Metabolomics: Open Access
ISSN: 2153-0769

De and Tagore, Metabolomics 2012, S:1
DOI: 10.4172/2153-0769.S1-005

Research Article Open Access

Automated Reconstruction of Metabolic Pathways of Homo Sapiens involved in the Functioning of GAD1 and GAD2 Genes based on Structural Grammars

Rajat K. De[1]* and Somnath Tagore[2]

[1]*Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700108, India*
[2]*Department of Biotechnology and Bioinformatics, Padmashree Dr. D.Y. Patil University, Navi Mumbai 400614, India*

Abstract

Modeling and analyzing the architecture of metabolic networks using various computational strategies can be successfully used for studying their internal metabolic dynamics as well as predicting missing links in diseased networks. In the present work, we have implemented our algorithm based on structural grammars, for automated metabolic pathway reconstruction and modeling in metabolic pathways responsible for coding genes responsible for the cause of Type 1 *Diabetes mellitus* (T1D) in *Homo sapiens*. We have especially implemented our algorithm for studying the metabolic pairs responsible for the functioning of GAD1 and GAD2 genes. We have also used the algorithm for automated reconstruction of glutamate metabolism, β-alanine metabolism, taurine & hypotaurine metabolism and butanoate metabolism pathway datasets. We have also used the algorithm for missing and multiple link prediction as well as nodal point analysis for all the four metabolic pathways with 90.4-100% accuracy.

Keywords: Combinatorial model; Lavenshtein distance; SMILES; Structural grammar; Type 1 *Diabates mellitus*

Introduction and Background

Metabolic pathway modeling is one of the most essential areas in the post-genomic era. Analyzing the key components of metabolic pathways, like, understanding the role of enzymes in catalyzing various biochemical reactions, gene expression ratios and characterizing their internal architecture, can be properly managed by modeling their attributes. A model can be also used to understand the flow of information within a metabolic network by means of performing internal and external perturbations. Computational models based on biological constraints can be built which are further used to relate the developed models with their biological behaviors [1]. These theoretical models can be further trained in order to analyze and simulate these complex networks of different organisms and tissues. Furthermore, computational models can also be used for simulating the function of metabolic pathways, thereby improving the understanding of the structure of cellular processes. Moreover, *in-silico* methods are also used for simulating biological processes and for their ability to handle large datasets, and identifying missing or incomplete links in disease pathways and for identifying nodal points in metabolic pathways. Furthermore, automated reconstruction is a process of building the complete metabolic network given some input metabolites and their constraints is a flourishing and fast developing domain [2].

In case of metabolic networks, pathway reconstruction is quite complex due to the fact that computational prediction of relations among the input metabolites is extremely difficult. For the same reason, it is extremely essential to find conditions for defining relations among given biological moieties. Our algorithm, based on structural grammars uses path mining concepts for linking these biological moieties and thereby reconstructing the complete metabolic pathway. In this case we have assumed the fact that for a given set of metabolites, the notion for creating a link is based on the probability that one would be converted to another. Furthermore, this conversion is highly possible if these metabolites are structurally similar. One of the fundamental aspects for this algorithm to work efficiently is the input metabolite format. For establishing possible links among metabolites, it uses the concept of topological indices that predicts the similarity and dissimilarity among them based upon some given inputs. Our algorithm uses SMILES strings as the input dataset format.

In this work, we have implemented our algorithm over some metabolic pathways responsible for functioning of some genes responsible for the cause of Type 1 *Diabetes mellitus* (T1D). *Diabetes mellitus* is a disorder characterized by uneven changes in the metabolism of carbohydrates, fats and proteins that are a result of defects in the secretion and action of insulin. The general accepted classification of *Diabetes mellitus* is Type 1 and Type 2. Type 1 categorizes the instances which are due to the destruction in pancreatic islet beta–cells as well as those liable to acidosis with an accumulation of ketone bodies. Similarly, Type 2 results from inadequate insulin secretion and resistance to insulin action. Furthermore, the effects of *Diabetes mellitus* also include failure and dysfunction of various organs, prevalence of foot ulcers, and amputation along with occurrences of cardiovascular and cerebrovascular diseases [3].

The role of genes in the development of Type 1 *Diabetes mellitus* (T1D) is also long due to its heterogeneity. But, till date scientists all the gene mutations that put a person at risk for T1D. Individuals suffering from T1D may be normal before the disease is clinically manifest metabolically, but detection of β–cells can be done. T1D is characterized by the presence of anti–glutamic acid decarboxylase (GAD), islet cell or insulin antibodies which identify the autoimmune processes that lead to β–cells destruction. Thus, T1D is also mediated

***Corresponding author:** Rajat K. De, Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700108, India, E-mail: rajat@isical.ac.in

Received March 16, 2012; **Accepted** April 19, 2012; **Published** April 21, 2012

Citation: De RK, Tagore S (2012) Automated Reconstruction of Metabolic Pathways of *Homo Sapiens* involved in the Functioning of GAD1 and GAD2 Genes based on Structural Grammars. Metabolomics S1:005. doi:10.4172/2153-0769.S1-005

by destruction of the β–cells of the pancreas which is quite variable, being rapid in some individuals and slow in others. Moreover, this rapidly progressive form is commonly observed in children, with some cases in adults. Furthermore, various markers of immune destruction, including autoantibodies to islet cells, insulin, GAD are present in 85–90 % of individuals with T1D. Till date, two mechanisms of onset for T1D have been proposed. The first one suggests that environmental factors trigger the autoimmune process, mostly in children before 10 years of age. It becomes evident after a long precursor period with gradual destruction of pancreatic β–cells. The second mechanism suggests that a super-antigen biochemical reaction results in rapid destruction of pancreatic β–cells within a few weeks to a month, leading to the onset of T1D [4]. The rest of the manuscript deals with explanation of the methodology of our algorithm, results and analysis and conclusion respectively.

Methodology

This section deals with explanation of the algorithm implemented by us for analyzing the metabolic pathways of *H. sapiens*. It gives an overview of our algorithm and its strategies for pathway analysis.

Automated pathway reconstruction algorithm

Our algorithm considers Simplified Molecular Input Line Entry System (SMILES) strings as input format for metabolites. It works with the notion that any metabolite can be compared on the basis of some structural features that can be identified based on their SMILES string notation, a 1D representation of a 3D metabolite. Due to the conversion of 3D to 1D format of metabolite, some basic properties like linkages among atoms are neglected. For this reason, this algorithm converts these input strings into some patterns for storing information about linkages in the form of bonds. On the basis of these patterns, five scoring schemes are implemented, namely, *weight score*, *comparison score*, *isomeric score*, *pattern comparison score*, and *association score* (Figure 1) [5].

The weight score calculates the overall weight contribution of atoms present in a metabolite as compared to its contribution in the overall metabolite weight as well as the weight contribution of metabolites in a reaction. For the same purpose, two categories of metabolites are taken into consideration, namely, pools (side metabolites of reactions such as ATP, ADP) and non-pools (main chain metabolites like glucose, galactose). Thus, for a particular reaction ATP + Pyruvate → ADP + Phosphoenolpyruvate, pools are ATP, ADP and non-pools are Pyruvate, Phosphoenolpyruvate. Our algorithm uses this weight score for linking the pools and non-pools, as this is quite cumbersome due to the absence of any past linking reference. The comparison score is calculated by comparing metabolites on the basis of the generated patterns for calculating similarity among them. Greater the similarity score, higher is the similarity among the compared metabolites. This score is calculated on the basis of the presence or absence of binary digits in the compared metabolite patterns [6].

The input for the algorithm is taken from biological databases, which may contain repetitive entries, resulting in unreliable results. For this reason, normalization is done for the datasets, resulting in removal of repetitive information by calculating the normalization score. Moreover, while converting SMILES into structural grammars, various anomalies may arise resulting in loss of information related to biological activities of metabolites. For removing this problem, our algorithm conserves the branching patterns among metabolites on the basis of the number of connectivities among the SMILES string patterns resulting in the calculation of symmetric difference score. It also distinguishes among metabolites having isomers by taking into consideration two fundamental properties of metabolites, namely, isomeric thresholds and isomeric count, whose values are chosen from PubChem Compound databases. The combination of normalization score, symmetric difference score, isomeric threshold and isomeric count gives rise to isomeric score. The pattern comparison score is calculated by taking into consideration four properties, metabolite class, metabolite order, Lavenshtein distance between compared metabolites and the pattern difference score. Metabolites are assigned five classes, namely, linear, cyclic, acyclic, branched and un-branched based on its structural linking pattern, calculated from the SMILES itself. Metabolite order is calculated by identifying occurrence number of metabolites in the input dataset. Lavenshtein distance between the metabolite pair is the minimum number of operations (insertion, deletion, or substitution of a single binary digit) needed for transforming a structural grammar one metabolite into another. The pattern difference score considers the number of patterns generated for each metabolite and the corresponding significant atoms in the patterns [7]. Furthermore, if the compared metabolites are of different classes, the algorithm calculates path number and polarity number along with two parameters from database that are topological polar surface area and logp3 value for calculating association score. The combined score is generated by accumulating all the five scores. In any automated pathway reconstruction strategy, metabolites participating in more than one reaction are possible. Thus, creating multiple links among metabolites need to be taken into consideration. For this purpose, we have assigned a threshold value for detecting such metabolites. Thus, only that final score is selected which is greater than a particular threshold value. If the score is not higher than the threshold, then the next best score is selected and the procedure is repeated (Figure 1).

Missing and multiple link identification

Identifying missing links in metabolic pathways deal with detecting those reaction links that are previously unknown or for which no proper evidence is available. Our algorithm calculates the similarity score among metabolites for the metabolite pair for which no possible

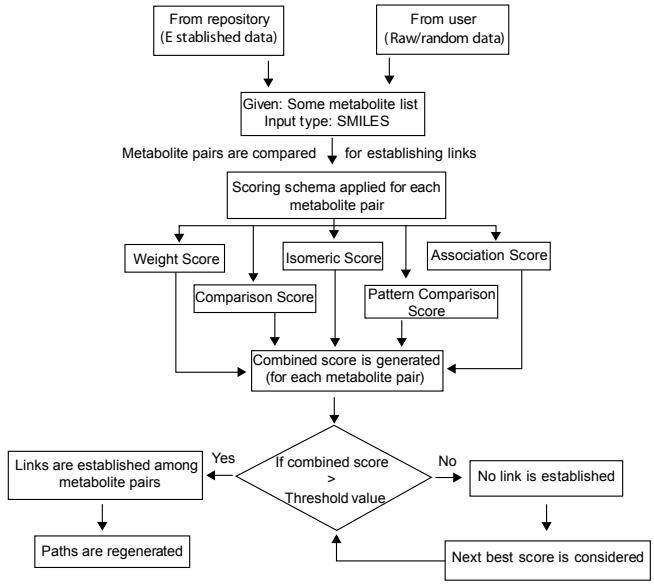

Figure 1: Steps of our algorithm.

reaction link has been identified previously for any possible similarity score between them (Figure 2). The given schematic diagram of a hypothetical metabolic pathway has 8 metabolites and 7 reaction links.

In the given Figure 2, a missing link has been hypothesized between metabolites m_7 and m_8. The algorithm calculates all the possible links for metabolites m_7 and m_8, on the basis of which it identifies whether based on the similarity score any possible link can be established between them. For establishing a link between metabolites m_7 and m_8, a good similarity score should be existent. It is also seen that metabolite m_1 has 4 reaction links. It uses a threshold value for identifying multiple links. Similarity scores greater than the threshold value are taken into consideration. In this case, the similarity scores for metabolic pairs, $m_1 - m_2$, $m_1 - m_3$, $m_1 - m_4$, $m_1 - m_5$ are taken into consideration [8].

Nodal point detection

Nodal points in metabolic networks are those metabolites and reaction links which are extremely essential for maintaining the overall structure of the network and can be targeted by external and/or internal agents during perturbations. In a metabolic network, metabolites which are highly connected can be detected and targeted for analyzing their role in behaving as nodal points. Figure 3 illustrates the hypothetical metabolic pathway discussed in Figure 2 showing nodal points. Reaction links adjacent to the highest connected metabolite can also be targeted for determining their role in metabolic pathway [9].

Results and Analysis

We implemented our algorithm to four metabolic pathways, namely, glutamate metabolism, β-alanine metabolism, taurine & hypotaurine metabolism and butanoate metabolism, responsible for the functioning of GAD1 and GAD2 genes involved in Type 1 *Diabetes mellitus* (T1D). We were interested in analyzing the reactions catalyzed by the enzymes which can be further analyzed for studying the expression analysis of GAD1 and GAD2. As discussed in the "Methodology" section, the algorithm can be used in three manners for analyzing the architectural complexity of metabolic pathways, we implemented the same in the above four metabolic pathways. This section has been explained under three sub-sections, namely, pathway reconstruction studies on metabolic pathways in T1D, studying missing and multiple reaction links in metabolic pathways for T1D and identifying nodal metabolites & reaction links.

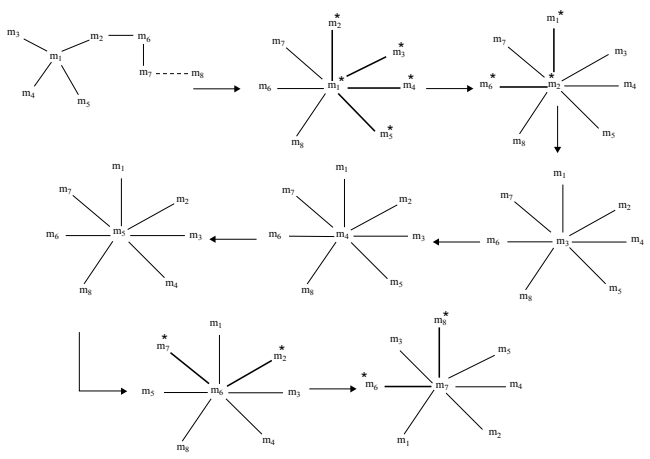

Figure 2: Hypothetical pathway showing identification of missing and multiple links.

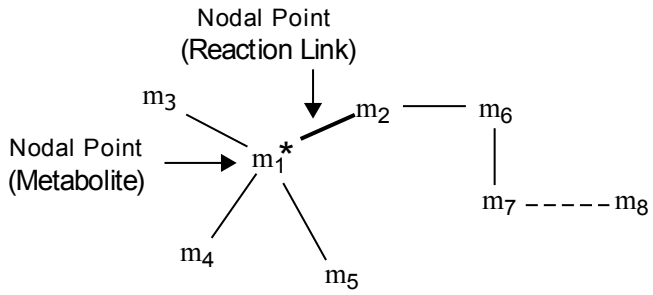

Figure 3: Hypothetical pathway showing nodal point detection.

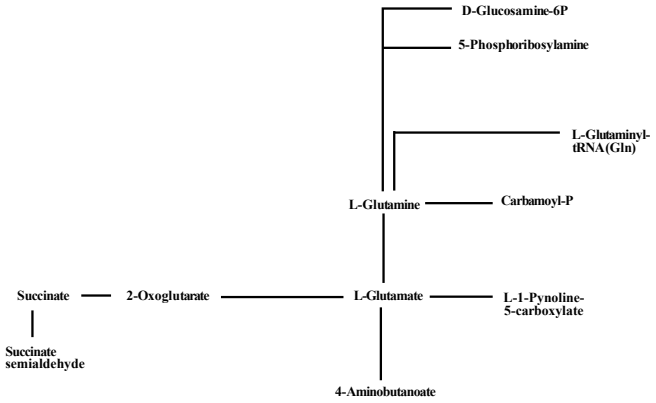

Figure 4: Glutamate metabolism in *H. sapiens* as generated by our algorithm.

Pathway reconstruction studies on metabolic pathways on T1D

Our algorithm reconstructs all the four metabolic pathways, given some list of metabolites. In glutamate metabolism, having 11 metabolites and 11 reaction links, the algorithm is able to reconstruct with 100% accuracy. The similarity scores among metabolite pairs are shown in Table 1. We observe that out of all the metabolite pairs, the algorithm is able to predict all the 11 reaction links. It is also able to predict the reaction link between L-glutamate and 4-aminobutanoate, responsible for the expression of GAD1 and GAD2 giving a very good score (0.75) as compared to other metabolite pairs (Table 1). The reconstructed metabolic pathway of glutamate metabolism in *H. sapiens* by our algorithm is illustrated in Figure 4.

Furthermore, in β-alanine metabolism there are 19 metabolites and 21 reaction links. In this case, the algorithm is able to reconstruct the metabolic pathways with 90.4% accuracy. The highest similarity scores among the metabolite pairs are represented in Table 2 (Supplementary Information). The algorithm is able to identify 20 correct and 1 incorrect reaction links and is able to identify the reaction link between β-alanine and L-aspartate which again plays an important role for the function of GAD1 and GAD2. The reconstructed metabolic pathway of β-alanine metabolism in *H. sapiens* by our algorithm is illustrated in Figure 5 (Supplementary Information). In case of taurine & hypotaurine metabolism, with 12 metabolites and 13 reaction links this algorithm is able to reconstruct it with 100% accuracy. Furthermore, the metabolite pairs having highest similarity scores are represented in Table 3 (Supplementary Information) [9].

Metabolite pair	Similarity Score
Succinate semialdehyde – Succinate	0.74
2-oxoglutarate	0.60
L-glutamate	0.54
4-aminobutanoate	0.48
L-1-pyrroline-5-carboxylate	0.42
L-glutamine	0.46
L-glutaminyl-tRNA(Gln)	0.41
5-phosphoribosylamine	0.38
D-glucosamine-6P	0.35
Succinate – 2-oxoglutarate	0.73
L-glutamate	0.62
4-aminobutanoate	0.55
L-1-pyrroline-5-carboxylate	0.50
L-glutamine	0.48
L-glutaminyl-tRNA(Gln)	0.46
5-phosphoribosylamine	0.41
D-glucosamine-6P	0.38
2-oxoglutarate – L-glutamate	0.74
4-aminobutanoate	0.64
L-1-pyrroline-5-carboxylate	0.60
L-glutamine	0.55
L-glutaminyl-tRNA(Gln)	0.50
5-phosphoribosylamine	0.47
D-glucosamine-6P	0.41
L-glutamate – 4-aminobutanoate	0.75
L-1-pyrroline-5-carboxylate	0.74
L-glutamine	0.73
L-glutaminyl-tRNA(Gln)	0.65
5-phosphoribosylamine	0.60
D-glucosamine-6P	0.55
4-aminobutanoate – L-1-pyrroline-5-carboxylate	0.65
L-glutamine	0.60
L-glutaminyl-tRNA(Gln)	0.55
5-phosphoribosylamine	0.48
D-glucosamine-6P	0.45
L-1-pyrroline-5-carboxylate – L-glutamine	0.56
L-glutaminyl-tRNA(Gln)	0.58
5-phosphoribosylamine	0.52
D-glucosamine-6P	0.50
L-glutamine – Carbamoyl-P	0.73
L-glutaminyl-tRNA(Gln)	0.74
5-phosphoribosylamine	0.72
D-glucosamine-6P	0.73
L-glutaminyl-tRNA(Gln) – 5-phosphoribosylamine	0.56
D-glucosamine-6P	0.59
5-phosphoribosylamine – D-glucosamine-6P	0.67

Table 1: Similarity scores for metabolite pairs in glutamate metabolism.

In this case our algorithm is able to predict 13 reaction links. It is also able to identify reaction links hypotaurine and 3-sulfino-L-alanine as well as taurine and L-cysteate that are involved in the functioning of GAD1 and GAD2 genes. The reconstructed metabolic pathway of taurine & hypotaurine metabolism in *H. sapiens* by our algorithm is illustrated in Figure 6 (Supplementary Information). Similarly, butanoate metabolism has 15 metabolites and 15 reaction links respectively, out of which our algorithm predicts 14 reaction links with 93.3% accuracy. Our algorithm also identifies L-glutamate and 4-aminobutanoate metabolite pair involved in functioning of GAD1 gene. The reconstructed metabolic pathway of butanoate metabolism in *H. sapiens* by the algorithm is illustrated in Figure 7 (Supplementary Information).

Studying missing and multiple reaction links in metabolic pathways on T1D

Prediction of missing and multiple reaction links in metabolic pathways is one of the most essential features of our algorithm. It is especially useful in those metabolic pathways which become infected and are involved in disease networks. The four selected metabolic pathways involved in T1D have been studied for identifying any possible missing and multiple reaction links. In case of glutamate metabolism, the algorithm predicts multiple links for L-glutamate (4 links), L-glutamine (5 links) based on their similarity scores (Table 5 in Supplementary Information). All the metabolite pairs, L-glutamate with 2-oxo-gluterate, 4-aminobutanoate, L-1-pyrroline-5-carboxylate, L-glutamine; L-glutamine with 2-oxo-glutaramate, L-glutamate, D-glucosamine-6P, carbamoyl-P, 5-phospho-ribosylamine, have similarity score within the threshold value range of 0.72-0.76. Similarly, in β-alanine metabolism, the metabolite for which multiple links were identified by this algorithm are L-aspartate (10 links), malonate semialdehyde (2 links) based on the similarity score (Tables 2 and 5 in Supplementary Information). The metabolite pairs of L-aspartate with N-carbamoyl- β-alanine, L-aspartyl-tRNA (Asp), β-alanine, L-alanine, adenylosuccinate, L-arginino-succinate, L-asparagine, oxaloacetate, N-acetyl-L-aspartate, D-aspartate; malonate semialdehyde with β-alanine, acetyl CoA respectively, have similarity score within the threshold value range of 0.68-0.71. Furthermore, in taurine & hypotaurine metabolism the metabolite pairs for which multiple links are established are taurine (7 links), hypotaurine (3 links) (Tables 3 and 5 in Supplementary Information). Moreover, taurine forms pairs with L-cysteate, taurocholate, sulfoacetaldehyde, taurocyamine, 5-glutamyl-taurine, hypotaurine, pyruvate whereas hypotaurine pairs with cysteamine, 3-sulfino-L-alanine and taurine respectively that have similarity score within the threshold value range of 0.75-0.78. Similarly, in butanoate metabolism, the metabolite pairs for multiple links predicted by our algorithm, are acetoacetyl CoA (5 links), acetyl CoA (3 links), crotonyl CoA (3 links) on the basis of their similarity score (Tables 4 and 5 in Supplementary Information). The metabolite pairs of acetoacetyl CoA with (S)-3-hydroxy butanoyl-CoA, (R)-3-hydroxy-butanoyl-CoA; acetyl CoA with pyruvate, acetoacetyl CoA, (S)-3-hydroxy-3-methylglutaryl CoA; crotonyl CoA with butanoyl-CoA, 3-butenoyl-CoA, (S)-3-hydroxybutanoyl-CoA having similarity score within the threshold value range of 0.58-0.61. Instances of missing links are not found in glutamate metabolism, β-alanine metabolism and taurine & hypotaurine metabolism respectively, with exception in butanoate metabolism where a missing link exists between 4-hydroxy butanoate and 3-butanoyl-CoA. Our algorithm successfully identifies the missing link with a similarity score of 0.60 [10].

Identifying nodal metabolites & reaction links

In metabolic pathways, nodal points are those metabolites which are involved in many useful reactions, whose presence is in various reactions and without which major reactions may not work properly. As discussed in "Methodology" section, our algorithm detects the nodal points based on their presence in the overall metabolic pathway. For instance, in case of butanoate metabolism the algorithm identifies L-glutamate; L-glutamine having 4 and 5 reaction links respectively. Also, the metabolite pairs L-glutamate and L-glutamine, L-glutamate and 4-aminobutanoate are chosen to be nodal reaction links due to the fact that both have highest similarity score. Similarly, in β-alanine metabolism the nodal points detected are β-alanine and L-aspartate. Furthermore, in taurine & hypotaurine metabolism, taurine and taurine-L-cysteate are identified as links. Lastly, in butanoate metabolism, acetoacetyl CoA and acetyl CoA are identified as reaction links [11].

Conclusion

We have presented here the use of our novel algorithm for analysis

of four metabolic pathways namely, glutamate metabolism, β-alanine metabolism, taurine & hypotaurine metabolism and butanoate metabolism, responsible for the expression of GAD1 and GAD2 genes involved in Type 1 *Diabetes mellitus* (T1D). We have specifically studied the architectural complexity of these metabolic pathways by implementing pathway reconstruction approach, studying missing and multiple reaction links and identification of nodal metabolites & reaction links. It is able to reconstruct glutamate metabolism, β-alanine metabolism, taurine & hypotaurine metabolism and butanoate metabolism with 100%, 90.4%, 100% and 93.3% accuracy respectively. In case of glutamate metabolism, the algorithm predicts multiple links in case of L-glutamate having 4 reaction links, L-glutamine having 5 reaction links, for β-alanine metabolism, β-alanine having 8 reaction links, malonate semialdehyde having 3 reaction links, spermadine having 3 reaction links, for taurine & hypotaurine metabolism, taurine having 6 reaction links, hypotaurine having 3 reaction links and in butanoate metabolism, acetoacetyl CoA having 5 reaction links, acetyl CoA having 3 reaction links, crotonyl CoA having 3 reaction links respectively. Furthermore, missing links were not found in glutamate metabolism, β-alanine metabolism and taurine & hypotaurine metabolism respectively, but in butanoate metabolism only. Moreover, our algorithm is also able to detect nodal points and reaction links in all the four metabolic pathways.

References

1. Klipp E, Herwig R, Kowald A, Wierling C, Lehrach H (2005) Systems biology in practice: concepts, implementation and application. John Wiley & Sons Inc., New York.
2. Kitano H (2001) Foundations of Systems Biology. MIT Press, Cambridge.
3. Holmes GK (2001) Coeliac disease and Type 1 diabetes mellitus - the case for screening. Diabet Med 18: 169-177.
4. Humphrey AR, McCarty DJ, Mackay IR, Rowley MJ, Dwyer T, et al. (1998) Autoantibodies to glutamic acid decarboxylase and phenotypic features associated with early insulin treatment in individuals with adult-onset diabetes mellitus. Diabetic Med. 15: 113-119.
5. Bonchev D (1983) Information Theoretic Indices for Characterization of Chemical Structures. Research Studies Press, Hertfordshire, UK.
6. García-Domenech R, Galvez J, de Julian-Ortiz JV, Pogliani L (2008) Some new trends in chemical graph theory. Chem Rev 108: 1127-1169.
7. Trinajstic N (1985) Mathematics and Computational Concepts in Chemistry. Horwood Publishers, Chichester.
8. Tada M, Shijima H, Nakamura M (2003) Smiles-type free radical rearrangement of aromatic sulfonates and sulfonamides: Syntheses of arylethanols and arylethylamines. Org Biomol Chem 1: 2499-2505.
9. Walker M, Turnbull DM (1997) Mitochondrial related diabetes: a clinical perspective. Diabet Med 14: 1007-1009.
10. Wang JH, Byun J, Pennathur S (2010) Analytical approaches to metabolomics and applications to systems biology. Semin Nephrol 30: 500-511.
11. Nica AC, Dermitzakis ET (2008) Using gene expression to investigate the genetic basis of complex disorders. Hum Mol Genet 17: R129-R134.