Wang et al. BMC Bioinformatics 2019, 20(Suppl 18):574
https://doi.org/10.1186/s12859-019-3127-4

BMC Bioinformatics

RESEARCH Open Access

Prioritizing candidate diseases-related metabolites based on literature and functional similarity

Yongtian Wang[1], Liran Juan[2], Jiajie Peng[3], Tianyi Zang[1*] and Yadong Wang[1*]

From Biological Ontologies and Knowledge bases workshop at IEEE BIBM 2018
Madrid, Spain. 3-6 December 2018

Abstract

Background: As the terminal products of cellular regulatory process, functional related metabolites have a close relationship with complex diseases, and are often associated with the same or similar diseases. Therefore, identification of disease related metabolites play a critical role in understanding comprehensively pathogenesis of disease, aiming at improving the clinical medicine. Considering that a large number of metabolic markers of diseases need to be explored, we propose a computational model to identify potential disease-related metabolites based on functional relationships and scores of referred literatures between metabolites. First, obtaining associations between metabolites and diseases from the Human Metabolome database, we calculate the similarities of metabolites based on modified recommendation strategy of collaborative filtering utilizing the similarities between diseases. Next, a disease-associated metabolite network (DMN) is built with similarities between metabolites as weight. To improve the ability of identifying disease-related metabolites, we introduce scores of text mining from the existing database of chemicals and proteins into DMN and build a new disease-associated metabolite network (FLDMN) by fusing functional associations and scores of literatures. Finally, we utilize random walking with restart (RWR) in this network to predict candidate metabolites related to diseases.

Results: We construct the disease-associated metabolite network and its improved network (FLDMN) with 245 diseases, 587 metabolites and 28,715 disease-metabolite associations. Subsequently, we extract training sets and testing sets from two different versions of the Human Metabolome database and assess the performance of DMN and FLDMN on 19 diseases, respectively. As a result, the average AUC (area under the receiver operating characteristic curve) of DMN is 64.35%. As a further improved network, FLDMN is proven to be successful in predicting potential metabolic signatures for 19 diseases with an average AUC value of 76.03%.

Conclusion: In this paper, a computational model is proposed for exploring metabolite-disease pairs and has good performance in predicting potential metabolites related to diseases through adequate validation. This result suggests that integrating literature and functional associations can be an effective way to construct disease associated metabolite network for prioritizing candidate diseases-related metabolites.

Keywords: Metabolite network, Collaborative filtering, Similarity of metabolites, Random walking with restart

* Correspondence: tianyi.zang@hit.edu.cn; ydwang@hit.edu.cn
[1]School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, People's Republic of China
Full list of author information is available at the end of the article

Background

While a gene-based approach has contributed to our knowledge on the genomic space of possible genes and proteins [1–4], it is increasingly understood that such an approach is far from sufficient because most cellular components work intricate networks of regulatory, metabolic, and protein interactions [5–8]. As the end products of cellular regulatory processes, metabolites can be considered as the ultimate response of biological systems to genetic or environmental changes [9]. In biological systems, metabolomics, which is an emerging area of research, can not only contribute to the discovery of metabolic signatures for disease diagnosis, but is very helpful to illustrate the underlying molecular disease-causing mechanisms [9–12]. Furthermore, metabolites are easier to be analyzed for recognizing diseases at the molecular level compared to genes, mRNA transcripts and proteins related to diseases, among which there are large quantities of intricate interactions. Therefore, metabolisms, as the final products of cellular regulatory processes, can be a significant factor to illustrate the disease-causing mechanisms.

Nowadays, the advanced technology is really helpful to researchers for studying diseases in the molecular level [13–17]. And more researchers have devoted their work to metabolomics for revealing more information about diseases. Breitling, R et al. [18] utilized Fourier transform mass spectrometry data to make prediction of metabolic networks. In 2010, Gao, J et al. [19] developed a plugin for visualizing and interpreting metabolomic data in human metabolic networks. Considering the global importance of metabolites and the unique character of metabolomic profile, Li Feng et al. [20] proposed a network-based method for metabolite pathway identification. In 2016, Sergushichev, AA et al. [21] presented a web-service for integrated transcriptional and metabolic network analysis, focusing on identification of the most changing metabolic subnetworks between two conditions of interest. Wang et al. [22] identified potential urinary biomarkers for early colorectal cancer detection utilizing NMR-based metabolomic techniques. Recently, Ohtana, Yuki et al. [23] made analysis of drug-endogenous human metabolite similarities and 3D-Structure similarity based network of Secondary Metabolites [24]. To figure out whether metabolite networks are reproducible across different populations, Iqbal, Khalid et al. [25] investigated similarity of metabolite networks in four German population-based studies (EPIC-Potsdam, EPIC-Heidelberg, KORA and CARLA). From the above it can be seen that researchers are paying more attention to metabolite research and metabolomics has developed rapidly.

As the link between genotypes and phenotypes, one metabolite is not always related to a sole disease, and the impact of certain disease spreads among functionally related metabolites in a network [26]. Thus, adjacent metabolites with functional associations in this network tend to relate to the same diseases or similar ones [6]. This suggests that the functional associations between metabolites can be measured by the similarities of diseases. Therefore, we aimed to identify more disease-related metabolites by analysis the metabolite and disease data.

Now there have been many methods to calculate medical terminology similarity [27–30]. But to our knowledge, no methods had been proposed to compute metabolite similarity based on collaborative filtering (CF) [31] with the functional similarities between diseases as weight. CF can effectively utilize associations among other similar members and discover potential but not yet found interests. It is able to finish personalized recommendation with high degree of automation. Thus, a disease associated metabolite network (DMN) can be built based on modified collaborative filtering, which takes advantage of the entire interaction network. However, relying entirely on metabolite-related diseases greatly limits the utility of the method because many metabolites still have very few or no associated diseases. To overcome this limitation, a new disease-associated metabolite network (FLDMN) is built by fusing functional associations and scores of literatures from STITCH database [32]. Finally, FLDMN is utilized to identify potential disease-related metabolites based on network random walk.

Materials and methods

To clarify the research that we do, the workflow of the computational model is shown in Fig. 1. First, we integrate information from Human Disease Ontology (DO) [33], Merged Disease vocabulary (MEDIC) [34] and Human Metabolome Database (HMDB) [35] to establish mapping between DO terms and metabolites. Next, we define a feature vector for each metabolite to calculate the similarities between metabolites. Given the associations between metabolites and diseases, disease functional similarities are calculated by FNSemSim [29] and added to the dimensions of this vector as relevance scores between diseases and metabolites. Based on functional associations between metabolites, a disease-related metabolite network (DMN) is built. Subsequently, extracting scores of literatures from STITCH, we build a new network of disease related metabolites (FLDMN). Random Walking with Restart (RWR) is applied in this new network to output the ranking of candidate disease-related metabolites. Therefore, the potential relationships between diseases and metabolites can be identified.

Data collection

Disease database

Merged Disease vocabulary (MEDIC) [34] from Comparative Toxicogenomics Database (CTD) [36] is a modified

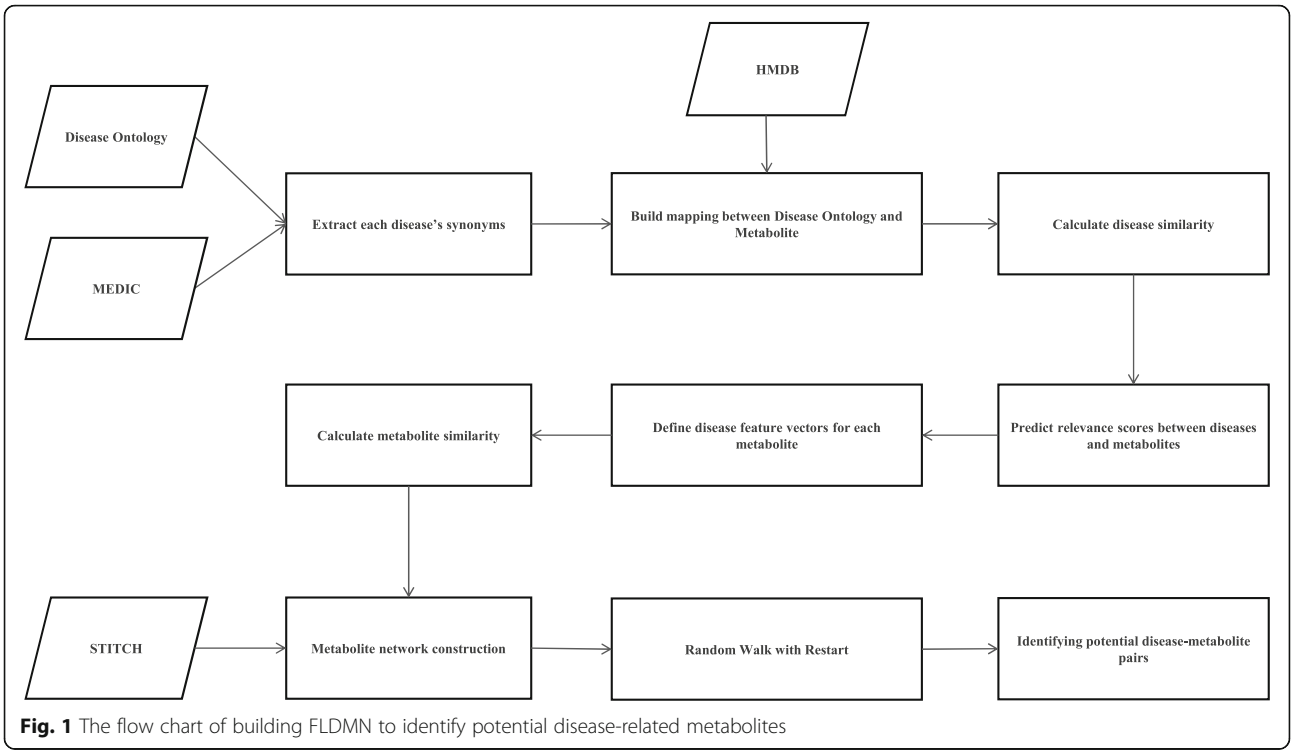

Fig. 1 The flow chart of building FLDMN to identify potential disease-related metabolites

subset of descriptors from "Diseases" category of Medical Subject Headings (MeSH) [37]. MEDIC is used to curate gene–disease and chemical–disease associations in CTD. In this study, we will use the "Synonyms" field and the "DiseaseName" field of MEDIC as a part of a combined vocabulary for mapping.

The Human Disease Ontology (DO) [33] is a community driven standards-based ontology that is focused on representing common and rare disease concept, which provides researchers with an open source ontology for the integration of biomedical data that is associated with human disease. The content of each disease in DO is a node, which has a parent-child relationship with others. All of these nodes are organized in a directed acyclic graph (DAG) with an 'IS_A' relationship. In this study, terms of DO will also be utilized as a part of a combined vocabulary. Finally, we use this combined vocabulary to annotate DO with metabolite-related diseases.

Human metabolome database

The Human Metabolome Database (HMDB) [35] is a freely metabolome database with detailed information about small molecule metabolites in the human body. Currently, HMDB involves 11,400 metabolites, which contains 835 disease associated metabolites with 825 diseases. In this study, we will use two different versions of HMDB, which released in Dec. 2017 and Apr. 2018, for constructing the metabolite network and extracting testing sets, respectively.

STITCH [32] is a database of known and predicted interactions between chemicals and proteins, which interactions includes direct and indirect associations. Currently, STITCH database contains 9,643,763 proteins from 2031 organisms. In this study, eligible interactions from STITCH are involved in the reconstruction of metabolic network.

Methods of the metabolite network construction

Mappings between diseases and metabolites

The xml file which contains information about metabolites can be found in HMDB web site. However, we find that these diseases in HMDB don't have any mapping with DO terms when this file is parsed. Therefore, we build mappings between DO terms and diseases in HMDB. As comprehensive disease corpuses, MEDIC and DO both contain abundant disease terms. First, we parse the HMDB file to get disease-related metabolites. Then we annotate DO entries with the terms from MEDIC and create a combined vocabulary of disease terms. Finally, mappings between DO terms and diseases in HMDB are built reference to this combined vocabulary.

Metabolite similarity calculation based on modified collaborative filtering

As one of the most successful technologies for recommender systems [38], collaborative filtering has been developed and improved over the past decade. In this

study, we define associations between metabolites based on modified collaborative filtering. In order to achieve this, each metabolite can be seen as a vector, which dimension is defined as the number of diseases. Through mapping diseases to metabolites, we obtain a set of initial vectors. Disease similarities are employed to predict the score of one dimension in a vector when there is no score in this dimension. Finally, we calculate similarities between vectors by cosine measure. The workflow for calculating metabolite similarity is shown in Fig. 2.

Metabolite-related diseases similarity After obtaining metabolite-related diseases, we calculate similarities between these diseases as inputs of further predicting relevance scores. In this paper, the method named FNSemSim [29], which we previously developed, is utilized to calculate disease similarities. This method measures the similarity of diseases by a fused gene functional network of HumanNet [39] and FunCoup [40]. Through assessment the method has good performance for calculating similarities between diseases.

Let a pair of gene sets $G_a = \{g_{a1}, g_{a2}, \ldots\}$ and $G_b = \{g_{b1}, g_{b2}, \ldots\}$ be related to disease d_a and d_b, respectively. The similarity between disease d_a and d_b is defined as follows:

$$DiseaseFunSim(G_a, G_b) = \frac{\sum_{1 \le i \le num(G_a)} R_{G_b}(g_{ai}) + \sum_{1 \le j \le num(G_b)} R_{G_a}(g_{bj})}{|G_a| + |G_b|}$$
$$g_{ai} \in G_a, g_{bj} \in G_b \quad (1)$$

where $|G_a|$ and $|G_b|$ respectively represents the numbers of genes related to disease d_a and d_b; and $R_G(g)$ represents the connection weights in the fused functional association network (see details in [29]). Finally, FNSemSim could be defined as follows:

$$FNSemSim(d_a, d_b) = DiseaseFunSim(G_a, G_b) * \frac{|G_a||G_b|}{|G_{MICA}||G_{MICA}|} \quad (2)$$

where $|G_a|$ and $|G_b|$ represent the size of two gene sets, G_a and G_b, related to disease d_a and d_b in Disease Ontology, respectively; $|G_{MICA}|$ represents the number of genes related to the most informative common ancestor of d_a and d_b. Finally, we normalize similarities between pair-wised diseases associated with metabolites.

Relevance scores between diseases and metabolites We utilize the similarities between diseases associated with metabolites to predict the relevance score of a

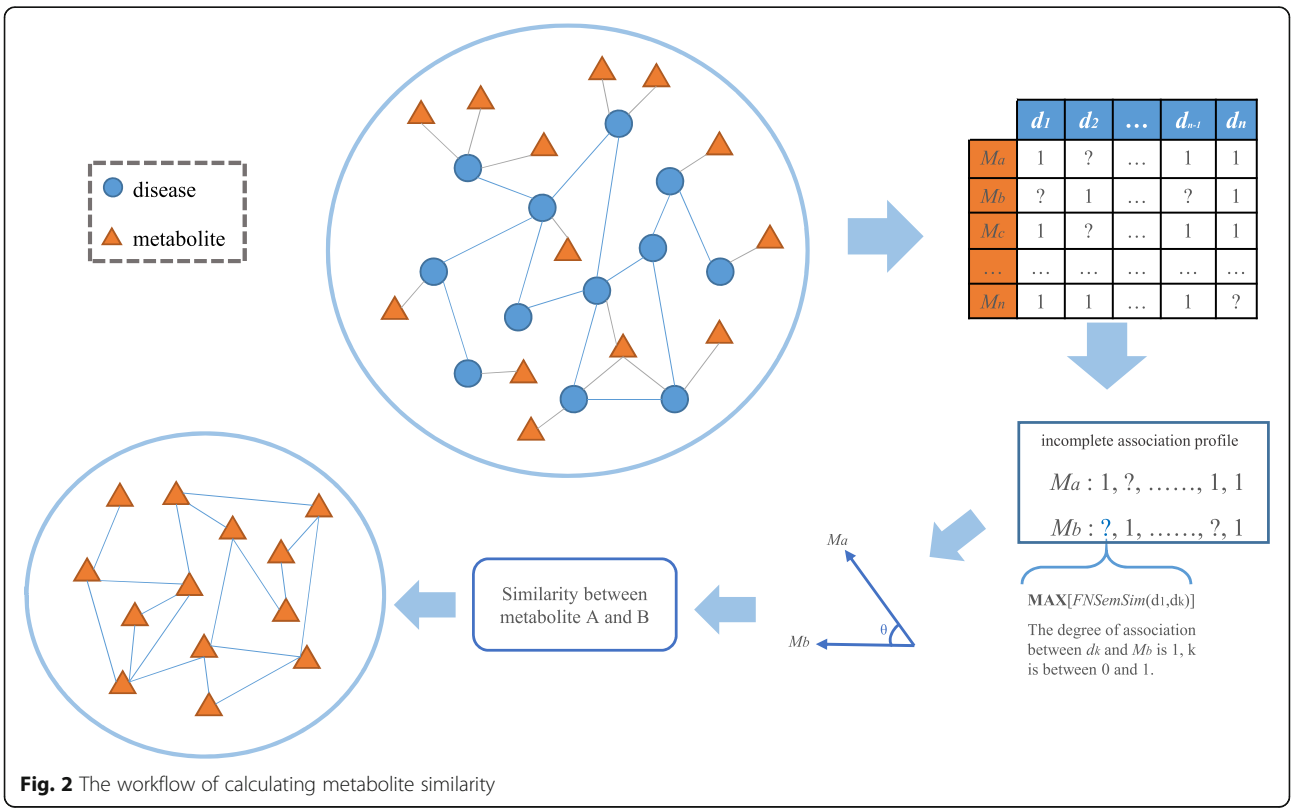

Fig. 2 The workflow of calculating metabolite similarity

disease that is not directly related to one metabolite. We define M and D as the set of metabolites and the set of related diseases, respectively. DR_m is defined as the set of diseases directly related to metabolite m. The predicted association score between disease d and metabolite m is defined as follows:

$$PA(d,m) = \begin{cases} MAX(FNSemSim(di,d)) & di \in DRm \text{ and } d \notin DRm \\ 1 & d \in DRm \end{cases} \quad (3)$$

where $m \in M$, $d \in D$, $DRm \subseteq D$ and $1 \le i \le |DRm|$; here, $|DRm|$ represents the number of diseases in the set of DR_m. We define a vector of each metabolite with $|D|$ dimension, respectively. $|D|$ represents the size of the disease set D. For each metabolite we can define its vector $\overrightarrow{m}$ as follows:

$$\overrightarrow{m} = \underset{m \in M,\ 1 \le k \le |D|}{(PA(d1,m)\ ,\ \cdots\cdots,\ PA(dk,m))} \quad (4)$$

where $|D|$ represents the size of the disease set D; $\overrightarrow{m}$ represents the score vector of metabolite m; and $PA(d\mathrm{k}, m)$ is the score between disease d_k and metabolite m. Now, we can obtain $|M|$ vectors of metabolites related to diseases.

Metabolite similarity Because each metabolite can be depicted by a multi-dimensional vector, we can find associations between metabolites in multi-dimensional space that is composed of metabolite-related diseases. Thus, we use cosine measure to calculate the similarity between any two vectors of metabolites. The association between metabolite m_1 and metabolite m_2 is defined as follows:

$$\mathrm{DMN}(\mathrm{m}_1, \mathrm{m}_2) = \frac{\sum_1^{\mathrm{n}} (PA_{1,i} \times PA_{2,\mathrm{i}})}{\sqrt{\sum_1^{\mathrm{n}} \mathrm{PA}_{1,i}^2} \times \sqrt{\sum_1^{\mathrm{n}} \mathrm{PA}_{2,i}^2}} \quad (5)$$

where $PA_{k,i}$ represents the association score between metabolite m_k and disease d_i in the i-th dimension of the vector $\overrightarrow{m}_{\mathrm{k}}$. The range of $DMN(m_1, m_2)$ is 0 to 1 because these values in all dimensions are positive numbers. Finally, we obtain all associations of pair-wised metabolites related to diseases and build a disease-associated metabolite network (DMN).

Metabolite network reconstruction In DMN, there exist only functional associations between disease-related metabolites, because all the links in this network are created by taking special phenotype as a measure. In view of this, we take metabolite related literatures as weight and extract text mining scores from STITCH to improve associations between metabolites in DMN. The combined weight of metabolite m_1 and metabolite m_2 is defined in Eq. 6.

$$\mathrm{FLDMN}(m_1, m_2) = 1\text{-}(1\text{-}DMN(m_1, m_2))(1\text{-}ST(m_1, m_2)) \quad (6)$$

where $ST(m_1, m_2)$ represents the text mining score of metabolite m_1 and metabolite m_2 in STITCH. The ranges of $ST(m_1, m_2)$ and $DMN(m_1, m_2)$ are both 0 to 1. Finally, we utilize the new associations between metabolites to reconstruct the network FLDMN.

Identifying novel candidate disease-related metabolites

The associations between a metabolite and its first neighbours are shown in FLMDN, but those between it and all the others in this network are ignored. To identify novel candidate disease-related metabolites by fully exploiting the global functional similarities of metabolites in this metabolite network, we employ RWR [41] to make relationship mining between any two metabolites in this network.

As a global optimization method, RWR can output more information about one metabolite and all the others in the network. In a network, the random walker starts from the root node and migrate to neighbouring nodes with the probabilities from that node to the others. After several iterations, the probabilities from the root node to all the other nodes in this network will become stable. Because RWR is a popular method based on graph structure, we do not repeat it here (see [41] for RWR details). Finally, we can obtain a rank for each metabolite in this network by RWR.

Results

Metabolites and diseases

We extracted 8704 disease terms from Disease Ontology (released in Nov. 2017) and calculated similarities between them. Those pair-wised diseases whose similarities are zero removed, there remain 3,801,586 associations among 4703 disease terms. Meanwhile, we found 1406 relationships between diseases and metabolites when DO terms are mapped to the diseases in HMDB (released in Dec. 2017) referring to the combined vocabulary of DO and MEDIC. Two hundred forty-eight diseases and 600 metabolites are totally contained.

We calculated 197,700 similarities among these 600 metabolites, and found that there were a large number of very weak associations and 25,709 irrelevant pair-wised metabolites in these results. To reduce the influence of noise on the network, we analyzed the distribution of metabolite associations with different similarities as thresholds. As we can see in Fig. 3, the number of associations is declining while the threshold is increasing. And when the threshold approaches

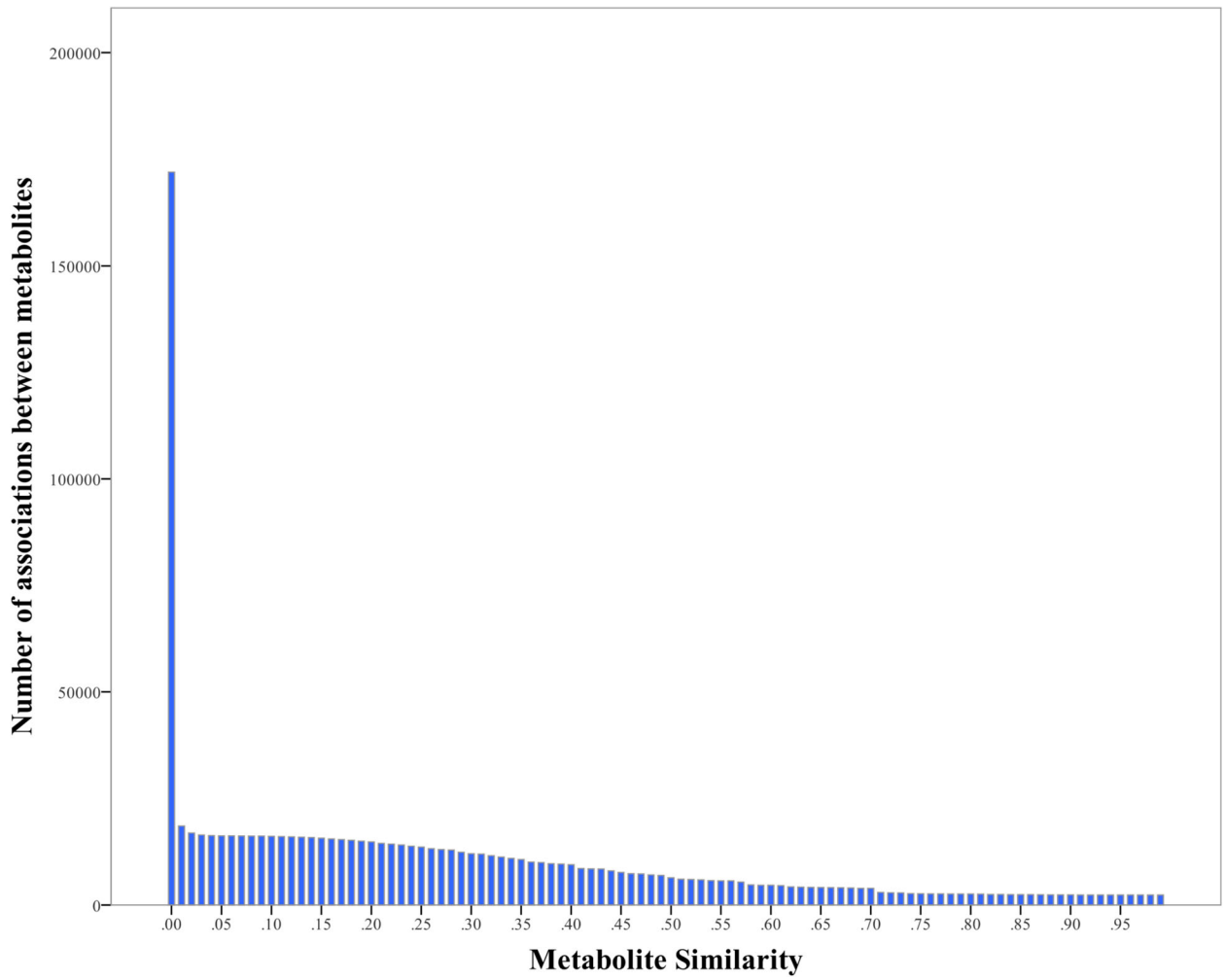

Fig. 3 The quantities distribution of metabolite associations with different similarities

0.01, the alternation in the number of associations levels off. Therefore, we filtered out 153,455 associations with 0.01 as a threshold. Finally, we built the network DMN with 18,536 associations among 587 metabolites that are associated with 245 diseases.

To improve associations between metabolites in DMN, we extracted text mining scores of pair-wised metabolites from STITCH. We obtained a network named ST_SUBNET composed of 28,292 associations among 485 metabolites from STITCH. Finally, FLDMN is built and contains 28,715 associations among 587 metabolites associated with 245 diseases.

Performance

To assess the performance of DMN and FLDMN, we performed a validation with 78 known disease metabolites associated with 19 diseases obtained from HMDB (released in Apr. 2018). But these known disease-related metabolites had no association with these 19 diseases in HMDB (released in Dec. 2017). The detailed statistics for evaluating disease-related metabolite networks are given in Table 1. For each disease, all of tested metabolites, which exist in the version of both 2017 and 2018 like other metabolites involving in the performance evaluation, only have associations with this disease in the version of 2018.

As a result, the average AUC (area under the receiver operating characteristic curve) of DMN for 19 diseases reached 64.35%. And FLDMN was proved to be successful in predicting novel metabolic signatures for 19 diseases with an average AUC value of 76.03%. Meanwhile, we also assessed ST_SUBNET to figure out whether the excellent performance of FLDMN is only due to ST_SUBNET. As shown in Fig. 4, the average AUC of ST_SUBNET reached 62.3% that was a little lower than DMN. This illustrates that the performance of FLDMN is the combined effect of DMN and ST_SUBNET. But AUC of ST_SUBNET doesn't mean that

Table 1 Statistics for evaluating disease-related metabolite network

Disease name	Disease Ontology	Test node	Positive group	Version 2017	Version 2018
L-2-hydroxyglutaric aciduria	DOID:0050574	2	4	2	4
medium chain acyl-CoA dehydrogenase deficiency	DOID:0080153	1	16	15	16
short chain acyl-CoA dehydrogenase deficiency	DOID:0080154	1	4	3	4
Crohn's colitis	DOID:0060192	5	8	3	16
cerebrotendinous xanthomatosis	DOID:4810	5	7	2	9
maple syrup urine disease	DOID:9269	6	23	17	24
abetalipoproteinemia	DOID:1386	3	4	1	4
celiac disease	DOID:10608	11	22	11	82
methylmalonic acidemia	DOID:14749	1	2	1	2
irritable bowel syndrome	DOID:9778	5	7	2	15
Fanconi syndrome	DOID:1062	1	2	1	5
citrullinemia	DOID:9273	6	8	2	8
inflammatory bowel disease 1	DOID:0110892	5	8	3	16
isovaleric acidemia	DOID:14753	2	11	9	12
type 2 diabetes mellitus	DOID:9352	1	27	27	27
aromatic L-amino acid decarboxylase deficiency	DOID:0090123	2	9	7	12
cholesterol ester storage disease	DOID:14502	1	2	1	2
congenital adrenal hyperplasia	DOID:0050811	15	18	3	28
Crohn's disease	DOID:8778	5	8	3	16

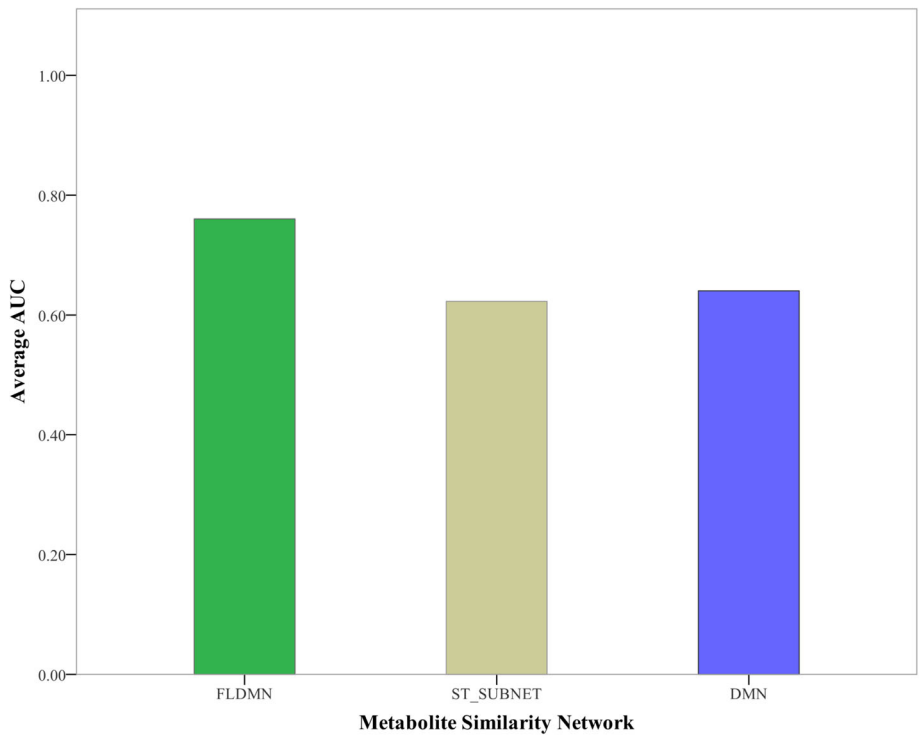

Fig. 4 Average AUC of three metabolite networks. The average AUC of FLDMN reaches 76.03%, while the average AUC of ST_SUBNET is 62.3% and DMN has an average AUC value of 76.03%

STITCH has an average performance because ST_SUBNET is only a small part of its.

We found that our method had outstanding performance on some diseases. For example, short chain acyl-CoA dehydrogenase deficiency (DOID:0080154) had an AUC of 97.68% in DMN, and there were 9 diseases whose AUC were more than 80% in FLDMN, as shown in Fig. 5. Among these 19 diseases, AUC of celiac disease (DOID:10608) in DMN was 48.1% while it reached 59.4% in FLDMN. As we can see in Table 1, the number of metabolites associated with celiac disease in HMDB (2018) was 71 more than that in 2017 version, while there were 22 positive samples for 11 test nodes. It implies that the relatively small number of positive samples could affect the result of predicting candidate metabolites related with celiac disease. In addition, the AUC of medium chain acyl-CoA dehydrogenase deficiency (DOID:0080153) was smaller in FLDMN than in DMN. Part of it may be the fact that some noise is introduced by ST_SUBNET. But in general, the performance of FLDMN is outstanding in predicting candidate disease-related metabolites.

Case study

We used Alzheimer's disease (DOID:10652) as one of case studies to further evaluate the performance of our computational model in predicting potential disease-related metabolites. First, we utilized the metabolite data from HMDB (released in Apr. 2018) to build FLDMN. We employed RWR and found that S-Adenosylhomocysteine (HMDB0000939) had a high score of 0.91 for Alzheimer's disease, which was ranked in top 3%. But the relationship between S-Adenosylhomocysteine and Alzheimer's disease was not included in HMDB (released in Apr. 2018). S-Adenosylhomocysteine has been demonstrated to be related to Alzheimer's disease [42]. L-Cysteine (HMDB0000574) was ranked in top 5% for Alzheimer's disease, which was a naturally occurring, sulfur-containingamino acid. It has been reported as a potentially metabolic intermediary of Alzheimer's disease [43]. Substance P (HMDB0001897), an 11-amino acid neuropeptide, was ranked in top 10% for Alzheimer's disease. Rosler, N et al. [44] have found that AD patients with late disease onset showed significantly higher values of Substance P than early onset patients.

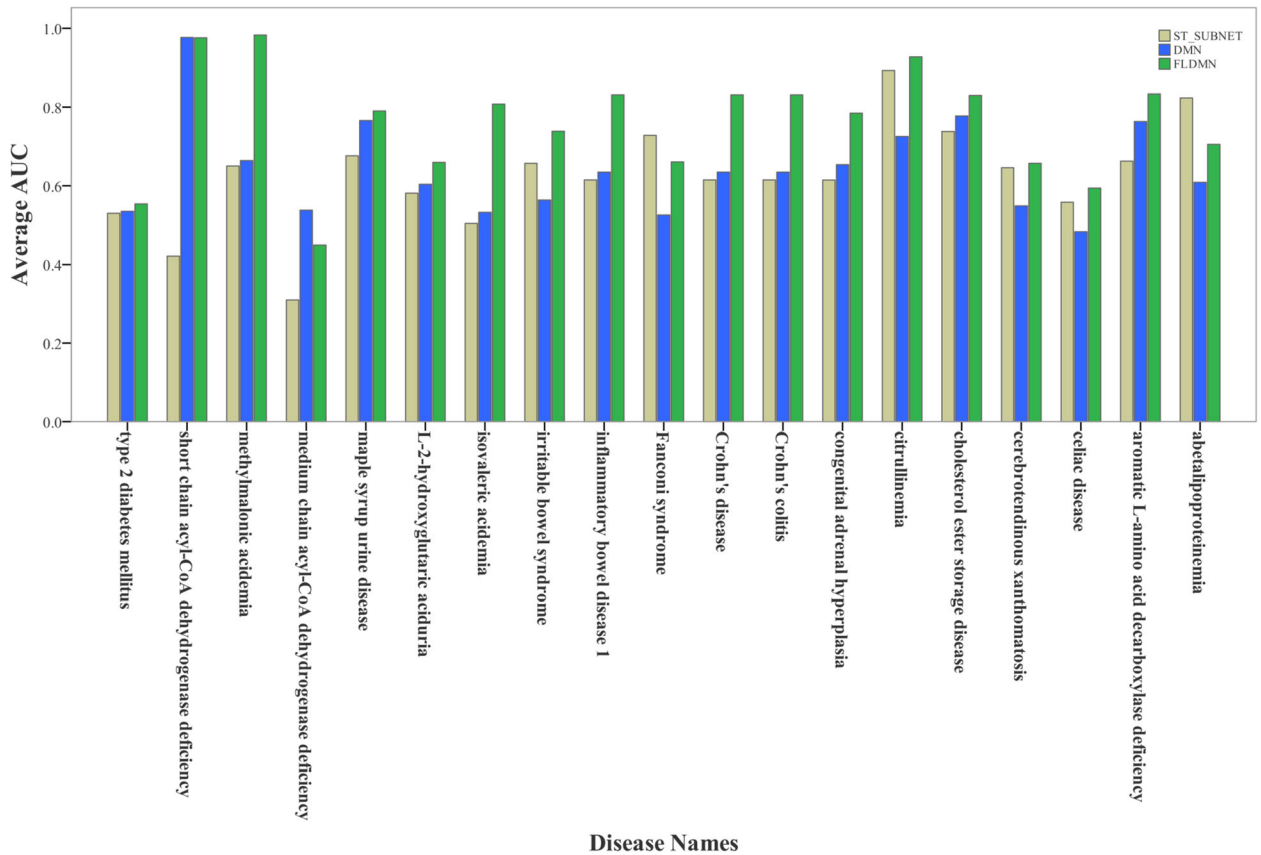

Fig. 5 Performances of three metabolite networks to predict candidate metabolites related with a given disease. ST_SUBNET, DMN and FLDMN are utilized to predict candidate metabolites for each of these 19 diseases, respectively. For a given disease, the three bars with different colours represent average AUC of the three metabolite networks, correspondingly

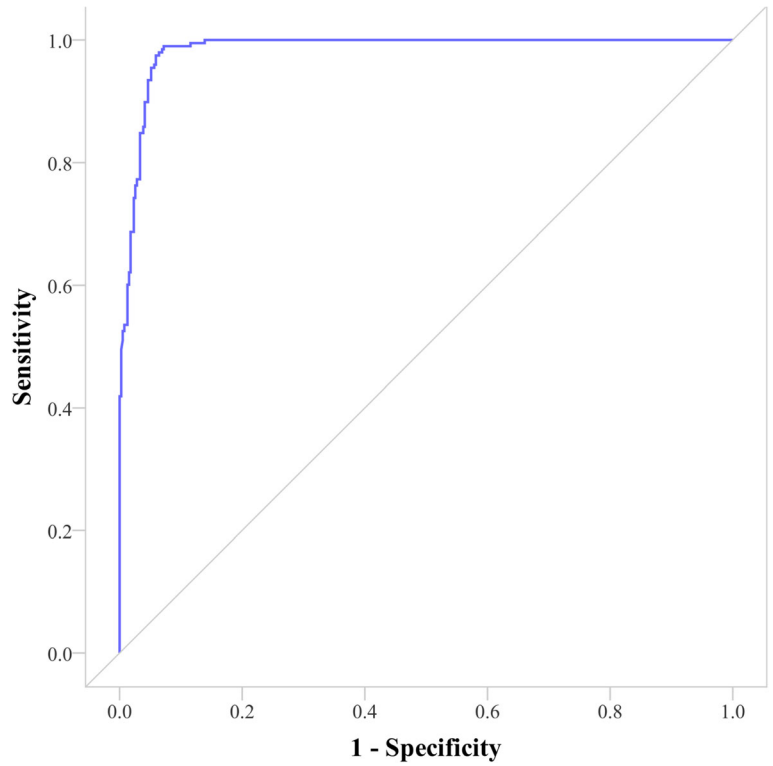

Fig. 6 ROC of L-Threonine based on FLDMN

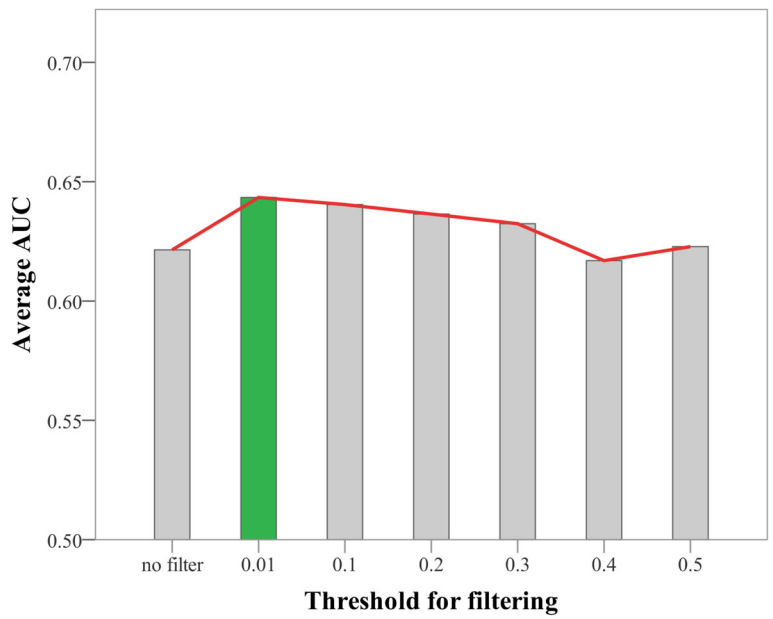

Fig. 7 Average AUC of DMN with different thresholds

We also found potential metabolites related to leukemia (DOID:1240). Putrescine (HMDB0001414), N8-Acetylspermidine (HMDB0002189) and N1-Acetylspermidine (HMDB0001276) were ranked in top 5% for leukemia, which were well documented to be associated with leukemia [45]. Type 1 diabetes mellitus (DOID:9744) is characterized by loss of the insulin-producing beta cells of the pancreatic islets, leading to insulin deficiency. We also applied this method to type 1 diabetes mellitus to find potential some metabolites. Pyruvic acid (HMDB0000243), 3-Hydroxyisovaleric acid (HMDB0000754), Dimethylamine (HMDB0000087) and Citric acid (HMDB0000094) were ranked in top 20% for type 1 diabetes mellitus, and these metabolites were reported in the study of type 1 diabetes mellitus [46].

Discussion

We identified candidate metabolites related with a certain disease in FLDMN and used AUC to measure its performance. We can also use FLDMN to prioritizing candidate disease-related metabolites for a certain metabolite. The AUC will be better. Because most of metabolites often associate with more than one diseases, positive samples will get more for a certain metabolite. Therefore, they can be tested first in the rank for this metabolite in FLDMN. Take L-Threonine (HMDB0000167) as an example. There are totally seven metabolites that have disease-related associations with it. As we can see in Fig. 6, its AUC was 98.44%. Subsequently, we used these above-mentioned test nodes as a target node respectively to rank candidate disease-related metabolites. The average AUC was 89.88%.

When DMN was built, the threshold was set as 0.01 to filter weak links. We did some experiments later to figure out whether the threshold was reasonable. There were seven networks with different thresholds to be constructed, respectively. As shown in Fig. 7, the average AUC of DMN with 0.01 as a threshold was outstanding. Therefore, there will be better results to use 0.01 as a threshold.

Conclusions

As the link between genotypes and phenotypes, metabolites can be used to explain the underlying molecular disease-causing mechanisms. For this purpose, we proposed a computational model to build a disease-related metabolite network and identified candidate metabolites related to diseases.

First, we used FNSemSim to calculate similarities of pair-wised diseases. Subsequently, we defined associations between metabolites by modified collaborative filtering and built a disease associated metabolite network (DMN). To improve these associations, a new disease associated metabolite network by fusing functional associations and scores of literatures (FLDMN) was constructed. Finally, we used RWR to prioritize candidate disease-related metabolites.

The results showed that our method was proved to be successful in predicting novel metabolic signatures for 19 diseases with an average AUC value of 76.03%. And it will be helpful for researchers in metabolomics. Take Alzheimer's disease and leukemia as examples, we found some unknown metabolites that were mapped to these diseases through our network.

Abbreviations

AUC: Area under the receiver operating characteristic curve; CF: Collaborative filtering; CTD: Comparative toxicogenomics database; DAG: Directed acyclic graph; DO: Disease ontology; HMDB: Human Metabolome Database; MEDIC: Merged disease vocabulary; MeSH: Medical Subject Headings; RWR: Random walking with restart; STITCH: Search tool for interactions of chemicals

Acknowledgements

Tianyi Zang and Yadong Wang are the corresponding authors. We thank them for their guidance. Thank Ling Wang, Rui Ma, Yanshuo Chu, Zhenxing Wang for their valuable suggestions on our work.

About this supplement

This article has been published as part of *BMC Bioinformatics Volume 20 Supplement 18, 2019: Selected articles from the Biological Ontologies and Knowledge bases workshop 2018*. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-20-supplement-18.

Authors' contributions

JL collected experimental data and WYT did data preprocessing. With the guidance of ZTY and WYD, WYT finished the algorithm design and validation. WYT and PJ were the major contributors in writing the manuscript. All authors have read and approved the final version of the manuscript.

Funding

The publication cost of this article was funded by the Major State Research Development Program of China [No: 2016YFC0901605, 2016YFC1201702–01], the National Natural Science Foundation of China (No: 61571152, 31601072), the National High-tech R&D Program of China (863 Program) [Nos: 2012AA02A601, 2015AA020101, 2015AA020108].

Availability of data and materials

All the datasets used in this paper could be downloaded from websites.

Ethics approval and consent to participate

Not Applicable.

Consent for publication

Not Applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, People's Republic of China. [2]School of Life Science and Technology, Harbin Institute of Technology, Harbin 150001, People's Republic of China. [3]School of Computer Science, Northwestern Polytechnical University, Xi'an, People's Republic of China.

Published: 25 November 2019

References

1. Pickrell JK, Al E. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010;464(7289):768–72.

2. Sun SQ, Zhu JQ, Mozaffari S, Ober C, Chen MJ, Zhou X. Heritability estimation and differential analysis of count data with generalized linear mixed models in genomic sequencing studies. Bioinformatics. 2019;35(3): 487–96.
3. Peng J, Wang X, Shang X. Combining gene ontology with deep neural networks to enhance the clustering of single cell RNA-Seq data. BMC Bioinformatics. 2019;20(8):284.
4. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5(7):621.
5. Zhu X, Gerstein M, Snyder M. Getting connected: analysis and principles of biological networks. Genes Dev. 2007;21(9):1010–24.
6. Lee DS, Park J, Kay KA, Christakis NA, Oltvai ZN, Barabási AL. The implications of human metabolic network topology for disease comorbidity. Proc Natl Acad Sci U S A. 2008;105(29):9880–5.
7. ME C, N K, M V, DE H: Interactome: gateway into systems biology. Hum Mol Genet. 2005, 14 Spec No. 2(suppl_2):R171–181.
8. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 2006;34(Database issue):D354.
9. Fiehn O. Metabolomics – the link between genotypes and phenotypes. Plant Mol Biol. 2002;48(2):155–71.
10. De PV. Metabonomics and systems biology. Methods Mol Biol. 2015;1277:245.
11. Holmes E, Wilson ID, Nicholson JK. Metabolic phenotyping in health and disease. Cell. 2008;134(5):714–7.
12. Nordström A, Lewensohn R. Metabolomics: moving to the clinic. J Neuroimmune Pharmacol. 2010;5(1):4–17.
13. Shao Y, Chen L, Lu R, Zhang X, Xiao B, Ye G, Guo J. Decreased expression of hsa_circ_0001895 in human gastric cancer and its clinical significances. Tumour Biol J Int Soc Oncodev Biol Med. 2017;39(4):1010428317699125.
14. Cheng L, Hu Y, Sun J, Zhou M, Jiang Q. DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function. Bioinformatics. 2018;34(11):1953–6.
15. Zhao ZJ, Shen J. Circular RNA participates in the carcinogenesis and the malignant behavior of cancer. RNA Biol. 2015;14(5):00.
16. Xia S, Feng J, Chen K, Ma Y, Gong J, Cai F, Jin Y, Gao Y, Xia L, Chang H. CSCD: a database for cancer-specific circular RNAs. Nucleic Acids Res. 2018; 46(Database issue):D925–9.
17. Peng J, Wang H, Lu J, Hui W, Wang Y, Shang X. Identifying term relations cross different gene ontology categories. BMC Bioinformatics. 2017;18(Suppl 16):573.
18. Rainer B, Shawn R, Dayan G, Stewart ML, Barrett MP. Ab initioprediction of metabolic networks using Fourier transform mass spectrometry data. Metabolomics. 2006;2(3):155–64.
19. Gao J, Tarcea VG, Karnovsky A, Mirel BR, Weymouth TE, Beecher CW, Cavalcoli JD, Athey BD, Omenn GS, Burant CF. Metscape: a Cytoscape plug-in for visualizing and interpreting metabolomic data in the context of human metabolic networks. Bioinformatics. 2010;26(7):971–3.
20. Feng L, Xu Y, Shang D, Yang H, Wei L, Han J, Sun Z, Yao Q, Zhang C, Ma J. MPINet: metabolite pathway identification via coupling of global metabolite network structure and Metabolomic profile. Biomed Res Int. 2014;2014(1):325697.
21. Sergushichev AA, Loboda AA, Jha AK, Vincent EE, Driggers EM, Jones RG, Pearce EJ, Artyomov MN. GAM: a web-service for integrated transcriptional and metabolic network analysis. Nucleic Acids Res. 2016;44(Web Server issue):W194–200.
22. Wang Z, Lin Y, Liang J, Huang Y, Ma C, Liu X, Yang J. NMR-based metabolomic techniques identify potential urinary biomarkers for early colorectal cancer detection. Oncotarget. 2017;8(62):105819–31.
23. O'Hagan S, Kell DB. Analysis of drug–endogenous human metabolite similarities in terms of their maximum common substructures. J Cheminformatics. 2017;9(1):18.
24. Ohtana Y, Abdullah AA, Altaf-Ul-Amin M, Huang M, Ono N, Sato T, Sugiura T, Horai H, Nakamura Y, Morita HA. Clustering of 3D-structure similarity based network of secondary metabolites reveals their relationships with biological activities. Mol Informatics. 2014;33(11–12):790–801.
25. Iqbal K, Dietrich S, Wittenbecher C, Krumsiek J, Kuhn T, Lacruz ME, Kluttig A, Prehn C, Adamski J, von Bergen M, et al. Comparison of metabolite networks from four German population-based studies. Int J Epidemiol. 2018;47(6):2070–81.
26. Yao Q, Xu Y, Yang H, Shang D, Zhang C, Zhang Y, Sun Z, Shi X, Feng L, Han J. Global prioritization of disease candidate metabolites based on a multi-omics composite network. Sci Rep. 2015;5:17201.
27. Peng J, Hui W, Shang X. Measuring phenotype-phenotype similarity through the interactome. BMC Bioinformatics. 2018;19(Suppl 5):114.
28. Peng J, Zhang X, Hui W, Lu J, Li Q, Liu S, Shang X. Improving the measurement of semantic similarity by combining gene ontology and co-functional network: a random walk based approach. BMC Syst Biol. 2018;12(2):18.
29. Wang Y, Juan L, Chu Y, Wang R, Zang T, Wang Y. FNSemSim: an improved disease similarity method based on network fusion. In: IEEE International Conference on Bioinformatics and Biomedicine; 2017. p. 630–3.
30. Peng J, Zhu L, Wang Y, Chen J: Mining relationships among multiple entities in biological networks. IEEE/ACM transactions on computational biology and bioinformatics 2019.
31. Ricci F, Rokach L, Shapira B. Introduction to recommender systems handbook. In: Recommender systems handbook. Springer; 2011. p. 1–35.
32. Damian S, Alberto S, Christian VM, Juhl JL, Peer B, Michael K. STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data. Nucleic Acids Res. 2016;44(Database issue):D380–4.
33. Kibbe WA, Arze C, Felix V, Mitraka E, Bolton E, Fu G, Mungall CJ, Binder JX, Malone J, Vasant D. Disease ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic Acids Res. 2015;43(Database issue):1071–8.
34. Davis AP, Wiegers TC, Rosenstein MC, Mattingly CJ. MEDIC: a practical disease vocabulary used at the Comparative Toxicogenomics Database. Database. 2012;2012:bar065.
35. Wishart DS, Feunang YD, Marcu A, Guo AC, Liang K, Vázquezfresno R, Sajed T, Johnson D, Li C, Karu N. HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Res. 2017;46(Database issue):D608–17.
36. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, King BL, McMorran R, Wiegers J, Wiegers TC, Mattingly CJ. The comparative Toxicogenomics database: update 2017. Nucleic Acids Res. 2017;45(D1):D972–8.
37. Lipscomb CE. Medical subject headings (MeSH). Bull Med Libr Assoc. 2000;88(3):265–6.
38. Herlocker JL. Evaluating collaborative filtering recommender systems. In: The adaptive web; 2011. p. 291–324.
39. Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res. 2011;21(7):1109.
40. Schmitt T, Ogris C, ELL S. FunCoup 3.0: database of genome-wide functional coupling networks. Nucleic Acids Res. 2014;42(Database issue):D380.
41. Tong H, Faloutsos C, Pan JY. Fast random walk with restart and its applications. In: International Conference on Data Mining; 2006. p. 613–22.
42. Linnebank M, Popp J, Smulders Y, Smith D, Semmler A, Farkas M, Kulic L, Cvetanovska G, Blom H, Stoffel-Wagner B. S-Adenosylmethionine Is Decreased in the Cerebrospinal Fluid of Patients with Alzheimeru2019s Disease. Neurodegener Dis. 2010;7(6):373–8.
43. Fonteh AN, Harrington RJ, Tsai A, Liao P, Harrington MG. Free amino acid and dipeptide changes in the body fluids from Alzheimer's disease subjects. Amino Acids. 2007;32(2):213.
44. Rosler N, Wichart I, Jellinger KA. Clinical significance of neurobiochemical profiles in the lumbar cerebrospinal fluid of Alzheimer's disease patients. J Neural Transm. 2001;108(2):231–46.
45. Lee SH, Suh JW, Chung BC, Kim SO. Polyamine profiles in the urine of patients with leukemia. Cancer Lett. 1998;122(1–2):1–8.
46. ŞTEFAN LI, Nicolescu A, Popa S, MOŢA M, Kovacs E, Deleanu C. 1H-NMR urine metabolic profiling in type 1 diabetes mellitus. Rev Roum Chim. 2010; 55(11–12):1033–7.

Publisher's Note