Prediction of compounds' biological function (metabolic pathway) based on compound similarity

Lei Chen[1], Ziliang Qian[2], Yudong Cai[3,*], Zhenbing Zeng[1], Wencong Lu[4]

[1]Shanghai Key Laboratory of Trustworthy Computing, East China Normal University, Shanghai, 200062, People's Republic of China;
[2]Department of Combinatorics and Geometry CAS-MPG Partner Institute for Computational Biology, Shanghai Institute for Biological Sciences, Chinese Academy of Sciences, Shanghai, 200031, People's Republic of China;
[3]Institute of systems biology, Shanghai University, Shanghai 200444, People's Republic of China;
[4]School of Materials Science and Engineering, Shanghai University, Shanghai, 200072, People's Republic of China.

Abstract. Correctly and efficiently mapping small molecules with great biological significance into the corresponding metabolic pathway is helpful both in basic researches on metabolic pathways and in drug discovery. In this research, we present a new method to the analysis of metabolic pathways. This method was divided into two steps: (1) Map small chemical molecules into 11 major metabolic pathway classes; (2) in each major pathway class, small chemical molecules are classified into specific pathways. 3836 compounds are selected for study. Both in step one and two, our method can provide a prediction sequence for each compound, indicating that our method can tackle the multi-class case. In details, we used compound similarity, a measurement to indicate how similar two compounds are, to make prediction. As a result, we obtained acceptable prediction rates both in step one and two, indicating that compound similarity can be used to classify small molecules efficiently.

Keywords: Compound, Metabolic pathway, Compound similarity, Jackknife cross-validation test

1. Introduction

It is a fundamental problem to understand the relationships between human gene and disease in the post genomic era. In order to understand the disease process clearer, many "omics" sciences, including genomics, transcriptomics, proteomics and metabonomics, have been used to provide helpful biological information. Among all "omics" sciences, some only investigate a biological system in some independent levels [1, 2]. However, metabonomics is much different, which aims to investigate a biological system as a whole. In details, metabonomics deals with the profiles of metabolites of integrated living systems [3]. Since living system is very complicated, especially within high complex living system such as humans [4], it is very difficult to release information about interactions between different components. However, analyzing metabolites is regarded as an efficient way to deal with this

* To whom correspondence may be addressed. Tel: 0086-21-66136132; Fax: 0086-21-66136109;
E-mail: cai_yud@yahoo.com.cn (Y.-D. Cai).

problem.

Metabolomics is an essential part of the metabonomics analysis, which focuses on all the metabolites with low molecular mass [2]. Hence, in recent years, small molecule has become one of the top stars in metabolic pathway analysis [5]. It is known that there are 11 major metabolic pathway classes [6] and in each major pathway class, it contains some specific metabolic pathways (see http://www.genome.jp/kegg/pathway.html). A metabolic pathway consists of a series of coupled, interconnecting chemical reactions. In recent years, researchers have made lots of effort to analyze the role of small molecule in metabolic pathways [7]. However, most of them used biochemical or physical experiments, which lead to the problem the speed of discovering of new small molecules is much faster than that of annotation, because more and more high-throughput equipment are being applied. Thus it is very important to correctly and efficiently to map small molecules with great biological significance into the corresponding metabolic pathway. Fortunately, efficient in silico approaches may provide useful hint, which is fast, automated and meets the need of high-throughput data processing.

Recently, Cai gave a machine learning method to map small molecules into the metabolic pathway based on functional group composition [8]. However, their method only map small molecules into 11 major pathway classes not the specific pathway and they can only tackle the mono-class case, i.e., a given small molecule is assumed to belong to one, and only one, pathway class. In addition, they can not discriminate a pair of isomeride, because they have the same functional group composition, i.e., they are same after they are converted into numeric vectors.

In this paper, we present a new prediction method based on compound similarity which can overcome the shortcomings in [8]. This method was divided into two steps: (1) Map small chemical molecules into 11 major metabolic pathway classes; (2) in each major pathway class, small chemical molecules are classified into specific pathways. 3836 compounds are employed in this study, which were collected from public available database KEGG compound (ftp://ftp.genome.jp/pub/kegg/pathway/map/). Among these compounds, some belong to more than one pathway class or pathway. We used compound similarity to obtain a prediction sequence for each compound in both step one and two. Thus our method can tackle the multi-class case. The compound similarity we used, which is a measurement to indicate how similar two compounds are, is proposed in [9] and can be obtained from http://www.genome.jp/ligand-bin/search_compound or http://pcal.biosino.org/workflowgallery/Search_KEGG_Compound/Search_KEGG_Compound.output.xml.html. Since the compound similarity is obtained by comparing two chemical structures of two compounds and a pair of isomeride have the different chemical structure, we deal with a pair of isomeride as two different compounds. As a result, we obtained acceptable prediction rates both in step one and two, which indicates that we can use compound similarity to classify small molecules efficiently.

2. Materials and methods

2.1 Dataset

The small chemical molecules were collected from public available database KEGG compound (ftp://ftp.genome.jp/pub/kegg/pathway/map/). After deleting compounds with no information about their similarity, 3836 compounds were obtained for our research dataset. According to the website http://www.genome.jp/kegg/pathway.html, there are 11 major metabolic pathway classes listed below:

1. Carbohydrate Metabolism
2. Energy Metabolism
3. Lipid Metabolism
4. Nucleotide Metabolism
5. Amino Acid Metabolism
6. Metabolism of Other Amino Acids
7. Glycan Biosynthesis and Metabolism
8. Biosynthesis of Polyketides and Nonribosomal Peptides
9. Metabolism of Cofactors and Vitamins
10. Biosynthesis of Secondary Metabolites
11. Xenobiotics Biodegradation and Metabolism

In each major pathway class, it contains some specific pathways (please see http://www.genome.jp/kegg/pathway.html). The number of pathways we studied in each major pathway class is listed in Table 1. There are some compounds and reactions that participate in each pathway. The distribution of 3836 compounds into 148 pathways can be found in supplement materials I.

Table 1: Number of pathways in each major pathway class

Major metabolic pathway class	**Number of pathways**
Carbohydrate Metabolism	17
Energy Metabolism	7
Lipid Metabolism	16
Nucleotide Metabolism	2
Amino Acid Metabolism	16
Metabolism of Other Amino Acids	9
Glycan Biosynthesis and Metabolism	8
Biosynthesis of Polyketides and Nonribosomal Peptides	9
Metabolism of Cofactors and Vitamins	12
Biosynthesis of Secondary Metabolites	26
Xenobiotics Biodegradation and Metabolism	26
Total pathways	148

Table 2 shows the distribution of 3836 compounds into 11 major metabolic pathway classes. Since some compounds belong to more than one pathway class, the total number of different compounds Ω is smaller than the total number of classified compounds $\tilde{\Omega}$. The relationship between these two is given by the following equation:

$$\tilde{\Omega} = \Omega + \sum_{n=2}^{\lambda} (n-1)\beta_n \qquad (1)$$

where λ is the number of total categories and β_n is the number of compounds belonging to exactly n categories. For instance, of the 3836 compounds in Table 2, 3503 $(=\beta_1)$ belong to exactly one

pathway class, 242 $(=\beta_2)$ belong to exactly 2 pathway classes, 52 $(=\beta_3)$ belong to exactly 3 pathway classes, 24 $(=\beta_4)$ belong to exactly 4 pathway classes, 10 $(=\beta_5)$ belong to exactly 5 pathway classes, 2 $(=\beta_6)$ belong to exactly 6 pathway classes, 2 $(=\beta_7)$ belong to exactly 7 pathway classes, 1 $(=\beta_8)$ belongs to exactly 8 pathway classes, and 0 belongs to exactly $n\ (=9,10,11)$ pathway classes. Substituting these numbers into Eq. (1), we have:

$$\tilde{\Omega} = 3836 + (2-1)\times 242 + (3-1)\times 52 + (4-1)\times 24 + (5-1)\times 10 + (6-1)\times 2 + (7-1)\times 2 + (8-1)\times 1 = 4323 \quad (2)$$

Table 2: The distribution of compounds into 11 major metabolic pathway classes

Major class of metabolic pathway	Number of compounds
Carbohydrate Metabolism	449
Energy Metabolism	110
Lipid Metabolism	555
Nucleotide Metabolism	144
Amino Acid Metabolism	567
Metabolism of Other Amino Acids	175
Glycan Biosynthesis and Metabolism	66
Biosynthesis of Polyketides and Nonribosomal Peptides	280
Metabolism of Cofactors and Vitamins	334
Biosynthesis of Secondary Metabolites	891
Xenobiotics Biodegradation and Metabolism	752
Total number of classified compounds $\tilde{\Omega}$	4323
Total number of different compounds Ω	3836

2.2 Compound Similarity

In [9], authors presented an efficient algorithm (SIMCOMP) to compare two chemical structures of compounds. Owing to the fact that chemical structure is a two-dimensional (2D) object, we can use a graph consisting of vertices (atoms) and edges (bonds) to represent this structure. The algorithm is based on detecting common subgraphs in two graphs. In this section, we introduce this algorithm briefly. For details, please see [9].

1. Let G_1 and G_2 be two graphs, which represent two compounds. For every vertex v in G_1 and G_2, there are two labels $p_1(v)$ and $p_2(v)$, where $p_1(v)$ and $p_2(v)$ represent the atom types and atom species of corresponding atom, respectively.

2. Construct the Association Graph $AG(V,E)$ of $G_1=(V_1,E_1)$ and $G_2=(V_2,E_2)$, where $V=V_1\otimes V_2$ and $(v_i,u_j)(v_s,u_t)\in E,$ if and only if (1) $v_iv_s\in E_1$ and $u_ju_t\in E_2$ or (2) $v_iv_s\notin E_1$ and $u_ju_t\notin E_2$. It is obvious that a common subgraph in G_1 and G_2 corresponds to a clique in AG. According to two labels on each vertex in G_1 and G_2, the weight of (v_i,u_j) is defined as:

$$w((v_i,u_j))=\begin{cases}1, \ if \ \ p_1(v_i)=p_1(u_j),\\ c, \ if \ \ p_1(v_i)\neq p_1(u_j) \ \ and \ \ p_2(v_i)=p_2(u_j),\\ 0, \ otherwise\end{cases} \tag{3}$$

where c is a constant between 0 and 1.

3. Search clique in AG based on Bron-Kerbosch algorithm [10]. The procedure outlined as follows: (1) stop the Bron-Kerbosch algorithm at the number of recursion steps r; (2) eliminate some simple connected common subgraphs whose cardinality is smaller than c; (3) extend the other simple connected common subgraphs greedily. As a result, a set of simple connected common subgraphs in G_1 and G_2, i.e., the set of cliques in AG, denoted by $CS(AG)$, is obtained.

4. Take the cliques in $CS(AG)$ with maximum number of vertices to constitute the set $MCS(AG)$. For each clique Q in $MCS(AG)$, calculate $w(Q)=\sum_{v\in Q}w(v)$. Then select the clique with largest weight, which is denoted by $CS_m(G_1,G_2)$.

5. Calculate the Jaccard coefficient [11, 12] $JC(G_1,G_2)$ as follows:

$$JC(G_1,G_2)=\frac{|G_1\cap G_2|}{|G_1\cup G_2|}\approx\frac{|CS_m(G_1,G_2)|}{|G_1|+|G_2|-|CS_m(G_1,G_2)|} \tag{4}$$

where $|G|$ is the cardinality of graph G. We can only obtain the approximate Jaccard coefficient, due to the fact that $|MCS(G_1,G_2)|$ is an approximate value. The value of $JC(G_1,G_2)$ is the similarity of two compounds c_1 and c_2, whose corresponding graphs are G_1 and G_2. In this research, we denote the similarity of two compounds c_1 and c_2 by $S(c_1,c_2)$.

The data of compound similarity of each compound can be obtained from the website

http://pcal.biosino.org/workflowgallery/Search_KEGG_Compound/Search_KEGG_Compound.output.xml.html or http://www.genome.jp/ligand-bin/search_compound. However, they only provide the similarity at least 0.4 for each compound. In this research, the similarity less than 0.4 between two compounds is considered to be zero. The data of compound similarity of 3836 compounds studied in this research can be found in supplement materials II.

2.3 Prediction Based on Compound Similarity (PCS)

Let CS be a training compound set and each compound belongs to one or more of m categories (pathway classes or pathways). CS_i $(1 \le i \le m)$ denotes the compounds belonging to i-th category. Then $\sum_{i=1}^{m} |CS_i|$ is the total number of classified compounds, which can be calculated by Eq. (1). For a given sample c, we calculate a vector $v_c = (v_1, v_2, \cdots, v_m)$, where $v_i = \max\{S(c,c') \mid c' \in CS_i\}$ for $i = 1,2,\cdots,m$. Since it is possible $S(c,c') = 0$ for each compound $c' \in CS_i$, v_c is not always a vector with each component greater than zero. We can consider that v_i is the score between sample c and the i-th category. Furthermore, we consider that sample c is more likely to belong to category with higher score. We rank the prediction categories of c into $(i_1, i_2, \cdots, i_n)$, where $1 \le i_k \le m$ and $i_l \ne i_k$ if $l \ne k$, if and only if (1) $v_{i_1} \ge v_{i_2} \ge \cdots \ge v_{i_n} > 0$ and (2) $v_j = 0$ for each $j \in \{1,2,\cdots m\} \setminus \{i_1, i_2, \cdots i_n\}$. Here, we use $'\ge'$ because some compounds belong to more than one categories. It is possible that there is no prediction sequence for some sample if $v_c = 0$ and the length of prediction sequences is different for two different samples.

2.4 Jackknife cross-validation test

The model is tested by jackknife cross-validation test [13]. In such test, every compound in the dataset is singled out in turn as the testing data and the other samples are used to train the prediction model. Thus every sample is tested exactly once.

3. Results and discussion

In this research, we used PCS to classify each compound by two steps:

1. 3836 compounds are mapped into 11major metabolic pathway classes;
2. In each major pathway class, compounds are mapped into specific metabolic pathways.

3.1 Results of PSC of Step One

In this step, 3836 compounds are mapped into 11 major metabolic pathway classes by PSC. Jackknife cross-validation test is employed to evaluate the predict accuracy. For each compound, we obtained a prediction pathway class sequence, which can be found in supplement materials III. Table 3 shows the result of compound C00044 used PSC and we can see that the true pathway classes are at the top of prediction sequence of C00044.

Table 3: Prediction for C00044 (true pathway classes: Nucleotide Metabolism, and Metabolism of Cofactors and Vitamins)

Rank	Predicted pathway class	Score
1	Nucleotide Metabolism	0.91
2	Metabolism of Cofactors and Vitamins	0.85
3	Xenobiotics Biodegradation and Metabolism	0.8
4	Biosynthesis of Secondary Metabolites	0.8
5	Energy Metabolism	0.8
6	Amino Acid Metabolism	0.7
7	Carbohydrate Metabolism	0.63
8	Glycan Biosynthesis and Metabolism	0.61
9	Lipid Metabolism	0.6
10	Metabolism of Other Amino Acids	0.59
11	Biosynthesis of Polyketides and Nonribosomal Peptides	0.47

Table 4: Success rate of prediction pathway class sequences by rank

Most likely pathway class											**Least likely pathway class**
Rank	1	2	3	4	5	6	7	8	9	10	11
Rates(%)	80.47	13.82	6.88	3.86	2.03	1.43	0.96	0.5	0.26	0.21	0

Table 5: Success rate by pathway class

Predicted pathway class	Number of correct/total	Rate (%)
Carbohydrate Metabolism	226/449	50.33
Energy Metabolism	9/110	8.18
Lipid Metabolism	442/555	79.64
Nucleotide Metabolism	71/144	49.31
Amino Acid Metabolism	289/567	50.97
Metabolism of Other Amino Acids	78/175	44.57
Glycan Biosynthesis and Metabolism	36/66	54.55
Biosynthesis of Polyketides and Nonribosomal Peptides	224/280	80
Metabolism of Cofactors and Vitamins	240/334	71.86
Biosynthesis of Secondary Metabolites	808/891	90.68
Xenobiotics Biodegradation and Metabolism	664/752	88.3

Table 4 shows the total prediction rates of prediction pathway class sequences by rank, which is defined as “the number of compounds with the predicted pathway class belonging to the true pathway

classes"/"the total number of different compounds". From table 4, we can see that the more the number of pathway classes to be identified, the less the success rate would be. In addition, the success rate of rank one reaches 80.47%, while the success rate of rank two is only 13.82%, which indicates that two compounds with high similarity may always belong to the same pathway class. Table 5 lists the success rate by pathway class of rank one, which is defined as "the number of correctly predicted compounds in each pathway class"/ "the number of compounds in each pathway class".

3.2 Results of PSC of Step Two

In this step, each compound in major pathway class is mapped into specific pathways by PSC. Jackknife cross-validation test is employed to evaluate the predict accuracy. For each compound, we obtained a prediction pathway sequence, which can be found in supplement materials IV. Table 6 shows the result of compound C11911, which belongs to pathway class Biosynthesis of Polyketides and Nonribosomal Peptides, and we can see that the true pathways are at the top of prediction sequence of C11911. Similar to section 3.1, we can obtain two tables for each major pathway class, which can be found in supplement materials V. The success rate of Biosynthesis of Polyketides and Nonribosomal Peptides are show in Table 7 and Table 8 for example.

Table 6: Prediction for C11911 (true pathway classes: Biosynthesis of 12-, 14- and 16-membered macrolides, and Polyketide sugar unit biosynthesis)

Rank	Predicted pathway class	Score
1	Polyketide sugar unit biosynthesis	0.92
2	Biosynthesis of 12-, 14- and 16-membered macrolides	0.92
3	Biosynthesis of vancomycin group antibiotics	0.82
4	Biosynthesis of ansamycins	0.64

Table 7: Success rate of prediction pathway sequences by rank

Most likely pathway									**Least likely pathway**
Rank	1	2	3	4	5	6	7	8	9
Rates(%)	89.29	11.79	2.5	0.36	0	0	0	0	0

Table 8: Success rate by pathway

Predicted pathway	Number of correct/total	Rate (%)
Biosynthesis of 12-, 14- and 16-membered macrolides	57/75	76
Polyketide sugar unit biosynthesis	20/27	74.07
Biosynthesis of ansamycins	22/24	91.67
Type I polyketide structures	9/21	42.86
Biosynthesis of siderophore group nonribosomal peptides	12/13	92.31
Nonribosomal peptide structures	12/14	85.71
Biosynthesis of vancomycin group antibiotics	13/19	68.42
Biosynthesis of type II polyketide backbone	17/19	89.47
Biosynthesis of type II polyketide products	88/89	98.88

From these tables, we can see that in each major pathway class, we can obtain the same conclusion in section 3.1, i.e., the success rate of rank one is much higher than that of rank two and the rate decreases with more pathways to be identified. Table 9 shows success rate of rank one in each major metabolic pathway class.

Table 9: The success rate of rank one in major pathway class

Major pathway class	Success rate of rank one (%)
Carbohydrate Metabolism	50.78
Energy Metabolism	60.91
Lipid Metabolism	85.59
Nucleotide Metabolism	88.89
Amino Acid Metabolism	64.73
Metabolism of Other Amino Acids	66.86
Glycan Biosynthesis and Metabolism	81.82
Biosynthesis of Polyketides and Nonribosomal Peptides	89.29
Metabolism of Cofactors and Vitamins	80.24
Biosynthesis of Secondary Metabolites	89.11
Xenobiotics Biodegradation and Metabolism	61.17

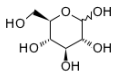
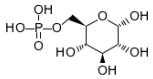
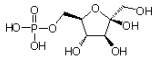
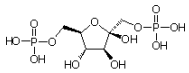
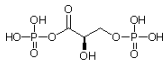
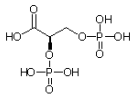
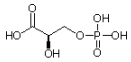
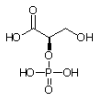
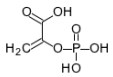
3.3 Discussion

The above results indicate that compounds with high similarity may always belong to the same pathway class and pathway. Take the classical glycolysis pathway for example. There are ten-step sequential reactions in this pathway, where one glucose is split and converted into two pyruvates. (http://www.genome.jp/kegg/pathway/map/map00010.html) [14]. 11 compounds and 10 reactions participate in the glycolysis pathway (See Table 10). From Table 10, we can see that two compounds with the coterminous position in the pathway may have similar structure, i.e., the compound similarity between them is high, which indicates that glucose is converted into pyruvate by degrees.

4. Conclusion

In this research, we try to predict the metabolic pathway of small chemical molecule by means of its compound similarity. There are two steps: (1) map small chemical molecules into major pathway classes, (2) then, map them into specific pathways. In each step, we can obtain a prediction sequence for each compound, which indicates that our method can tackle multi-class case. In addition, since our method is based on compound similarity, we can deal with a pair of isomeride as two different compounds. At last, we obtain acceptable results, which indicate that a small chemical molecule may belong to the pathway class and pathway which its most similar compound belongs to. Therefore, our method can be used for mapping a new compound into corresponding pathway efficiently.

Table 10: The 11 compounds in the Glycolysis / Gluconeogenesis pathway

Compound	Compound ID	Compound's structure
D-Glucose	C00031	
	↓ Reaction 1	
alpha-D-Glucose 6-phosphate	C00668	
	↕ Reaction 2	
beta-D-Fructose 6-phosphate	C05345	
	↓ Reaction 3	
beta-D-Fructose 1,6-bisphosphate	C05378	
	↕ Reaction 4	
D-Glyceraldehyde 3-phosphate	C00118	
	↕ Reaction 5	
3-Phospho-D-glyceroyl phosphate	C00236	
	↕ Reaction 6	
2,3-Bisphospho-D-glycerate	C01159	
	↕ Reaction 7	
3-Phospho-D-glycerate	C00197	
	↕ Reaction 8	
2-Phospho-D-glycerate	C00631	
	↕ Reaction 9	
Phosphoenolpyruvate	C00074	
	↓ Reaction 10	
Pyruvate	C00022	

References

1. Nicholson, J.K., Connelly, J., Lindon, J.C., Holmes, E.: Metabonomics: a platform for studying drug toxicity and gene function. Nat Rev Drug Discov 1,153–161 (2002)
2. Nicholson, J.K., Wilson, I.D.: Opinion: understanding 'global' systems biology: metabonomics and the continuum of metabolism. Nat Rev Drug Discov 2, 668–676 (2003)

3. Nicholson, J.K., Lindon, J.C., Holmes, E.: 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. Xenobiotica 29, 1181–1189 (1999)
4. Nicholson, J.K., Holmes, E., Lindon, J.C.,Wilson, I.D.: The challenges of modeling mammalian biocomplexity. Nat Biotechnol 22, 1268–1274 (2004)
5. Vastrik, I., D'Eustachio, P., Schmidt, E., Joshi-Tope, G., Gopinath, G., Croft, D. et al: Reactome: a knowledge base of biologic pathways and processes. Genome Biol. 8, R39 (2007)
6. Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., Hirakawa, M.: From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 34(database issue) D354–D357 (2006)
7. Burkart, M.D.: Metabolic engineering—a genetic toolbox for small molecule organic synthesis. Org. Biomol. Chem. 1, 1–4 (2003)
8. Cai, Y.D., Qian, Z., Lu, L., Feng, K.Y., Meng, X., Niu, B., Zhao, G.D., Lu, W.C.: Prediction of compounds' biological function (metabolic pathway) based on functional group composition, Mol. Divers. 12, 131-137 (2008)
9. Masahiro, H., Yasushi, O., Susumu, G., Minoru, K.: Development of a Chemical Structure Comparison Method for Integrated Analysis of Chemical and Genomic Information in the Metabolic Pathways. J. AM. CHEM. SOC. 125, 11853-11865 (2003)
10. Bron, C., Kerbosch, J.: Algorithm 457: Finding all cliques of an undirected graph. Commun. ACM 16, 575-577 (1973)
11. Jaccard, P.: The distribution of the flora of the alpine zone. New Phytol 11, 37-50 (1912)
12. Watson, G.A.: An algorithm for the single facility location problem using the Jaccard metric. SIAM J. Sci. Stat. Comput. 4, 748-756 (1983)
13. Chou, K. C., Zhang, C. T.: Critical Reviews in Biochemistry and Molecular Biology, 30 (4), 275-349 (1995)
14. Trudy, M.: Biochemistry: an introduction. 2nd edn, McGraw-Hill Companies, Inc (1999)