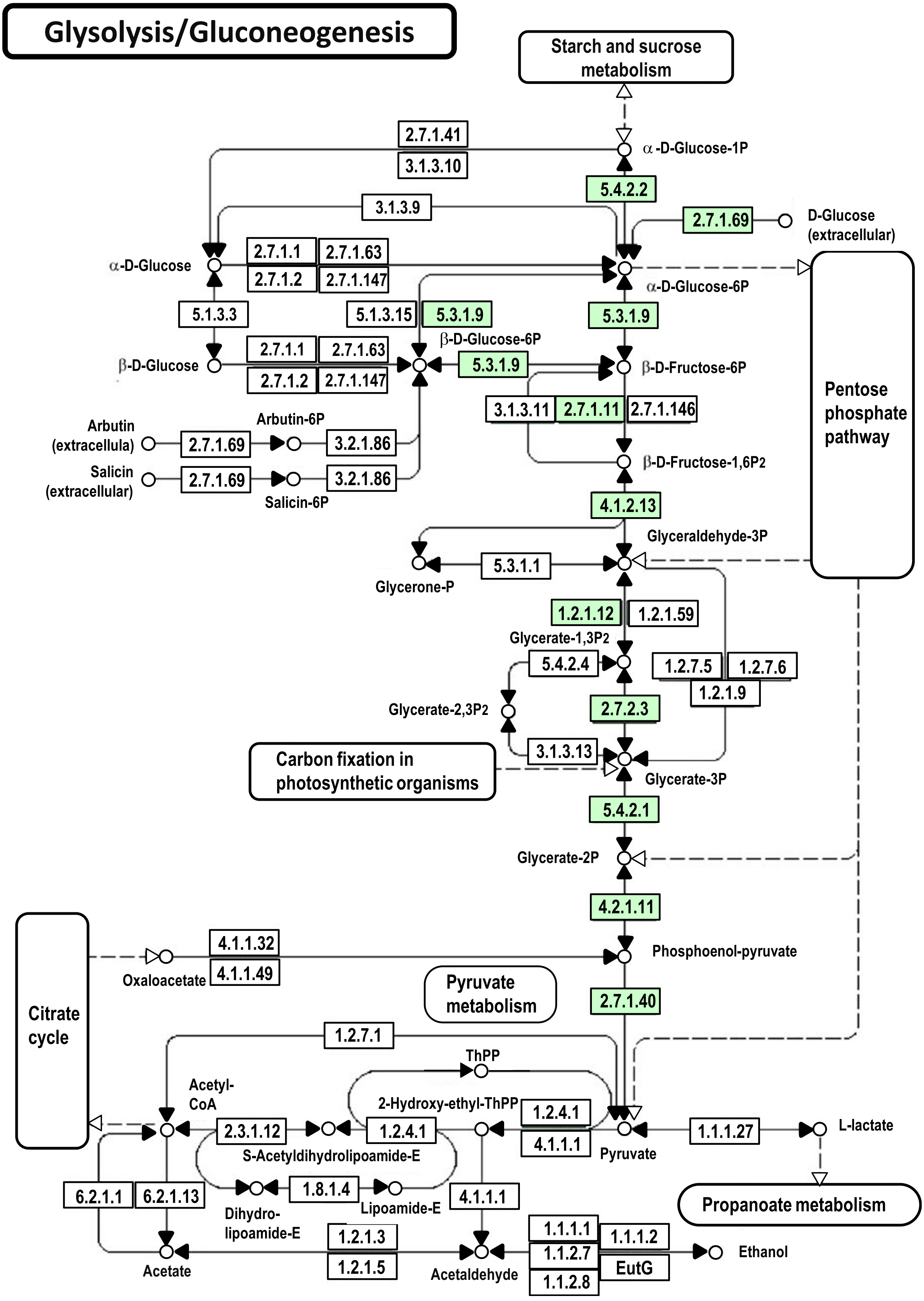

Fig. 2 Schematic representation of a catabolic (glycolysis) (from http://www.genome.jp/dbget-bin/www_bget?pathway+hsa00010) (**a**) and an anabolic pathway (histidine biosynthesis) (from http://www.genome.jp/kegg/pathway/map/map00340.html) (**b**) in the gamma-proteobacterium *Escherichia coli* K12

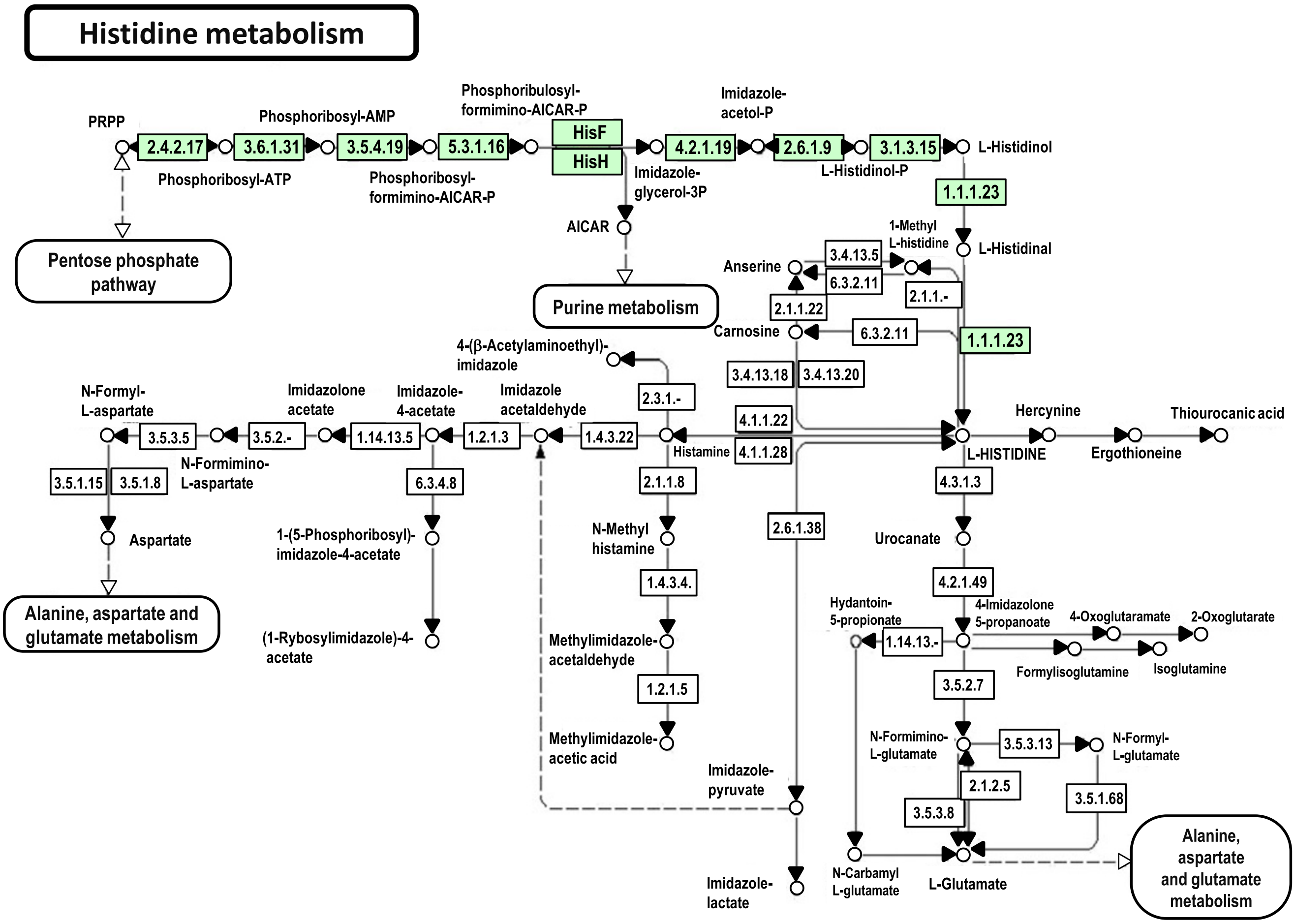

Fig. 2 (continued)

primordial soup containing different organic molecules (many of which are used by extant life forms). This soup of nutrient compounds was available to the early heterotrophic organisms, so they had to do a minimum of biosynthesis. An experimental support to this proposal was obtained in 1953 when Miller (1953) and Urey showed that amino acids and other organic molecules are formed under atmospheric conditions thought to be representative of those on the early Earth. The first living systems probably did stem directly from the primordial soup and evolved relatively fast up to a common ancestor, usually referred to as Last Universal Common Ancestor (LUCA), an entity representing the divergence starting-point of all the extant life forms on Earth (Fig. 3). If we assume that life arose in a prebiotic soup containing most, if not all, of the necessary small molecules, then a large potential availability of nutrients on the primitive Earth can be surmised, providing both the growth and energy supply for a large number of ancestral organisms. We can imagine the existence of an "early floating living world" constituted of primordial cells that might have looked like "soap bubbles" embedding one or more informational molecules and performing a limited number of metabolic reactions. These bubbles were able to divide, to interact with each other, and to fuse and share their genomes and metabolic abilities, giving rise to progressively complex living beings. If this scenario is correct, that is that primordial organisms were heterotrophic and had no need for developing new and improved metabolic abilities since most of the required nutrients were available, we can go back to the two questions that can be addressed, that is, *why* and *how* did primordial cells expand their metabolic abilities and genomes?

The answer to the first question is rather intuitive. Indeed, the increasing number of early cells thriving on primordial soup would have led to the depletion of essential nutrients, imposing a progressively stronger selective pressure that, in turn, favored (in a Darwinian sense) those microorganisms that had become capable of synthesizing those molecules whose concentration was decreasing in the primordial soup. Hence, the origin and the evolution of basic metabolic pathways represented a crucial step in molecular and cellular evolution since it rendered the primordial cells less dependent on exogenous sources of nutrients (Fig. 4).

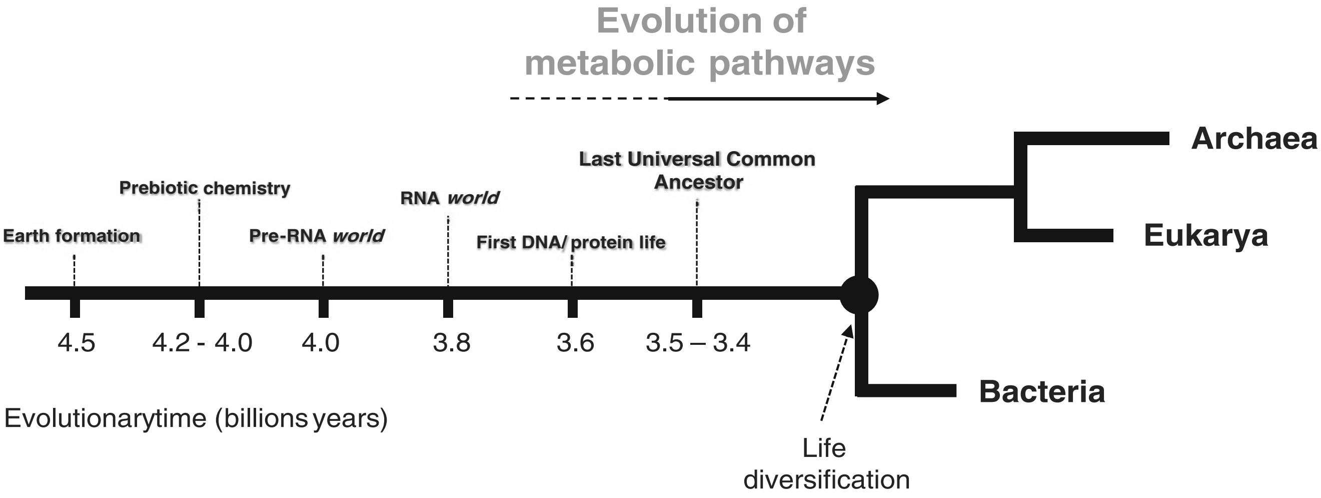

Fig. 3 Tentative evolutionary time line from the origin of Earth to the diversification of life

But how did the expansion of genomes occur? The following section will focus on the molecular mechanisms that guided this transition, i.e., the expansion and the refinement of ancestral metabolic routes, leading to the structure of the extant metabolic pathways.

The Role of Duplication and Fusion of DNA Sequences in the Evolution of Metabolic Pathways in Early Cells

Since ancestral cells probably contained small chromosomes and consequently possessed limited coding capabilities, it is

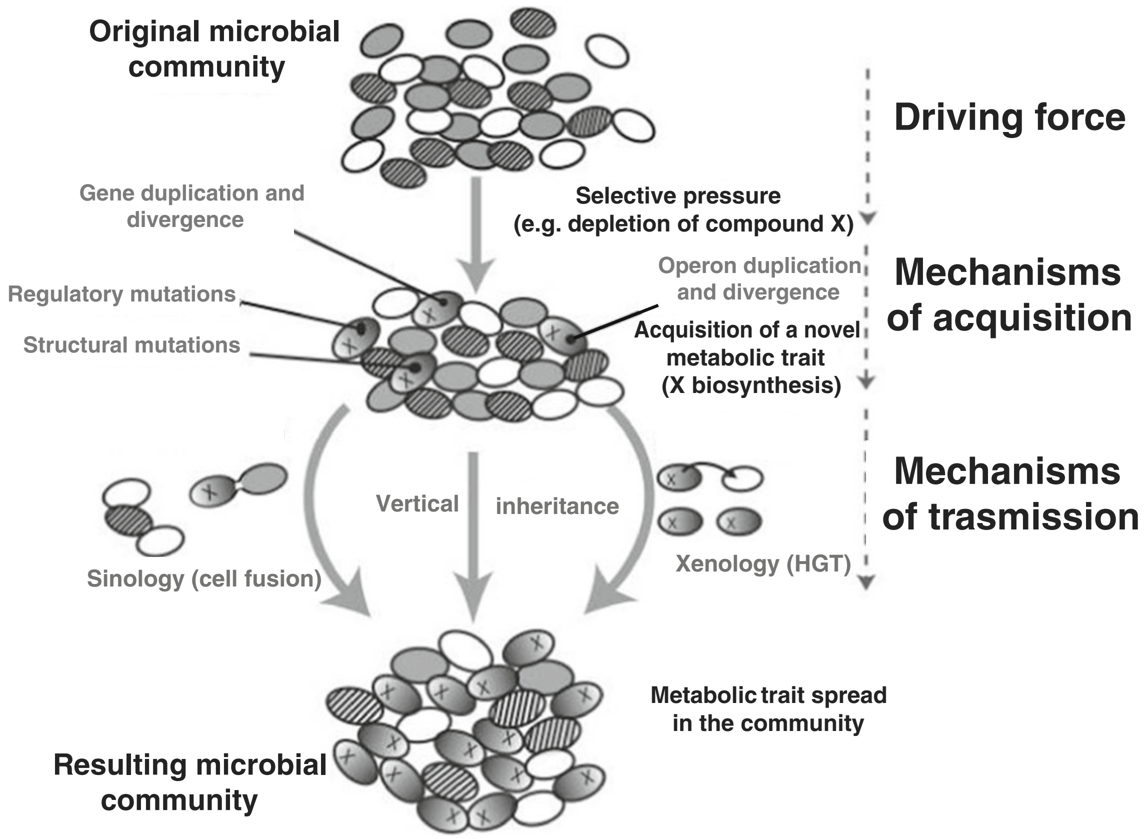
Original microbial community
Driving force
Gene duplication and divergence
Selective pressure (e.g. depletion of compound X)
Regulatory mutations
Structural mutations
Operon duplication and divergence
Acquisition of a novel metabolic trait (X biosynthesis)
Mechanisms of acquisition
Vertical
inheritance
Mechanisms of trasmission
Sinology (cell fusion)
Xenology (HGT)
Metabolic trait spread in the community
Resulting microbial community

Fig. 4 Schematic representation of an ancestral cell community with a selective pressure allowing for the acquisition and spreading of a new metabolic trait (modified from Fondi et al. 2009a)

plausible to imagine that their metabolism could count on a limited number of enzymes. Hence, how could the ancestral cells fulfill all their metabolic tasks possessing such a restricted enzyme repertoire? A possible (and widely accepted) explanation is that these ancestral enzymes possessed broad substrate specificity, allowing them to catalyze several different chemical reactions (see below). Hence, the hypothetical ancestral metabolic network (Fig. 5) was probably composed of a limited number of nodes (enzymes) that were highly interconnected (i.e., participated in different, although linked, biological processes). On the contrary, network models of extant metabolisms reveal remarkably complex structures (Fig. 5); thousands of different enzymes form well-defined routes that transform many distinct molecules in an ordered fashion and with a predefined output.

The Starter Types and Explosive Expansion of Metabolism in the Early Cells

Different molecular mechanisms may have been responsible for the expansion of early genomes and metabolic abilities. Data obtained in the last decade clearly indicate that a very large proportion of the gene set of different (micro)organisms is the outcome of more or less ancient gene duplication events predating or following the appearance of the LUCA and involving ancestral genes, referred to as the *starter types*, a term first coined by Lazcano and Miller (1994), that underwent (many) duplications. These findings strongly suggest that the duplication and divergence of DNA sequences of different size represents one of the most important forces driving the evolution of genes and genomes during the early evolution of life. Indeed, this process may allow the formation of new genes from pre-existing ones. However, there are a number of additional mechanisms that could have increased the rate of metabolic evolution, including the modular assembly of new proteins by gene fusion events and horizontal gene transfer, the latter permitting the transfer of entire metabolic routes or part thereof.

Gene Duplication

The importance of gene duplication for the development of metabolic innovations was first discussed by Lewis (1951) and later by Ohno (1970) and has been recently confirmed by the comparative analysis of complete sequences of archaeal, bacterial, and eukaryal genomes. Genes descending from a common ancestor via a duplication event are called paralogs, and they may undergo successive duplications leading to a *paralogous gene family* (Fig. 6). Paralogous genes often catalyze different, although similar, reactions.

Fate of duplicated genes The structural and/or functional fate of duplicated genes is an intriguing issue that has led to the proposal of several classes of evolutionary models accounting for the possible scenarios emerging after the appearance of a paralogous gene pair.

Structural fate Duplication events can generate genes arranged in tandem or scattered at different loci within the genome (Fani 2004; Li and Graur 1991). If an in-tandem duplication occurs, at least two different scenarios for the structural evolution of the two copies can be depicted: (1) the two genes undergo an evolutionary divergence, becoming paralogs; and (2) the two genes fuse, doubling their original size forming an elongated gene (see below). Moreover, if the two copies are not arranged in tandem, they may either (1)

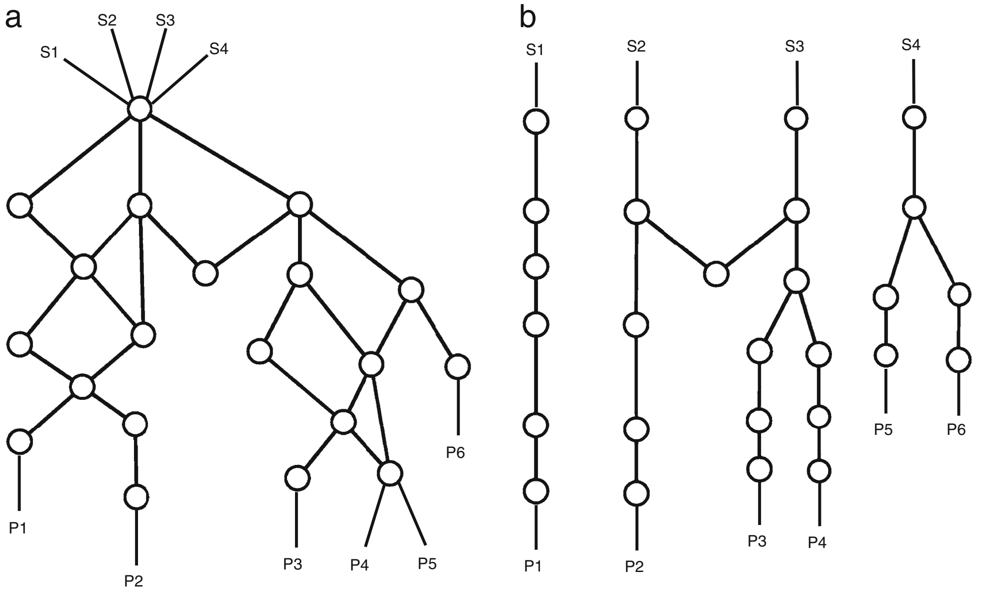

Fig. 5 Schematic representation of **a** an ancestral metabolic network and **b** an extant one. Nodes and links represent enzymes and catalytic reactions, respectively. *S* substrate, *P* product

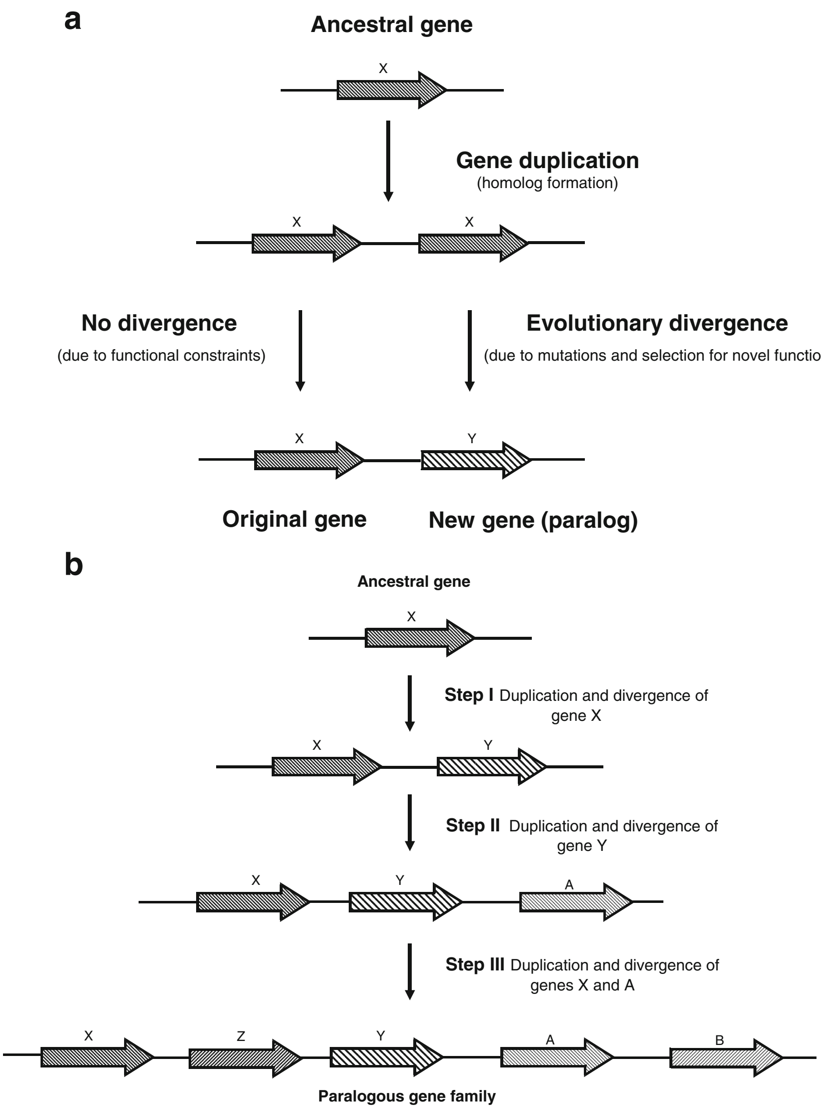
a
Ancestral gene
X
Gene duplication
(homolog formation)
X
X
No divergence
(due to functional constraints)
Evolutionary divergence
(due to mutations and selection for novel function)
X
Y
Original gene
New gene (paralog)
b
Ancestral gene
X
Step I Duplication and divergence of gene X
X
Y
Step II Duplication and divergence of gene Y
X
Y
A
Step III Duplication and divergence of genes X and A
X
Z
Y
A
B
Paralogous gene family

Fig. 6 Schematic representation of the formation of paralogous genes (**a**) and paralogous gene family (**b**) (modified from Fani 2004)

become paralogous genes; or (2) one copy may fuse to an adjacent gene, with a different function, giving rise to a mosaic or chimeric gene that potentially may evolve to perform other metabolic role(s). Tandem duplications of DNA stretches are often the result of an unequal crossing-over between two DNA molecules, but other processes, such as replication slippage, may be invoked to explain the existence of tandemly arranged paralogous genes. The presence of paralogous genes at different sites within a microbial genome might be the result of ancient activity of transposable elements and/or duplication of genome fragments as well as whole-genome duplications (Fani 2004).

Functional fate The functional fate of the two (initially) identical gene copies originating from a duplication event depends on the further modifications (evolutionary divergence) that one (or both) of the two redundant copies accumulates during evolution. It can be surmised, in fact, that after a gene duplicates, one of the two copies becomes dispensable and can undergo several types of mutational events, mainly substitutions, that, in turn, can lead to the appearance of a new gene, harboring a different function in respect to the ancestral coding sequence (Fig. 6). On the other hand, duplicated genes can also maintain the same function in the course of evolution, thereby enabling the production of a large quantity of RNAs or proteins (gene dosage effect).

Operon Duplication

DNA duplications may also concern entire clusters of genes involved in the same metabolic pathways and transcribed from a promoter into a polycistronic mRNA, i.e., entire operons or part thereof. Thus, we can imagine that if an entire operon A, responsible for the biosynthesis of amino acid A, duplicates giving rise to a couple of paralogous operons, one of the copies (B) may diverge from the other and evolve in such a way that the encoded enzymes catalyze reactions leading to a different amino acid, B. If this event actually occurs, it might provoke a (rapid) expansion of the metabolic abilities of the cell and the increase of its genome size (Fani and Fondi 2009).

Once acquired, metabolic innovations might have been spread rapidly between microorganisms through horizontal gene transfer mechanisms.

Gene Fusion

In addition to gene duplication, another route of gene evolution is the fusion of independent cistrons leading to bi- or multi-functional proteins (Brilli and Fani 2004b; Xie et al. 2003). Gene fusions that have been disclosed in genes of many metabolic pathways provide a mechanism for the physical association of different catalytic domains or of catalytic and regulatory structures (Jensen 1976). Fusions frequently involve genes coding for proteins that function in a concerted manner, such as enzymes catalyzing sequential steps within a metabolic pathway (Yanai et al. 2002). Fusion of such catalytic centers likely promotes the channeling of intermediates that may be unstable and/or in low concentration. The high fitness of gene fusions can also rely on the tight regulation of the expression of the fused domains. Even though gene fusion events have been described in many prokaryotes, they may have a special significance among nucleated cells, where the very limited number, if not the complete absence, of operons does not allow the coordinate synthesis of proteins by polycistronic mRNAs.

Gene Duplication and Fusion Acting Together: Gene Elongation

It is generally accepted that ancestral protein-encoding genes would have been relatively short sequences encoding simple polypeptides likely corresponding to functional and/or structural domains. These "mini-genes" may increase their size through a mechanism called *gene elongation*, that is the increase in gene size, which represents one of the most important steps in the evolution of complex genes from simple ones (Fani 2004). A gene elongation event can be the outcome of an in-tandem duplication of a DNA sequence. Then, if a deletion of the intervening sequence between the two copies occurs followed by a mutation converting the stop codon of the first copy into a sense codon (Fig. 7), this results in the elongation by fusion of the ancestral gene and its copy. Hence, the new gene is constituted of two paralogous moieties (modules). In principle, each module or both of them might undergo further duplication events, leading to a gene constituted by more repetitions of amino acid sequences. This type of duplication has occurred in many genes and its biology might rely on (1) the improvement of the function of a protein by increasing the number of active sites and/or (2) the acquisition of an additional function by modifying a redundant segment.

Hypotheses on the Origin and Evolution of Metabolic Pathways

As discussed in the previous sections, the emergence and refinement of basic biosynthetic pathways allowed primitive organisms to become increasingly less dependent on exogenous sources of chemical compounds accumulated in the primitive environment as a result of prebiotic syntheses. But how did these metabolic pathways originate and evolve? And what is the role that the molecular mechanisms described above (gene elongation, duplication, and/or fusion) played in the assembly of metabolic routes? How the major metabolic pathways actually originated is still an open question, but several different theories have been suggested to account for the establishment of metabolic routes. All these ideas are based on gene duplication. Two of them are discussed in the following paragraphs.

The Retrograde Hypothesis (Horowitz 1945, 1965)

The first attempt to explain in detail the origin of metabolic pathways was made by Horowitz (1945), who suggested that biosynthetic enzymes had been acquired via gene duplication that took place in the reverse order found in current pathways. This idea, also known as the retrograde hypothesis, has intuitive appeal and states that if the contemporary biosynthesis of compound "A" requires the sequential transformation of precursors "D," "C," and "B" through the corresponding enzymes, the final product "A" of a given metabolic route was the first compound used by the primordial heterotrophs (Fig. 8). In other words, if compound "A" was essential for the survival of primordial cells, when "A" became depleted from the primitive soup, this should have imposed a selective pressure allowing the survival and reproduction of those cells that had become able to perform the transformation of a chemically related compound "B" into "A" catalyzed by enzyme "a" that would have led to a simple, one-step pathway. The selection of variants having a mutant "b" enzyme related to "a" via a duplication event and capable of mediating the transformation of molecule "C" chemically related into "B" would lead into an increasingly complex route, a process that would continue until the entire pathway was established in a backward fashion, starting with the synthesis of the final product, then the penultimate

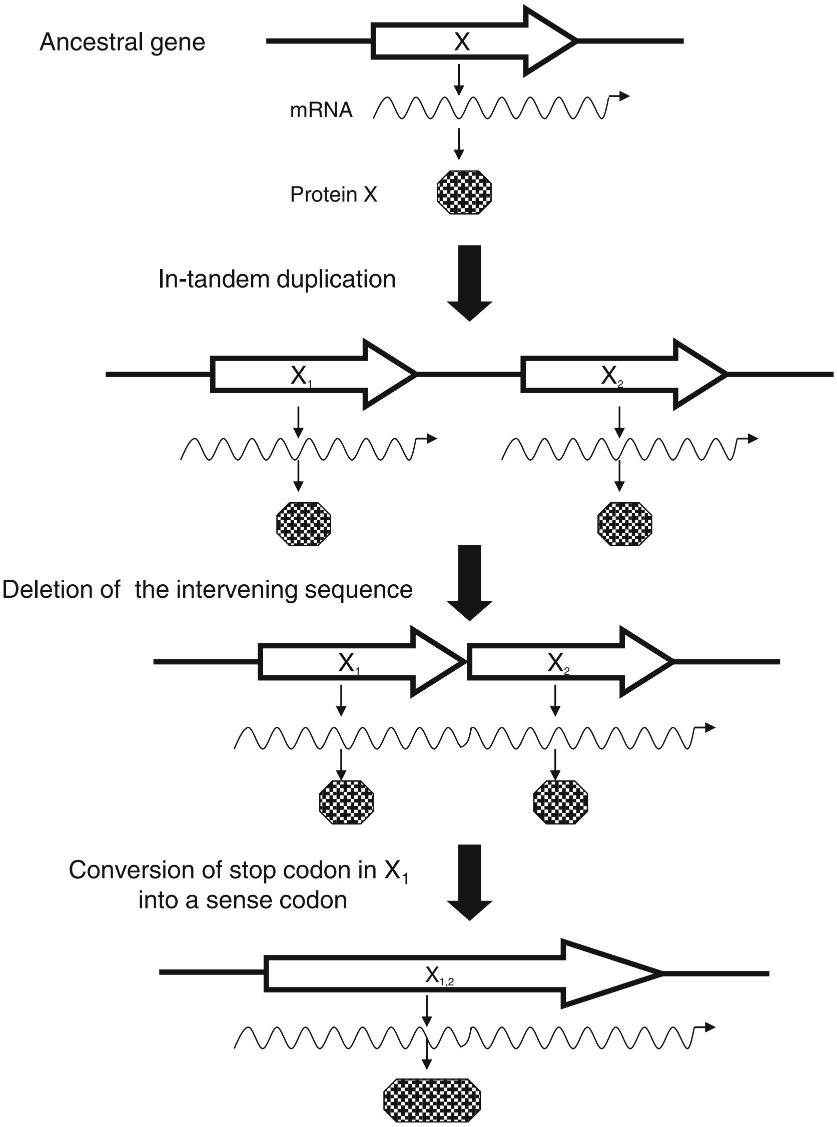
Ancestral gene
X
mRNA
Protein X
In-tandem duplication
X_1
X_2
Deletion of the intervening sequence
X_1
X_2
Conversion of stop codon in X_1 into a sense codon
$X_{1,2}$

Fig. 7 Gene elongation: the duplication of an ancestral gene and the subsequent fusion of the two homologs to produce a longer protein (modified from Fani 2004)

pathway intermediate, and so on down the pathway to the initial precursor (Fig. 9). Twenty years later, the discovery of operons prompted Horowitz to restate his model, arguing that it was supported also by the clustering of genes, which could be explained by a series of early tandem duplications of an ancestral gene; in other words, genes belonging to the

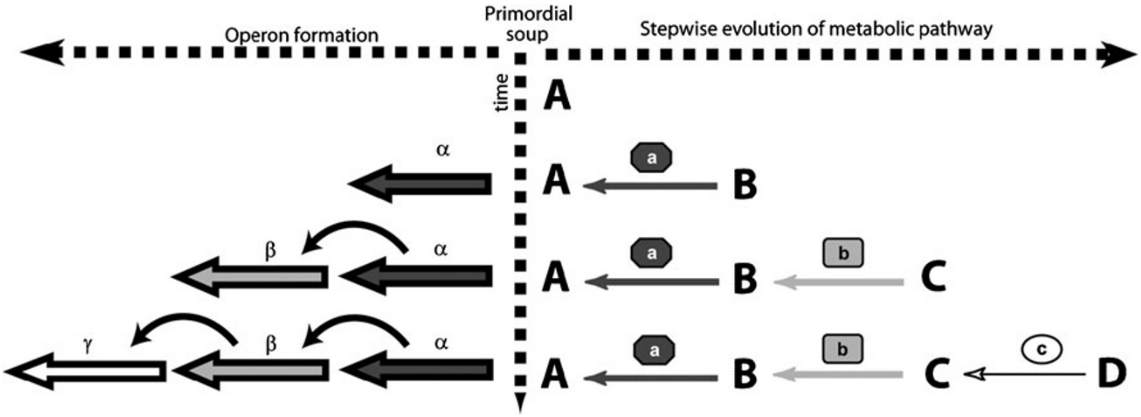

Fig. 8 Schematic representation of the Horowitz hypothesis on the origin and evolution of metabolic pathways (modified from Fondi et al. 2009a)

same operon and/or to the same metabolic pathway should have formed a paralogous gene family.

The Patchwork Hypothesis (Ycas 1974; Jensen 1976)

Gene duplication has also been invoked in another model, the so-called patchwork hypothesis (Ycas 1974; Jensen 1976), according to which metabolic pathways may have been assembled through the recruitment of primitive enzymes that could react with a wide range of chemically related substrates. Such relatively slow, non-specific enzymes may have enabled primitive cells containing small genomes to overcome their limited coding capabilities. Figure 9 shows a schematic three-step model of the patchwork hypothesis: (a) an ancestral enzyme E0 endowed with low substrate specificity is able to bind to three substrates (S1, S2, and S3) and catalyze three different, but similar, reactions; (b) a duplication of the gene encoding E0 and the subsequent divergence of one of the two copies leads to the appearance of enzyme E2 with an increased and narrowed specificity; and (c) a further gene duplication event, followed by evolutionary divergence, leads to E3. In this way, the ancestral enzyme E0 belonging to a given metabolic route is "recruited" to serve other novel pathways.

The patchwork hypothesis is also consistent with the possibility that an ancestral pathway may have had a primitive enzyme catalyzing two or more similar reactions on related substrates of the same metabolic route and whose substrate specificity was refined as a result of later duplication events.

In this way, primordial cells might have expanded their metabolic capabilities. Additionally, this mechanism may have permitted the evolution of regulatory mechanisms coincident with the development of new pathways (Fani 2004; Lazcano et al. 1995).

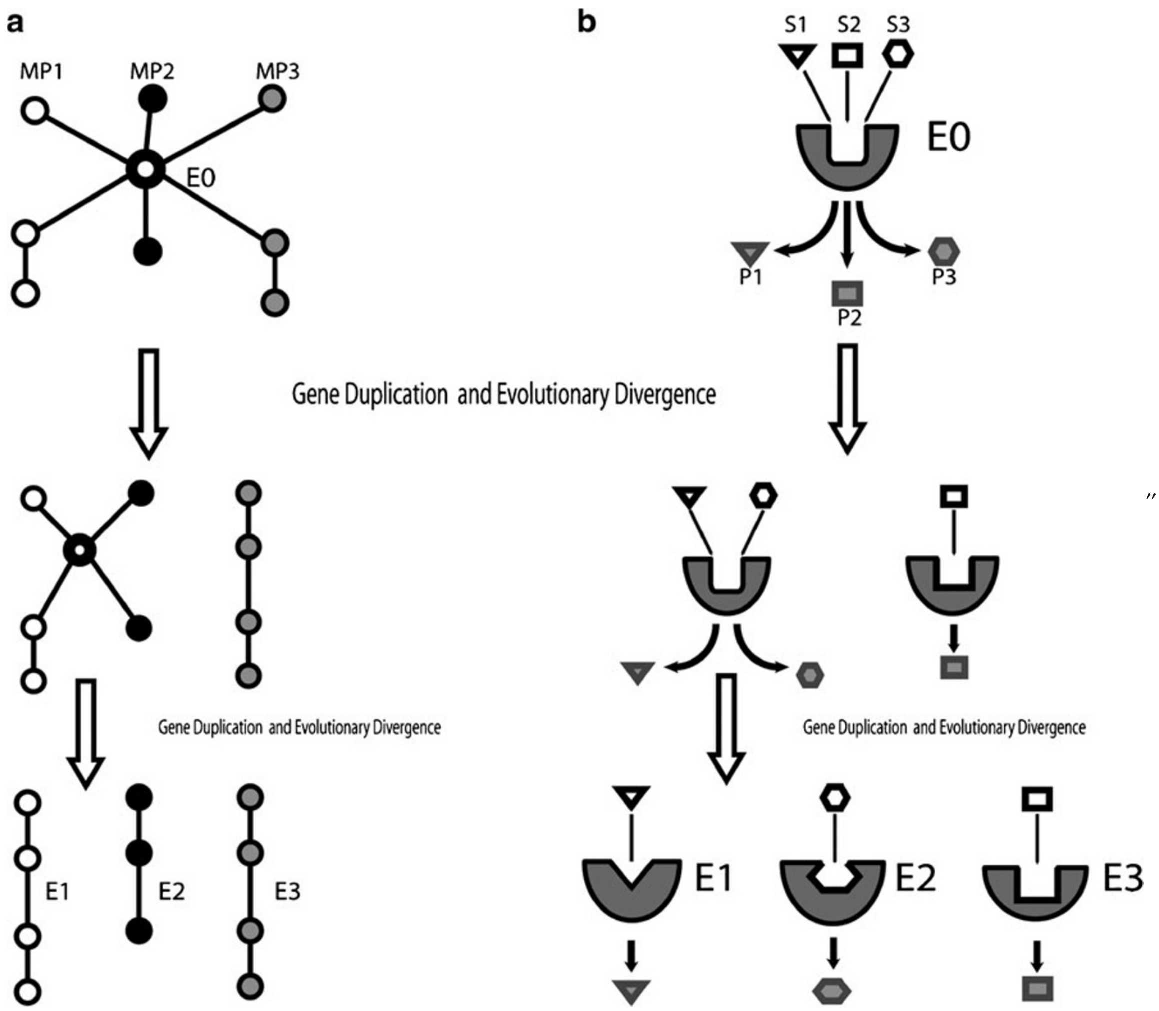
a
MP1
MP2
MP3
E0
Gene Duplication and Evolutionary Divergence
Gene Duplication and Evolutionary Divergence
E1
E2
E3
b
S1
S2
S3
E0
P1
P2
P3
Gene Duplication and Evolutionary Divergence
E1
E2
E3

Fig. 9 The patchwork hypothesis on the origin and evolution of metabolic pathways. **a** The origin of enzymes with narrowed specificity from an ancestral unspecific one. **b** Hypothetical overall structure of the metabolic pathways (MP) in which enzymes (*E0*, *E1*, *E2*, *E3*) are involved

The Reconstruction of the Origin and Evolution of Metabolic Pathways

How can the origin and evolution of metabolic pathways be studied and reconstructed? By assuming that useful hints may be inferred from the analysis of metabolic pathways existing in contemporary cells (Peretò et al. 1998), important insights into the evolutionary development of microbial metabolic pathways can be obtained by (1) the use of bioinformatic tools that allow the comparison of gene and genomes from organisms belonging to the three cell domains (Archaea, Bacteria, and Eukarya), and (2) laboratory studies in which new substrates are used as carbon, nitrogen, or energy sources. These are the so-called directed-evolution experiments in which a microbial (typically bacterial) population is subjected to a (strong) selective pressure that leads to the establishment of new phenotypes capable of exploiting different substrates (Clarke 1974; Mortlock and Gallo 1992). By assuming that the processes involved in acquiring new metabolic abilities are comparable to those found in natural populations, directed-evolution experiments can provide useful insights in early cellular evolution (Fani 2004).

Histidine Biosynthesis: A Paradigm for the Study of the Origin and Evolution of Metabolic Pathways

Histidine biosynthesis is one of the best-characterized anabolic pathways. There is a large body of genetic and biochemical information, including operon structure, gene expression, and an increasingly larger number of sequences available for this route. This pathway has been extensively studied, mainly in the two enterobacteria *Escherichia coli* and *Salmonella typhimurium*. In all histidine-synthesizing organisms, the pathway is unbranched and includes several unusual reactions. Moreover, it consists of nine intermediates and of eight distinct proteins that are encoded by eight genes, *hisGDC(NB)HAF(IE)*, with three of them (*hisD*, *hisNB*, and *hisIE*) coding for bifunctional enzymes (Alifano et al. 1996). In the two enterobacteria, the eight genes are arranged in a compact operon (Fig. 10).

Histidine biosynthesis is a *metabolic crossroad* and plays an important role in cellular metabolism, being interconnected to both the de novo synthesis of purines and to nitrogen metabolism. The connection to purine biosynthesis results from an enzymatic step catalyzed by imidazole glycerol phosphate synthase, a heterodimeric protein composed by one subunit each of the *hisH* and *hisF* products (Alifano et al. 1996). Chemical and biological data suggest that histidine was present in the primordial soup and that this biosynthetic route is ancient. It has also been suggested that histidine-containing small peptides could have been involved in the prebiotic formation of other peptides and nucleic acid molecules, once these monomers accumulated in primitive tidal lagoons or ponds (Fani and Fondi 2009 and references therein). If primitive catalysts required histidine, then the eventual exhaustion of the prebiotic supply of histidine and histidine-containing peptides imposed a selective pressure favoring those microorganisms capable of synthesizing histidine. Hence, this metabolic pathway might have been assembled long before the appearance of the LUCA (Brilli and Fani 2004a, b; Fani et al. 1994, 1995; Alifano et al. 1996; Fondi et al. 2009b), but once the entire pathway was assembled, it underwent major rearrangements during evolution, as suggested by the wide variety of different clustering strategies of *his* genes that has been documented.

How the *his* pathway originated remains an open question, but the analysis of the structure and organization as well as the phylogenetic analyses of the *his* genes in (micro)organisms belonging to different phylogenetic archaeal, bacterial, and eukaryal lineages reveals that different molecular mechanisms played an important role in shaping this pathway. Actually, an impressive series of well-documented duplication (Fani et al. 1994), elongation (Fani et al. 1994) and fusion (Brilli and Fani 2004a, b; Fani et al. 2007) events has shaped this pathway. Therefore, the histidine biosynthetic pathway represents an excellent model for understanding the molecular mechanisms driving the assembly and refinement of metabolic routes.

The Refinement and Expansion of Metabolic Abilities Through a Cascade of Gene Elongation and Duplication Events: *hisA* and *hisF*

Two of the histidine biosynthetic genes, *hisA* and *hisF*, are exceptionally interesting from an evolutionary viewpoint. They code for a [*N*-(5′-phosphoribosyl) formimino]-5-aminoimidazole-4-carboxamide ribonucleotide (ProFAR) isomerase and a cyclase, respectively, which catalyze two central and sequential reactions (the fourth and fifth ones) of the pathway (Fig. 10). The comparative analysis of the HisA and HisF proteins from different archaeal, bacterial, and eukaryotic (micro)organisms reveals that they are paralogous and share a similar internal organization into two paralogous modules half the size of the entire sequence (Fani et al. 1994). According to the model proposed, the first duplication involved an ancestral module (half the size of the present-day *hisA* gene) and led by a gene elongation event to the ancestral *hisA* gene which, in turn, underwent a duplication that gave rise to the *hisF* gene (Fig. 11). Since the overall structure of the *hisA* and *hisF* genes are the same in all known organisms, it is likely that they were part of the genome of the LUCA and that the two duplication events

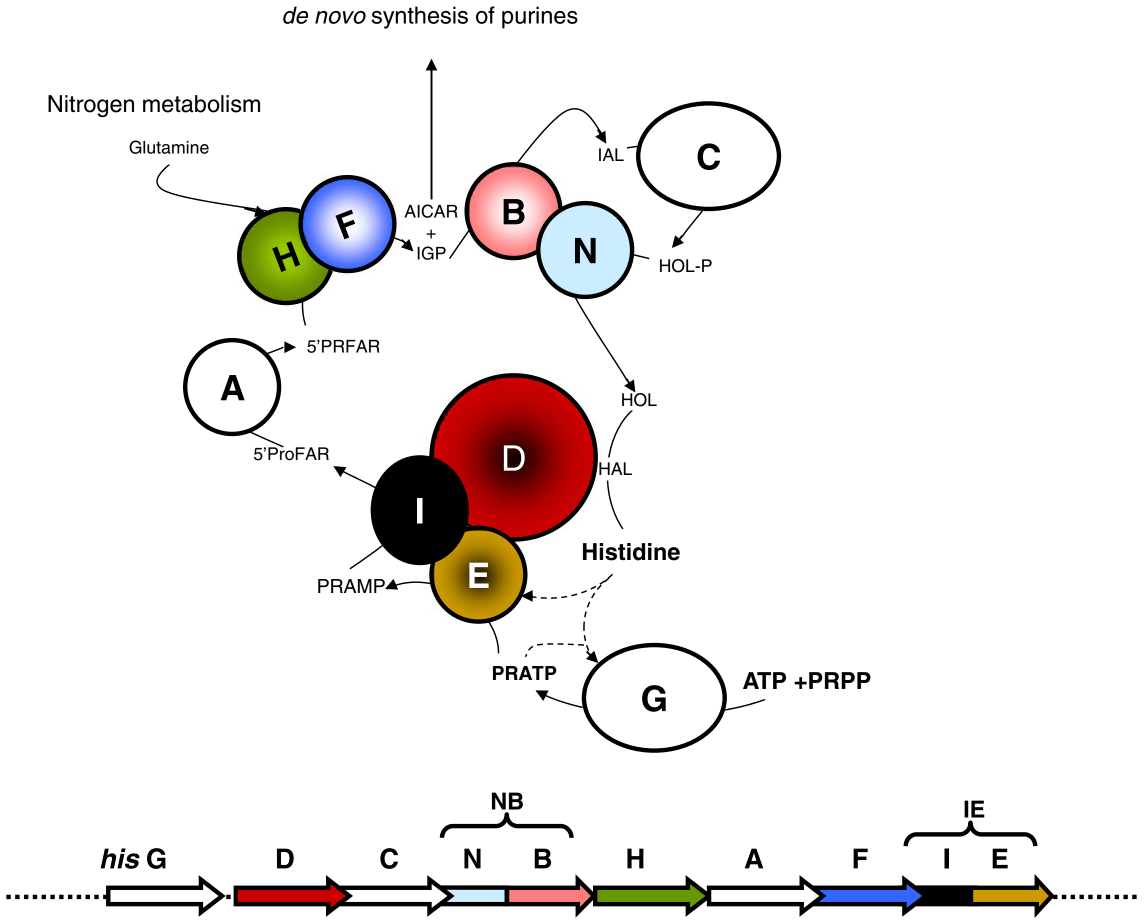

Fig. 10 Schematic representation of the histidine biosynthetic pathways and the organization of *his* gene in *Escherichia coli*. Genes and proteins in color are those involved in fusion events (modified from Fani et al. 2007)

occurred long before its appearance. The biological significance of the *hisA–hisF* structure relies on the structure of the encoded enzymes; indeed, they contain a triose phosphate isomerase (TIM) $_8(\beta\alpha)$ barrel-like fold (Copley and Bork 2000). The barrel structure is composed of eight concatenated (β-strand)-loop-(α-helix) units. The β-strands are located at the interior of the protein, forming the staves of a barrel, whereas the α-helices pack around them facing the exterior. The model proposed predicts (Fani et al. 2007) that the ancestral gene coded a half-barrel, which might assemble to form a functional enzyme by homo-dimerization. The elongation event leading to the ancestor of *hisA*/*hisF* genes resulted in the covalent fusion of two half-barrels, producing a protein whose function was refined and optimized by mutational changes; once assembled, the "whole-barrel gene" underwent gene duplication, leading to the ancestor of *hisA* and *hisF*. The possibility of an even older gene-elongation event involving (β/α)-mers smaller than the $_4$(β/α) units of the ancestral "half-barrel" precursor was recently investigated (Fani et al. 2007) by an extensive analysis of all the available HisA and HisF sequences. Data obtained supports an evolutionary model suggesting that the extant *hisA*/*hisF* structure could have arisen by two sequential gene elongation events, each of which doubled the length of the ancestral gene and the number of (β/α)-modules in the product. Thus, the ancestor of the present-day HisA/HisF TIM barrels would be the result of a cascade of (at least) two consecutive gene elongations (Fig. 11). Therefore, *hisA* and *hisF* represent a paradigmatic example of how evolution works at both the molecular and functional levels and represent a crossroad of different molecular mechanisms and hypotheses on the origin of metabolic pathways. Indeed, they are the result of gene elongation (duplication and fusion) and gene duplication events, which finally led to a large paralogous gene family (TIM barrels). Besides, these structural events are strongly linked to the function performed by the enzymes. The ancestral enzyme might have catalyzed different, even though similar, reactions in different metabolic routes (i.e., tryptophan and histidine biosynthesis), as well as two sequential steps in the same biosynthetic route (the biosynthesis of histidine) completely fitting the Jensen's idea (1976). Lastly, *hisA* and *hisF* also supported, even though partially, the Horowitz idea on the origin and evolution of metabolic pathways since they catalyze two sequential steps in the same biosynthetic route,

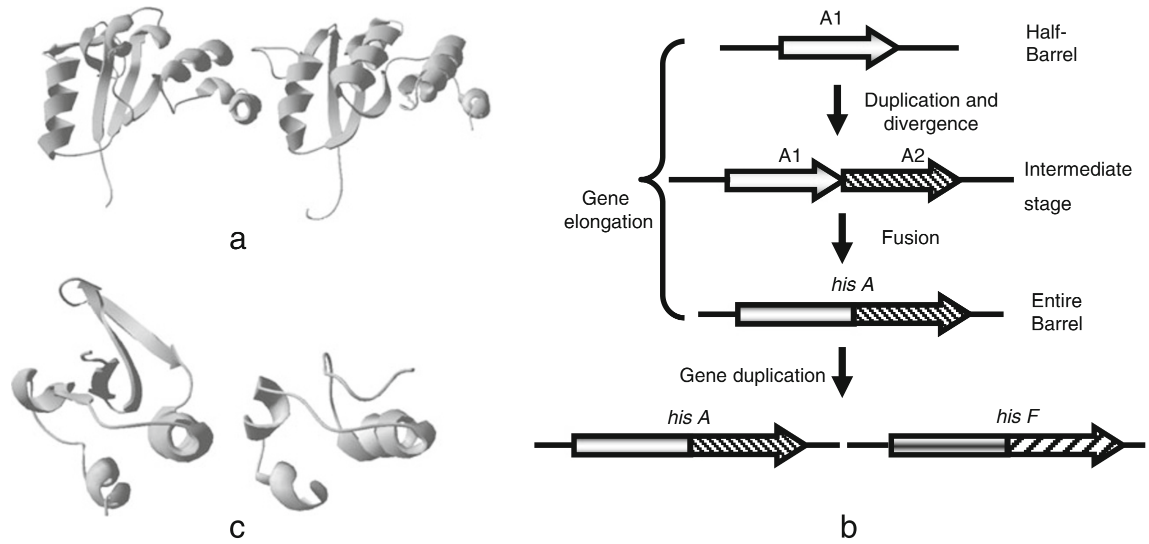

Fig. 11 Evolutionary model proposed for the origin and evolution of *his*A and *his*F (**b**). The first and the second quarters and two single (β/α) modules of the HisA protein from *Thermotoga maritima* are reported in (**a**) and (**c**), respectively

are paralogous, and are arranged in tandem in the same operon.

Gene Fusion in the Assembly of Histidine Biosynthesis

It has been recognized that (at least) seven (*hisD*, *N*, *B*, *H*, *F*, *I*, and *E*) out of the ten *his* biosynthetic genes (*hisGDCNBHAFIE*) underwent different single or multiple fusions in diverse prokaryotic and eukaryal phylogenetic lineages, demonstrating that gene fusion represents one of the most important routes for the evolution of *his* genes. Recently (Fani et al. 2007), the amino acid sequences of all the available His proteins have been analyzed for (1) gene structure, (2) phylogenetic distribution, (3) timing of appearance, (4) horizontal gene transfer, (5) correlation with gene organization, and (6) biological significance. Data obtained allowed the reconstruction of the evolutionary history of three interesting gene fusions. Quite interestingly, it has been demonstrated that fusion events involving different histidine biosynthetic genes that gave rise to genes coding for bifunctional or multifunctional enzymes, such as *hisNB*, *hisIE*, and *hisHF*, occurred in different evolutionary timescales and in different (micro)organisms, and that they have very different phylogenetic distributions (see below).

The whole body of data permitted the depiction of a likely scenario for the origin and evolution of histidine biosynthetic genes. According to the model proposed (Fani et al. 2007; Fondi et al. 2009b) on the basis of the available data, it has been suggested that the complete histidine biosynthetic pathway was assembled long before the appearance of the LUCA, which possessed mono-functional *his* genes. Concerning the organization of these genes in LUCA, it is not still possible to establish if they were (1) scattered throughout its genome, (2) organized in a single more-or-less compact operon, or (3) exhibited a mixed organization (i.e., some scattered genes or organized in more mini-operons).

However, it is quite clear that after the divergence from LUCA, the organization of histidine biosynthetic genes underwent several different rearrangements.

Concerning the structure of *his* genes, the only "universal" gene fusion concerns *hisA* and *hisF* genes, which are the outcome of a cascade of (at least) two gene elongation events followed by a paralogous gene duplication. This suggests that the two elongation events as well as the paralogous duplication event leading to *hisA* and *hisF* are very ancient, i.e., they predate the appearance of LUCA. During the early steps of molecular evolution, *hisA* and its copies underwent multiple duplication events leading to a paralogous gene family. The fusion between *hisI* and *hisE* occurred more than once in Bacteria, indicating a phenomenon of convergent evolution. Moreover, this gene might have been horizontally transferred (Fani et al. 2007). The *hisNB* fusion is a relatively recent evolutionary event that occurred in the γ-branch of proteobacteria. This fusion was parallel to the introgression of *hisN* into an already formed and more or less compact *his* operon. Having once occurred, the fusion was fixed and transferred to other proteobacteria and/or CFB group along with the entire operon or part thereof. The fusions involving *hisH* and *hisF* were found only in two bacteria.

Conclusions

Metabolic pathways of the earliest heterotrophic organisms arose during the exhaustion of the prebiotic compounds present in the primordial soup.

In the course of molecular and cellular evolution, different mechanisms and different forces might have concurred in the emergence of new metabolic abilities and the shaping of metabolic routes. However, duplication of DNA regions represents a major force of gene and genome evolution. The evidence for gene elongation, gene duplication, and operon duplication events sugggests, in fact, that the ancestral forms of life might have expanded their coding abilities and their genomes by "simply" duplicating a small number of minigenes (the starter types) via a cascade of duplication events involving DNA sequences of different size. In addition to this, gene fusion also played an important role in the construction and assembly of chimeric genes.

The dissemination of metabolic routes between microorganisms might be facilitated by horizontal transfer events. The increasing frequency of protein phylogenies that are in conflict with the conventional universal tree (Brown and Doolittle 1997) and the finding that the horizontal transfer of genetic information is pervasive among microbial lineages and that it may occur across different phylogenetic kingdoms (Gogarten et al. 1996; Lazcano and Miller 1996) indicates that this mechanism played a major role in shaping genome architectures and in fostering genetic adaptation and evolution. The horizontal transfer of entire metabolic pathways or part thereof might have had a special role during the early stages of cellular evolution.

There are many different schemes that can be proposed for the emergence and evolution of metabolic pathways, depending on the available prebiotic compounds and the available enzymes previously evolved. Even though most data coming from the analysis of completely sequenced genomes and directed-evolution experiments strongly support the patchwork hypothesis, we do not think that all the metabolic pathways arose in the same manner. In our opinion, the different schemes might not be mutually exclusive. Thus, some of the earliest pathways may have arisen from the Horowitz scheme, some from the semi-enzymatic proposal and later ones from Jensen's enzyme recruitment hypothesis. However, other ancient pathways, including histidine biosynthesis, might be assembled using (at least) two different schemes (Horowitz and Jensen).

References

Alifano P, Fani R, Lió P, Lazcano A, Bazzicalupo M, Carlomagno MS, Bruni CB. Histidine biosynthetic pathway and genes: structure, regulation and evolution. Microbiol Rev. 1996;60:44–69.

Brilli M, Fani R. Molecular evolution of *hisB* genes. J Mol Evol. 2004a;58:225–37.

Brilli M, Fani R. The origin and evolution of eukaryal *HIS7* genes: from metabolon to bifunctional proteins? Gene. 2004b;339:149–60.

Brown JR, Doolittle WF. Archaea and the prokaryote-to-eukaryote transition. Microbiol Mol Biol Rev. 1997;61:456–502.

Clarke PH. The evolution of enzymes for the utilization of novel substrates. Cambridge: Cambridge University Press; 1974.

Copley RR, Bork P. Homology among (betaalpha)(8) barrels: implications for the evolution of metabolic pathways. J Mol Biol. 2000;303:627–41.

Fani R. Gene duplication and gene loading. In: Microbial evolution: gene establishment, survival, and exchange. Washington, DC: ASM; 2004

Fani R, Fondi M. Origin and evolution of metabolic pathways. Phys Life Rev. 2009;6:23–52.

Fani R, Chiarelli I, Liò P, Bazzicalupo M. The evolution of the histidine biosynthetic genes in prokaryotes: a common ancestor for the *hisA* and *hisF* genes. J Mol Evol. 1994;38:489–95.

Fani R, Lió P, Lazcano A. Molecular evolution of the histidine biosynthetic pathway. J Mol Evol. 1995;41:760–74.

Fani R, Brilli M, Fondi M, Lió P. The role of gene fusions in the evolution of metabolic pathways: the histidine biosynthesis case. BMC Evol Biol. 2007;7 Suppl 2:S4.

Fondi M, Emiliani G, Fani R. Origin and evolution of operons and metabolic pathways. Res Microbiol. 2009a;160:502–12.

Fondi M, Emiliani G, Liò P, Gribaldo S, Fani R. The evolution of histidine biosynthesis in Archaea: insights into *his* genes structure and organization in LUCA. J Mol Evol. 2009b;69:512–26.

Gogarten JP, Hilario E, Olendzenski L. Gene duplications and horizontal gene transfer during early evolution. In: Roberts DML, Sharp P, Alderson G, Collins MA, editors. Evolution of microbial life. Cambridge: Cambridge University Press; 1996. p. 1996.

Holliday GL, Fischer JD, Mitchell BO, Thornton JM. Characterizing the complexity of enzymes on the basis of their mechanisms and structures with a bio-computational analysis. FEBS J. 2011; 278:3835–45.

Horowitz NH. On the evolution of biochemical syntheses. Proc Natl Acad Sci USA. 1945;31:153–7.

Horowitz NH. The evolution of biochemical syntheses—retrospect and prospect. In: Bryson V, Vogel HJ, editors. Evolving genes and proteins. New York: Academic; 1965. p. 15–23

Jensen RA. Enzyme recruitment in evolution of new function. Annu Rev Microbiol. 1976;30:409–25.

Lazcano A, Miller SL. How long did it take for life to begin and evolve to cyanobacteria? J Mol Evol. 1994;34:546–54.

Lazcano A, Miller SL. The origin and early evolution of life: prebiotic chemistry, the pre-RNA world, and time. Cell. 1996;85:793–8.

Lazcano A, Diaz-Villagomez E, Mills T, Orò J. On the levels of enzymatic substrate: implications for the early evolution of metabolic pathways. Adv Space Res. 1995;15:345–56.

Lewis EB. Pseudoallelism and gene evolution. Cold Spring Harb Symp Quant Biol. 1951;16:159–74.

Li WH, Graur D. Fundamentals of molecular evolution. Sunderland: Sinauer; 1991.

Miller SL. Production of amino acids under possible primitive earth conditions. Science. 1953;117:528–9.

Mortlock RP, Gallo MA. Experiments in the evolution of catabolic pathways using modern bacteria. In: Mortlock RP, Gallo MA, editors. The evolution of metabolic functions. Boca Raton: CRC; 1992

Ohno S. Evolution by gene duplication. Berlin: Springer; 1970.

Oparin AI. Proiskhozhdenie zhizny. Moscow: Izd. Moskovhii RabochiI; 1924.

Peretò J, Fani R, Leguina JI, Lazcano A. Enzyme evolution and the development of metabolic pathways. In: Cornish-Bowden A, editor. New beer in an old bottle: Eduard Buchner and the growth of biochemical knowledge. Valencia: Universitat de Valencia; 1998. p. 173–98.

Xie G, Keyhani NO, Bonner CA, Jensen RA. Ancient origin of the tryptophan operon and the dynamics of evolutionary change. Microbiol Mol Biol Rev. 2003;67:303–42.

Yanai I, Wolf YI, Koonin EV. Evolution of gene fusions: horizontal transfer versus independent events. Genome Biol. 2002;3.

Ycas M. On earlier states of the biochemical system. J Theor Biol. 1974;44:145–60.