Application of Graph Theory to Evaluate Chemical Reactions in Cells

Sachiyo Aburatani¹, Yuichi Kokabu², Ryota Teshima², Teppei Ogawa², Michihiro Araki³, Tomokazu Shirai⁴

¹ Computational Bio Big Data Open Innovation Lab. (CBBD-OIL), National Institute of Advanced Industrial Science and Technology (AIST), AIST Tokyo Waterfront Main Bldg., 2-3-26, Aomi, Koto-ku, Tokyo 135-0064, Japan
² Bioscience Department, Mitsui Knowledge Industry Co. Ltd., Atago Green Hills MORI Tower, 2-5-1 Atago, Minato-ku, Tokyo 105-6215, Japan
³ Graduate School of Medicine, Kyoto University, 54 Kawahara-cho, Shogoin, Sakyoku, Kyoto 606-8507, Japan
⁴ Center for Sustainable Resource Science, RIKEN, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan

E-mail: s.aburatani@aist.go.jp and tomokazu.shirai@riken.jp

Abstract. Chemical reactions occur in cells for survival and adaptation to various conditions. After these chemical reactions, the reactants and products are often sequentially modified through metabolic pathways. In this study, we defined new features to evaluate the possibility of such inferred metabolic pathways. We focused on the main chain structure of a compound as a non-directional graph, and developed a method to define the similarity between these main chain structure graphs. In this study, we defined four features: 1) the number of main chain graph nodes, 2) the graphical density of the main chain graph, 3) the chemical density of the main chain, and 4) the graph centrality of the reaction group in the main chain graph. We defined the main chain structures of about 16,000 chemical compounds, and calculated the values of the four features by the defined equations for each compound. Finally, we calculated the correlation coefficients between all chemical compound pairs from the four defined features. A comparison of the similarities of the main chain graphs between known chemical reactions revealed that our defined features are suitable for detecting the possible reactions.

1. Introduction

Synthetic biology has recently entered a new era for designing new biological systems to obtain useful biological functions, by the combination of several types of scientific fields. The development of new empirical techniques and computational approaches for the design, analysis, and construction in the synthesis of a novel compound is currently an important theme in synthetic biology [1,2]. From a biological viewpoint, designing and analyzing new biological systems by computational approaches and constructing and evaluating the designed systems by empirical techniques are an efficient strategy. For the design and analysis of a new biological system, several types of computational or mathematical approaches have been applied. In previous investigations, the M-path and BioProV programs were developed to infer new biosynthetic pathways [3], and those approaches are considered...
to be suitable for establishing novel and efficient biosynthetic systems for generating compounds in microorganisms.

Designing a new biosynthetic pathway for the construction of novel compounds in a microorganism is a problem-solving task, in which the optimal enzyme reactions should be selected. Biosynthetic pathways from the starting material to the target molecule usually include over ten sequential reaction steps, and thus a computational approach is required to identify the possible and optimal reactions in the design step.

To improve the accuracy of computational approaches for designing novel biosynthetic pathways, new evaluation indices for the chemical reactions that can occur in cells are useful. In this study, we developed a new index combined with graph theory, to evaluate the chemical reactions in cells. We feature the structural formula of the compound as an undirected graph. The compounds in cells can be defined as simple or multiple graphs, and we can calculate the similarity of the structures between these graphs. We utilized these graph similarities to evaluate the possible enzymic reactions in cells.

2. Definition

2.1. Data Processing
For the evaluation, the structural formulae of all compounds within the metabolic pathways were compiled from the KEGG database (https://www.genome.jp/kegg/) [4]. We accumulated 16,042 compounds, with their names, structural formulae, molecular formulae, and structural shifts. In cells, the chemical reactions involving the conversion of one compound to another compound are sequentially executed, and catalyzed by specific enzymes. To design a new biosynthetic pathway, the conversion between the two compounds must be evaluated.

2.2. Definition of the Main Chain
We focused on the main chain structures of compounds to detect the similarity between two compounds that are connected by enzymic reactions. Generally, the International Union of Pure and Applied Chemistry (IUPAC) nomenclature for organic chemistry is considered to be the best systematic method for naming compounds [5]. Since some compounds have very complicated structures, in terms of a graphical viewpoint, the main chain structures defined by the IUPAC nomenclature methods were not suitable for this analysis. Thus, we newly defined the main structures of all compounds as follows:

a) The max carbon chain of the compound was defined as the main chain without an aromatic or other ring, except for b) and c).
b) If the number of max carbon chain that bonds with an aromatic ring was less than 1, then the aromatic ring of the compound was defined as the main chain.
c) If a non-aromatic ring has more carbon molecules than the other carbon chain(s) within the compound, then the non-aromatic ring was defined as the main chain.
d) For a polycyclic compound, the polycyclic structure was defined as one large aromatic or other type of ring.
e) If at least one aromatic ring was included in the polycyclic structure, then the polycyclic structure was defined as one aromatic ring.

2.3. Definition of Structural Features
To evaluate the similarity between the structures of the compounds, we defined several types of features for the defined main chain, as a non-directional graph. From the graph structure viewpoints, we shed light on the number of nodes within the main chain and the density of the graph structure of the main chain. Furthermore, we defined the new indices: the bond density of the main chain, and the point centrality for the carbon molecule that bonds to the functional group of the compound.
2.3.1. Graph Structural Index

The number of nodes is an important feature of the graph. We defined “Length of Main Chain” as the index from the number of carbon molecules within the main chain, as follows:

\[ L_i = N_i \]  

Here, \( N_i \) is the number of carbon molecules in the defined main chain.

One graph was composed of some nodes and some edges connected between nodes. Generally, the graph constructed by edges connected between all node pairs is considered as the complete model. In contrast, the graph with no edges between any node pairs is called an independent model. The filling rate of the existing edges by all possible edges can be defined as the “Graph Density”, expressed by

\[ D_i = \frac{2m}{N_i(N_i - 1)} \]  

where \( m \) is the number of connections between node pairs within the main chain. In this case, a double bond between two carbon molecules was counted as one connection. Since a double bond between a node pair was not considered as two edges, but just one edge between the nodes in the graph structure, we defined the index “Graph Density” by the presence or absence of connections between two nodes.

The node features in the graph were defined as the point centrality. We utilized several types of point centralities: “Closeness Centrality”, “Degree Centrality”, and “Eigenvector Centrality” [6,7,8,9]. These point centralities were calculated for each carbon molecule within the main chain. In graph theory, the total amount of the shortest distances from one node to the other nodes is referred to as the status [10]. The normalized Closeness Centrality of one node was defined by the inverse of the node’s status, as follows:

\[ \text{norm } C_c(i) = \frac{N - 1}{\sum_{j=1}^{N} d_{ij}} \]  

Here, \( d_{ij} \) is the shortest path from node \( i \) to node \( j \), and \( N \) is the number of carbon molecules within the main chain graph. Thus, \( \sum_{j=1}^{N} d_{ij} \) means the status of node \( i \). Since the Closeness Centrality depends on the number of nodes within a graph, it was normalized by \( N-1 \).

In the graph structure, a node connected with other nodes can be considered as a key role node in the graph. To evaluate the degree of the key role for each node, we utilized the normalized Degree Centrality as follows:

\[ \text{norm } C_d(i) = \frac{\sum_{j=1}^{N} a_{ij}}{N - 1} \]  

where \( a_{ij} \) is the existence of a connection between node \( i \) and node \( j \). If a connection exists between nodes \( i \) and \( j \), then the value of \( a_{ij} \) is one. If there is no connection between node \( i \) and node \( j \), then the value of \( a_{ij} \) is zero. In this case, the type of connection was not considered, and only the existence or absence of a connection was examined. If multiple connections exist between node \( i \) and node \( j \), then the value of \( a_{ij} \) is also defined as one.

To evaluate the connected nodes for each node, we utilized the Eigenvector Centrality. The connections with the other nodes were considered to have the same weight in both the Closeness Centrality and Degree Centrality. In the Eigenvector Centrality, the connection with a node that has many connections with other nodes was considered to be important, and added higher weight. The connection with a node that rarely connected with other nodes had lower weight added. The Eigenvector Centrality was defined as follows:
\[ C_r(i) = \frac{1}{\lambda} \sum_{j=1}^{N} a_{ij} C_r(j) \]  
\[ G(C_c) = \left[ \sum_{i=1}^{N} \left[ \max \left( \text{norm}_{-}C_c - \text{norm}_{-}C_c(i) \right) \right] \right] \left( \frac{N^2 - 3N + 2}{2N - 3} \right)^{-1} \]  
\[ G(C_d) = \sum_{i=1}^{N} \left[ \max \left( \text{norm}_{-}C_d - \text{norm}_{-}C_d(i) \right) \right] \frac{N^2 - 3N + 2}{N^2 - 3N + 2} \]  

where \( a_{ij} \) is the same as (4), and \( \lambda \) is the maximum eigen value for matrix \( A \). Matrix \( A \) is an \( N \times N \) matrix representing the connections between nodes.

To evaluate the graph centralization, we utilized the Graph Centrality, given by

\[ I_I m = \frac{2b_N}{4N} \]  

Here, \( G(C_c) \) is the Graph Centrality corresponding to the Closeness Centrality, and \( G(C_d) \) is the Degree Centrality.

### 2.3.2 Chemical Index

The Chemical Index is defined by the chemical features of the main chain of the compound. We defined a new index, “Bond Density”, given by

\[ B_i = \frac{2b_N}{4N} \]  

Here, \( b_N \) is the number of bonds between the carbon molecule pairs within the main chain. Simply, a single bond between two carbon molecules is defined as \( b_N = 1 \), and a double bond between carbon molecules is defined as \( b_N = 2 \). In our definition of the main chain, some aromatic rings and non-aromatic rings were defined as the main chain of the compound. In this case, we counted the number of bonds between the carbon molecule pairs and divided it by the number of carbon molecules. In an aromatic ring, \( b_N \) was defined as 1.5. The coefficient of the denominator in equation (7) is the valence electron number of a carbon molecule. Since two valence electrons are shared by two neighboring carbon molecules to form a covalent bond between the two carbon molecules, the coefficient of the numerator was set to 2.

### 3. Calculation

#### 3.1. Similarity by simple indices

To evaluate the graph similarities between all compound pairs, we calculated the standard scores for the indices, given by

\[ z(I_x) = \frac{I_x - \left( \sum_{c=1}^{m} I_x^c \right) m^{-1} \left( \sum_{c=1}^{m} \left( I_x^c \right)^2 - \left( \sum_{c=1}^{m} I_x^c \right) m^{-1} \right) m^{-1}}{m^{-1}} \]  

where \( I_x \) represents a defined index \( x \), \( I_x^c \) is the value of index \( x \) for compound \( c \), and \( m \) is the number of compounds. All index values were calculated by their standard scores. First, we calculated the Pearson’s correlation from only the simple indices, Length of Main Chain, Graph Density, and Bond Density. The correlation matrix between the compounds is displayed in Figure 1. The correlation matrix between all 16,042 compounds is displayed in Figure 1a, and the correlation between only 200 compounds is displayed in Figure 1b. The order of the compounds corresponds to their IDs in the KEGG Database. Since the absolute values of the correlations between compounds are very high in Figure 1, it was too difficult to discern the graph similarities from only the simple indices.
3.2. Structural similarity by defined indices

In cells, enzymic reactions are sequentially executed along metabolic pathways. To design novel enzymic reactions from a compound structure, the structural changes of compounds by known enzymic reactions should be detected as high similarity. Meanwhile, unknown structural changes of compounds should be detected as low similarity. In Figure 1, the similarities calculated by the information about the components for the main chain graph could not distinguish between the known/unknown structural changes. To increase the sensitivity of detecting possible structural changes of the components, we calculated the structural similarities from all of the indices that were defined in this study.

We selected a typical enzymic reaction to evaluate our defined indices, from alpha-D-glucose to D-glyceraldehyde 3-phosphate and dihydroxyacetone phosphate. This pathway is categorized as glycolysis, and beta-D-fructose 1,6-bisphosphate is decomposed to D-glyceraldehyde 3-phosphate and dihydroxyacetone phosphate. The first four compounds, alpha-D-glucose, alpha-D-glucose 6-phosphate, beta-D-fructose 6-phosphate, and beta-D-fructose 1,6-bisphosphate, do not have aromatic rings as the main chain. In contrast, in the last two compounds, the ring is opening in dihydroxyacetone phosphate and D-glyceraldehyde 3-phosphate. The structural similarity calculated by the simple indices is displayed in Figure 2a. In Figure 2a, the main chain structures with a ring or without any rings were well divided. However, the similarity between alpha-D-glucose and beta-D-fructose 1,6-bisphosphate was also high, even though these two compounds are not connected in the metabolic pathway.
Figure 2. Similarity matrices corresponding to metabolic pathways. a) Matrix calculated by only simple indices. b) Matrix calculated by defined indices. The similarity score between beta-D-fructose 6-phosphate and beta-D-fructose 1,6-bisphosphate was 0.23.

The similarity matrix between these six compounds, calculated from the simple indices and our defined indices, is displayed in Figure 2b. In this figure, the structural similarities between the six compounds are well defined. The structure of alpha-D-glucose was similar to the structures of alpha-D-glucose 6-phosphate and beta-D-fructose 6-phosphate, but was not similar to the structure of beta-D-fructose 1,6-bisphosphate. The compounds with ring structures were well divided from the compounds without ring structures. Almost all neighboring compound pairs had high similarities, except for beta-D-fructose 6-phosphate and beta-D-fructose 1,6-bisphosphate. This figure shows that the structure of beta-D-fructose 1,6-bisphosphate is different from those of all of the other compounds.

4. Conclusion
In cells, structural shifts of compounds occur by enzymic reactions along metabolic pathways. The similarity scores calculated from our defined indices were well reflected by the distances on the metabolic pathways. Thus, our newly defined indices based on graph theory are considered to be suitable for detecting possible structural shifts in cells. To calculate the similarity scores between all compound pairs, the reactive carbon molecules connected with functional groups must be defined for all compounds. Furthermore, the distance between two compounds on the metabolic pathway must be calculated. We are in the process of detecting the reactive carbon molecules for each compound, and will be able to calculate the similarity scores for all compound pairs in the near future.

Acknowledgements
This study was supported by funding from New Energy and Industrial Technology Development Organization.

References
[1] Noda S, Shirai T, Oyama S and Kondo A 2016 Metab. Eng. 33 119
[2] Noda S, Shirai T, Mori Y, Oyama S and Kondo A 2017 Nat. Commun. 8 1153
[3] Araki M, Cox RS 3rd, Makiguchi H, Ogawa T, Taniguchi T, Miyaoku K, Nakatsui M, Hara K Y and Kondo A 2015 Bioinformatics 31(6) 905
[4] Hattori M, Okuno Y, Goto S and Kanehisa M 2003 J. Am. Chem. Soc. 125(39) 11853
[5] Panico R, Richer J C and Powell W H 1994 A Guide to IUPAC Nomenclature of Organic Compounds (IUPAC Chemical Data) (New Jersey: Wiley-Blackwell)
[6] Hage P and Harary F 1995 Soc. Netw. 17(1) 57
[7] Harary F 1969 Graph Theory (Boston: Addison-Wesley)
[8] Freeman L C 1979 Soc. Netw. 1(3) 215
[9] Koschützki D 2008 Network Centralities. Analysis of Biological Networks ed. Junker B H and Schreiber F (New Jersey: John Wiley & Sons) pp. 65-84
[10] Harary F 1959 Sociometry 22(1) 23