EMBnet.journal
Bioinformatics in Action

Chanumolu et al. (2021) EMBnet.journal 26, e949
http://dx.doi.org/10.14806/ej.26.0.949

Technical Notes

KEGG2Net: Deducing gene interaction networks and acyclic graphs from KEGG pathways

Sree Chanumolu[1], **Mustafa Albahrani**[1], **Handan Can**[1], **Hasan Otu**[1]✉

[1]Department of Electrical and Computer Engineering, University of Nebraska-Lincoln, Lincoln, NE, United States

Competing interests: SC none; MA none; HC none; HO none

Abstract

The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database provides a manual curation of biological pathways that involve genes (or gene products), metabolites, chemical compounds, maps, and other entries. However, most applications and datasets involved in omics are gene or protein-centric requiring pathway representations that include direct and indirect interactions only between genes. Furthermore, special methodologies, such as Bayesian networks, require acyclic representations of graphs. We developed KEGG2Net, a web resource that generates a network involving only the genes represented on a KEGG pathway with all of the direct and indirect gene-gene interactions deduced from the pathway. KEGG2Net offers four different methods to remove cycles from the resulting gene interaction network, converting them into directed acyclic graphs (DAGs). We generated synthetic gene expression data using the gene interaction networks deduced from the KEGG pathways and performed a comparative analysis of different cycle removal methods by testing the fitness of their DAGs to the data and by the number of edges they eliminate. Our results indicate that an ensemble method for cycle removal performs as the best approach to convert the gene interaction networks into DAGs. Resulting gene interaction networks and DAGs are represented in multiple user-friendly formats that can be used in other applications, and as images for quick and easy visualisation. The KEGG2Net web portal converts KEGG maps for any organism into gene-gene interaction networks and corresponding DAGs representing all of the direct and indirect interactions among the genes.

Introduction

The KEGG pathway database provides hundreds of manually curated maps that involve molecular interactions between gene products, compounds, maps, DNA, RNA, and other molecules (Kanehisa and Goto, 2000). The maps are categorised under seven groups, such as "Metabolism" or "Environmental Information Processing," which are stored in proprietary files in XML format, called KGML. There exist numerous approaches that process the KEGG pathway maps, such as KEGGtranslator (Wrzodek *et al.*, 2011), KEGGParser (Arakelyan and Nersisyan, 2013), CyKEGGParser (Nersisyan *et al.*, 2014), KEGGgraph (Zhang and Wiemann, 2009), KEGGconverter (Moutselos *et al.*, 2009), and graphite (Sales *et al.*, 2012) among others (Wrzodek *et al.*, 2013). These approaches convert KGML files to other formats (*e.g.*, SBML, BioPAX) to be used in applications for data visualisation (*e.g.*, Cytoscape) or graph-theoretic analysis (*e.g.*, MATLAB®, Bioconductor).

Although these approaches are extremely useful, they do not provide a representation that only includes genes deduced from the KEGG pathways considering all of the direct and indirect gene interactions. However, when analysing experimental datasets that only involve the genes (or gene products) in the context of KEGG pathways (*e.g.*, transcriptomic or proteomic data), an interaction network that only involves these molecules is required. Among the existing tools, KEGGgraph provides a *"genesOnly"* parameter that results in a gene-oriented graph; but that approach only deduces the direct gene interactions provided in the maps. graphite also represents gene-only networks, but it does not take into account all of the compounds between the genes to obtain the exhaustive set of indirect interactions – some compounds based on their identity or localisation are ignored. Indeed, there is a need for an approach that recovers all of the direct and indirect gene-gene interactions from a KEGG pathway that can be used in

Article history
Received: 14 September 2020
Accepted: 13 December 2020
Published: 24 February 2021

Chanumolu et al. (2021) EMBnet.journal 26, e949
http://dx.doi.org/10.14806/ej.26.0.949

downstream analysis involving data coming only from these molecules.

KEGG pathways include cycles that may be problematic in analysis approaches, such as Bayesian networks (BNs), which use directed acyclic graphs (DAGs) (Friedman *et al.*, 2000; Isci *et al.*, 2014; Isci *et al.*, 2011; Korucuoglu *et al.*, 2014; Liu *et al.*, 2016). None of the existing approaches that process KEGG pathways provide DAGs as their output. Furthermore, there has been no study that compares different cycle removal methods in the context of the fitness of biological data to the resulting DAGs.

In light of these observations and perceived needs, we developed KEGG2Net, a web resource that converts KEGG pathways into gene interaction networks involving all of the direct and indirect relations between the genes that can be deduced from the map. KEGG2Net offers four alternative methods to convert the resulting gene interaction networks into DAGs. This paper also provides a comparative assessment of the cycle removal methods via their fitness to the data obtained from the original gene interaction network deduced from KEGG.

Implementation

Given the KGML file for a pathway map, the engine parses the file to obtain an adjacency matrix that represents all of the interactions (relation, reaction, *etc.*) between all of the node types (compound, map, gene, *etc.*) defined in the file. In this graph, if there exists a path between two genes that contain non-gene nodes only, then an indirect relation between the two genes is established by placing an edge between them. Next, nodes that are not genes are removed from the adjacency matrix. This way, the resulting graph represents all of the direct and indirect interactions between the genes that can be deduced from the KEGG pathway map.

The resulting gene network may contain cycles that are removed using four different methods: a depth-first search (DFS) (Suominen and Mader, 2014), a greedy local heuristic to the minimum feedback arc set (MFAS) problem (Eades *et al.*, 1993), and two graph-hierarchy-based methods where the hierarchy is inferred either through PageRank (PR) (Page *et al.*, 1999) or an ensemble method (EN) (Sun *et al.*, 2017) based on the TrueSkillTM (Herbrich *et al.*, 2007) and social agony (Gupte *et al.*, 2011) metrics. The DFS-based approaches use fast, simple heuristics to remove back edges, MAFS-based approaches try to minimise the number of edges removed, and hierarchy-based methods define a hierarchy in the graph first and then devise an edge removal strategy that prioritises the maintenance of the defined hierarchy as much as possible.

The input to KEGG2Net is the KGML files for the pathways that belong to the organism selected by the user. The output of KEGG2Net consists of the gene interaction networks deduced from the pathways and four DAGs per network where the cycles are removed by the aforementioned four algorithms. The networks and DAGs are represented as adjacency matrices and simple interaction files (a.k.a. SIF or .sif format) for use in the downstream analysis by other software and for visualisation purposes.

In order to compare the accuracy of different cycle removal methods and to provide a sample output using graph images for visualisation purposes, we applied the KEGG2Net approach on the 335 available human KEGG pathways. The KEGG2Net workflow adopted in this paper is shown in Figure 1.

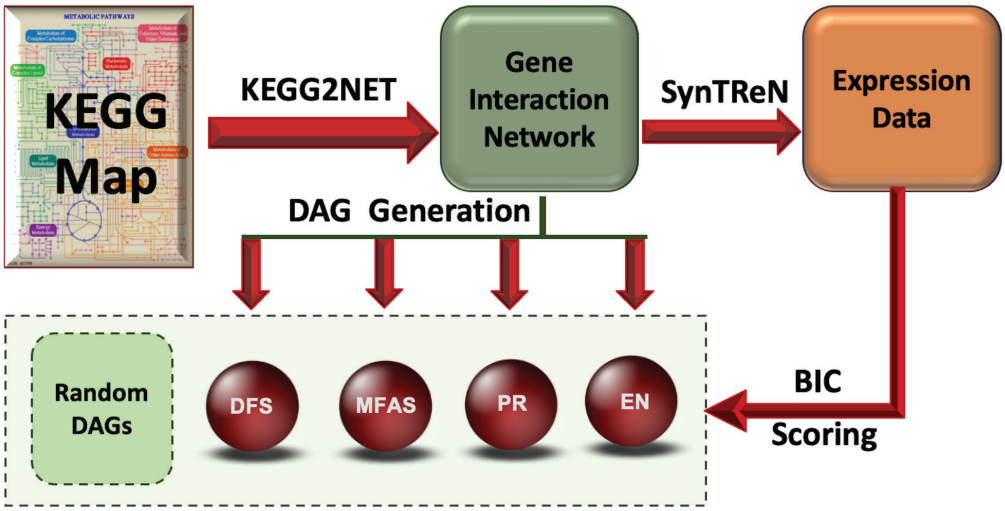

Figure 1. KEGG2Net workflow and directed acyclic graph (DAG) generation.
Each KEGG pathway is converted into a gene interaction network where only the genes in the pathway are represented, and all direct and indirect interactions among the genes are preserved. For each network, four DAGs (using four alternative cycle removal methods depth-first search (DFS), minimum feedback arc set (MFAS), PageRank (PR), and ensemble (EN)) and 1,000 random DAGs for each of the four DAGs are generated. The random DAGs follow the node and edge statistics of their corresponding DAG. For each gene interaction network, synthetic gene expression data is generated using SynTRen and the fitness of the four DAGs (and the corresponding 4x1,000 = 4,000 random DAGs) is assessed using Bayesian Information Criterion (BIC) scoring.

Chanumolu et al. (2021) EMBnet.journal 26, e949
http://dx.doi.org/10.14806/ej.26.0.949

Out of the 335 pathways, we considered only the networks with six or more non-isolated nodes, which left us with 280 networks. Of these networks, 150 did not have any cycles. For each of the remaining 130 that were cyclic, a synthetic gene expression data fitting the graph topology was generated using SynTReN (v. 1.2) (Van den Bulcke *et al.*, 2006). The expression data was processed to be used by BN scoring methods as previously described (Isci *et al.*, 2011). The four cycle removal methods were applied to the 130 networks with cycles generating four DAGs per network, which were scored based on the Bayesian Information Criterion (BIC) in the *bnlearn* R package (Scutari, 2010) using the processed synthetic expression data. For each of the four DAGs per network, we generated 1,000 random DAGs (with the same number of edges and nodes as the original DAG) and obtained their BIC scores based on the same processed synthetic expression data to assess the goodness of the original DAG's score.

Results

Our web portal provides the gene interaction networks obtained for all of the 335 human KEGG pathways. For the networks that have cycles, we also list the DAGs obtained using the four methods. The gene interaction networks and the DAGs obtained from them are represented by adjacency matrices, SIF format files, and graph images. KEGG2Net can be used for all of the organisms listed in KEGG where the user can download the relevant network and DAG files via our web portal.

One direct way to assess the performance of different cycle removal methods is to compare the number of edges removed by each method. However, this approach does not infer the degree to which the topology and the dependency structure among the nodes of the network are preserved in the DAGs. For this purpose, we first generated synthetic gene expression data that follow the regulatory dynamics explained by the gene interaction network deduced by KEGG2Net. We then scored each of the four DAGs with this expression data using BIC scoring, where a higher score indicated a better fit. Finally, for each of the four DAGs that result from the given network, we generated 1,000 random DAGs (i.e., 4,000 random DAGs per network) where the random DAGs had the same number of nodes and links as the DAG they were associated with. This exercise was repeated for all of the 130 networks, and the DAG statistics were compared.

The complete set of results for our simulations are given in the Supplementary Data, which involves the 130 KEGG2Net gene interaction networks with six or more non-isolated nodes and a cycle. We provide the original pathway's ID, name, number of edges, number of nodes, the number of edges removed by the four cycle removal methods, the rank of each method for each pathway based on the edges removed, and the rank of the score for each method (among themselves and their 1,000 random DAGs).

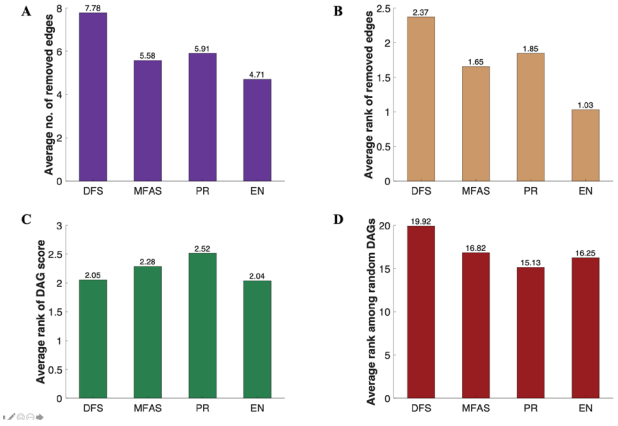

Figure 2. Directed acyclic graph (DAG) statistics. Based on the four cycle removal methods, depth-first search (DFS), minimum feedback arc set (MFAS), PageRank (PR), and ensemble (EN), applied on 130 gene interaction networks deduced from pathways using KEGG2Net, **A** average number of edges removed; **B** average rank of the method (per network) based on the number of edges removed (ascending); **C** average rank of the method's DAG score (per network) among the four methods (descending); **D** average rank of the method's DAG score (per network) among the 1,000 random DAGs that has the same numbers of edges and nodes as the DAG (descending). For parts **C** and **D**, Bayesian Information Criterion (BIC) scoring is used to assess the fitness of the DAGs to the synthetic data generated based on the gene interaction network (obtained by KEGG2Net).

These results, summarised in Figure 2, showed that on average, the EN method required the least number of edges to be removed and provided the DAG with the highest BIC score.

On average, the EN method removed 4.71 edges per pathway; and it was the method that required the least number of edges to be removed in 127 out of 130 networks. The EN method accomplished an average rank of 1.03 in all networks, where 1 represented the rank that removed the minimum number of edges.

The average rank of the EN score among the four DAGs was also the highest, where it attained an average rank of 2.04. In other words, on average, the EN DAG provided the best topology that fit the synthetic expression data. The only category where the EN method was outperformed was its average rank among the 1,000 random DAGs. On average, 16.25 random DAGs for a given network performed better than the EN DAG, whereas this number was 15.13 for the PR DAG, the only method that beat the EN approach in this category. Given the comprehensive evaluation summarised in Figure 2, we recommend EN as the method of choice for DAG generation.

Discussion

In this work, we provide a web resource that converts KEGG pathways into gene interaction networks representing all of the direct and indirect interactions between the genes. We also provide four alternative

Technical Notes

ways of converting the resulting graph into a DAG. Our results showed that the EN method finds the best DAG that explains the underlying hierarchy and dependency structure defined in the interaction network with the minimum number of edge removals. Our web portal lists the graph structure and the four DAGs for each of the KEGG human pathways. The KEGG2Net web resource can be used to obtain the networks and DAGs for any organism listed in the KEGG database.

Key Points

- Deduces only the gene-gene interactions from a KEGG pathway considering all direct and indirect interactions.
- Provides networks that are ready to be applied to transcriptomic or proteomic data for system-level analysis.
- Using multiple alternative methods, converts networks to directed acyclic graphs (DAGs) that can be used in methodologies such as Bayesian networks, which require DAGs.
- Generates multiple output formats that can be directly used in different network visualisation or analysis software.
- Reports a comparative analysis of cycle removal methods for biological pathways.

Funding

The research reported in this publication was supported by the National Library of Medicine (NLM) of the National Institutes of Health (NIH) under award number **R21LM012759**. The funding body was not involved in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

References

1. Arakelyan A and Nersisyan L (2013) KEGGParser: parsing and editing KEGG pathway maps in Matlab. Bioinformatics **29**(4): 518-519 http://dx.doi.org/10.1093/bioinformatics/bts730
2. Eades P, Lin X and Smyth WF (1993) A fast and effective heuristic for the feedback arc set problem. Inform Process Lett **47**(6): 319–323 http://dx.doi.org/10.1016/0020-0190(93)90079-O
3. Friedman N, Linial M, Nachman I and Pe'er D (2000) Using Bayesian networks to analyze expression data. J Comput Biol **7**(3-4): 601-620 http://dx.doi.org/10.1089/106652700750050961
4. Gupte M, Shankar P, Li J, Muthukrishnan S and Iftode L (2011) Finding Hierarchy in Directed Online Social Networks. Proceedings of the 20th International Conference on World Wide Web (WWW '11): 557–566 http://dx.doi.org/10.1145/1963405.1963484
5. Herbrich R, Minka T and Graepel T (2007) TrueSkill™: A Bayesian Skill Rating System. In: Advances in Neural Information Processing Systems (NIPS), Scholkopf P.B., Platt J.C., Hoffman T. (eds.) pp. 569–576.
6. Isci S, Dogan H, Ozturk C and Otu HH (2014) Bayesian network prior: network analysis of biological data using external knowledge. Bioinformatics **30**(6): 860-867 http://dx.doi.org/10.1093/bioinformatics/btt643
7. Isci S, Ozturk C, Jones J and Otu HH (2011) Pathway analysis of high-throughput biological data within a Bayesian network framework. Bioinformatics **27**(12): 1667-1674 http://dx.doi.org/10.1093/bioinformatics/btr269
8. Kanehisa M and Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res **28**(1): 27-30 http://dx.doi.org/10.1093/nar/28.1.27
9. Korucuoglu M, Isci S, Ozgur A and Otu HH (2014) Bayesian pathway analysis of cancer microarray data. Plos One **9**(7): e102803 http://dx.doi.org/10.1371/journal.pone.0102803
10. Liu F, Zhang SW, Guo WF, Wei ZG and Chen L (2016) Inference of Gene Regulatory Network Based on Local Bayesian Networks. PLoS Comput Biol **12**(8): e1005024 http://dx.doi.org/10.1371/journal.pcbi.1005024
11. Moutselos K, Kanaris I, Chatziioannou A, Maglogiannis I and Kolisis FN (2009) KEGGconverter: a tool for the in-silico modelling of metabolic networks of the KEGG Pathways database. BMC Bioinformatics **10**(324) http://dx.doi.org/10.1186/1471-2105-10-324
12. Nersisyan L, Samsonyan R and Arakelyan A (2014) CyKEGGParser: tailoring KEGG pathways to fit into systems biology analysis workflows. F1000Res **3**(145) http://dx.doi.org/10.12688/f1000research.4410.2
13. Page L, Brin S, Motwani R and Winograd T, (1999). "The PageRank citation ranking: Bringing order to the web", Technical Report Stanford InfoLab.
14. Sales G, Calura E, Cavalieri D and Romualdi C (2012) graphite - a Bioconductor package to convert pathway topology to gene network. BMC Bioinformatics **13**(20) http://dx.doi.org/10.1186/1471-2105-13-20
15. Scutari M (2010) Learning Bayesian Networks with the bnlearn R Package. Journal of Statistical Software **35**(3): 1-22 http://dx.doi.org/10.18637/jss.v035.i03
16. Sun J, Ajwani D, Nicholson PK, Sala A and Parthasarathy S (2017) Breaking Cycles In Noisy Hierarchies. Proceedings of the 2017 ACM on Web Science Conference: 151-160 http://dx.doi.org/10.1145/3091478.3091495
17. Suominen O and Mader C (2014) Assessing and Improving the Quality of SKOS Vocabularies. Journal on Data Semantics **3**(1): 47–73 http://dx.doi.org/10.1007/s13740-013-0026-0
18. Van den Bulcke T, Van Leemput K, Naudts B, van Remortel P, Ma H *et al.* (2006) SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. BMC Bioinformatics **7**(43) http://dx.doi.org/10.1186/1471-2105-7-43
19. Wrzodek C, Buchel F, Ruff M, Drager A and Zell A (2013) Precise generation of systems biology models from KEGG pathways. BMC Syst Biol **7**(15) http://dx.doi.org/10.1186/1752-0509-7-15
20. Wrzodek C, Drager A and Zell A (2011) KEGGtranslator: visualizing and converting the KEGG PATHWAY database to various formats. Bioinformatics **27**(16): 2314-2315 http://dx.doi.org/10.1093/bioinformatics/btr377
21. Zhang JD and Wiemann S (2009) KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor. Bioinformatics **25**(11): 1470-1471 http://dx.doi.org/10.1093/bioinformatics/btp167