Article - Engineering, Technology and Techniques

The Relationship between Functional and Evolutionary Correlations of Enzyme Reactions in Metabolic Network Evolution

Hao Luo[1]
https://orcid.org/0000-0003-3145-7030

Tie Shen[2]
https://orcid.org/0000-0001-7313-3494

Xiaoyao Xie[2]*
https://orcid.org/0000-0003-4545-8480

[1]Guizhou University, College of Computer Science and Technology, Guiyang, Guizhou, China; [2]Guizhou Normal University, Key Laboratory of Information and Computing Science Guizhou Province, Guiyang, Guizhou, China.

*Correspondence: xyx@gznu.edu.cn; Tel.: +86-851-86702278

HIGHLIGHTS

- A method of computing the correlation connection network is first proposed.

- A concept of the mean functional concentration ratios of the enzyme reactions is proposed.

- The functional concentration ratio and the evolutionary correlation are related.

Abstract: The evolution of species is inevitably accompanied by the evolution of metabolic networks to adapt to different environments. The metabolic networks of different species were collected from the Kyoto Encyclopedia of Genes and Genomes (KEGG) website, and some enzyme reactions with the highest occurrence frequency in all species were found and are reported in this paper. The correlation coefficients of whether the enzyme reactions appear in all species were calculated, and the corresponding evolutionary correlation connection networks were calculated according to different correlation coefficient thresholds. These studies show that, as the evolutionary correlation of enzyme reactions increases, the weighted average of the mean functional concentration ratios of the enzyme reactions also increases, indicating that the functional concentration ratio of enzyme reactions has a certain correlation with the evolutionary correlation. The work presented in this paper enhances our understanding of the characteristics and general rules of metabolic network evolution.

INTRODUCTION

Metabolic networks [1-6] are the basis of energy conversion and material replacement, which are the basis of biological adaptation to the environment. In recent years, metabolic network alignment [7,8], robustness analyses of metabolic networks [9] and the determination of minimum metabolic networks[10] have become popular research topics. Additionally, metabolic networks are of great importance to the evolution and adaptability of various species. As species evolve and adapt to the environment, their metabolic networks must similarly evolve to adapt to different environments. Therefore, studying the evolution of metabolic networks is necessary for the analysis and understanding of species evolution and their adaptation to the environment.

By collecting the metabolic networks of each species, the pan-metabolic network [11] of these species was constructed, and the evolution of each enzyme reaction [12,13] was studied in this paper. We calculated the correlation connection network of the enzyme reactions in evolution and obtained the following results:

(1) The enzyme reactions with the highest occurrence frequency are fundamentally important.

(2) Certain enzyme reactions have a very high evolutionary correlation. The enzyme reactions with this phenomenon often have special metabolic functions and can adapt to specific environments, such that they can co-evolve.

(3) The correlation of enzyme reactions in evolution has a certain relationship with the functional concentration ratio of the enzyme reactions. The functional relevance is related to the adaptation of species to the environment, indicating that the adaptation of species to the environment affects the evolutionary process of the metabolic network. Considering the evolution of metabolic networks from the perspective of the network structure while ignoring the environment and function of enzyme reactions provides an incomplete understanding.

MATERIAL AND METHODS

Data source

The Kyoto Encyclopedia of Genes and Genomes (KEGG) database [14,15] was established in 1995 by the Kanehisa Laboratory of the Bioinformatics Center of Kyoto University, Japan. This database is among the most commonly used bioinformatics databases worldwide. In August 2018, we downloaded the Extensible Markup Language (XML) files of 5,492 organisms from the KEGG website to construct a pan-metabolic network.

Construction of the pan-metabolic network

In these 5,492 XML files, reactions in the same reaction tag may have different cofactors, and the corresponding reaction number is different. We considered the first reaction number in the reaction tag representative. Each reaction tag was found in the 5,492 XML files. Then, the reactions and metabolites in the reaction tag were extracted, and a pan-metabolic network was constructed, as shown in Figure 1.

Solving the weighted average of the mean functional concentration ratios of enzyme reactions

First, from the XML files of 5,492 organisms, all 1,898 reactions that are not repeated were extracted, and each reaction was assessed to determine whether it appears in each organism. If a reaction occurs in an organism, a value of one is assigned, and if a reaction does not appear in the organism, the assigned value is zero. By parity of reasoning, a 0-1 matrix of 1,898 rows and 5,492 columns was constructed.

Then, the reactions that occur in all species were found, and after programming, four reactions that occur in all species are found, as shown in the first four items in Table 1.

To facilitate the subsequent calculations, we removed the four reactions shown in Table 1 and obtained 1,894 reactions. In these reactions, the corresponding number of reactions is repeated. If the reaction corresponding to these repeated numbers is a reversible reaction, only the forward reaction is retained, and if it is not a reversible reaction, only one of the reactions is retained. Therefore, after such processing, 1,881 reactions were obtained, i.e., a 0-1 matrix of 1,881 rows and 5,492 columns was obtained.

We transposed the 0-1 matrix of 1,881 rows and 5,492 columns and calculated a correlation coefficient matrix A of 1,881×1,881. Let each element of the matrix be a real number d_{ij}, $1 \le i \le 1,881, 1 \le j \le 1,881$. The thresholds were set as 0.5, 0.7, 0.9, 0.95 and 0.98. Considering the case of a threshold of 0.5, if $d_{ij} \ge 0.5$,

$d_{ij} = 1$; if $d_{ij} < 0.5$, $d_{ij} = 0$, and i $\neq$ j. The diagonal element of the correlation coefficient matrix was set to zero because the correlation between the same enzyme reaction was not considered. Thus, matrix A was transformed into a 0-1 matrix B of 1,881×1,881.

Thus, if each reaction is regarded as a node of the undirected graph, matrix B is the adjacency matrix of the undirected graph. An undirected graph G is set to correspond to this adjacency matrix. According to the adjacency matrix of the undirected graph G, the connected components[16,17] of undirected graph G were obtained after the calculation.

The connected components of the undirected graphs, with thresholds of 0.5, 0.7, 0.9, 0.95 and 0.98, were calculated. The number of enzyme reactions in the connected components were sorted, and the first 16 connected components with a number of enzyme reactions greater than or equal to three and thresholds of 0.5 and 0.98 are shown in Figure 2 and Figure 3, respectively. Furthermore, the reaction numbers in the undirected graph connected components obtained according to different thresholds were extracted and saved.

First, we found the website of the KEGG database, and then we found the webpage with the URL https://www.kegg.jp/kegg/tool/map_pathway1.html. Taking the connected components with a number of enzyme reactions greater than or equal to three and a threshold of 0.5 as an example, the reaction numbers in one of the connected components were input into the list box on the webpage, and the corresponding search result was obtained. This result is a list of metabolic pathways, including metabolic pathways with specific functions.

The reactions in each connected component of the undirected graph were investigated one-by-one. Query results with fewer than two entries were not considered. Let the number of other reactions with specific functional metabolic pathways be R_i, except for "map01100 Metabolic pathways", according to the query result, $0 < i \leq (L_j - 1)$, $L_j \geq 2$. L_j is the total number of items queried, and $1 \leq j \leq C.C$ is the total number of connected components. The total number of reactions in each connected component of the undirected graph is T_j. The proportion of the reaction number R_i to the total number T_j is determined, i.e., the average value V_j of the functional concentration ratios. The calculation formula is shown in (1) as follows:

$$V_j = \frac{\sum_{i=1}^{L_j-1} R_i}{T_j(L_j - 1)} \tag{1}$$

Next, the weighted average of the functional concentration ratio average values of V_j in the enzyme reactions in the connected component was calculated and hereinafter referred to as the weighted average value W. The calculation formula is shown in (2) as follows:

$$W = \frac{\sum_{j=1}^{C} T_j \cdot V_j}{S} \tag{2}$$

In this equation, S represents the sum of the number of enzyme reactions in all connected components with a threshold of 0.5.

Similarly, the weighted average value was calculated according to the thresholds of 0.7, 0.9, 0.95 and 0.98. As shown in Figure 4(A), a line chart can be drawn by using the threshold as the abscissa and the weighted average value calculated according to each threshold as the ordinate.

Enzyme reactions with extremely high evolutionary correlations and functional correlations

During the process of solving the concentration ratio of the enzyme reaction functions, we obtained the 46 metabolic network graphs corresponding to 46 connected components with a number of enzyme reactions greater than or equal to three according to the threshold of 0.98, and the enzyme reactions in these metabolic networks have extremely high evolutionary correlations. Furthermore, we found that among the 46 connected components, the averages of the functional concentration ratios of the enzyme reactions in 16 connected components are one. Among the 16 connected components, we considered nine connected components and constructed the nine metabolic network graphs corresponding to these connected components, as shown in Figure 5.

RESULTS

Pan-metabolic network

Using the data obtained from the KEGG database, we removed the repeated reactions and metabolites and obtained 1,898 reactions and 1,692 metabolites. A 1,692×1,898 stoichiometric matrix of the pan-metabolite network was constructed by sequentially determining whether each metabolite was a substrate or a product in each reaction. The pan-metabolic network is shown in Figure 1.

The largest connected sub-metabolic network contains 1,741 reactions and 1,485 metabolites. There are also many separate sub-metabolic networks that are not connected to this main metabolic network, mainly because some metabolic networks are not completely reconstructed in the KEGG network. In a given species, there is a missing case of the corresponding metabolic network, resulting in partial metabolic network separation.
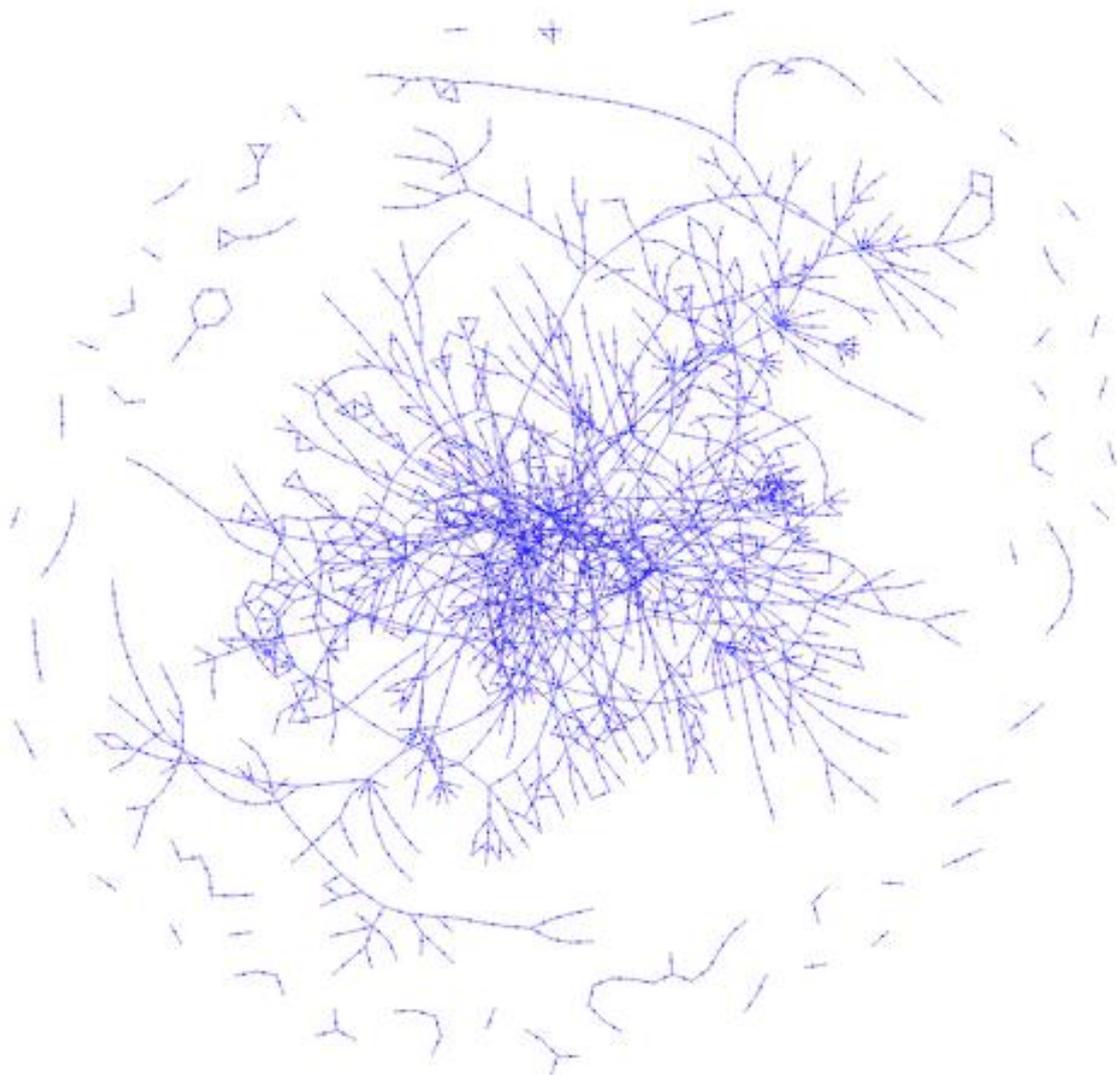
Figure 1. Pan-metabolic network

Conservation of enzyme reactions

The conservation of different enzyme reactions differs across evolutionary trajectories; thus, the frequency with which a given enzyme reaction occurs in different species also differs, and the conservation of these reactions can be understood by studying their occurrence frequency in species. We found that the occurrence frequency of some enzyme reactions, i.e., the number of species in which a given enzymatic reaction occurred, was more than 97%; these reactions include four enzyme reactions with reaction numbers R00377, R00378, R00375 and R00376. These reactions appear in all species shown in Table 1.

Table 1. Enzyme reactions with an occurrence frequency in all species greater than 97%.

Reaction number	Reaction name
R00377	Deoxycytidine triphosphate:DNA deoxynucleotidyltransferase (DNA-directed)
R00378	Deoxythymidinetriphosphate:DNA deoxynucleotidyltransferase (DNA-directed)
R00375	Deoxyadenosine 5'-triphosphate:DNA deoxynucleotidyltransferase (DNA-directed)
R00376	Deoxyguanosine 5'-triphosphate:DNA deoxynucleotidyltransferase (DNA-directed)
R00443	UTP:RNA uridylyltransferase
R00442	CTP:RNAcytidylyltransferase (DNA-directed); CTP:RNAcytidylyltransferase (RNA-directed)
R00435	ATP:polynucleotideadenylyltransferase; ATP:RNAadenylyltransferase
R00441	GTP:RNAguanylyltransferase (DNA-directed); GTP:RNAguanylyltransferase (RNA-directed)
R00124	ATP:ADP phosphotransferase
R01137	ATP:dADP phosphotransferase
R00330	ATP:GDP phosphotransferase
R01857	ATP:dGDP phosphotransferase
R00127	ATP:AMP phosphotransferase
R01547	ATP:dAMP phosphotransferase
R00616	ATP:thiamine-diphosphate phosphotransferase
R00158	ATP:UMP phosphotransferase
R00199	ATP:pyruvate,water phosphotransferase
R01061	D-glyceraldehyde-3-phosphate:NAD+ oxidoreductase (phosphorylating)
R01518	D-phosphoglycerate 2,3-phosphomutase
R01015	D-glyceraldehyde-3-phosphate aldose-ketose-isomerase
R00658	2-phospho-D-glycerate hydro-lyase (phosphoenolpyruvate-forming)
R00209	pyruvate:NAD+ 2-oxidoreductase (CoA-acetylating); pyruvate dehydrogenase complex
R01056	D-ribose-5-phosphate aldose-ketose-isomerase
R01512	ATP:3-phospho-D-glycerate 1-phosphotransferase
R00945	5,10-Methylenetetrahydrofolate:glycine hydroxymethyltransferase

The full names of the abbreviations in Table 1 are as follows:
Deoxyribonucleic acid (DNA), ribonucleic acid (RNA), uridine triphosphate (UTP), cytidine triphosphate (CTP), adenosine triphosphate (ATP), guanosine triphosphate (GTP), adenosine diphosphate (ADP), deoxyadenosine diphosphate (dADP), guanosine diphosphate (GDP), deoxyguanosine diphosphate (dGDP), adenosine monophosphate (AMP),deoxyadenosine monophosphate (dAMP), uridine monophosphate (UMP), nicotinamide adenine dinucleotide (NAD+), coenzyme A (CoA).

These high-frequency enzyme reactions occur so commonly because their functions are important and cannot be replaced. Most species need these reactions; thus, these enzyme reactions are conserved in evolution. These enzyme reactions are mainly related to DNA replication and extension, such as the deoxycytidine triphosphate:DNA deoxynucleotidyltransferase reaction, which plays a role in the bases infiltration of DNA chains; reactions related to core energy metabolism, especially transfer reactions of high-energy phosphate bonds between high-energy molecules, such as ATP: GDP phosphotransferase; reactions related to RNA formation, such as UTP: RNA uridylyltransferase; and conservative carbon metabolism reactions, such as D-glyceraldehyde-3-phosphate: NAD+ oxidoreductase.

The function of some enzyme reactions may be highly dependent on the environment to which they are adapted. Since specific strains are generally found in a specific environment, these enzyme reactions are redundant and useless in many species and are not preserved in evolution; thus, these reactions occur less frequently. For instance, the glutathione disulfide: 2,6-dichlorohydroquinone (chlorinating) reaction, which participates in the degradation of chlorocyclohexane and chlorobenzene, occurs in *shewanella* sp. WE21, *burkholderiaoklahomensis*C6786, and *sphingopyxis* sp. LPB0140, and its occurrence frequency is only 0.054625%. The violaxanthin-neoxanthin isomerase (epoxide-opening) reaction, which functions in the biosynthesis of carotenoids, occurs in *solanum lycopersicum* (tomato), *solanum pennellii*, and *solanum tuberosum* (potato), and its occurrence frequency is only 0.054625%. The beta-D-glucose:oxygen 1-oxidoreductase reaction is a part of the pentose phosphate pathway that occurs in *chondruscrispus*(carragheen) and *emiliania huxleyi*, and its occurrence frequency is only 0.036417%. The o-xylene, ferredoxin: oxygen oxidoreductase (methyl-hydroxylating) reaction, which participates in the degradation of xylene, occurs in *stenotrophomonas* sp. WZN-1, and its occurrence frequency is only 0.018208%.

Construction Results of the Enzyme Reaction Evolutionary Correlation Networks

We constructed a 0-1 matrix with species as the samples and the enzyme reactions in species as the variables. Using this 0-1 data set, we calculated the correlation coefficients between different enzyme reactions.

An evolutionary correlation network with a threshold of 0.5 was obtained by linking the enzyme reactions whose correlation coefficients are greater than or equal to 0.5 with undirected edges and expressing the enzyme reactions as nodes. The connected components with a number of nodes greater than or equal to three were obtained according to the evolutionary correlation network. After the calculation, there were 23 connected components, and 1,611 enzyme reactions were obtained.

By sorting the number of enzyme reactions in the connected components, the first 16 connected components were obtained, and the number of nodes in each connected component was greater than or equal to three as shown in Figure 2.
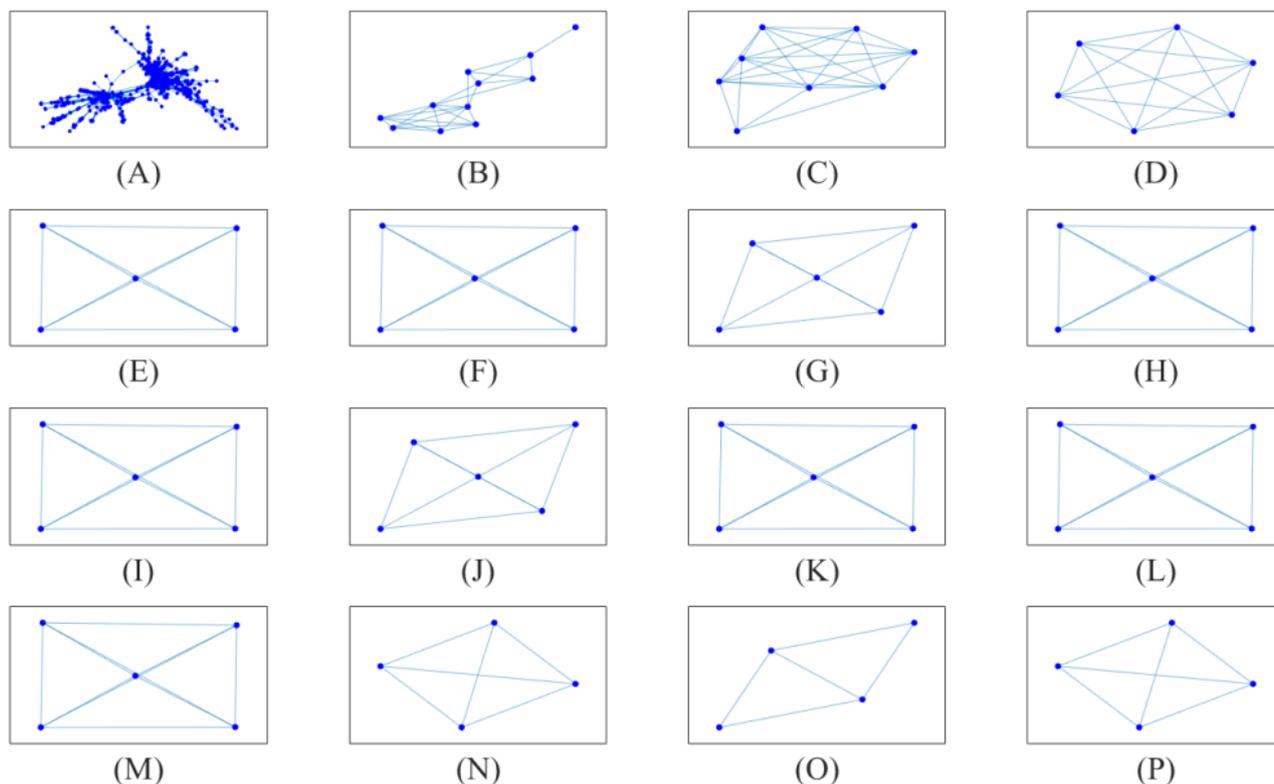
Figure2. (A) to (P), respectively, represent graphs of the first 16 connected components with the largest number of enzyme reactions at a threshold of 0.5.

According to the evolutionary correlation network with the corresponding threshold of 0.98, the connected components with a number of nodes greater than or equal to three were obtained. After the calculation, there were 46 connected components and 313 enzyme reactions. By sorting the number of enzyme reactions in the connected components, the first 16 connected components was obtained, and the number of nodes in each connected component was greater than or equal to three, as shown in Figure 3.
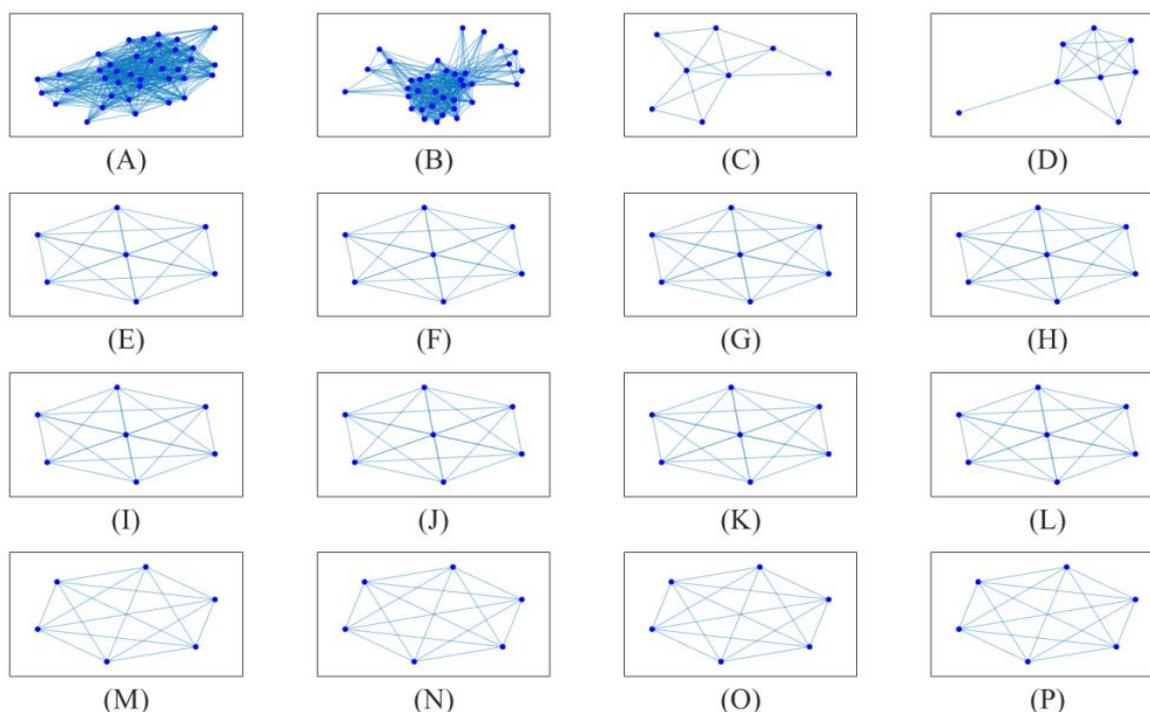
Figure 3. (A) to (P), respectively, represent graphs of the first 16 connected components with the largest number of enzyme reactions at a threshold of 0.98.

Such correlation coefficients actually reflect the distribution of enzyme reactions throughout the biological world and the evolutionary correlations of the enzyme reactions. There is coevolution between enzyme reactions with high correlation coefficients and independent evolution between enzyme reactions with low correlation coefficients. As the threshold increased, the total number of enzyme reactions in the connected component with a number of enzyme reactions greater than or equal to three decreased. Obviously, highly coevolutionary enzyme reactions are only a small part. In addition, after the threshold was raised, the previously large related network was divided into small networks with stronger correlations between different enzyme reactions.

There were cases in which the correlation coefficient between the enzyme reactions is negative, which is often due to different enzyme reactions adapting to opposite environments. A specific enzyme reaction occurs in one environment, while the other does not appear. However, the number of negative correlation coefficients with large absolute values is relatively small; thus, we did not construct the corresponding network.

There are cases in which the correlation coefficient between the enzyme reactions is almost zero, indicating that these enzyme reactions have completely independent evolutionary relationships, and there are many evolutionary relationships in this case.

The relationship between evolutionary correlation and functional correlation

According to the calculation method described in section 2.3, the relationship between the weighted average of the mean functional concentration ratios and evolutionary correlation based on each threshold can be obtained, as shown in Figure 4 (A).
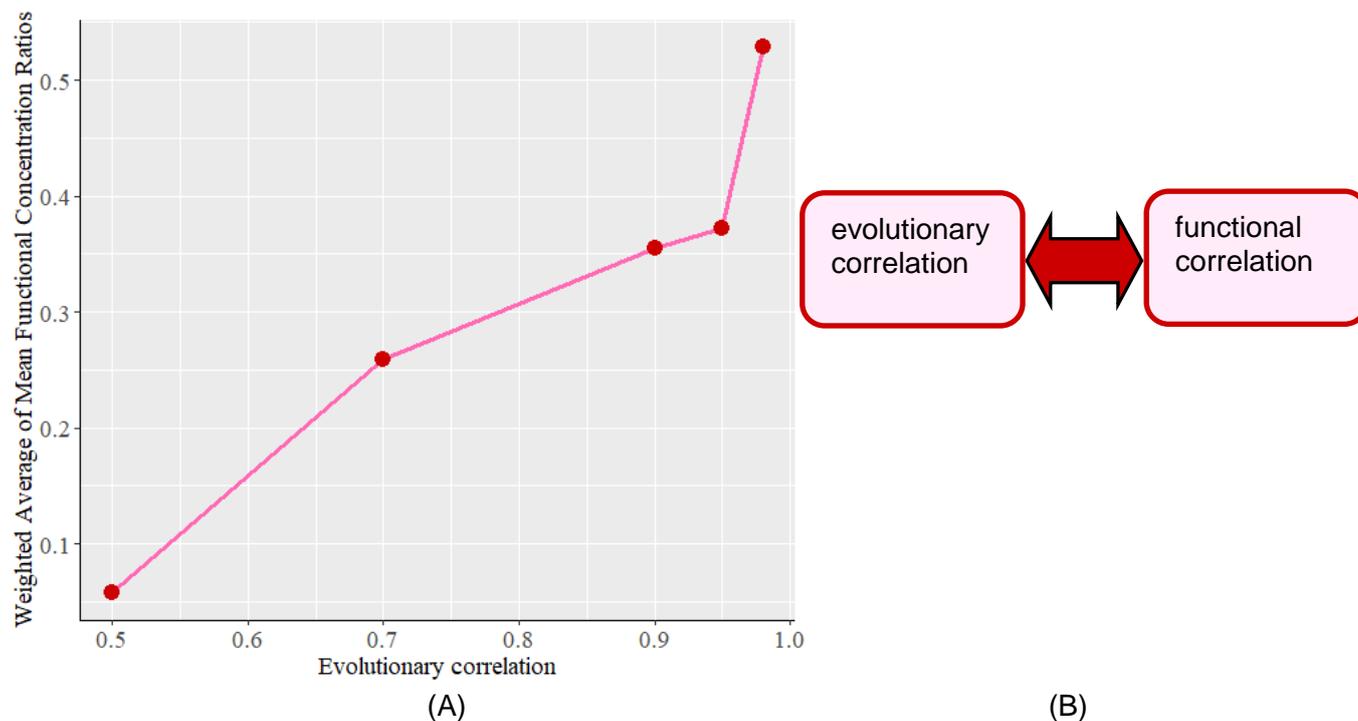
Figure 4. Line chart and schematic diagram of the relationship between evolutionary correlation and functional correlation. (A) Line chart; (B) Schematic diagram.

The functional correlation of the enzyme reactions increases with the evolutionary correlation of the enzyme reactions, as shown in Figure 4(A). In addition, the relationship between the evolutionary and functional correlations is shown in Figure 4(B).

As shown in Figure 4(B), evolutionary correlation and functional correlation are highly correlated. The extent to which metabolic functions adapt to the environment is an important driving factor of metabolic network evolution.

Visualization result of enzyme reactions with extremely high evolutionary correlations and functional correlations

During the process of solving the functional concentration ratio of the enzyme reactions, we identified some enzyme reactions with extremely high evolutionary correlations and functional correlations, and these enzyme reactions belong to a total of 16 connected components. Among the 16 connected components, we considered nine connected components and generated metabolic network graphs corresponding to these connected components, as shown in Figure 5.
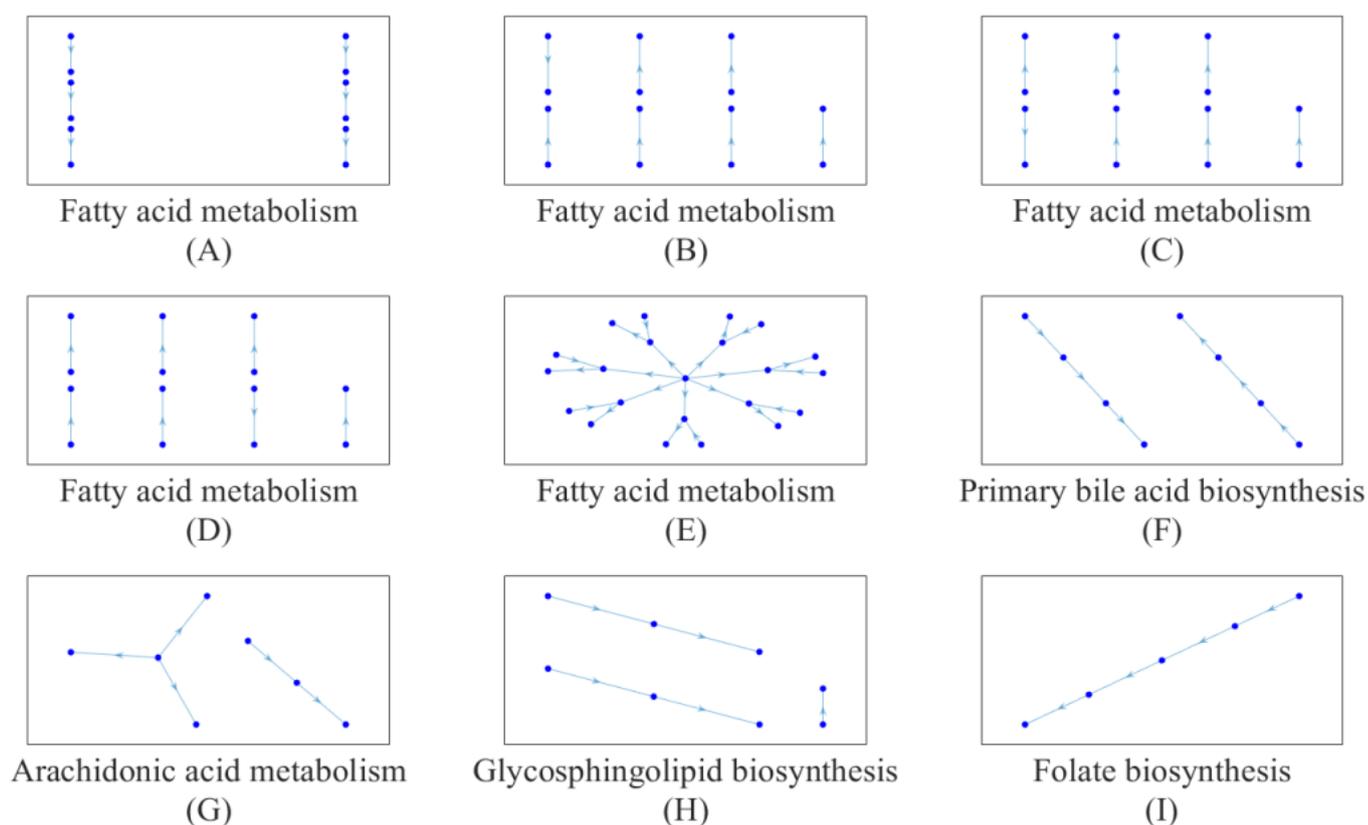
Figure 5. Metabolic network graphs corresponding to nine connected components containing enzyme reactions with extremely high evolutionary correlations and functional correlations. The (A) metabolic network to (E) metabolic network all belong to fatty acid metabolism. The (F) metabolic network belongs to primary bile acid biosynthesis. The (G) metabolic network belongs to arachidonic acid metabolism. The (H) metabolic network belongs to glycosphingolipid biosynthesis. The (I) metabolic network belongs to folate biosynthesis.

In Figure 5, the description below each panel indicates that all enzyme reactions in the panel belong to the metabolic pathways described in the corresponding description. For example, as shown in Figure 5(A), the enzyme reactions in the metabolic network belong to fatty acid metabolism.

DISCUSSION

As shown in Table 1, these enzymatic reactions occur very frequently in all species and are very important for the evolution of species. The occurrence frequency of each reaction among species also reflects the evolutionary importance of these enzyme reactions.

The evolutionary correlation network constructed by the evolution correlation coefficients changes as the threshold of the correlation coefficient changes. As the threshold increases, the largest connected component is divided into additional but smaller connected components due to the removal of some correlative edges. Thus, in the evolutionary correlation network at the threshold 0.5, there is a large component, while in the evolutionary correlation network at the threshold 0.98, this large component is dispersed into a series of smaller components that are responsible for the corresponding metabolic functions.

To study the relationship between evolutionary correlation and functional correlation, as shown in Figure 4(A), we constructed a line chart of the relationship between the evolutionary correlation and the weighted average of the mean functional concentration ratios. As shown in Figure 4(A), as the threshold increases, the weighted average of the mean functional concentration ratios of the enzyme reactions also gradually increases. Thus, as the evolutionary correlation of the enzyme reactions increases, their functional correlation also gradually increases, suggesting that evolutionary correlation and functional correlation are highly correlated.

Figure 5 shows metabolic network graphs corresponding to connected components containing some enzyme reactions with extremely high evolutionary correlations and functional correlations. These components were calculated according to the threshold of 0.98. Nine metabolic networks correspond to the connected components in the Figure. As shown in Figure 5(A), there are six enzyme reactions in the metabolic network. Based on the analysis, all six enzyme reactions are simultaneously involved in fatty acid

metabolism. As shown in Figure 5(I), again, there are four enzyme reactions in the metabolic network, all of which are simultaneously involved in folate synthesis. Because the functions of these enzyme reactions are specific to the adaptations required for a given environment, these reactions are highly correlated with each other. All enzyme reactions in each connected component belong to a specific functional pathway.

CONCLUSION

Through the research in this paper, the following results were obtained:

(1) The occurrence frequency of enzyme reactions in various species is partially due to the conservation of enzyme reactions during evolution. Obviously, this is related to the function of the enzyme reactions. The more basic the function of the enzyme reaction, the more highly conserved it is.

(2) The correlation in the evolution of enzyme reactions has a certain relationship with the functional concentration ratio of the enzyme reactions themselves. In this relationship, as the evolutionary correlation increases, the weighted average of the mean functional concentration ratios of the enzyme reactions also increases.

(3) The construction of the evolutionary correlation network tends to be decentralized as the threshold value increases.

Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

1. Yilmaz LS, Walhout AJ. Metabolic network modeling with model organisms. Curr Opin Chem Biol. 2017;36:32-9.
2. Wang H, Marcišauskas S, Sánchez BJ, Domenzain I, Hermansson D, Agren R, et al. RAVEN 2.0: A versatile toolbox for metabolic network reconstruction and a case study on Streptomyces coelicolor. PLoS Comput Biol. 2018;14(10):e1006541.
3. Hartleb D, Fritzemeier CJ, Lercher MJ. Automated high-quality reconstruction of metabolic networks from high-throughput data. bioRxiv. 2018:282251.
4. Ramon C, Gollub MG, Stelling J. Integrating -omics data into genome-scale metabolic network models: principles and challenges. Essays Biochem. 2018;62(4):563-74.
5. Yu G, Rui B, Fan X, Xie X. Visual FTBL Generator: Visual Generation Tool for 13C-FLUX File. Braz Arch Biol Technol. 2019;62:e19180071.
6. Shen F, Sun R, Yao J, Li J, Liu Q, Price ND, et al. OptRAM: In-silico strain design via integrative regulatory-metabolic network modeling. PLoS Comput Biol. 2019;15(3):e1006835.
7. Zhou W, Liu J, Xing H, Zhang Z, Xie X, Jing F. Comparisons Between Different Methods in Measuring Enzyme Similarity for Metabolic Network Alignment. Braz Arch Biol Technol. 2016;59:e16150429.
8. Shen T, Zhang Z, Chen Z, Gu D, Liang S, Xu Y, et al. A genome-scale metabolic network alignment method within a hypergraph-based framework using a rotational tensor-vector product. Sci Rep. 2018;8(1):16376.
9. Jiang D, Zhou S, Chen YP. Compensatory ability to null mutation in metabolic networks. Biotechnol Bioeng. 2009;103(2):361-9.
10. Jiang D, Zhou S, Liu H, Chen YP. Inferring Minimal Feasible Metabolic Networks of *Escherichia coli.* Appl Biochem Biotechnol. 2010;160(1):222-31.
11. Raymond J, SegrèD. The Effect of Oxygen on Biochemical Networks and the Evolution of Complex Life. Science. 2006;311(5768):1764-7.
12. Hold C, Billerbeck S, Panke S. Forward design of a complex enzyme cascade reaction. Nat Commun. 2016;7:12971.
13. Ribeiro AJM, Holliday GL, Furnham N, Tyzack JD, Ferris K, Thornton JM. Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. Nucleic Acids Res. 2018;46(D1):D618-23.
14. Kanehisa M, Goto S. KEGG: Kyoto Encyclopaedia of Genes and Genomes. Nucleic Acids Res. 2000;28(1):27-30.

15. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. 2014;42(Database issue):D199-205.
16. Kryven I, Duivenvoorden J, Hermans J, Iedema PD. Random Graph Approach to Multifunctional Molecular Networks. Macromol Theor Simul. 2016;25(5):449-65.
17. Passat N, Naegel B, Kurtz C. Component-Graph Construction. J Math Imaging Vis. 2019;61(6):798-823.