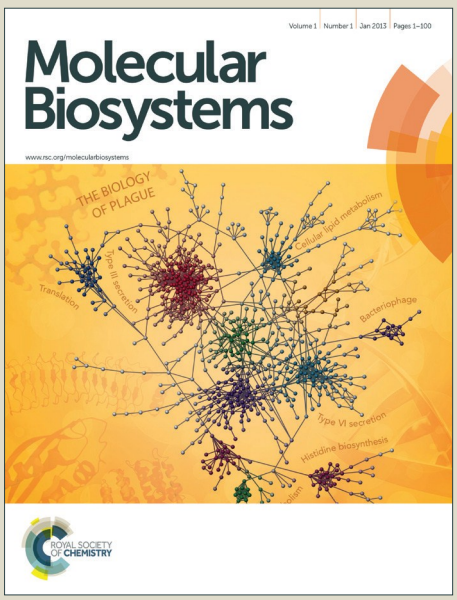
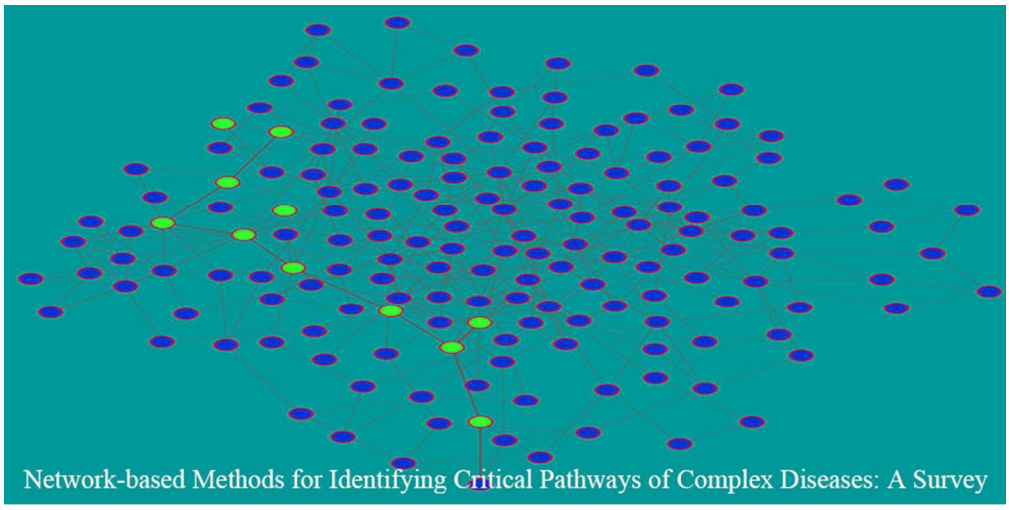
Molecular BioSystems

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems



80x39mm (300 x 300 DPI)

Molecular BioSystems

ARTICLE

Network-based Methods for Identifying Critical Pathways of Complex Diseases: A Survey

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

Qiaosheng Zhang[a,b], Jie Li[a,*], Hanqing Xue[a], Leilei Kong[c], Yadong Wang[a,*]

Biological pathways play important role in the development of complex diseases, such as cancers, a group of multifactorial complex diseases, are generally caused by mutation of multiple genes or dysregulation of pathways. It has become one of the most important issues to analyze pathway through combining multiple types of high-throughput data, such as genomics and proteomics, to understand the mechanism of complex diseases. Currently several network-based pathway analysis methods were proposed. In the overview, we reviewed seven major network-based pathway analysis methods and enumerate their benefits and limitations from an algorithmic perspective to provide a reference for the next generation of pathway analysis methods. Finally, we discuss the challenges that the next generation of methods face.

1 Introduction

Complex diseases, such as diabetes, cancers, heart diseases, hypertensive diseases, nerve system diseases, and so on, do not follow Mendel's law, are likely to be associated with the effects of multiple genes, proteins and biological pathways, which are different from single-gene diseases.[1] A biological pathway which plays an important role in understanding the mechanisms of complex diseases, improving clinical treatment, discovering drug target and biomarker, is a series of actions among molecules (including genes, gene products and compounds etc.) in a cell that leads to a certain product or a change in the cell.[2-5] During the past 10 years, several pathway knowledge databases are built, such as KEGG, BioCyc, MetaCyc, Reactome, RegulonDB and PantherDB.[6-11] The establishment of these knowledge bases laid the foundation for studying pathways and pathways' roles in the development of complex diseases. In addition, with the still-ongoing development of high-throughput sequencing technology for which the cost per reaction is falling dramatically. A large number of related-pathway omics data is growing exponentially.[12] Pathway-related knowledge databases and omics data contain a wealth of disease-related knowledge and information, such as information of the related-pathway genes, molecule interactions in the same pathway, topology structure of pathways, gene expression, and so on.

Not only a variety of data can be integrated together to identified the significant pathways of complex diseases, but also the significant pathways can be mapped to meaning biological process for better understanding the mechanisms of complex diseases. So how to effectively use these knowledge and data to build the model of pathway analysis to implement the interpretation of complex diseases, and to accelerate the understanding of complex diseases, drug development is an urgent issue. Researchers proposed several approaches to identify the critical pathways associated with complex diseases.[13-14] These methods can be divided into three categories: 1) Pathway-based gene set enrichment analysis; 2) Pathway-based functional class clustering and scoring approaches; 3) Network-based pathway approaches. For the first two categories of methods, only the quantity or gene expression information in a pathway is used, topology information available from pathway databases is ignored. In fact, genes or proteins are not independent; they perform a variety of functions or tasks through their interactions or connections. To take advantage of the pathway topology information to build pathway analysis model which reflecting the law of life activity, the third category of method is proposed, which includes SPIA (Signaling Pathway Impact Analysis), PARADIGM (Pathway Recognition Algorithm using Data Integration on Genomic Models), PathOlogist (Identification of Key Processes Underlying Cancer Phenotypes Using Biologic Pathway Analysis), Active Modules (Discovering regulatory and signaling circuits in molecular interaction networks), AMBIENT (Active Modules for Bipartite Networks), GIGA (Graph-based iterative Group Analysis enhances microarray interpretation) and nexus (Network—cross(X)-species—Search) etc.[15-21] Although network-based methods are widely used now, there are still some issues to be addressed with the further accumulation of high-throughput biological data. In this paper, we elaborated the characteristics of these methods to provide a reference for the development of better analysis ones.

[a] Harbin Institute of Technology.
[b] Heilongjiang Bayi Agricultural University.
[c] Heilongjiang Institute of Technology.
*Corresponding Author: jieli@hit.edu.cn
† Electronic Supplementary Information (ESI) available: Table S1.
See DOI: 10.1039/x0xx00000x

2 Network-based pathway analysis approaches

Although the different network-based pathway analysis approaches employ different data or networks, generally they have the similar analysis framework (Fig. 1) which includes the following steps: analyze differentially expressed genes, extract network topology, build the model of scoring approaches and identify the critical pathway. Here, we introduce the above seven kinds of network-based pathway analysis methods separately.

2.1 SPIA method

The change of a gene expression value is influenced by two factors: its own change and the change of other genes and molecules which are associated with the gene. Similarly, the change of pathway's state is also influenced by two factors: the change of molecular components in the pathway and the change of the interaction of molecular components in or close the pathway. Based on the above view, Tarca et al.[15] proposed SPIA method, in which these two factors are considered.

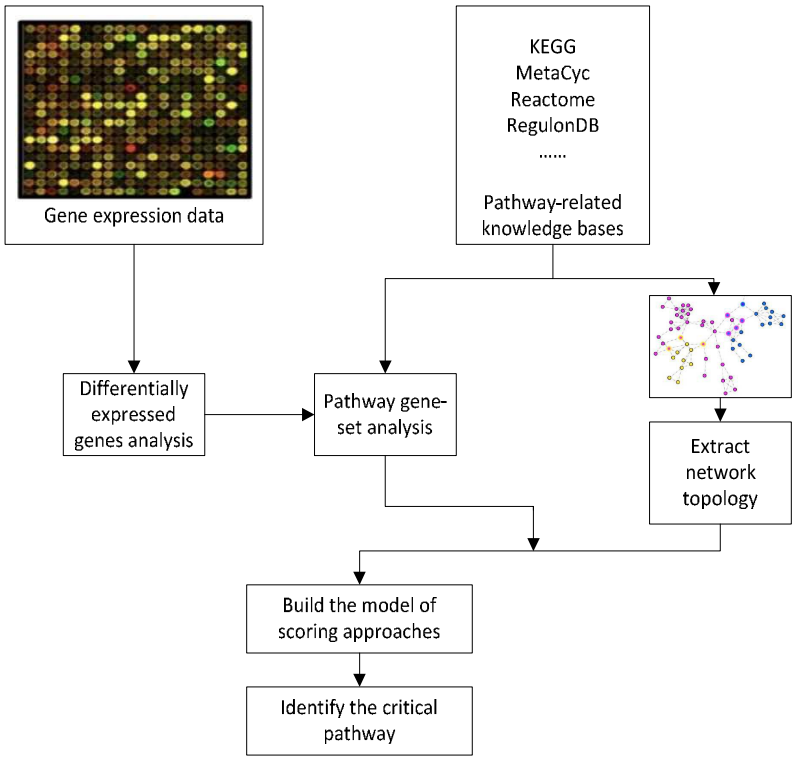


Fig. 1 The framework of network-based pathway analysis approaches

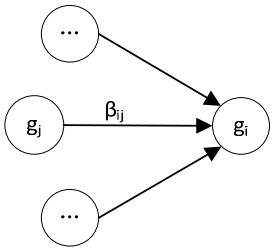


Fig. 2 The influence of gene g_j (j=1, 2, …) on its target gene g_i

Primarily, Tarca et al. defined the factor β_{ij} to quantify the strength of interaction between gene g_j and g_i (one target gene of g_j) (Fig. 2). The sign of β_{ij} represents the type of interaction: +1 for induction (activation), -1 for repression and inhibition, as described by each pathway. Then, they used $\Delta E(g_i)$ to represent the signed normalized measured expression change of the gene g_i. Finally, Tarca et al. put the two parts into a perturbation factor $PF(g_i)$ to describe the change of g_i expression value and the influence of directly upstream genes of g_i on g_i, normalized by the number of downstream genes of g_j, denoted by $N_{ds}(g_j)$.

$$PF(g_i) = \Delta E(g_i) + \sum_{j=1}^{n} \beta_{ij} \cdot \frac{PF(g_j)}{N_{ds}(g_j)}$$

Thus, for a pathway which includes n genes, it can be represented using a perturbation matrix *PF*.

$$PF = \begin{pmatrix} PF(g_1) \\ PF(g_2) \\ \cdots \\ PF(g_n) \end{pmatrix} = \begin{pmatrix} \Delta E(g_1) \\ \Delta E(g_2) \\ \cdots \\ \Delta E(g_n) \end{pmatrix} + \begin{pmatrix} \frac{\beta_{11}}{N_{ds}(g_1)} & \frac{\beta_{12}}{N_{ds}(g_2)} & \cdots & \frac{\beta_{1n}}{N_{ds}(g_n)} \\ \frac{\beta_{21}}{N_{ds}(g_1)} & \frac{\beta_{22}}{N_{ds}(g_2)} & \cdots & \frac{\beta_{2n}}{N_{ds}(g_n)} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\beta_{n1}}{N_{ds}(g_1)} & \frac{\beta_{n2}}{N_{ds}(g_2)} & \cdots & \frac{\beta_{nn}}{N_{ds}(g_n)} \end{pmatrix} \begin{pmatrix} PF(g_1) \\ PF(g_2) \\ \cdots \\ PF(g_n) \end{pmatrix}$$

In order to identify pathways with significant change in different diseases, Tarca et al. designed a score P_G to test the change of the *i*th pathway.

$$P_G = P_{NDE}(i) \cdot P_{PERT}(i) - P_{NDE}(i) \cdot P_{PERT}(i) \cdot \ln P_{NDE}(i) \cdot P_{PERT}(i) \quad (*)$$

Where $P_{NDE} = P(X \geq N_{DE} | H_0)$, X is a random variable that represents the number of differentially expressed genes in a pathway, N_{DE} is the number of differentially expressed genes in a specific pathway. $P_{PERT} = P(T_A \geq t_A | H_0)$, T_A is a random variable that represents the influence of upstream genes of the pathway, $t_A = \sum_i Effect(g_i)$, $Effect(g_i) = \sum_{j=1}^{n} \beta_{ij} \cdot \frac{PF(g_i)}{N_{ds}(g_i)}$, H_0 represents the null hypothesis.

Equation (*) was employed to identify critical pathway from data sets with 12 colorectal cancer samples and 10 normal samples.[22] Experimental results show that only the SPIA method find the colorectal cancer pathway which is significantly related with colorectal cancer, compared with GSEA.[23] This is possible due to additional evidence P_{PERT} which allows SPIA to find the colorectal cancer pathway, which is one of the outstanding characteristics of the method. In addition, SPIA has the increased sensitivity compared with GSEA, as well as improves specificity and better pathway ranking compared with ORA (Over-Representation Analysis).[13] Drawback is that only simple topology information is used and prior knowledge is not fully utilized.

2.2 PARADIGM

It is seen from the central dogma of molecular biology that explains that DNA codes for RNA, which codes for proteins. DNA is the molecule of heredity that passes from parents to offspring. It contains the instructions for building RNA and proteins, which make up the structure of the body and carry out most of specific biological functions. The central dogma of molecular biology illustrates the transmission of genetic information from DNA to RNA and then to proteins. This process involves many levels of information, therefore, if it can be used to analyze biological pathways, we can more accurately describe the activities of the pathway, and the results of the analysis are also more able to explain the pathogenesis of the disease from a biological point of view, but also to integrate more information. To take advantage of the central dogma of molecular biology, Vaske et al.[16] proposed PARADIGM method, in which pathway and its related information are transformed into discrete probability factor graph model used to identify significant pathway.

In PARADIGM method, to represent a biological pathway Vaske et al. defined five types of biological entities including protein-coding genes, small molecules, complexes, gene families and abstract processes to describe the transcription process of pathway components. The states of entities in a cell are described using variables and interactions between entities are represented using factors. Thus a pathway can be converted into a factor graph.

In PARADIGM method, for every variable x_i (the *i*th entity), the corresponding factor is $\varphi(X_i)$, where $X_i = \{x_i\} \cup \{\text{Parents}(x_i)\}$ and parents(x_i) refers to all the parents of x_i in factor graph; each entity take on one of three states corresponding to activated, nominal or deactivated relative to a control level and encoded as 1, 0 or -1 respectively, and the expected state of factor $\varphi(X_i)$: $\varphi_i(x_i, \text{Parents}(x_i))$ is specified as:

$$\varphi_i(x_i, Parents(x_i)) = \begin{cases} 1-\varepsilon & x_i \text{ is the expected state from Parents}(x_i) \\ \frac{\varepsilon}{2} & \text{otherwise} \end{cases}$$

Where ε is appointed to 0.001.

Based on above definitions, PARADIGM produces a matrix of integrated pathway activities (IPAs) to assess the significance of pathway:

$$IPAs = \begin{pmatrix} A_{11} & \cdots & A_{1n} \\ \cdots & A_{ij} & \cdots \\ A_{m1} & \cdots & A_{mn} \end{pmatrix}$$

Where A_{ij} represents the inferred activity of entity *i* in patient sample *j*. The matrix can then be used in place of the original constituent datasets to identify associations with clinical outcomes.

PARADIGM was compared with the SPIA[15] on the breast cancer data.[24] Only PARADIGM successfully identified the AKT1-related PI3K signaling pathway which is significantly associated with breast cancer.[25] This is due to PARADIGM's power to integrate part of the pathway-level interactive information and identify altered activities in cancer-related pathways compared to a competing pathway activity inference approach called SPIA. In addition, the PARADIGM algorithm

can be used to infer hidden quantities by combining multiple '-omics' data, and provide clues about the possible mechanisms underlying the differences in observed survival. The disadvantage of the PARADIGM is that it requires analysts who have enough biological knowledge to draw out entities with biological significance. Sedgewick et al. further improve PARADIGM method using Naive Bayesian assumption to reduce the computational complexity.[26]

2.3 PathOlogist

In order to take advantage of information of gene expression profiles and pathway topology, Efroni et al.[17,27] proposed PathOlogist method to analyze pathways. Two descriptive metrics: activity and consistency, are defined in PathOlogist. Activity scores provide a measure of how likely the interactions are to occur while consistency scores determine whether these interactions follow the logic of the defined network structure.

In PathOlogist, each gene is assumed to have two alternative states: "up" and "down" which follow gamma distribution γ_u and γ_d respectively, and the overall distribution of gene expression is considered to be a mixture of the two gamma distributions: γ_m.

$$\gamma_m = \eta_1\gamma_u + \eta_2\gamma_d, \quad \eta_1 + \eta_2 = 1$$

Where $\gamma_u = f(x|a_u,b_u) = \frac{1}{b_u^{a_u}\Gamma(a_u)} x^{a_u-1}e^{\frac{x}{b_u}}$, $\gamma_d = f(x|a_d,b_d) = \frac{1}{b_d^{a_d}\Gamma(a_d)} x^{a_d-1}e^{\frac{x}{b_d}}$, η_1 and η_2 are mixture coefficients, and $\eta_1 + \eta_2 = 1$, here, a_u (a_d) is shape parameter, b_u (b_d) is scale parameter.

For genes A, B and C in a given pathway (Fig. 3), Efroni et al. compute the prior probabilities of two states of gene A using following formulas:

$$P_A(Up) = \frac{N_u}{N_u + N_d}, \quad P_A(Down) = \frac{N_d}{N_u + N_d}$$

Where N_u and N_d are the number of genes in the "up" and "down" groups respectively. $P_B(Up)$, $P_B(Down)$, $P_C(Up)$ and $P_C(Down)$ can be get using similarly calculate. Then, activity and consistency scores are computed using the following process (Fig. 4).

The probability of the interaction being "active" is

$$Activity = p(A) \times p(B)$$

The probability of the interaction consistency is

$$Consistency = [Activity \times p(C)] + [(1 - Activity) \times (1 - p(C))]$$

The above calculation is scores of an interaction corresponding genes A, B and C.

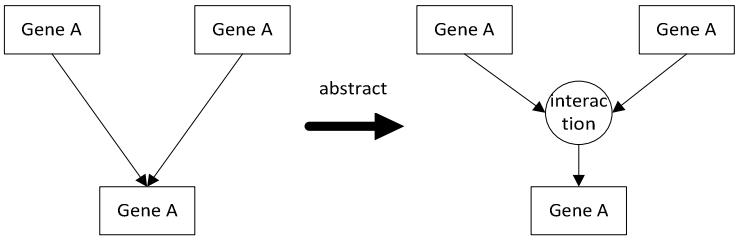


Fig. 3 The construction of gene causal logic model

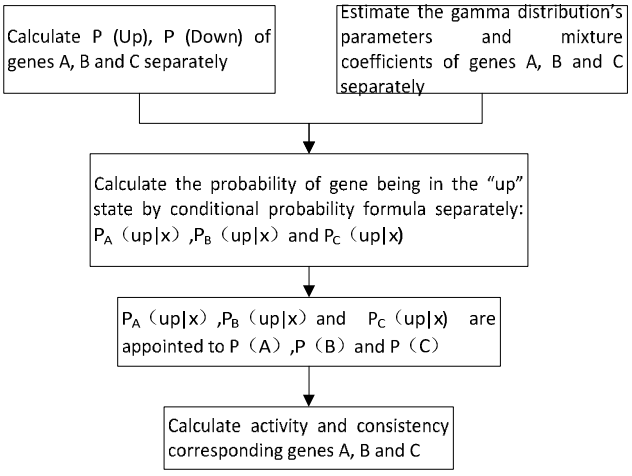


Fig. 4 The flow diagram of PathOlogist

For a pathway which includes more interactions, the corresponding average values of interaction scores are used as pathway's activity and consistency scores. Thus, for each sample, there are two scores which are used to assess the behavior of each pathway.

PathOlogist was compared with SPIA[15]and PARADIGM[16] on the data with 377 glioblastoma multiforme (GBM) tumor samples and 10 unmatched normal samples. Efroni et al. found that Only PathOlogist successfully identified some pathways, such as the RAC1, CDC42, FAS and PDGFR signaling pathways which are significantly associated with GBM[28-29] and the three methods also identified common pathways such as the Pi3k signaling pathway and the histone deacytelase (HDAC) signaling pathway etc. The advantage of the PathOlogist is that both a quantitative and qualitative analysis of pathway behavior can be done in a format accessible to both laboratory researchers and informatics analysts through activity and consistency. Of course, there are some limitations to the PathOlogist that it can only be used to analyze established pathways. Additionally, small subsets of interactions that have real association to a clinical feature may be overshadowed.

2.4 Active Modules

Gene expression values often have significantly changes with the development of complex diseases. In order to explain the intrinsic mechanism of these changes, Ideker et al.[18] proposed a approach to find active modules through combining gene expression value and protein networks which topology structures are from the existing knowledge data bases, such as KEGG, BioCyc, TCGA and so on. When they mapped the identified modules to pathways and got significant pathway. Active score of a subnetwork T with k genes is computed as follow under single condition.

1. Compute p-value of the *i*th gene: p_i (i=1, 2,…, k) according to its expression values in different tissues.

2. Compute Z-score of the *i*th gene: $z_i = \Phi^{-1}(1-p_i)$, where Φ^{-1} is the normal inverse cumulative distribution function.

3. Compute the score of subnetwork T of k genes: $z_T = \frac{1}{\sqrt{k}}\sum_{i\in T} z_i$.

4. Obtain the score of subnetwork T of k genes: $s_T = \frac{(z_T - \mu_k)}{\sigma_k}$, where μ_k and σ_k are average value and standard deviation of scores of subnetworks from random gene sets with the k size.

One gene may be measured over multiple conditions. In this case, the score of subnetwork T is caculated under different conditions respectively, the specific steps is as follow:

1. Calculate z_T of the subnetwork T under the condition of the *j*th (j=1, 2, … m) conditions respectively and get the corresponding active scores: $z_{T(1)}$, $z_{T(2)}$, … $z_{T(m)}$.

2. Sort from highest to lowest: $z_{T(1)}$, $z_{T(2)}$, … $z_{T(m)}$.

3. Compute the probability that at least j of the m conditions have scores above $z_{T(j)}$: $P_{T(j)} = \sum_{h=j}^{m}\binom{m}{h}(p_z)^h(1-p_z)^{m-h}$ and convert $P_{T(j)}$ into $r_{T(j)} = \Phi^{-1}(1-p_{T(j)})$, where $p_z = 1-\Phi(z_{T(j)})$.

4. Choose the maximum of $r_{T(j)}$ (j=1, 2 ,…m) as new score of subnetwork T: $r_T^{\max} = \max_j(r_{T(j)})$.

5. Calibrate $r_T^{\max}$ against the background distribution $s_T = \frac{(r_T^{\max} - \mu_k)}{\sigma_k}$, where μ_k and σ_k are average value and standard deviation of scores of subnetworks from random set of genes with the k size. s_T is just the score of the subnetwork T.

The proposed method was employed to identify active modules from yeast data containing 362 protein-protein and protein-DNA interactions.[30] Experimental results show that the many subnetworks with higher score have striking overlap with well-known pathways described in the yeast study. For example, one of subnetworks with higher score includes the path GAL3—GAL80—GAL4—GAL1,7,10, which is the core of the galactose-induction pathway.[31] The advantage of the Active Modules is that subnetwork under certain conditions is not required to be predefined gene sets and pathways. This method is also simple and intuitive, and is a milestone of many methods developed based on its principle. A need for further improvement is that priori knowledge is not taken full advantage.

2.5 AMBIENT

Bryant et al.[19] extended Active Modules to bipartite network in which both metabolites and reactions are considered as two types of nodes. Other methods, such as PathExpress, KEGG spider and PathWave also take the similar strategies.[32-34] TAMBIENT is better to overcome the influence of currency metabolites and isozymes in analyzing metabolic subnetworks and avoids the loss of useful information due to arbitrary classification of compounds.

For a metabolic network which has been converted into a metabolite-reaction bipartite network according to AMBIENT, the score of its subnetwork m (or module which is composed of a subset r^m of reactions and a subset c^m of metabolites) is given by

$$S(m) = \ln(q)\left(\sum_i s(r_i^m) - \alpha\sum_j w(c_j^m)\right)$$

where $S(m)$ is score of subnetwork m , $q = |r^m| + |c^m|$, $|r^m|$ is the number of the subset r^m, $|c^m|$ is the number of the subset c^m, $s(r_i^m)$ is the score of the *i*th reaction in the module m , $w(c_j^m)$ is the weight (the degree in the original network) of the *j*th metabolite in the module, α is a balance factor.

Bryant et al. employed AMBIENT to analyze yeast data[35-36] and successfully found that the TCA cycle and associated respiratory chain complexes are the most significantly affected parts of metabolism. In addition to identifying the important parts of metabolism discovered by other methods, such as

GiGA, AMBIENT also finds several pathways which are undiscovered by previous methods.

One of the most advantages of AMBIENT is that metabolic network is converted into a bipartite network and the score of module is computed based on reaction-metabolite rather than protein-protein relationship. Another advantage is that AMBIENT tends to find larger subnetwork which might otherwise be hidden due to individual interconnecting low-scoring nodes or due to the lack of experimental data. A need for further improvement is that the run time needs to reduce.

2.6 GIGA

Without the need for rich prior knowledge, whether active sub-networks can be found by greedy algorithms on the interactive network only using statistical methods combined with gene expression data. Based on this purpose, Breitling et al. proposed the GIGA algorithm[20] which is an extension of the iGA[37] based on graph structure. The topology structure comes from the GeneOntology annotations (GO network) and one where the evidence comprises enzyme substrates (metabolic network).

The principle of GIGA algorithm is explained by Fig. 5 as follows:

1. Extract evidence network from the GO term network (Fig. 5-1);

2. Convert evidence network to a simple network. Genes which share a common annotation are connected (Fig. 5-2);

3. Rank genes in descending order according to the change of their expression value: Gene4→Gene1→Gene2→Gene3→ Gene6→Gene5 (Fig. 5-3), the corresponding serial numbers of these genes are NO. 1→2→3→4→5→6;

4. Find gene with local minima, i.e. the serial number of gene is lower than its neibor's (Fig.5-4), such as gene 1 and gene 4;

5. Iterative expansion of subgraph from one of the local minima nodes (Fig. 5-5). The neighbor node of No.1 with highest rank (gene 3, No.4) is included (Fig. 5-6), which leads to the additional inclusion of genes 1 (rank 2) and 2 (rank 3) (Fig. 5-7), Gene 6 (rank 5) is added to subnetwork(Fig. 5-8). The last gene is included (Fig. 5-9). For each of the subgraphs a p-value is calculated as follow:

$$p = \prod_{i=0}^{n-1} \frac{m-i}{N-i}$$

Where N is the number of total genes in the graph, n is the number of genes in the subgraph, and m represents the maximum rank;

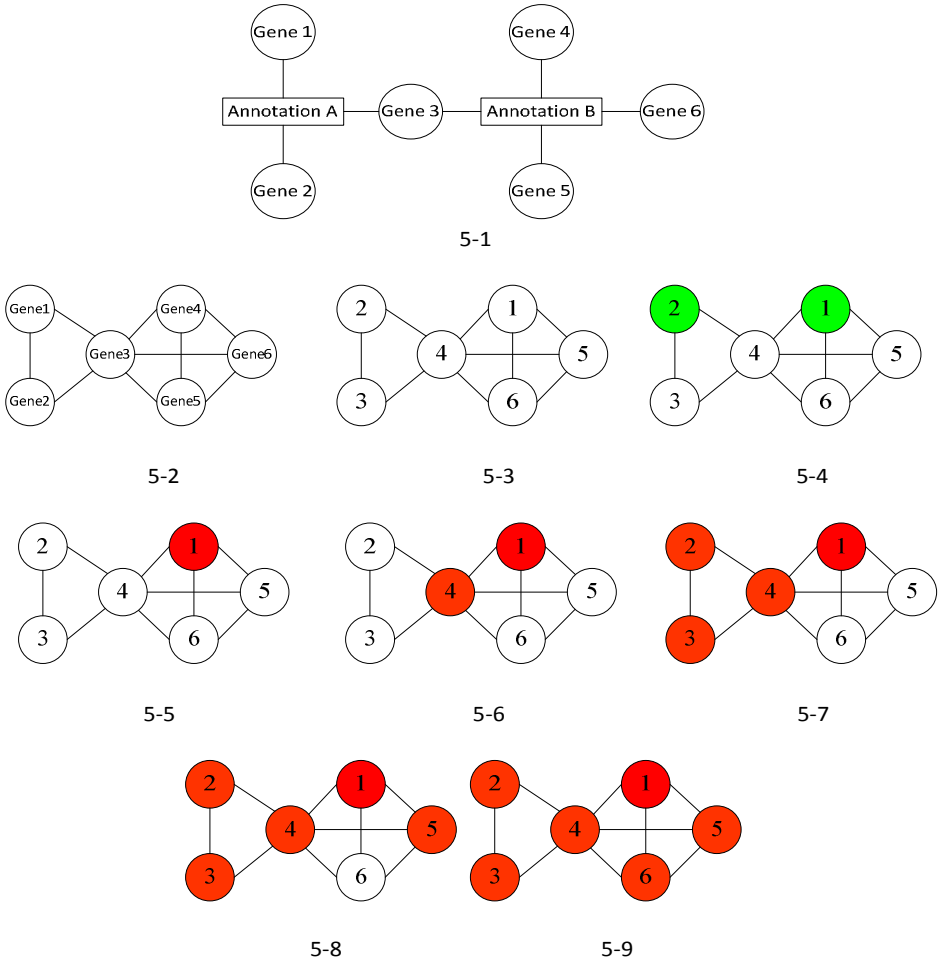


Fig. 5 The specific calculation process of the GIGA algorithm

6. Subnetworks are sorted according to their p-values. The smaller p-value, the more significantly the corresponding sub-network changes.

Breitling et al. found the significance of the TCA cycle using the GIGA algorithm in the YEASTNET dataset,[35] namely GIGA also has the same discovery as AMBIENT algorithm. This shows that although the GIGA method is simple, it is effective in the analysis of the pathway. The advantages of the GIGA method are that it does not require too much prior knowledge and it is simple, robust and cost saving. The disadvantage is that the method is time-consuming and the interpretation of experimental results needs more biological background.

2.7 neXus

Biological network alignment is an important approach in the study of organisms' structure, function and evolution.[38] The relationship of structure and function between biological networks from different species could be found by biological network alignment. Through knowledge transfer, different species can provide each other with some prior knowledge. This also helps to discover the conserved subnetwork structure. Based on the above principle, Deshpande et al. first proposed the neXus algorithm[21] which was used to find the conserved active subnetworks through comparing interaction networks from multiple species respectively. The detail process of finding subnetwork from two networks is following: First, the growth of subnetwork starts with a seed node which comes from the intersection of the human differential genes and mouse differential genes; Then, subnetworks are simultaneously grow in both species from seed genes by adding nearby genes. Growth of each subnetwork is constrained by two parameters: a minimum network activity score and a minimum clustering coefficient constraint. Subnetwork growth is stopped when either the clustering coefficient constraint or the minimum network score constraint is not satisfied. This process is repeated for all differentially expressed genes. Finally, more than one conserved subnetworks are found.

The neXus algorithm can be used to find the valuable conserved active subnetworks and species-specific networks [39-41] from two different biological networks. It also can be used to identify significant pathway through mapping the conserved active subnetworks to the associated pathways. The advantage of the neXus algorithm is that the approach can be readily extended to discover conserved subnetworks across more than two species and also is an effective method to improve sensitivity and specificity by the cross-species network alignment. However, it requires several days to run to get experiment results due to the complexity of the algorithm.

At last, we formatted ours analysis in a table S1 (ESI†) provided a comparison among these methods to pinpoint their limitations and performance.

3 Discussion and future perspectives

Complex diseases, such as cancer, high blood pressure, heart disease, diabetes, nervous system disease and so on, are produced by a variety of factors, which developments often involve multiple pathways. The network-based pathway analysis method is an effective means to understand functions of pathways and identify critical pathways of complex diseases. The post-genomic era is coming after the completion of the human genome project. Advance in high-throughput sequencing and gene/protein profiling techniques generate multiple omics data such as genomics, granscriptomics, proteinomics, metabolomics and phenomics etc. The rapid growth of these omics data provides opportunities to study the roles of pathways in complex disease at various molecular levels. These data could be effectively integrated into biological network models. So in the post-genomic era, network-based pathway analysis has become one of main tasks to gain insight into the underlying mechanism about the changes of differentially expressed genes and proteins. On the other hand, these omics data also lead to some challenges to analyze pathway through biological network. Therefore, the one of main directions which we will struggle for is to establish various models which could integrate multilevel data to accurately describe dynamic response of pathways under multiple conditions and help us understand the mechanism of complex diseases.

Biological network or pathway is dynamics and this means that their topology structure and state of nodes is often change under different conditions. One of main challenges in pathway analysis is how to get high precision pathway topology structure which can accurately describe the dynamics changes of cell. Based on the dynamical network biomarker (DNB), Chen et al. proposed a method to predict the critical transition of diseases.[42] A dynamic model based on the structural output controllability of complex Networks is developed to identify effective drug targets.[43] Wu et al. developed a theoretic framework for studying transitions between two specific states of directed complex networks.[44] These studies promoted the development of pathway analysis.

Pathway plays an important role in the development of complex diseases, it's critical to find key pathway for accurate diagnosis,[45] prediction[46], precise treatment and interpretation of complex diseases. The increasing availability of high-throughput biological data of complex diseases and the development of various biological networks provided the better conditions to build accurate pathway analysis models, but there is still a lack of multiresolution knowledge bases to support the accurate pathway analysis. To the best of our knowledge, due to lack of abundant pathway knowledge bases, most of pathway analysis results is incomplete, unreliable or inaccurate. So it's an urgent task to build accurate, multiresolution pathway knowledge bases with detailed organs, tissues, cell types under multiple conditions.

In the studies of traditional drug development, researchers always focus on a single gene or protein target to design corresponding experiments, so they cannot know overall interactions between drugs and organisms. With the sharp

development of modern biology theory and experimental technology, it is known that organisms often exhibit physical functions by functional pathways consisting of compositions like kinases and transcription factors executing certain interactions in order, which might be more complicated when multiple pathways make up a biological regulation network by cross-talks to each other. Thus the focus of biological researchers turns to the identification of potential targets from a biological regulation network, which initiates a new era of drug development. How to find the pathway associated with drug using molecular network and develop new pathway analysis models for the discovery of potential therapeutic targets and new drugs has become the core issue of current network pharmacology research and is one of the important research directions of the pathway analysis.

In order to analyze pathways, a variety of tools are developed based on different biological data. But these tools are developed to solve a specific problem faced by pathway analysis; there is still a lack of comprehensive software used to analyze pathways from different aspects. Currently there is an urgent need to develop a new pathway analysis platform to better service to the study of complex diseases and drug development.

Most of pathway analysis methods only consider pathway separately and ignore the interaction of pathways. Hence, relationship analysis between pathways is also a direction in the future.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (61471147), China National 863 High-Tech Program (2015AA020101), Natural Scientific Research Innovation Foundation in Harbin Institute of Technology (KMQQ5750010615) and Harbin Innovation Talents of Special Funds (2014RFQXJ090).

References

1 W. Jin, P. Qin, H. Lou, L. Jin and S. Xu, Human molecular genetics, 2011, ddr599.
2 U. D. Akavia, O. Litvin, J. Kim, F. Sanchez-Garcia, D. Kotliar, H. C. Causton, P. Pochanard, E. Mozes, L. A. Garraway and D. Pe'er, Cell, 2010, 143, 1005-1017.
3 K. A. Hoadley, C. Yau, D. M. Wolf, A. D. Cherniack, D. Tamborero, S. Ng, M. D. Leiserson, B. Niu, M. D. McLellan and V. Uzunangelov, Cell, 2014, 158, 929-944.
4 A. M. Sonabend, M. Bansal, P. Guarnieri, L. Lei, B. Amendolara, C. Soderquist, R. Leung, J. Yun, B. Kennedy and J. Sisti, Cancer research, 2014, 74, 1440-1451.
5 M. S. Carro, W. K. Lim, M. J. Alvarez, R. J. Bollo, X. Zhao, E. Y. Snyder, E. P. Sulman, S. L. Anne, F. Doetsch and H. Colman, Nature, 2010, 463, 318-325.
6 M. Kanehisa and S. Goto, Nucleic acids research, 2000, 28, 27-30.
7 R. Caspi, H. Foerster, C. A. Fulcher, P. Kaipa, M. Krummenacker, M. Latendresse, S. Paley, S. Y. Rhee, A. G. Shearer and C. Tissier, Nucleic acids research, 2008, 36, D623-D631.
8 P. D. Karp, M. Riley, S. M. Paley and A. Pellegrini-Toole, Nucleic acids research, 2002, 30, 59-61.
9 G. Joshi-Tope, I. Vastrik, G. Gopinath, L. Matthews, E. Schmidt, M. Gillespie, P. D'EUSTACHIO, B. Jassal, S. Lewis and G. Wu, 2003.
10 A. M. Huerta, H. Salgado, D. Thieffry and J. Collado-Vides, Nucleic Acids Research, 1998, 26, 55-59.
11 P. D. Thomas, M. J. Campbell, A. Kejariwal, H. Mi, B. Karlak, R. Daverman, K. Diemer, A. Muruganujan and A. Narechania, Genome research, 2003, 13, 2129-2141.
12 R. Sharan and T. Ideker, Nature biotechnology, 2006, 24, 427-433.
13 P. Khatri, M. Sirota and A. J. Butte, PLoS Comput Biol, 2012, 8, e1002375.
14 P. Creixell, J. Reimand, S. Haider, G. Wu, T. Shibata, M. Vazquez, V. Mustonen, A. Gonzalez-Perez, J. Pearson and C. Sander, Nature methods, 2015, 12, 615-621.
15 A. L. Tarca, S. Draghici, P. Khatri, S. S. Hassan, P. Mittal, J.-s. Kim, C. J. Kim, J. P. Kusanovic and R. Romero, Bioinformatics, 2009, 25, 75-82.
16 C. J. Vaske, S. C. Benz, J. Z. Sanborn, D. Earl, C. Szeto, J. Zhu, D. Haussler and J. M. Stuart, Bioinformatics, 2010, 26, i237-i245.
17 S. Efroni, C. F. Schaefer and K. H. Buetow, PloS one, 2007, 2, e425.
18 T. Ideker, O. Ozier, B. Schwikowski and A. F. Siegel, Bioinformatics, 2002, 18, S233-S240.
19 W. A. Bryant, M. J. Sternberg and J. W. Pinney, BMC systems biology, 2013, 7, 26.
20 R. Breitling, A. Amtmann and P. Herzyk, BMC bioinformatics, 2004, 5, 100.
21 R. Deshpande, S. Sharma, C. M. Verfaillie, W.-S. Hu and C. L. Myers, PLoS computational biology, 2010, 6, e1001028.
22 Y. Hong, K. S. Ho, K. W. Eu and P. Y. Cheah, Clinical Cancer Research, 2007, 13, 1107-1114.
23 A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub and E. S. Lander, Proceedings of the National Academy of Sciences of the United States of America, 2005, 102, 15545-15550.
24 S. F. Chin, A. E. Teschendorff, J. C. Marioni, Y. Wang, N. L. Barbosa-Morais, N. P. Thorne, J. L. Costa, S. E. Pinder, M. A. van De Wiel and A. R. Green, Genome Biol, 2007, 8, R215.
25 X. Ju, S. Katiyar, C. Wang, M. Liu, X. Jiao, S. Li, J. Zhou, J. Turner, M. P. Lisanti and R. G. Russell, Proceedings of the National Academy of Sciences, 2007, 104, 7438-7443.
26 A. J. Sedgewick, S. C. Benz, S. Rabizadeh, P. Soon-Shiong and C. J. Vaske, Bioinformatics, 2013, 29, i62-i70.
27 S. I. Greenblum, S. Efroni, C. F. Schaefer and K. H. Buetow, BMC bioinformatics, 2011, 12, 133.
28 V. M. Zohrabian, B. Forzani, Z. Chau, R. Murali and M. Jhanwar-Uniyal, Anticancer research, 2009, 29, 119-123.
29 D. B. Hoelzinger, L. Mariani, J. Weis, T. Woyke, T. J. Berens, W. McDonough, A. Sloan, S. W. Coons and M. E. Berens, Neoplasia, 2005, 7, 7-16.
30 T. Ideker, V. Thorsson, J. A. Ranish, R. Christmas, J. Buhler, J. K. Eng, R. Bumgarner, D. R. Goodlett, R. Aebersold and L. Hood, Science, 2001, 292, 929-934.
31 D. Lohr, P. Venkov and J. Zlatanova, The FASEB Journal, 1995, 9, 777-787.
32 N. Goffard, T. Frickey and G. Weiller, Nucleic acids research, 2009, 37, W335-W339.
33 A. V. Antonov, S. Dietmann and H. W. Mewes, Genome Biol, 2008, 9, R179.
34 G. Schramm, S. Wiesberg, N. Diessl, A.-L. Kranz, V. Sagulenko, M. Oswald, G. Reinelt, F. Westermann, R. Eils and R. König, Bioinformatics, 2010, 26, 1225-1231.

35 J. L. DeRisi, V. R. Iyer and P. O. Brown, Science, 1997, 278, 680-686.
36 M. J. Herrgård, N. Swainston, P. Dobson, W. B. Dunn, K. Y. Arga, M. Arvas, N. Blüthgen, S. Borger, R. Costenoble and M. Heinemann, Nature biotechnology, 2008, 26, 1155-1160.
37 R. Breitling, A. Amtmann and P. Herzyk, BMC bioinformatics, 2004, 5, 34.
38 E. Almaas, Journal of Experimental Biology, 2007, 210, 1548-1558.
39 Q.-L. Ying, J. Nichols, I. Chambers and A. Smith, Cell, 2003, 115, 281-292.
40 I. G. M. Brons, L. E. Smithers, M. W. Trotter, P. Rugg-Gunn, B. Sun, S. M. C. de Sousa Lopes, S. K. Howlett, A. Clarkson, L. Ahrlund-Richter and R. A. Pedersen, Nature, 2007, 448, 191-195.
41 R.-H. Xu, R. M. Peck, D. S. Li, X. Feng, T. Ludwig and J. A. Thomson, Nature methods, 2005, 2, 185-190.
42 L. Chen, R. Liu, Z.-P. Liu, M. Li and K. Aihara, Scientific reports, 2012, 2.
43 L. Wu, Y. Shen, M. Li and F.-X. Wu, NanoBioscience, IEEE Transactions on, 2015, 14, 184-191.
44 F.-X. Wu, L. Wu, J. Wang, J. Liu and L. Chen, Scientific reports, 2014, 4.
45 J. Li, X. Tang, J. Liu, J. Huang and Y. Wang, Pattern Recognition, 2008, 41, 1975-1984.
46 J. Li, A. E. Lenferink, Y. Deng, C. Collins, Q. Cui, E. O. Purisima, M. D. O'Connor-McCourt and E. Wang, Nature communications, 2010, 1, 34.