Thought Graph: Generating Thought Process for Biological Reasoning

Chi-Yang Hsu
University of Texas at Austin
Austin, Texas, USA
ch52669@utexas.edu

Kyle Cox
University of Texas at Austin
Austin, Texas, USA
kylecox@utexas.edu

Jiawei Xu
University of Texas at Austin
Austin, Texas, USA
jiaweixu@utexas.edu

Zhen Tan
Arizona State University
Tempe, Arizona, USA
ztan36@asu.edu

Tianhua Zhai
University of Pennsylvania
Philadelphia, Pennsylvania, USA
tianhua.zhai@pennmedicine.upenn.edu

Mengzhou Hu
University of California San Diego
La Jolla, California, USA
mhu@health.ucsd.edu

Dexter Pratt
University of California San Diego
La Jolla, California, USA
depratt@health.ucsd.edu

Tianlong Chen
The University of North Carolina at Chapel Hill
Chapel Hill, North Carolina, USA
tianlong@cs.unc.edu

Ziniu Hu
California Institute of Technology
Pasadena, California, USA
acbull@caltech.edu

Ying Ding
University of Texas at Austin
Austin, Texas, USA
ying.ding@ischool.utexas.edu

ABSTRACT

We present the Thought Graph as a novel framework to support complex reasoning and use gene set analysis as an example to uncover semantic relationships between biological processes. Our framework stands out for its ability to provide a deeper understanding of gene sets, significantly surpassing GSEA by 40.28% and LLM baselines by 5.38% based on cosine similarity to human annotations. Our analysis further provides insights into future directions of biological processes naming, and implications for bioinformatics and precision medicine. Here's our Github Code.

CCS CONCEPTS

• **Applied computing** → **Life and medical sciences**.

KEYWORDS

large language model, natural language processing, semantic web biological process, gene ontology, bioinformatics

ACM Reference Format:
Chi-Yang Hsu, Kyle Cox, Jiawei Xu, Zhen Tan, Tianhua Zhai, Mengzhou Hu, Dexter Pratt, Tianlong Chen, Ziniu Hu, and Ying Ding. 2024. Thought Graph: Generating Thought Process for Biological Reasoning. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion), May 13–17, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3589335.3651572

WWW '24 Companion, May 13–17, 2024, Singapore, Singapore.

ACM ISBN 979-8-4007-0172-6/24/05
https://doi.org/10.1145/3589335.3651572

1 INTRODUCTION

The systematic study of human disease necessitates an in-depth understanding of the links between diseases, drugs, phenotypes, genes, and biological processes [4]. Analyzing gene sets that share common biological functions, locations, or regulatory mechanisms can reveal patterns in gene behavior across health and disease states, contributing to the advancement of precision medicine for cancer treatment [7]. Yet, the task of identifying biological processes from gene sets is fraught with challenges. Individual genes often display weak signals, and when strong signals are present, they rarely converge on a singular biological theme [7]. This complexity is compounded when different research groups studying the same biological systems arrive at vastly divergent conclusions.

In response to these challenges, our paper introduces the **Thought Graph** framework that aims to address two critical aspects: firstly, it adopted a Tree-of-Thought (ToT) [11] architecture to facilitate thought expansion, ensuring inclusive yet precise coverage of biological processes across varying specificity levels. Thought expansion is strategically directed with the assistance of a voter LLM, which guides the decision-making for future steps. This design aims to mitigate the potential discrepancies in human annotations encountered by researchers, yet ensure the quality of the generated processes. Second, our framework prioritizes the integration of domain-specific external knowledge bases to understand the semantics of connections within the Thought Graph. Consequently, it creates semantic relationships like "is-a" and "part-of" among

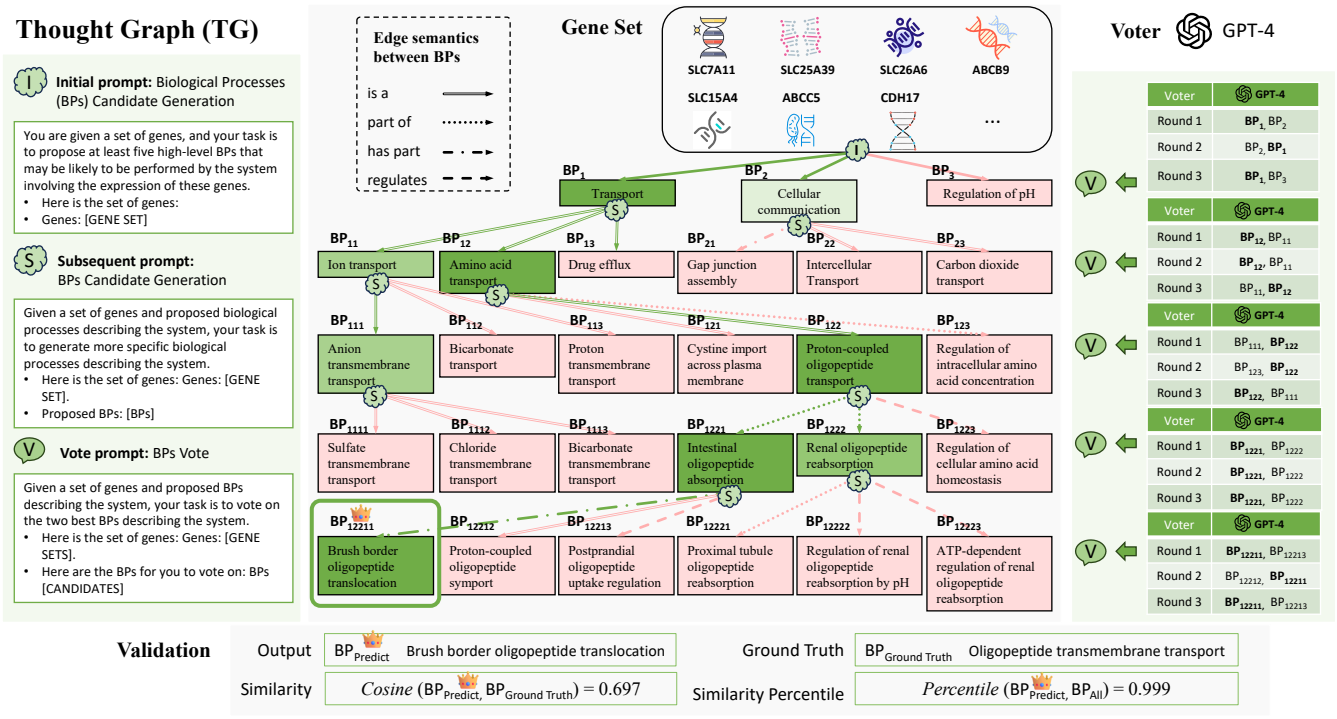

Figure 1: The flowchart presents the application of the Thought Graph to the Gene Ontology (GO) database. First, Thought Graph uses a gene set and initial prompt to generate three Biological Processes (BPs). Then, a voter evaluates and selects the best BP (dark green) and second best BP (light green), which are more accurately descriptive of the gene set. Each chosen BP, along with a subsequent prompt, is utilized to generate two additional, more specific BPs. This procedure is conducted recursively until Thought Graph has reached five layers. Finally, a voter chooses the final answer from the last layer.

various thought steps. This strategy not only facilitates complex decision-making processes but also ensures a more nuanced and interconnected understanding of biological systems, facilitating data interoperability and knowledge integration. Our novel contributions can be summarized as follows:

(1) We propose Thought Graph as a complex reasoning framework that generates diverse yet precise entities to tackle potential annotations discrepancies in biological processes.
(2) Thought Graph can generate thought graphs with edge semantics by recalling external knowledge (e.g., Gene Ontology) to build rich semantics among thought steps.
(3) We have successfully applied Thought Graph in biological process generation with significant improvement compared to SOTA methods, surpassing GSEA by 40.28% and LLM baselines by 5.38% in cosine similarity score, and identified the optimal steps of complex reasoning by balancing specificity and accuracy.

2 RELATED WORK

2.1 LLM Reasoning

Prompt strategies attempt to decompose a complicated problem into a sequence of smaller sub-problems so that the problem becomes more manageable [12]. One popular line of study is the Chain-of-Thought (CoT) [9] series, structuring prompts to encourage the LLM to step through its reasoning process, such as Least-to-Most prompting [12], and Self-Consistency with CoT (CoT-SC) [8]. However, these prompting strategies only utilize linear reasoning paths and struggle in tasks that require exploration and strategic lookahead. Alternatively, Tree of Thoughts (ToT) [11] and Graph of Thoughts (GoT) [3] excel in these sorts of tasks. LLM-based prompting frameworks' effectiveness is hindered by inherent limitations such as self-bias and hallucination. To address this, through in-context learning, our work introduces the semantics of edges within our Thought Graph (TG), offering structural information.

2.2 Knowledge Graph for LLM Reasoning

LLMs exhibit limitations in integrating new knowledge and occasionally generate hallucinations. A survey [1] on knowledge-graph-based knowledge augmentation in LLMs reveals using knowledge graphs (KGs) as a source of external information has promising results in reducing hallucinations. For example, MindMap [10] has developed a prompt pipeline enabling LLMs to comprehend and integrate KG input with their implicit knowledge. In our approach, we give LLM examples from the gene ontology knowledge graph to enable the edge semantics.

2.3 LLM Reasoning in Biomedical Domain

With the rise of LLMs, recent studies explore LLMs' application in various biomedical tasks. The gene set biological process was

formulated by [5] as inputting a gene set to an LLM and outputting a biological process name that is predominant in the system and correctly describing the function of the gene set. It's challenging because it requires the LLM to accurately understand and interpret complex biological concepts, including the nuanced roles of genes in various cellular contexts and their interactions within intricate biological networks. Although their results [5] have shown that GPT-4 provides better biological process names than the conventional Gene Set Enrichment Analysis (GSEA) [7], the performance is still far from perfect.

3 METHODOLOGY

3.1 Problem Formulation

Given a gene set $X = \{x_1, x_2, ..., x_n\}$, where each x_i is a gene, the objective $G = F(X)$ is to design a framework F to generate a tree structure graph $G = (N, E)$ that represents the terms (e.g., biological processes or pathways) associated with the genes in X. In this graph, N is the set of nodes, and E is the set of edges between these nodes.

3.2 Infrastructure of Thought Graph

Our framework Thought Graph adapts ToT [11] as a graph generator to generate a curated tree graph G, named Thought Graph. Thought Graph contains terms as the nodes N and their dependencies as edges E. ToT uses self-reflection to prune and only explore relevant paths. The result, after exploration, is a graph Thought Graph that illustrates the reasoning path and a final answer chosen selected from the last layer of the graph as the term that best describes the gene set $X = \{x_1, x_2, ..., x_m\}$.

3.2.1 Thoughts expansion. Thought Graph process with n steps proceeds in a breadth-first fashion to generate a tree of depth n. At each step, the process expands the tree by generating a set of candidate nodes. The first step generates a set of general "high-level" terms that describe the gene set, and subsequent steps iterate on the candidate terms by proposing more specific but related terms.

Step 1 (Initial Expansion). The first step is unique from all subsequent steps because its task is to generate the initial set of k candidate terms $T_i = t_1^1, \ldots, t_k^1$, where t_j^i denotes the term j from layer i. This set of candidate terms is generated with an "initial prompt" that takes the gene set as input: $T_i \sim p_\theta(t^1_{1\ldots k} \mid x_1 \ldots, x_n)$.

Subsequent Steps (Recursive Expansion). In step i, we use a Voter (V) to examine and vote across the candidate terms T_i: $V(p_\theta, T_i)(t_j^i) = 1[t_j^i = t_j^i *]$, where a good term $t_j^i * \sim p_\theta^{vote}(t_j^i * |B)$ is based on comparing the candidate terms T_i in the vote prompt, and select two best terms. For each selected term from the previous step, t_j^{i-1} and gene set X are added to the "subsequent prompt" for the LLM generates k new terms: $\{t_1^i, \ldots, t_k^i\} \sim p_\theta(t^i_{1\ldots k} \mid x_{1\ldots n}, t_j^{i-1})$. This process will be conducted recursively for $n-1$ times (minus the initial expansion). For the final layer, $t_1^n \sim t_k^n$ are presented to the LLM to choose the final answer.

3.3 Thought Graph

The Thought Graph output provides a representation of the stepwise reasoning process and integrates edge and node semantics for domain-specific context. Each node $n_i \in N$ is a unique biological process, arranged hierarchically to reflect varying levels of specificity. The edges E represent the relationships between these processes. Specifically, we use four pre-defined relations from the Gene Ontology (GO): *is a*, *part of*, *has part*, and *regulates*. These relations establish a hierarchy where, for instance, if A *is a* subtype of B, A is deemed more specific than B. This approach helps to elucidate the nuanced relationships between different biological processes, as detailed in the GO database.[1]

4 EXPERIMENT & EVALUATION

4.1 Data Collection

The GO database [2] forms the basis of our study. We specifically use a dataset compiled by Hu et al. [5] from the Biological Process branch of Gene Ontology consisting of 12,214 human gene sets, each annotated with a biological process name and description. Due to constraints in financial and computational resources, we randomly select 100 samples from this dataset for evaluation.

4.2 Baselines and Model Description

Our evaluation framework includes one domain-specific tool and five Large Language Model (LLM) baselines. GSEA (gene set enrichment analysis) [7] is a statistical method for associating the expression of groups of genes with biological processes. Our LLM baselines involve different approaches. Input-Output (IO) Prompting with zero-shot and zero-shot-9 prompts generate one and nine unique terms for a single gene set, respectively, with no examples, while few-shot includes five question-answer examples. Chain-of-Thought (CoT) employs the two top pathways from Thought Graph for detailed step-by-step prompting. The approach by Hu et al. [5] integrates expert-curated prompts with specific guidelines that solicit post-hoc critical analysis. For all LLM instances, we use GPT-4 (*gpt-4-1106-preview*) in Chat Completion mode with temperature 0.7. In Thought Graph, we set the number of steps to five and vote on two samples at each step to proceed.

4.3 Evaluation Methods

We use two evaluation metrics: cosine similarity and similarity percentile. Cosine similarity measures the semantic similarity of the predicted term to ground-truth term from 0 (no similarity) to 1 (identical). We calculate similarity using embeddings from SapBERT [6], a masked language model trained to model medical entity relations. After calculating the similarity between the predicted and ground-truth terms, we also calculate the similarity between the predicted term and all 12,214 terms in our dataset to form a null distribution. The percentile score is the percentile of the similarity between the predicted term and the ground-truth term in our null distribution. We also include the proportion of similarity percentiles greater than 99% as a proxy for accuracy.

Among the nine nodes that receive positive votes (indicated as green nodes in Fig. 1), the one with the highest similarity score is selected as the best score (b), while the score of the node predicted by Thought Graph is recorded as the predicted score (p). To establish a fair baseline comparison, we implemented IO zero-shot-9 to generate nine answers, and select the best of these for evaluation.

[1] https://geneontology.org/docs/ontology-relations/

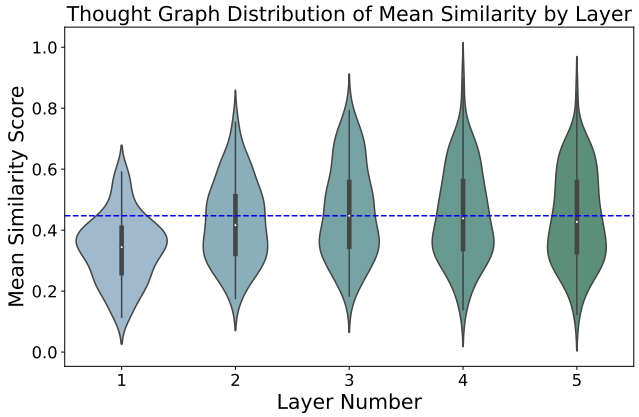

Figure 2: The distribution of the mean similarity score at each layer using Thought Graph (p). The blue line denotes the median of layer 3.

Method	Similarity	Percentile	Percentile > 99%
GSEA	24.78%	52.00%	17%
IO zero-shot	45.75%	77.00%	27%
IO zero-shot-9 (b)	59.68%	91.42%	61%
IO few-shot	48.73%	81.85%	32%
CoT	28.83%	43.71%	0%
Hu et al. [5]	52.31%	84.44%	43%
Thought Graph (p)	48.53%	80.90%	42%
Thought Graph (b)	**65.06%**	**95.05%**	**65%**

Table 1: Mean cosine similarity, mean cosine similarity percentile, and proportion percentile above 99% of a domain-specific tool and seven LLM methods on 100 GO data samples.

4.4 Performance Evaluation

Overall Performance: Table 1 indicates that Thought Graph (b) achieves the top performance in both cosine similarity (65.06%) and similarity percentile (95.05%). In particular, we want to posit that IO zero-shot learning emphasizes coverage across a wide range of biological process names (diversity), while the CoT focuses on an in-depth exploration of these names (specificity), whereas our framework is designed to balance both. Thought Graph (b) outperforms IO zero-shot-9 (b) and CoT, indicating that depth without breadth, or vice versa, is insufficient. Thought Graph and other LLM baselines outperform GSEA, and we also noticed GSEA cannot provide any terms for 26% of the time, highlighting the advantage of the LLMs. In addition, Thought Graph (p) scores lower than few-shot and Hu et al. baselines. This may be the result of our decision to constrain the final answer to the last layer. However, that Thought Graph (b) outperforms all baselines, including zero-shot-9, assures us that our approach to generating candidate sets of terms is promising, and that it is adept at generating a correct answer, but further optimization is needed.

Thought Graph Analysis: Layer-by-layer analysis in Fig.2 demonstrates increasing performance from layers 1 to 3, followed by a decrease in layers 4 and 5. This trend suggests a trade-off between specificity and accuracy, with layer 3 the optimal level by a small margin. While the performance at layer 1 is lower, this is largely because our initial prompt specifically requests "high-level" terms, and only generates three of them. As expected, the variance in mean similarity scores increases with the number of layers, as deeper layers explore deeper and more distant parts of the ontology, but stabilize after layer 3. In the latter layers, more specific terms are often voted out in favor of more accurate, general terms, demonstrating the ability of the voting mechanism to dynamically moderate specificity. Though our results reflect a modest sample size, layer 3 emerges as an early candidate for the optimal depth.

5 CONCLUSION

Thought Graph represents an advancement in the field of gene ontology and bioinformatics. Integrating gene set analysis with semantic graphs allows for a more nuanced and comprehensive understanding of biological processes. The effectiveness of the Thought Graph in mapping complex gene interactions and functions has been demonstrated, showing its potential to outperform existing methods. This novel method not only enhances the accuracy of gene set analysis but also opens avenues for research in understanding genetic influences on various BPs. Future work can expand on this foundation, exploring broader applications and measuring uncertainty in complex reasoning.

6 ACKNOWLEDGEMENT

We thank the support from NIH (OTA-21-008, R01LM014306-01) and NSF (NSF 2303038, NSF 2333703).

REFERENCES

[1] Garima Agrawal, Tharindu Kumarage, Zeyad Alghami, and Huan Liu. 2023. Can Knowledge Graphs Reduce Hallucinations in LLMs? : A Survey. arXiv:2311.07914 [cs.CL]
[2] M. Ashburner, C.A. Ball, Judith Blake, David Botstein, Heather Butler, and J. Cherry. 2000. Gene ontology: Tool for the unification of biology. *The Gene Ontology Consortium. Nat Genet* 25 (01 2000), 25–29.
[3] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2023. Graph of Thoughts: Solving Elaborate Problems with Large Language Models. arXiv:2308.09687 [cs.CL]
[4] Payal Chandak, Kexin Huang, and Marinka Zitnik. 2023. Building a knowledge graph to enable precision medicine. *Scientific Data* 10, 1 (2023), 67.
[5] Mengzhou Hu, Sahar Alkhairy, Ingoo Lee, Rudolf T. Pillich, Robin Bachelder, Trey Ideker, and Dexter Pratt. 2023. Evaluation of large language models for discovery of gene set function. arXiv:2309.04019 [q-bio.GN]
[6] Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-Alignment Pretraining for Biomedical Entity Representations. arXiv:2010.11784 [cs.CL]
[7] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* 102, 43 (2005), 15545–15550. https://doi.org/10.1073/pnas.0506580102 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.0506580102
[8] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. arXiv:2203.11171 [cs.CL]
[9] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903 [cs.CL]
[10] Yilin Wen, Zifeng Wang, and Jimeng Sun. 2023. MindMap: Knowledge Graph Prompting Sparks Graph of Thoughts in Large Language Models. arXiv:2308.09729 [cs.AI]
[11] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. arXiv:2305.10601 [cs.CL]
[12] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. arXiv:2205.10625 [cs.AI]