Influence of pathway topology and functional class on the molecular evolution of human metabolic genes

Ludovica Montanucci[1], Hafid Laayouni[1,2], Begoña Dobon[1], Kevin L. Keys[1,3], Jaume Bertranpetit[1]*, Juli Peretó[4]

1 IBE Evolutionary Biology Institute (CSIC- UPF), Universitat Pompeu Fabra, PRBB, Barcelona, Catalonia, Spain, **2** Bioinformatics Studies, ESCI-UPF, Barcelona, Catalonia, Spain, **3** UCLA David Geffen School of Medicine, Los Angeles, California, United States of America, **4** Institute of Systems Biology, Department of Biochemistry and Molecular Biology, University of Valencia, Valencia, Spain

* jaume.bertranpetit@upf.edu

Abstract

Metabolic networks comprise thousands of enzymatic reactions functioning in a controlled manner and have been shaped by natural selection. Thanks to the genome data, the footprints of adaptive (positive) selection are detectable, and the strength of purifying selection can be measured. This has made possible to know where, in the metabolic network, adaptive selection has acted and where purifying selection is more or less strong and efficient. We have carried out a comprehensive molecular evolutionary study of all the genes involved in the human metabolism. We investigated the type and strength of the selective pressures that acted on the enzyme-coding genes belonging to metabolic pathways during the divergence of primates and rodents. Then, we related those selective pressures to the functional and topological characteristics of the pathways. We have used DNA sequences of all enzymes (956) of the metabolic pathways comprised in the HumanCyc database, using genome data for humans and five other mammalian species. We have found that the evolution of metabolic genes is primarily constrained by the layer of the metabolism in which the genes participate: while genes encoding enzymes of the inner core of metabolism are much conserved, those encoding enzymes participating in the outer layer, mediating the interaction with the environment, are evolutionarily less constrained and more plastic, having experienced faster functional evolution. Genes that have been targeted by adaptive selection are endowed by higher out-degree centralities than non-adaptive genes, while genes with high in-degree centralities are under stronger purifying selection. When the position along the pathway is considered, a funnel-like distribution of the strength of the purifying selection is found. Genes at bottom positions are highly preserved by purifying selection, whereas genes at top positions, catalyzing the first steps, are open to evolutionary changes. These results show how functional and topological characteristics of metabolic pathways contribute to shape the patterns of evolutionary pressures driven by natural selection and how pathway network structure matters in the evolutionary process that shapes the evolution of the system.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Metabolism is the set of enzymatic reactions that allows the synthesis, degradation and transformation of the biochemical components necessary for the maintenance and reproduction of a cell. Understanding the evolution of a system whose functioning arises from the interplay of many cellular components, is important both, for understanding the biology of the cell and for unraveling general principles of evolution of complex biological systems.

The origin and evolution of metabolic pathways is a difficult problem and several ideas have been proposed (reviewed in [1]). Among them, the patchwork model has gained a general acceptance. It proposes the evolution of enzymes from broader to narrower substrate specificities through gene duplication and the cooption of metabolic functions by the diverse pathways [2,3]. Nevertheless, the ability to contrast different models has been limited by the fact that all of them predate the current availability of complete genome sequences from the three domains of life [4]. Nowadays, complete genome sequences and subsequent reconstructions of genome-scale metabolic networks for many organisms have been used to test some of the predictions of evolutionary models. In the context of those systemic studies, the patchwork model exhibits a higher explicative power [5–10].

A full understanding of the evolution of metabolism also requires the understanding of the functional evolution of the enzymes. This can be achieved by investigating the selective pressures that have been acting on the genes that code for the enzymes (metabolic genes), and trying to understand their evolutionary dynamics in relation to the molecular systems they participate. An interesting approximation is the study of the selective pressures over the network structure of molecular systems; this can be achieved through the study of the relationship between parameters of the evolutionary histories of the enzyme-coding genes and the topological properties of their gene products within a network. Connectivity, which is the number of links of a node in a network, in a metabolic network represents the number of metabolic interactions and it is an initial measure for the topological description of each node.

Selective pressures are at the base of understanding adaptation and two main types have to be distinguished. By one hand purifying selection, which is the force that eliminates genetic variants that impair the function and, by the other, positive selection, in which one (or several) variants show a better adaptation and the frequency in the population will increase, reaching eventually fixation.

These evolutionary pressures may be detected and measured through dN/dS when comparing the genomes from different species. A negative correlation between connectivity and the evolutionary rates have been reported in the Drosophila and yeast genome-scale metabolic networks. In these networks, highly connected genes have been shown to evolve at slower rates, indicating selective constraints acting on them [11,12]. Hence connectivity has an effect on evolutionary rates, with higher connected genes under stronger purifying selection. In mammal genomes, the negative correlation between dN/dS and degree centrality is only found in four sub-cellular compartments while in other three a negative correlation is found between dN/dS and betweenness centrality [13].

This same approach, which couples the molecular evolution of genes with the knowledge of the interaction networks of their gene-products, has been also applied to study specific metabolic pathways, and the influence of the local network structure (that is the structure of only the metabolic pathway) is studied. This may reveal local strategies of adaptation that may be found different from the constraints imposed by the whole metabolic network, as well as it may shed light on network constraints specific to only specific pathways. In these works, the analysis at the network level initially seeks differences in the strength of the selection on genes located upstream versus those located downstream in the pathway, in order to detect whether

the position of a gene along the pathway may constrain gene evolution. In plant biosynthetic pathways, it has been found that upstream genes tend to evolve slower than those downstream due to a stronger selective constraint [14–18]. The analyzed pathways are secondary metabolite biosynthesis, with an inherent directionality given by the consecutive steps of the biosynthetic process. Further, many of the analyzed pathways are organized into branched structures: one or few initial substrates are processed into many final outputs. In such branched pathways with no loops, upstream genes are more likely to be above branch points and hence to be involved in the synthesis of more products than downstream genes [14]. These branched structures are common to many biosynthetic metabolic pathways. A proposed explanation for the observed gradient in selective pressures is that upstream genes are under stronger purifying selection because they are more pleiotropic than those downstream, affecting a greater number of end products. When the evolution of the genes of the N-glycosylation pathway during the divergence of primates was analyzed, an opposite trend emerged, with genes at the downstream position of this metabolic pathway being more constrained than upstream ones [19]. In a non-metabolic pathway, phototransduction, proteins which are topologically central in the signaling pathway, were found to be more constrained in their evolution; proteins peripheral to the pathway have experienced a relaxation of selective pressures [20].

All these studies suggest the idea that a part of the variation of evolutionary rates in metabolic pathways can be accounted for by the structure of their functional network, both the local metabolic pathway and the global whole-cell metabolic network. However, in the case of single metabolic pathways, different patterns have been found for different pathways and different species sets and no general pattern has emerged from these few cases. A large number of pathways and an overall vision of the metabolic network should be analyzed to reveal whether there exist general patterns in their evolutionary history and to better understand the distribution of selective pressures in the network, both in terms of constraints or adaptations. Here, we address the relationship between the topology of the whole set of human metabolic pathways during the time of divergence of primate and rodents and relate it to the evolutionary behavior of each gene in terms of natural selection (purifying and adaptive). Its relationship may help understand the distribution of evolutionary forces within complex biomolecular networks.

Methods

Data set

Pathways. The data set is composed of the metabolic pathways comprised in the Human-Cyc database, release 18.1. The number of human pathways present in HumanCyc 18.1 is 325. Of these, 13 pathways were not considered in the present study because they are signaling or protein modification pathways. Two pathways (morphine biosynthesis and melatonin degradation III) could not be used because their reactions are not associated to any annotated human gene. The final total number of considered metabolic pathways for the analysis is then 310. Of these, 275 are base pathways (comprised of reactions only) while 35 are super-pathways, which are comprised of one or more pathways, plus possible additional reactions. These super-pathways were treated separately in the analysis because they add redundant information. A full list of the pathways can be found in S1 Table.

The election of HumanCyc may requires justification. Even if there are other sources of reconstruction of human metabolism [21–23], in our case the choice of the HumanCyc database was precisely driven by the fact that it annotates pathways, and the study of the influence of the local topology (pathway topology) on the evolution of metabolic genes was precisely the question motivating this work. Our main objective was to gather the best possible functional,

expert-curated annotations within a pathway classification to be further manually curated and our examination indicated that HumanCyc had the best quality collection of pathways.

Reactions. The total number of enzymatic reactions comprised in the 310 pathways that are associated to at least one annotated gene is 879.

Genes. The total number of genes that encodes enzymes that take part in the annotated pathways is 956. Genes were functionally classified by assigning them to the functional class of the pathway in which they participate. Two different classification schemes were considered since there are different criteria according to which pathways can be assigned to distinct functional groups. The first classification scheme used, the ontology-based scheme, relies on the HumanCyc Pathway-ontology that aims at classifying pathways into a tree-based structure (as a gene ontology). We used the top level of the tree (the classes immediately below the root of the tree, which is "Pathway") as classes to assign pathways. We used 7 of the 9 parent classes just after the root of the HumanCyc pathways ontology, excluding the two classes for non-metabolic pathways, "Macromolecule Modifications" and "Signal transduction pathways". The basic criteria of this ontology rely on the metabolic mode of the pathway, for example biosynthesis versus degradation. The second classification schema, compound-based, is based on the kind of compounds that are primarily transformed in the pathway, for example nucleotide versus fatty acids (see classification in S1 Table).

Computation of evolutionary rates

Evolutionary rates were estimated during the divergence of mammals and rodents. For each human gene, its orthologous sequences were derived from the following species: chimpanzee, orangutan, gorilla, mouse and rat. Multiple sequence alignments of the coding regions were downloaded from Ensembl (release 75). When absent, orthologous sequences were predicted, if possible, through a similarity search of the human gene sequence against the genome assembly, followed by subsequent gene prediction by GeneWise [24], in a procedure described in [19]. In case of predicted orthologues, multiple sequence alignments were obtained through T-coffee with default options by aligning protein sequences and then back-translating to genomic sequence. The same procedure was adopted for incomplete or bad quality sequences.

Evolutionary rates were computed using the codeml program of the PAML package [25]. Two likelihood ratio tests between pairs of nested model (M1a versus M2a and M7 versus M8) were carried out to detect positive selection events. The overall strength of purifying selection on each gene was estimated though a unique dN/dS over the entire tree and sequence length (model M0). Each maximum-likelihood estimation, including likelihood ratio tests, was carried out 5 times, each with 3 different initial dN/dS values: 0.1, 1 and 2 to check for stability of the results to repeated runs and different initial conditions. Final results of the likelihood ratio tests were corrected through a False Discovery Rate (FDR) method [26]. Positive selection was inferred when either one or both of the two likelihood ratio tests was significant after correcting for multiple testing. Given the shallow divergence considered, non-branch-specific models (M0 averaged dN/dS and site-specific positive selection tests) provide the best estimation of the overall selective pressure acting on each gene, given that for low number of closely related species branch-specific estimations that lack the power to provide stable estimations.

Given the strong impact of alignment errors in generating spurious signals of positive selection, the alignments corresponding to genes with $P < 0.05$ in the likelihood ratio test for positive selection were inspected visually. Alignment regions containing evident errors (usually contiguous positions in a sequence of the alignment with a suspiciously high number of differences often resulting from sequencing or assembly errors) were manually masked and then evolutionary rates and positive selection tests were then recomputed.

Computation of topological parameters

Building the reaction graph. Metabolic pathways were derived from HumanCyc. The classical representation of metabolic pathways, also used in HumanCyc, is through the substrate graph, in which nodes represent metabolites and edges represent reactions that transform metabolites. Here we represent pathways as reaction graphs [27], in which nodes represent reactions and edges link consecutive reactions. Consecutive reactions within a pathway were derived from the PREDECESSORS field of the pathways.dat flat-file of the Human-Cyc database. The direction provided in this field was used to build directed graphs, in which edges are arrows having a direction that goes from the preceding to the following reaction. The graphs have been constructed through Python scripts from the flat-files downloaded from HumanCyc 18.1.

Centrality measures. Topological measures within the directed graphs built for each pathway independently were computed through build-in functions of the NetworkX Python package. Four centrality measures were computed: closeness, betweenness, and in-degree and out-degree centrality. Centrality measures for the corresponding undirected graphs were also computed.

Loops. A Boolean variable was derived for each pathway to detect whether the pathway contains loops. This computation has been achieved through the *simple_cycles()* function of the NetworkX package applied to the directed graph. This variable was only computed for pathways of more than one reaction.

Top/Bottom position. For each pathway, nodes were assigned to three different classes depending whether they are at the beginning of the pathway (in-degree = 0, "upstream"), at the end (out-degree = 0, "downstream") or in any other position in between ("intermediate"). The uppermost positions (top positions), are the first reaction steps of the pathway, corresponding to initial reactions. At bottom positions there are the reactions that produce the final products of the pathway. Reactions in any other position in between of these two are assigned to the same class (intermediate). Beside these three classes, a fourth class has been added for nodes (reactions) that become isolated when the substrate graph is transformed into the corresponding reaction graph. This is the case for pathways for which the reaction graph comprises more than one connected component. These reactions have the feature of directly catalyzing the end products of the pathway starting from the initial substrates, thus being at the beginning (first step) and at the end (last step) at the same time.

Statistical analyses

To evaluate the importance of relationship between evolutionary estimates and the different descriptive network properties, we performed a multivariate analysis, through an automated linear modeling routine implemented in SPSS software. The automated linear modeling created a single standard model to explain the relationship between fields. The adopted linear modeling also included codon bias, protein length and CG content as explanatory variables. Before the linear modeling missing values were replaced. No variable selection method was used and all variables entered into the model were assessed. The automated linear modeling allows also for detecting outliers. However, outliers were not excluded of the analysis. To compare groups, non-parametric methods were used. To correct for multiple testing, False Discovery Rate (FDR) methods were applied [26].

Genes considered under positive selection were compared to the whole set of metabolic genes for several centrality measures using a permutation test. For every centrality measure, the mean score of the genes under positive selection was compared to the distribution of the

mean scores of a set of randomly selected genes from the whole set of metabolic genes with the same sample size, using a permutation test (10.000 permutations).

Results

Data set description

The total number of pathways used for the analysis is 310, with 275 base pathways and 35 super pathways (see Methods). A full list of the pathways can be found in S1 Table. Of the 275 base pathways, 30 comprise only one reaction, while the remaining 245 pathways comprise a number of reactions that ranges from 2 to 30. The majority of the pathways (208 out of 275 base pathways and 28 out of 35 super pathways) show no loop structures. They are characterized by a non-feedback topology, with either linear or branched structures. For 27 of the 35 super-pathways, their reaction set is fully comprised in base pathways, so excluding them from the analysis would not result in loss of reactions and genes.

The total number of genes associated to the reactions in the pathways is 956. 943 out of the 956 gens participate in base pathways, while 13 genes (*NAGK, GGCT, SRR, TYW4, GGT1, EBP, SC5DL, DHCR7, DHCR24, CCBL1, OPLAH, CNDP2, CCBL2*) are associated to reactions that are uniquely present in super pathways. Of the 956 genes, 335 encode proteins that carry out their enzymatic activity within protein complexes while 621 genes encode proteins that are themselves the functional enzymes.

The relationship between enzymatic activities and enzymes is not one-to-one: on the one hand, each gene encodes an enzyme (or an enzymatic subunit) that may carry out more than one catalytic activity and, on the other hand, the same catalytic activity can be served by more than one enzyme (isoenzymes) that are encoded by different genes. Within this scenario, 71% of the genes (677 out of 956) code for enzymes (or enzymatic subunits) that carry out only one metabolic function, 26% of the genes (247 out of 956) are associate to a number of reactions between 2 and 5, and the remaining 3% of the genes (32 out of 956) are associated to more than 5 reactions. The genes whose encoded enzymes are associated to the biggest number of reactions are: *FASN* (fatty acid synthase), *CYP2A6* (a type of cytochrome P450), *UGT2B11* (a type of UDP-glucuronosyltransferase), *SCMOL* (a methylsterol monooxygenase) *DIO3* (deiodinase, iodothyronine, type III), *ALOX5* (arachidonate 5-lipoxygenase) and *ALDH3A2* (fatty aldehyde dehydrogenase).

Purifying and positive selection in metabolic genes

The evolutionary rates have been computed for 927 genes (for 29 it was not possible due to lack of or poor quality of one or more of the orthologous sequences) considering the full set of six species (human, chimpanzee, orangutan, gorilla, mouse and rat): the synonymous evolutionary rate (*dS*), the non-synonymous evolutionary rate (*dN*) and their ratio (*dN/dS*), which provides the direction of the action of the selection and an overall estimation of the strength of the purifying selection that have acted on each gene. The distribution of the evolutionary rates can be seen in Fig 1, which shows that the main force that has shaped the evolution of metabolic genes during the mammal evolution has been purifying selection, with all *dN/dS* values lower than 0.5.

Beside an estimation of purifying selection, for each one of the 927 genes we performed two likelihood ratio tests to look for sequence signature of positive selection. After multiple test correction, using false discovery rate, we found that only six genes show a significant P-value for the likelihood ratio test. The genes and their p values for the M7 versus M8 test are: CYP2E1 (P = 0.0000005, corrected = 0.00043), HDC (P = 0.0000217, corrected = 0.01006), CES1 (P = 0.0000471, corrected = 0.01455), DPM2 (P = 0.0001017, corrected = 0.02356),
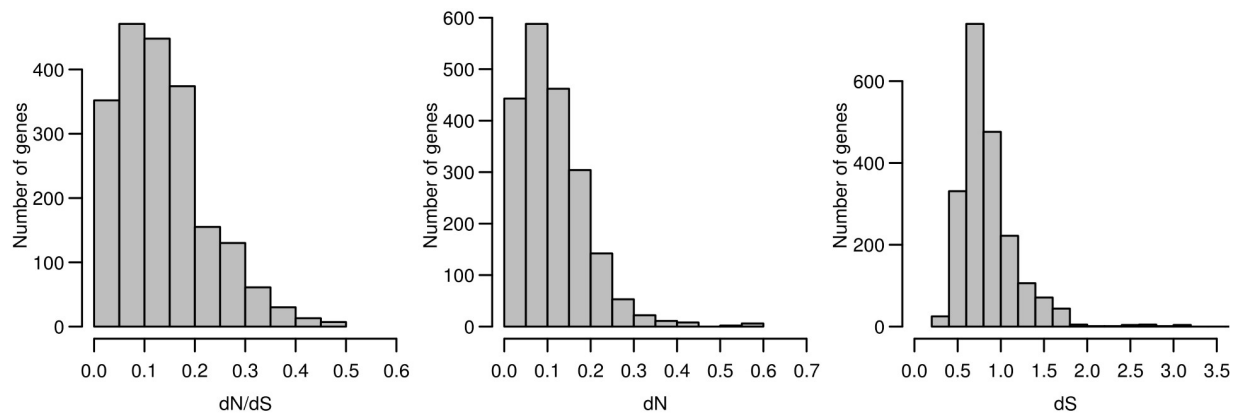
Fig 1. Distribution of the evolutionary rates (*dN/dS*, *dN* and *dS*) for the 927 metabolic genes.

SPAM1 (P = 0.0001575, corrected = 0.02919) and AKR1C1 (P = 0.0002451, corrected = 0.03787).

CYP2E1 is a member of the cytochrome 450 family of enzymes involved in the inactivation of drugs and xenobiotics. The HDC gene codes for histidine decarboxylase, which converts L-histidine into histamine, a biogenic amine involved in different physiological processes such as neurotransmission, gastric acid secretion, inflammation and regulation of circadian rhythm. CES1 codes for a carboxylesterase involved in the hydrolysis of various xenobiotics and drug clearance in liver. DPM2 encodes for the regulatory subunit of the dolichol-phosphate mannose (DPM) synthase complex whose main function is recognition on the cellular surface. SPAM1encodes for a hyaluronidase located on the human sperm surface that enables sperm to penetrate through the hyaluronic acid-rich envelope of the oocyte. AKR1C1 belongs to a superfamily of aldo-keto reductases and catalyzes the reaction of progesterone to the inactive form 20-alpha-hydroxy-progesterone.

The quality control of the alignments which involve the removal of bad quality regions from the computation of the evolutionary rates might have led to an underestimation of the positive selection events, however it guarantees the reliability of the events found.

Functional classes

Beside the evolution of specific genes under positive selection, we investigated different levels of selective constraint between functional classes by comparing evolutionary rates of genes participating in pathways performing different functions. A Kruskal-Wallis (KW) test was used to test differences in evolutionary rates between genes belonging to different functional classes of pathways. For both ontology-based and compound-based classification, we find significant differences among different functional classes in the evolutionary rates *dN*, and *dN/dS* (with *dS* close to significance).

When we consider the ontology-based classification (S1 Fig), differences in *dN/dS* reveals relaxed constraints (high *dN/dS* values) for external routes ("Detoxification"), and strong constraints (low *dN/dS)*, for routes of the core metabolism ("Generation of precursor metabolites"). We also find that biosynthesis routes are more constrained than degradation ones. These differences in the strength of purifying selection can reflect the need of making the biosynthesis of very specific metabolites, but degradation of a broad range of external compounds, most likely unknown and toxic.

When we consider the compound-based classification (S2 Fig), values of dN/dS separate those with highest values, Steroid, Secondary metabolism and Detoxification, and those with the lowest, mainly Glycolysis/TCA/PentoseP, which corresponds to the inner core of the metabolic network. In particular, the detoxification class, the one with the highest dN/dS, contains pathways with enzymes able to recognize a broad range of metabolites. Thus, both classification schemes show that genes participating in peripheral routes have evolved under relaxed constraint, while stronger selective constraint has acted on the genes whose encoded enzymes have roles within the central metabolism.

From a conceptual point of view, traditionally metabolism has been regarded as a series of layers of complexity, from central pathways involved in basic energy transactions and the generation of intermediate precursors, followed by a tier of almost universal biosynthetic pathways that use a few of those precursors to generate biomass components, and, finally, some processes connected to central intermediates that generate a diversity of metabolites usually related to behavioral or environmental cues, and showing a more restricted phylogenetic distribution, also known as secondary metabolism. Theoretical approaches based on graph theory support this classical image of metabolic organization [28,29]. Thus, we grouped the functional classes into three main groups according to the layer of the global metabolism in which they operate: the inner core of intermediary metabolism comprises classes of Glycolysis/TCA/PentoseP and Polysaccharides; the second layer of intermediary metabolism with Membrane Lipids, Nucleotide, Fatty acid/TAG, Cofactor, Fatty Acid/hormone, and Amino acid; and the outer layer of metabolism comprises the classes of Steroid, Secondary Metabolism and Detoxification. When we compare the evolutionary rates of these classes through a KW test we find highly significant differences with a clear gradient of selective constraints in which central routes of the core metabolism are more constrained (lower dN and dN/dS) than the ones in the second layer which, in turn, are more constrained than those of the more peripheral layer (Fig 2 and S3 Fig). Pairwise comparisons hold significant after correcting for multiple testing ($P < 0.001$). Fig 2 clearly shows the steep gradient of selective pressures with genes of the central metabolism under the stronger selective constraint and genes of the peripheral routes under relaxed selective constraint.

Presence/Absence of isoenzymes

Given that roughly 36% of the reactions are catalyzed by more than one enzyme (different enzymes, encoded by different genes, able to catalyze the same reaction or isoenzymes) we
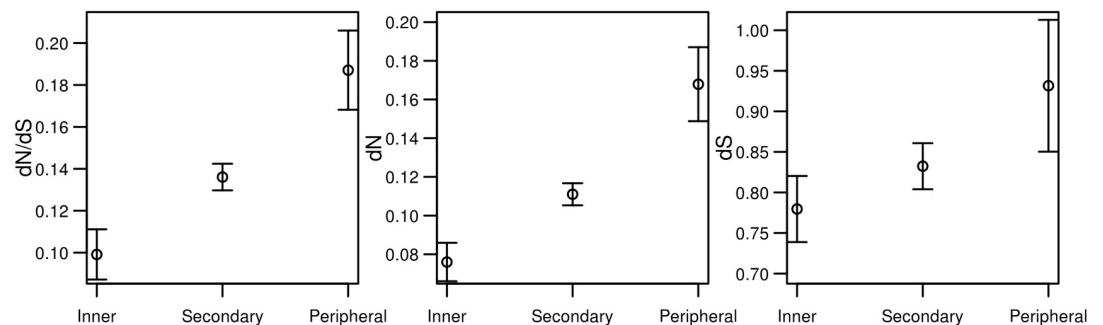
Fig 2. Graph representing *dN/dS*, *dN* and *dS* among genes belonging to different functional classes. Class 1 comprises the inner metabolism (Glycosis/TCA/PentoseP, Polysaccharides), class 2 comprises the second layer (Membrane Lipids metabolism, nucleotide metabolism, Fatty acid/TAG, Cofactor, Fatty Acid/hormone, and Aminoacid) while class 3 comprises the outer layer of cell metabolism (Steroid, Secondary Metabolism and Detoxification).

https://doi.org/10.1371/journal.pone.0208782.g002

Table 1. Results for the three independent runs of automated linear modeling with evolutionary measures as dependent variable for the 275 base pathways. The following variables have been considered by the model: 3 gene sequence properties, effective number of codons (ENC), gene length (Length) and CG content (CG); and 4 topological centralities: closeness, betweenness and in-degree and out-degree centralities. Beta coefficients and P-values (between brackets) are reported; significant values according to a threshold of 0.05 are presented in bold.

	ENC	Length	CG	Closeness	Betweenness	In-degree	Out-degree
dN/dS	**0.003** **(<0.0001)**	**-0.000 (<0.0001)**	0.002 (0.958)	-0.056 (0.071)	-0.017 (0.225)	**-0.035** **(<0.0001)**	0.038 (0.246)
dN	0.000 (0.901)	**-0.000 (<0.0001)**	0.022 (0.529)	-0.019 (0.502)	-0.021 (0.111)	**-0.030** **(<0.0001)**	0.011 (0.703)
dS	**-0.025 (<0.0001)**	**-0.000 (<0.0001)**	**0.549** **(<0.0001)**	0.021 (0.866)	0.054 (0.339)	-0.019 (0.573)	-0.037 (0.779)

https://doi.org/10.1371/journal.pone.0208782.t001

compared the selective pressures that have acted on genes that encode for unique enzymes, to that of genes that encode for enzymes for which isoenzymes exist. The existence of isoenzymes indicates that alternative proteins could be recruited to perform the same metabolic function, whereas the presence of a unique enzyme is critical for their specific metabolic function to be served. Interestingly, we find no difference in selective constraint between these two classes of genes. A Mann-Whitney U test shows no significant differences: P = 0.066 for *dN/dS*, P = 0.109 for *dN* and P = 0.668 for *dS*. This result shows that sequence properties of isoenzymes do not differ from those of the rest of the enzymes; they cannot be considered simply as redundant.

Evolutionary rates and topological properties

In order to investigate whether and how the organization into metabolic networks imposes constraints on the evolution of the enzyme-coding genes, we carried out automated linear modeling to reveal possible linear relationships between evolutionary rates and topological features of reactions (Table 1). Among the topological features that can be associated to each node and that summarize aspects of the node's position within the network, we first considered four centrality measures: (i) in-degree centrality, which is indicative of the number of incoming links pointing to a node; (ii) out-degree centrality, which is indicative of the number of outgoing links stemming from a node; (iii) betweenness centrality, which is indicative of the importance of a node in linking parts of the networks; and (iv) closeness centrality, which is indicative of whether a node lies in the central or peripheral part of the network. In this analysis, pathways with the category "one-reaction" are excluded, given that topological measures by definition cannot be computed. In the multivariate statistical analysis we included three sequence features of the genes that are known to influence evolutionary rates and thus have to be taken into account when analyzing variations in evolutionary rates. These three sequence features are: (i) ENC, the effective number of codons, which quantifies codon bias and is correlated with expression levels; (ii) length of the coding sequence (CDS) of the gene; and (iii) CG content of the CDS.

As seen in Table 1, synonymous substitution rates show significant linear relationships only with sequence features (codon bias, sequence length and CG content), while no significant correlation is found with any of the topological parameters. This implies that, as expected, neutral evolution (evolution at synonymous sites) is only affected by the nucleotide composition of the sequence itself and not by the position of the gene in the network. On the contrary, in the case of functional evolution (evolution at non-synonymous sites and *dN/dS*) a significant linear relationship is found with topological parameters, showing that the evolutionary rates suffer the influence of the position and the role of gene products within the pathways. In particular a highly significant negative correlation is found between both *dN* and *dN/dS* and in-

degree centrality, meaning that genes that have high in-degree centralities are highly constrained in their evolution with a strong purifying selection, while genes that have low in-degree centralities are free to evolve under relaxed constraints and accumulate non-synonymous substitutions at a faster rate. This result holds for pathways of different topologies regardless of whether the structure presents loops or not.

Position along the pathway: Top/Intermediate/Bottom classification

In order to analyze the possible relationship of topological centrality measures and the position within the specific metabolic pathways, we took advantage of the physiological directionality of metabolic pathways and we generated a new categorical variable to account for position. Three different classes of reactions have been defined according to their position within the pathway: reactions that catalyze the first enzymatic step in the pathway (top position); reactions that catalyze the last step in the pathway (bottom position); all the in-between steps are assigned to same class (intermediate). A fourth class has been introduced for reactions that catalyze the first and last step at the same time, being, for example the only reaction of one branch of a branched pathway that directly converts the initial substrates into the final products. Genes are assigned to the class of the reaction catalyzed by the enzyme they encode. It should be noted that this classification on position is correlated with in-degree centrality: nodes at top positions are those with no incoming links (in-degree equal to 0) and those at bottom positions are defined solely on the basis of out-degree (out-degree equal to 0). This new variable, even if correlated with in-degree, encodes different positional information that is not fully captured by the in-degree centrality.

We analyzed whether genes whose encoded enzymes catalyze reactions belonging to these four classes have evolved under different selective pressures. Evolutionary rates for these four classes are shown in Fig 3a and S4a Fig. The statistical significances of their differences have been assessed through Kruskal-Wallis, which show that there are significant differences between the four considered classes in dN/dS (P = 0.001) and dN (P = 0.001) while no significant differences are found in dS (P = 0.151). Both cases of dN/dS and dN pairwise comparisons show that the intermediate class has a lower dN and dN/dS than the top (0.022 for dN/dS and 0.010 for dN) and one-step (0.010 for dN/dS and 0.005 for dN) classes. This means that genes whose encoded enzymes catalyze reactions at top positions have undergone more relaxed evolution and faster non-synonymous substitution rates than genes at intermediate position. Genes belonging to the one-step class have the highest evolutionary rates.

Given that the last three classes (top, intermediate, and bottom) constitute an ordered variable of the position along the pathway, we carried out a trend test to test whether a linear relationship exists of the evolutionary rates and these positional classes. A significant linear relationship is found for dN for the three classes of top-intermediate-bottom with P-value of 0.008. Non-significant P-values for the linear relationship are instead found for dN/dS (0.072) and for dS (0.481). This means that a statistically significant gradient in the rate of non-synonymous substitution is found along metabolic pathways with genes in top positions having experienced faster non-synonymous substitution rates.

Given that the top/bottom position implies directionality for the pathway, we repeated the analysis in the subset of 208 pathways that have no loops thus have a marked directional structure (Fig 3b and S4b Fig). As before, KW tests show that there are significant differences between dN/dS (P = 0.023) and dN (P = 0.005) between the four classes considered, while no significant differences are found in dS (P = 0.300). When looking at pairwise comparisons we find that dN/dS of genes of the one-step class have a significantly higher dN/dS than those belonging to the intermediate (P = 0.035) and to the bottom (P = 0.019) class, and the dN of
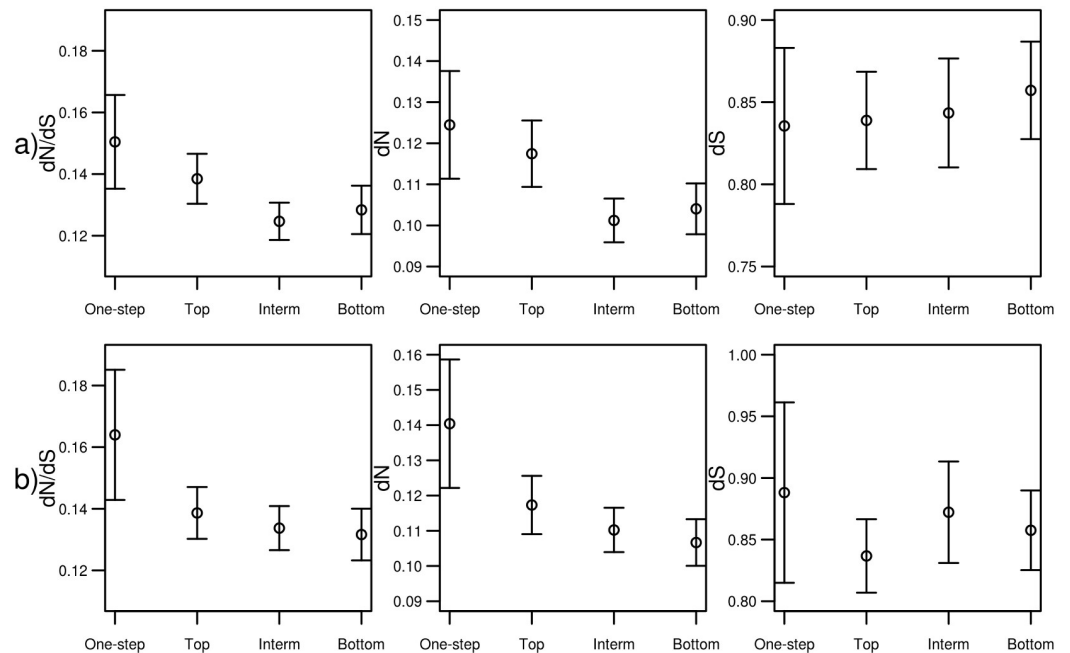
Fig 3. Representation of ω (*dN/dS), dN* and *dS* among genes belonging to positions within the pathway. a) Graph representing *dN/dS, dN* and *dS* among genes whose position within the pathway belongs to four classes (one-step, top, intermediate, bottom positions) for the 275 base pathways. **b)** The same for the 208 base pathways with no loops. Dots show the mean ± 2 standard error (SE).

genes belonging to the one-step class are significantly different from *dN* of the remaining three classes: top (P = 0.034), intermediate (P = 0.006), bottom (P = 0.004).

In summary, the stringent KW test clearly highlighted differences between the rates of functional evolution of the genes belonging to the one-step class and those belonging to the other classes, pointing to their relaxed evolutionary constraint. This implies a gradient of non-synonymous substitution rates along metabolic pathways with genes at top-positions allowed to evolve at a faster rate than genes at intermediate and bottom positions, and with genes at intermediate and bottom positions being more tightly constrained to fix non-synonymous substitutions at a slower rate.

Topological measures of positive selected genes

We measured the differences in the mean of the four centrality measures (in-degree, out-degree, closeness and betweenness) between genes with signatures of positive selection and the whole set of metabolic genes. We tested two sets: genes under positive selection with and without multiple testing correction. When only the six genes under positive selection after multiple testing correction were considered, no statistical significance is observed, even though genes under positive selection show higher out-degree and higher closeness than the average (0.33 vs. 0.22 for the out-degree and 0.38 vs. 0.28 for the closeness). Given the small number of positively selected genes (six genes), this same analysis was also repeated with all genes (49 genes) that have a p-value smaller than 0.05 in the positive selection test before the multiple test correction. The aim of this analysis is to test if there is any difference in the centrality measures values for the genes that belong to the tail of the positive selection test distribution. When these two groups are compared, higher out-degree (one tail permutation test, P = 0.0084) and higher closeness (one tail permutation test, P = 0.0116,) values are found for positively selected

genes with an increase of 32% in the average value of out-degree and of 21% in closeness in genes under positive selection compared to the whole set of metabolic genes. This result shows that the tail of positive selection distribution is enriched by genes with higher out degree and higher closeness.

These results indicate that genes encoding enzymes with a greater number of reactions that make use of their products in the human metabolic pathways are more likely to present signals of positive selection than those with fewer enzymes using their products. Genes under positive selection have also higher closeness than the average and hence shorter path lengths to other nodes in the pathway.

Discussion

Linking the action of natural selection in the evolution of genes to the network structure and topology is an interesting approach to understand the constraints that the network structure may have on the evolution of complex molecular systems such as metabolism. Here, we have carried out a comprehensive molecular evolutionary study of human metabolism by investigating the selective pressures that acted on the enzyme-coding genes during the divergence of primates and rodents, and their relationships with functional and topological features of the pathways that constitute the system. Extensive studies have made possible the reconstruction of the biochemical pathways that constitute the metabolism; here we use this information to investigate the influence of the local network topology of the metabolic pathways in its evolutionary behavior by analyzing the distribution on the network of selective forces, be they in form of innovations (positive or adaptive selection) or in the strength of conservation (purifying selection).

The analysis of individual metabolic pathways instead of the whole metabolic network has several advantages: i) it allows the study of the influence onto the evolution of metabolic genes of their local relevant environment, that is, the context of the gene products in which the metabolic task is achieved; ii) it allows to separately study the functional units responsible for the different metabolic tasks, classify and compare them, and study the distribution of selective pressures within each functional unit; these functional units are particularly relevant because are likely to approximate the "molecular phenotypes" targeted by selection; and iii) it allows an intermediate analysis between the single enzyme and the whole metabolic network in the line of considering the hierarchical and topological structure of the pathways. However, the partitioning of the metabolic network into pathways is a somehow arbitrary process, the most arbitrary decision being the definition of pathway boundaries. Criteria for the definition of a pathway have been analyzed and compared [30,31] and the best collection of pathways available for human metabolism is comprised in the HumanCyc database [31,32]. In HumanCyc, the criteria used for pathway definition are clearly stated and uniformly applied to the whole database [30]. Importantly, the HumanCyc/MetaCyc approach of defining pathway boundaries through a multi-organism meta-metabolism implicitly introduces phylogenetic information and thus, pathways defined therein represent the best approximation of the functional units targeted by selection.

The final dataset under study was composed of 927 genes, whose products are integrated in 310 pathways. The species that have been considered are human, chimpanzee, gorilla, orangutan, mouse and rat, and thus the analysis embraces the divergence time of both primates and rodents. The tools for detecting positive selection and for measuring the strength of purifying selection (see Methods) are based on the amino acid impact of nucleotide changes.

The detection of genes that underwent positive selection, and that are thus at the base of innovative changes, have resulted in a very small number of genes, six out of the 927 genes.

These genes are: CYP2E1, a member of the cytochrome 450 family; HDC, a histidine decarboxylase; CES1, a carboxylesterase; DPM2, a subunit of the dolichol-phosphate mannose synthase complex; SPAM1, a hyaluronidase; and AKR1C1 an aldo-keto reductase. Two of the six genes that show sequence signature of positive selection, CYP2E1 and CES1, encode for detoxification enzymes and contribute to the solubility of molecules that must be expelled from the cell as fast as possible. Therefore, response to xenobiotic molecules has likely been a target of adaptive selection during primate and rodent divergence.

The main selective force in metabolism is the maintenance of the system through purifying selection; we have calculated this strength for each gene and analyzed the context according to biochemical and network properties. We observed a steep gradient of selective constraints that goes from genes serving functions within the inner core of intermediary metabolism (the most constrained), through those of a second layer of intermediary metabolism, to the outer peripheral layer of metabolism (*i.e.* secondary metabolism), which shows the most relaxed selective constraint. We have shown that the stronger selective constraint has acted on the genes whose encoded enzymes have roles within the inner core of metabolism; pathways comprised in this inner layer are involved in the transformations of small precursor metabolites for cell maintenance. These pathways are the oldest, the more phylogenetically conserved [33] and are enriched in enzymes exhibiting more substrate specificity [34]. Accordingly, we have found that enzymes participating in more peripheral routes are evolutionarily less constrained and more plastic, having experienced faster functional evolution. Generalist enzymes, able to cope with a vast diversity of possible small molecular structures, populate pathways of this outer layer. The strategy of adopting generalist enzymes at the outer interface of metabolism may be a more efficient strategy than to develop a specific enzyme for each type of possible metabolite that may be present in the environment. Indeed, a global analysis of kinetic parameters of several thousands of known enzymes showed that central metabolism enzymes perform better in terms of catalytic constants (*i.e.* higher k_{cat} and k_{cat}/K_M) than secondary metabolism enzymes [35]. In line with the authors' suggestion, we have been able to prove that there is a stronger selective pressure on central metabolic enzymes, probably due to the need of maintaining the catalytic parameters that allow higher fluxes in central pathways, in comparison to those operating at lower fluxes and less specificity in secondary metabolism [35].

Isoenzymes are of special interest because they provide material for metabolic evolutionary innovation through sub- or neo-functionalization [36]. Here we have found no differences in evolutionary rates between isoenzymes and the rest of the enzymes; no differences in their connectivities were found in the global metabolic network of *E.coli* [37]. Thus, both the sequence and the topological properties of isoenzymes do not differ from those of the rest of the enzymes. Selective constraint acting with the same strength onto these two classes of genes suggests that what may seem like alternative enzymes for the same metabolic function, are, indeed, equally "essential", and the functional degeneracy is only apparent. Isoenzymes are likely to be essential to their function either through regulation, differential expression in time (different developmental stages), or in space, displaying their function in different cellular compartments or tissues. The study of evolutionary pressures over gene sequences has clearly pointed to lack of functional degeneracy for isoenzymes, given that redundancy leaves clear detectable footprints in terms of acceleration of substitution rates, as for example in the case of paralogous gene copies just after the duplication event [36].

To assess the influence of system properties on gene evolution we have used the local topology of the metabolic network. Among the many possible ways of representing metabolic pathways through graph structures, we have encoded pathways as reaction graphs, a type of representation in which nodes represent reactions; hence they can be directly associated to the genes that catalyze the reactions. In this representation edges represent metabolites (reaction

substrate and products) and here we have considered their direction. So, in our representation edges are indeed arrows and the resulting graph is a directed graph. By encoding metabolic pathways through directed graphs, we are able to take into account the physiological direction of the reactions in the cell.

Even in absence of specific information about the metabolic pathways for all the considered species, the relatively shallow phylogenetic divergence considered and the well-known conservation of metabolism among mammals ensures the use of a unique structure of pathways and function of all enzymes. The influence of the local network topology over gene's evolution was investigated for both positive and purifying selection.

The influence of the local network topology over gene's evolution had been previously investigated in few cases of specific metabolic pathways. The comprehensive analysis carried out here allowed revealing that this influence is pervasive and general patterns can be found. When positive selection has been considered we have found that positively selected genes have higher out-degree centralities than non-adaptive genes. Genes with high out-degree are not involved in the production of the final products of the metabolic task and, at the same time, are those whose products are subsequently transformed by a high fraction of different reactions in the pathway. Here we see that genes in these positions are preferentially targeted by adaptive evolution.

When purifying selection has been considered, in-degree centrality has been identified as the strongest topological factor constraining metabolic gene's sequence evolution. Genes characterized by higher in-degree connectivities have evolved under stronger purifying selection. Within the context of biochemical reaction graphs in which each node represents a biochemical reaction, in-degree reflects the number of reactions that directly precede the given one, that is, the number of reactions that produce metabolites that are then taken as substrates by the given reaction. This means that in-degree centrality is higher for reaction with many incoming links and lower for reactions with few incoming links.

Besides the encoding of topological information through centrality measures we also used topological features that encode the position along the pathway (top, intermediate and bottom) of each enzyme. When this topological information was considered for plant biosynthetic pathways [14–18] a progressive relaxation of selective constraint along metabolic pathways was found. Here we found an opposite gradient when the whole set of human metabolic pathways were analyzed during the divergence of primates and rodents, with genes at top positions being under more relaxed constraint and genes at bottom position being under stronger selective constraint. Thus, relaxed evolutionary forces at top positions allow broader evolutionary changes while, at the bottom, strong selective constraint narrows the allowed changes in functional variation, according to a funnel like model. This funnel like distribution of selective pressures along positions in the pathway is a general pattern found throughout the considered metabolic pathways. This result is consistent with the results obtained in the N-glycosylation pathway [19] where genes at downstream positions of the pathways were found to be under stronger selective constraint than genes located at upstream (top) positions in the pathway which in turn had undergone more relaxed evolution. For metabolic routes that are directly connected with external environment, enzymes at top positions are usually more generalists in response to the need of using a broad diversity of possible metabolic material, while those at the bottom positions are more specialized, in response to the need of producing very specific products. It can be speculated that the results obtained here may reflect the relevance of accuracy in the synthesis of final metabolic products.

The comprehensive analysis of the whole set of human metabolic pathways revealed that both adaptive and purifying selection are not evenly distributed among the genes encoding the enzymes involved in metabolic pathways. Adaptive selection has targeted a small number of

genes during the divergence of primates and rodents and adaptive genes are mainly involved in detoxification functions. Purifying selection has been a pervasive selective force dominating the evolution of metabolic genes during the divergence of primates and rodents. It has acted with different strengths according to the layer of metabolism over which it acts, with the inner core of metabolism being strongly conserved and with little or no room left for evolutionary innovation. From these results, it is tempting to conclude that it is less likely to innovate on pathways that were established in the early (*i.e.* prokaryotic) stages of evolution and that are involved in the synthesis of a small set of metabolic precursors linking the synthesis and degradation of essential biomolecules, namely sugars, lipids, amino acids and nucleotides. A more relaxed selection has been found for enzymes that manage higher levels of chemodiversity. This is the case of detoxification of xenobiotics or the biosynthesis of a wide spectrum of secondary metabolites that, by definition, are not directly involved in the survival of the organism, but rather in its ecological and behavioral traits.

Supporting information

S1 Table. List of the 310 pathways, their codes, common names and their classifications.
(DOC)

S1 Fig. Graph representing *dN/dS*, *dN* and *dS* among genes belonging to different functional classes according to ontology-based classification.
(DOCX)

S2 Fig. Graph representing *dN/dS*, *dN* and *dS* among genes belonging to different functional classes according to compound-based classification.
(DOCX)

S3 Fig. Boxplots representing ω (*dN/dS), dN* and *dS* among genes belonging to different functional classes.
(DOCX)

S4 Fig. Boxplots representing ω (dN/dS), dN and dS among genes according to the position in the pathway.
(DOCX)

Acknowledgments

Author Contributions

Conceptualization: Ludovica Montanucci, Jaume Bertranpetit, Juli Peretó.

Data curation: Ludovica Montanucci, Hafid Laayouni, Begoña Dobon, Kevin L. Keys, Juli Peretó.

Formal analysis: Ludovica Montanucci, Begoña Dobon, Kevin L. Keys.

Funding acquisition: Jaume Bertranpetit.

Investigation: Ludovica Montanucci, Hafid Laayouni, Begoña Dobon, Kevin L. Keys, Jaume Bertranpetit.

Methodology: Begoña Dobon, Juli Peretó.

Resources: Hafid Laayouni, Juli Peretó.

Software: Ludovica Montanucci.

Supervision: Jaume Bertranpetit.

Validation: Jaume Bertranpetit.

Writing – original draft: Ludovica Montanucci, Hafid Laayouni, Jaume Bertranpetit, Juli Peretó.

Writing – review & editing: Ludovica Montanucci, Hafid Laayouni, Jaume Bertranpetit, Juli Peretó.

References

1. Peretó J. Origin and evolution of metabolisms. Gargaud M, Lopez-Garcia P, Martin H, editors. Origins and Evolution of Life. An Astrobiological Perspective. Cambridge: Cambridge University Press; 2011.

2. Yčas M. On earlier states of the biochemical system. J Theor Biol. 1974; 44: 145–160. https://doi.org/10.1016/S0022-5193(74)80035-4 PMID: 4207200

3. Jensen RA. Enzyme Recruitment in Evolution of New Function. Annu Rev Microbiol. Annual Reviews; 1976; 30: 409–425. https://doi.org/10.1146/annurev.mi.30.100176.002205 PMID: 791073

4. Lazcano A, Díaz-Villagómez E, Mills T, Oró J. On the levels of enzymatic substrate specificity: Implications for the early evolution of metabolic pathways. Adv Sp Res. 1995; 15: 345–356. https://doi.org/10.1016/S0273-1177(99)80106-9

5. Alves R, Chaleil RA., Sternberg MJ. Evolution of Enzymes in Metabolism: A Network Perspective. J Mol Biol. 2002; 320: 751–770. https://doi.org/10.1016/S0022-2836(02)00546-6 PMID: 12095253

6. Light S, Kraulis P. Network analysis of metabolic enzyme evolution in Escherichia coli. BMC Bioinformatics. 2004; 5: 15. https://doi.org/10.1186/1471-2105-5-15 PMID: 15113413

7. Díaz-Mejía J, Pérez-Rueda E, Segovia L. A network perspective on the evolution of metabolism by gene duplication. Genome Biol. 2007/02/27. BioMed Central; 2007; 8: R26. https://doi.org/10.1186/gb-2007-8-2-r26 PMID: 17326820

8. Fani R, Fondi M. Origin and evolution of metabolic pathways. Phys Life Rev. 2009; 6: 23–52. https://doi.org/10.1016/j.plrev.2008.12.003 PMID: 20416849

9. Grassi L, Tramontano A. Horizontal and vertical growth of S. cerevisiae metabolic network. BMC Evol Biol. 2011; 11: 301. https://doi.org/10.1186/1471-2148-11-301 PMID: 21999464

10. Peretó J. Out of fuzzy chemistry: from prebiotic chemistry to metabolic networks. Chem Soc Rev. The Royal Society of Chemistry; 2012; 41: 5394. https://doi.org/10.1039/c2cs35054h PMID: 22508108

11. Vitkup D, Kharchenko P, Wagner A. Influence of metabolic network structure and function on enzyme evolution. Genome Biol. 2006/05/09. BioMed Central; 2006; 7: R39–R39. https://doi.org/10.1186/gb-2006-7-5-r39 PMID: 16684370

12. Greenberg AJ, Stockwell SR, Clark AG. Evolutionary constraint and adaptation in the metabolic network of Drosophila. Mol Biol Evol. 2008; 25: 2537–2546. https://doi.org/10.1093/molbev/msn205 PMID: 18799713

13. Hudson CM, Conant GC. Expression level, cellular compartment and metabolic network position all influence the average selective constraint on mammalian enzymes. BMC Evol Biol. BioMed Central Ltd; 2011; 11: 89. https://doi.org/10.1186/1471-2148-11-89 PMID: 21470417

14. Rausher MD, Miller RE, Tiffin P. Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. Mol Biol Evol. 1999; 16: 266–274. https://doi.org/10.1093/oxfordjournals.molbev.a026108 PMID: 10028292

15. Lu Y, Rausher MD. Evolutionary Rate Variation in Anthocyanin Pathway Genes. Mol Biol Evol. 2003; 20: 1844–1853. https://doi.org/10.1093/molbev/msg197 PMID: 12885963

16. Rausher MD, Lu Y, Meyer K. Variation in Constraint Versus Positive Selection as an Explanation for Evolutionary Rate Variation Among Anthocyanin Genes. J Mol Evol. 2008; 67: 137–144. https://doi.org/10.1007/s00239-008-9105-5 PMID: 18654810

17. Livingstone K, Anderson S. Patterns of Variation in the Evolution of Carotenoid Biosynthetic Pathway Enzymes of Higher Plants. J Hered. 2009; 100: 754–761. https://doi.org/10.1093/jhered/esp026 PMID: 19520763

18. Ramsay H, Rieseberg LH, Ritland K. The Correlation of Evolutionary Rate with Pathway Position in Plant Terpenoid Biosynthesis. Mol Biol Evol. 2009; 26: 1045–1053. https://doi.org/10.1093/molbev/msp021 PMID: 19188263

19. Montanucci L, Laayouni H, Dall'Olio GM, Bertranpetit J. Molecular Evolution and Network-Level Analysis of the N-Glycosylation Metabolic Pathway Across Primates. Mol Biol Evol. 2011; 28: 813–823. https://doi.org/10.1093/molbev/msq259 PMID: 20924085

20. Invergo BM, Montanucci L, Laayouni H, Bertranpetit J. A system-level, molecular evolutionary analysis of mammalian phototransduction. BMC Evol Biol. 2013; 13: 52. https://doi.org/10.1186/1471-2148-13-52 PMID: 23433342

21. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, et al. Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci. 2007; 104: 1777–1782. https://doi.org/10.1073/pnas.0610772104 PMID: 17267599

22. Ma H, Sorokin A, Mazein A, Selkov A, Selkov E, Demin O, et al. The Edinburgh human metabolic network reconstruction and its functional analysis. Mol Syst Biol. Nature Publishing Group; 2007; 3: 135. https://doi.org/10.1038/msb4100177 PMID: 17882155

23. Thiele I, Swainston N, Fleming RMT, Hoppe A, Sahoo S, Aurich MK, et al. A community-driven global reconstruction of human metabolism. Nat Biotechnol. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2013; 31: 419–425. https://doi.org/10.1038/nbt.2488 PMID: 23455439

24. Birney E. GeneWise and Genomewise. Genome Res. Cold Spring Harbor Laboratory Press; 2004; 14: 988–995. https://doi.org/10.1101/gr.1865504 PMID: 15123596

25. Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol. 2007; 24: 1586–1591. https://doi.org/10.1093/molbev/msm088 PMID: 17483113

26. Storey JD. A direct approach to false discovery rates. J R Stat Soc Ser B (Statistical Methodol. 2002; 64: 479–498. https://doi.org/10.1111/1467-9868.00346

27. Montañez R, Medina MA, Solé R V., Rodríguez-Caso C. When metabolism meets topology: Reconciling metabolite and reaction networks. BioEssays. 2010; 32: 246–256. https://doi.org/10.1002/bies.200900145 PMID: 20127701

28. Guimerà R, Nunes Amaral L a. Functional cartography of complex metabolic networks. Nature. 2005; 433: 895–900. https://doi.org/10.1038/nature03288 PMID: 15729348

29. Noor E, Eden E, Milo R, Alon U. Central Carbon Metabolism as a Minimal Biochemical Walk between Precursors for Biomass and Energy. Mol Cell. 2010; 39: 809–820. https://doi.org/10.1016/j.molcel.2010.08.031 PMID: 20832731

30. Caspi R, Dreher K, Karp PD. The challenge of constructing, classifying, and representing metabolic pathways. FEMS Microbiol Lett. 2013/06/27. 2013; 345: 85–93. https://doi.org/10.1111/1574-6968.12194 PMID: 23746312

31. Romero P, Wagg J, Green ML, Kaiser D, Krummenacker M, Karp PD. Computational prediction of human metabolic pathways from the complete human genome. Genome Biol. 2005; 6: R2. https://doi.org/10.1186/gb-2004-6-1-r2 PMID: 15642094

32. Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res. 2014; 42: D459–D471. https://doi.org/10.1093/nar/gkt1103 PMID: 24225315

33. Peregrin-Alvarez JM. The Phylogenetic Extent of Metabolic Enzymes and Pathways. Genome Res. Cold Spring Harbor Laboratory Press; 2003; 13: 422–427. https://doi.org/10.1101/gr.246903 PMID: 12618373

34. Nam H, Lewis NE, Lerman JA, Lee D-H, Chang RL, Kim D, et al. Network Context and Selection in the Evolution to Enzyme Specificity. Science (80-). 2012; 337: 1101–1104. https://doi.org/10.1126/science.1216861 PMID: 22936779

35. Bar-Even A, Noor E, Savir Y, Liebermeister W, Davidi D, Tawfik DS, et al. The Moderately Efficient Enzyme: Evolutionary and Physicochemical Trends Shaping Enzyme Parameters. Biochemistry. American Chemical Society; 2011; 50: 4402–4410. https://doi.org/10.1021/bi2002289 PMID: 21506553

36. Innan H, Kondrashov F. The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet. Nature Publishing Group; 2010; 11: 97–108. https://doi.org/10.1038/nrg2689 PMID: 20051986

37. Light S, Kraulis P, Elofsson A. Preferential attachment in the evolution of metabolic networks. BMC Genomics. 2005; 6: 159. https://doi.org/10.1186/1471-2164-6-159 PMID: 16281983