大規模言語モデルを用いた文献からの知識抽出と細胞内ネットワークの数理モデルのデータ駆動的な抽出

荒金 究
大阪大学 蛋白質研究所

1. はじめに

細胞は内的なゲノム情報と外的な環境情報を統合し、増殖、分化、細胞死などの多様な運命制御を行う。このような振る舞いを可能にするのは、細胞内のタンパク質をはじめとした生体分子の相互作用からなる細胞内ネットワークである。遺伝子変異などの要因でこのネットワークの制御構造が乱れると、細胞の運命制御に乱れが生じ、がんなどの疾患の発症につながることが知られている[1]。そのため、このような細胞内ネットワークの動態を理解することが細胞の運命決定や制御機構の解明に必要不可欠であると考えられる。

しかしながら、細胞内ネットワークは大きく複雑であり、一つの変化が全体に及ぼす影響を予測することは非常に困難である。この問題を解決する手法として、ネットワーク中の分子間の反応を常微分方程式でモデル化し、シミュレーション解析を通じてネットワークの振る舞いを理解する数理モデリングが用いられてきた。数理モデリングはネットワーク内のそれぞれの分子の時間的動態を定量的に議論できるようになり、より細胞内の分子メカニズムに迫ったモデルを立てることが可能になる。また、過去には疾患のような細胞の好ましくない状態を抑制あるいは脱するために最適な薬剤標的の予測・発見にもつながった[2]。

数理モデリングは強力なアプローチである一方で、モデルの構築に膨大な事前知識を要するというボトルネックを抱えている。これは、観察された現象を説明可能かつ生物学的に有意義な予測が可能なモデルを作るためには、扱っている細胞の種類や観察条件などの生物学的な文脈に基づいてモデルに含める遺伝子や相互作用を選択しなければならないためである。従来は、研究者が蓄えた生物学的知識や経験を基に、そして関連する論文やデータベース等を調査することで数理モデル構築は行われてきた。

こうした背景から、注目すべき細胞内ネットワークをデータ駆動的に構築するスケーラブルな手法の開発が期待される。そこで本研究では、深層学習技術を活用し、これまで人の手によって行われていた文献情報からの知識抽出を自動化し、細胞の動的な振る舞いを説明可能な数理モデルをデータ駆動的に構築する手法を開発することを目指す。

2. 文脈依存的な情報抽出パイプラインの構築

上記の課題に取り組むために、与えられた生物学的な文脈に即した遺伝子間相互作用を文献から抽出するパイプラインを構築した（図 1）。以下では、パイプラインの各段階の説明を行う。

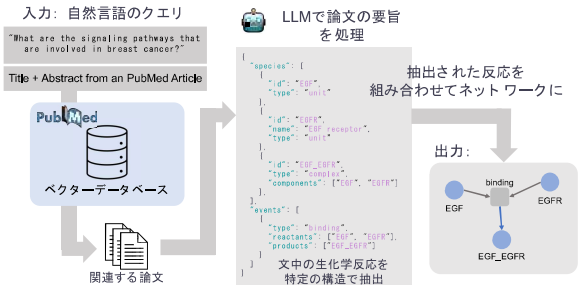

図 1：情報抽出パイプラインの概要

まず、文脈依存的な情報抽出を実現するために、与えられた文脈に関連する論文を検索可能なシステムを構築した。これには、テキストをそれが表す意味や文脈に基づいて高次元空間に埋め込むことができる Sentence Transformer [3]という自然言語モデルを用いて、生物医学系の論文データベースである PubMed の処理すること

で、入力となる自然言語のクエリに対して文脈的に近い論文を検索可能なベクターデータベースを作成した。これにより、特定の文脈に関連する論文のみから情報抽出を行うことが可能になる。

次に、ベクターデータベースから取得された論文から、目的の情報を抽出する手法を構築した。ChatGPT をはじめとする大規模な生成言語モデルは、特定のタスクを解く際に、それに特化した学習を行わずとも、入力となるプロンプトで指示するだけで高い性能を示すことが知られている[4]。そこで、OpenAI の gpt-3.5-turbo に対して、検索された論文のアブストラクトに含まれる遺伝子間相互作用や生化学反応を、決まった構造で抽出するようなプロンプトを与えることで、文献から計算機上で扱いやすい構造で情報抽出を行うことを試みた。

最後に、このようにして抽出されたネットワークを統合し、最終的な出力ネットワークを得る。ここでは、抜き出された遺伝子等の分子の名前を、Gilda[5]というソフトウェアを用いてデータベースの ID に紐付けて標準化を行った。紐づけられた ID によって、論文によって異なる表記がされている場合でも遺伝子の同一性を認識することができるようになり、ネットワークの統合を行うことが可能になる。

このパイプラインを用いて得られる出力ネットワークの例を図 2 に示す。これは、EGFR パスウェイの数理モデリングに関する論文[6]を入力とし、ベクターデータベースから近傍 100 件の論文を処理した際の結果である。実際に、EGFR を中心として、他の関連遺伝子との相互作用や下流の遺伝子の活性化反応を含むようなネットワークが抽出されることが確認できた。

しかしながら、出力されたネットワークには、化学量論的に誤った反応を含むことがある。そのため、このネットワークを数理モデルとして用いる場合は、そのような反応を修正、あるいは取り除くような処理が必要である。

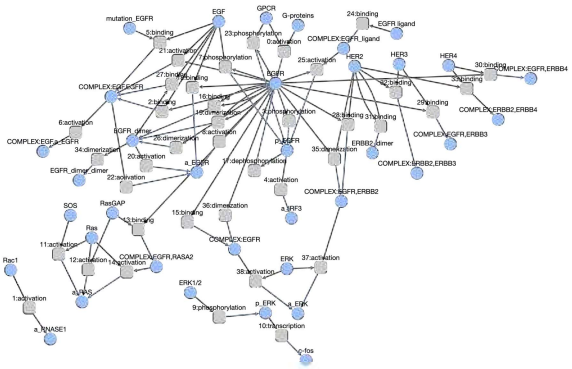

図 2：出力ネットワークの例

3. ドメイン知識の有無による影響の検証

一般に、特定のドメインに特化したタスクを言語モデルに解かせる際には、そのドメインのコーパスで学習されたモデルの方が高いパフォーマンスを示すことが知られている。実際に生物や医学の知識が関わるようなタスクにおいては、PubMed で事前学習されたモデルが良い性能を示すことが確認されている[7], [8]。そこで、上記のパイプラインで用いた Sentence Transformer のドメイン知識の有無がベクターデータベースの性能に及ぼす影響を検証した。

比較には、元々用いていた一般的なコーパスで事前学習されたモデルに加えて、PubMed で事前学習された二つのモデルを用いた（表 1）。また、これらのモデルを用いて作成されたベクターデータベースの性能を定量的に比較するために、(1) Title-abstract matching と (2) MeSH term matching という二つの評価タスクを考案した。

(1) では、論文のタイトルを埋め込んだ際に、同じ論文の要旨から作られたベクトルが近傍に埋め込まれていたかを評価した。PubMed の論文 20 万件を用いて、評価指標として Mean Reciprocal Rank (MRR) と上位 10 件までの Normalized Discounted Cumulative Gain (nDCG@10) を用いた。(2) では、PubMed の論文にトピック等に基づいて付与される Medical Subject Headings (MeSH) term を用いて、入力として与えられた論文と、その近傍に埋め込まれた論文の MeSH Term の類似性を評価した。

表 1：比較した Sentence Transformer モデル

Model name	Pretraining dataset type	Fine-tuning dataset type	References
sentence-transformers/multi-qa-mpnet-base-dot-v1	Wikipedia, BooksCorpus, and other corpora [9]	Collection of Q&A corpora from various domains	[3], [10]
NeuML/pubmedbert-base-embeddings	PubMed abstracts, PubMedCentral full texts[8]	PubMed title-abstract pairs & similar title pairs	[11]
pritamdeka/S-PubMedBert-MS-MARCO	PubMed abstracts, PubMedCentral full texts[8]	MS-MARCO Q&A dataset	[12], [13]

ここでは入力論文として[6]を使用し、上位 500 件の論文と入力論文の MeSH term の一致度を F1 スコアで評価した。

これらの評価タスクでの各モデルのスコアを表 2 にまとめる。平均して、ドメイン知識を有するモデルの方が評価タスクでより良い性能を示すことが確認できた。特筆すべき点として、PubMed で事前学習された二つのモデルのうち、直接 PubMed の title-abstract pair でファインチューニングされたモデルよりも Q&A データセットで学習されたものの方が突出して良い性能を示すことがわかった。これは、文意を反映した文章埋め込みを学習させるには、Q&A がより適したデータセットであることを示唆する結果であると考えられる。

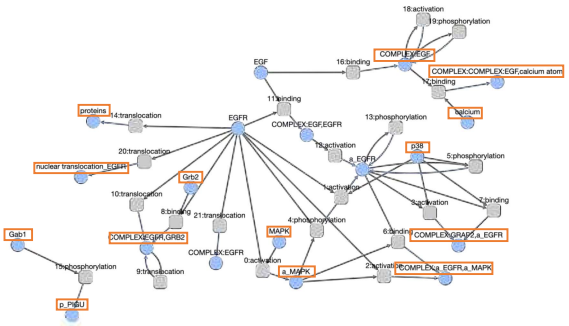

図 3：ドメイン知識を有するモデルを用いた際の出力ネットワーク

最後に、評価タスクで最も性能が良かったモデルのベクターデータベースを用いて、図 2 と同条件で得られるネットワークの出力結果の比較を行った（図 3）。元々のモデルを用いて得られたネットワークと比較して、より小さいネットワークが得られた。これはベクターデータベースから取得される論文の関連性が上がったことにより、ノイズと捉えられる関連性が低い反応が含まれにくくなったためではないかと考えられる。しかし、これを確かめるためには厳密なネットワークの評価指標が必要になる。

4. おわりに

今回の手法を発展させる上で、解決すべき課題がいくつか考えられる。一つは、論文から抽出されるネットワークの正確性を評価する必要がある。これには、あらかじめ正解のわかっているデータセットを作成し、それを用いて言語モデルの性能評価を行う必要がある。

他にも、出力されたネットワークの評価指標についても考える必要がある。これには、入力された生物学的文脈に対する出力ネットワークの妥当性や、化学反応ネットワークとしての妥当性などの要素を盛り込む必要があると考えられる。しかし、どのようなネットワークが妥当であるかは細かい実験条件等によって大きく異なることもある。そのため、言語モデルのみを用いて上述の問題を全て解決する手法はあまり現実的ではない可能性がある。そのため、あらかじめ多くの相互作用を網羅した大きなネットワークを用意し、そこに文脈情報を反映して、

表 2：ベクターデータベースの性能比較

Model name	Title-abstract		MeSH term
	MRR	**nDCG@10**	**Average F1 Score**
sentence-transformers/multi-qa-mpnet-base-dot-v1	0.257	0.108	0.246
NeuML/pubmedbert-base-embeddings	0.382	0.168	0.260
pritamdeka/S-PubMedBert-MS-MARCO	0.431	0.200	0.353

関連性が高いと考えられるサブネットワークを抽出するようなアプローチも考えられる。このようなアプローチを採用すれば、入力された文脈に対して「骨組み」のようなネットワークを提案し、そこから周辺のノードやエッジを追加・削除して編集できるようなインタラクティブなツールの開発にも繋げることが可能である。

参考文献

(1) M. A. Clarke and J. Fisher, Nat. Rev. Cancer, **20**, 343-354, (2020).

(2) B. Schoeberl, et al., npj Syst. Biol. Appl., **3**, 1-17, (2017).

(3) N. Reimers and I. Gurevych, Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 3982-3992, (2019).

(4) T. Brown, et al., Proc. of the 34th International Conference on Neural Information Processing Systems, 1877-1901, (2020).

(5) B. M. Gyori, et al., Bioinf. Adv., **2**, 1-5, (2022).

(6) B. N. Kholodenko, et al., J. Biol. Chem., **274**, 30169-30181, (1999).

(7) J. Lee, et al., Bioinformatics, **36**, 1234-1240, (2020).

(8) Y. Gu, et al., ACM Trans. Comput. Healthcare, **3**, 1-23, (2021).

(9) K. Song, et al., Proc. of the 34th International Conference on Neural Information Processing Systems, 16857-16867, (2020).

(10) https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-dot-v1

(11) https://huggingface.co/NeuML/pubmedbert-base-embeddings

(12) P. Deka, et al., J. Data Intell., **3**, 474-505, (2022).

(13) https://huggingface.co/pritamdeka/S-PubMedBert-MS-MARCO