テキストマイニングによる PubMed・PubMed Central からの遺伝子ネットワークの抽出

荒金究 [1]　井元宏明 [1]　岡田眞里子 [2]
[1] 大阪大学大学院理学研究科　[2] 大阪大学蛋白質研究所
{k.arakane,himoto,mokada}@protein.osaka-u.ac.jp

概要

生物学においては，細胞内のタンパク質の生化学反応ネットワークを数理モデルとして記述し，計算機を用いてその時空間動態を再現することでその系に関する洞察を得るシステム生物学という分野がある．しかしモデルを作るには再現したい現象に対する知識を得るために文献情報を大量に収集する必要があり，また知識の偏りによって重要な因子を見落とす可能性があるなどの問題点がある．これらを解決するために，我々は公開されている論文データベースからデータ駆動的にモデルを構築することを目指している．本稿では，論文中に出現する生物学的な固有表現の出現頻度や共起頻度情報を用いて疾患など特定の生命現象に関連するタンパク質ネットワークを抽出する手法と，既存の知識データベースから数理モデルを自動生成する手法を紹介する．

1　はじめに

システム生物学の分野では，シグナル伝達経路というタンパク質の生化学反応系を連立常微分方程式（ODE）を用いて数理モデル化し，計算機を用いてその動態を再現するという手法が用いられている [1]．そのようなシミュレーション解析を通じて，その経路の中で重要な働きをする因子を特定したり，それまで知られていなかった制御機構を予測したりすることで新たな知見を得ることが可能である．

しかしながら，注目する現象を計算機上で再現できるモデルを構築するためには数多くの文献に当たり，シグナル伝達経路を構成するタンパク質やそれらの相互作用についての情報を収集することが必要である．このような人の手によるキュレーションをベースとした従来のモデル構築手法は多大な時間を要し，なおかつバイアスを排し切ることができない．そのため自然言語処理を活用し，データ駆動的に論文データベースからシグナル伝達経路の数理モデルを自動的に構築する手法が確立すれば，バイアスを取り除きつつ，モデル構築からシミュレーション解析，仮説生成というサイクルを高速化し，より多くの発見をもたらす事ができる．

本稿では，そのようなデータ駆動型の数理モデル構築を目的として，論文中の遺伝子や疾患名などの固有表現の出現頻度や共起頻度を用いたタンパク質間相互作用（Protein-Protein Interaction; PPI）ネットワークの抽出と，それを用いて特定の疾患に関連するネットワークを効率よく抽出する手法，そして最後に知識データベースを用いた自動的な数理モデルの構築手法を紹介する．

2　研究手法

以下より，本稿で紹介する研究手法の各ステップの詳細について記す．

2.1　固有表現抽出と共起頻度解析

まず最初に，遺伝子や疾患名といった生物学的な固有表現同士の共起頻度解析を行うため，PubTator central[2]（以降 PubTator）データベースを利用した．PubTator では，機械学習モデルを用いて PubMed, PubMed Central に登録されている生物医学系の論文中に存在する生物学的な固有表現を標識し，なおかつそれらを適切なデータベースの項目に紐づけたデータが公開されている．またこのデータの大きな特徴として標識された各固有表現が，例えばタンパク質名であればタンパク質のデータベースなど，適切なデータベースの項目と紐づけられていることが挙げられる．そのため，（後述する PPI データベースのような）異なるデータベースの情報と組み合わせることも容易である．よって，今回はそれを利用した共起頻度解析を行った．

共起頻度解析は，PubTator から得たデータセット

中に出現する全ての固有表現に対し行なった．この時，同一の文中に二つ以上の固有表現が出現した場合を一度の共起と見做した．

2.2 PPI ネットワークの構築と重みづけ

ある細胞で発現するタンパク質間の相互作用を網羅的に含んだ PPI ネットワークの中には，その細胞内で働くシグナル伝達経路の情報が含まれていることが期待される．そのため，PPI ネットワークから，共起頻度など自然言語処理により得られた情報を用いて，疾患などの特定の現象との関連性が高いサブネットワークを抽出するタスクを考えた．

まず，ネットワークの構築のために OmniPath[3] という PPI データベースを用いる．OmniPath は，タンパク質だけでなく DNA や RNA などの様々な生体分子間の相互作用のデータベースを統合したデータベースである．最終的なネットワークの大きさを制限するために，OmniPath に登録されている PPI のうちいくつかの条件[1)]を満たすものを抜き出し，それらを組み合わせたネットワークを重みづけを行う対象とした．最終的に 1142 のノードと 1993 のエッジからなるネットワークが構築された．

ここで，特定の生命現象との関連性をネットワークのノードやエッジの重みとして表現したい．そのため，まず注目する現象を表すキーワードを与えた．キーワードは共起頻度解析中に出現した（遺伝子名や病名などの）固有表現に紐づけられる．その後 PubTator の中の論文のうち，与えられたキーワードを含む論文を抜き出し，再度共起頻度解析を行う．ノードにはそれが表すタンパク質の出現頻度，エッジにはそれが結ぶタンパク質同士の共起頻度が，それぞれに関してネットワーク全体で正規化された値を紐づける．同様の操作を PubTator データセット全体から算出した出現頻度や共起頻度の値を用いて行い，論文にフィルタをかける前後で対応するノードとエッジ同士の重みを比較する．これらの操作を通じて，与えられたキーワードの文脈において強調されるノード（固有表現）とエッジ（共起）の値が大きくなるように重みづけることができる．

2.3 重みに応じたサブネットワーク抽出

前節で特定の生命現象に重みづけられたネットワークにおいては，前述の通り注目する現象との関連性が高いタンパク質や相互作用により大きい重みが紐づけられていることが期待される．単純に重みに関して閾値を設定し抽出するネットワークに含めるノードやエッジを選択することも可能だが，シグナル伝達経路の中に閾値に満たない重みを持つノード・エッジが多数混在する可能性があるため，この手法ではシミュレーション解析が行えるような適切なサイズの連続した経路情報を取り出すことができない．

そこで，この問題を解決するために，グラフ理論における Prize-Collecting Steiner Tree (PCST) 問題 [4] の応用を考える．PCST は，ノードの重み（prize）を最大化しつつ，エッジの重み（cost）を最小化する木を求める組合せ最適化問題である．そこでこの PCST 問題を解くアルゴリズムを自分たちのネットワークに適用した．サイズが大きくなりがちな生物ネットワークに PCST アルゴリズムを適用するという考えは新しいものではない [5, 6, 7, 8]．しかしながら，PCST 問題を解くことで得られたネットワークをシミュレーション解析に用いる試みは初めてと思われる．そこで，本研究では巨大なネットワークに対しても高速に解を計算できるヒューリスティックアルゴリズム [9] を利用した．また，問題を解く際にはノードとエッジに紐づけられた重みを変換した値を用いた．

2.4 数理モデルへの変換

次に，PPI ネットワークを数理モデルに変換しシミュレーション解析を行うために，本研究室で開発した生化学シミュレーション解析ソフト BioMASS[10] を利用した．BioMASS は，シグナル伝達経路における生化学反応を自然言語に近い形式（中間言語）で記述することで，これを自動的に ODE モデルに変換する Text2Model と呼ばれる機能を有する（図 1）．

ここで見られるように，シグナル伝達経路においてはタンパク質同士が結合し複合体を形成したり，化学修飾を通じてタンパク質を活性化・不活性化したりといった制御関係の方向性を有する反応が起こる．数理モデル化にはこのような反応の連鎖の方向性を記述することが必要になる．

しかしながら，前節までに得られた PPI ネットワークから，シグナル伝達経路の数理モデルを得ることは難しい．これは，図 1 における `EGF_ErbB1` や `RafP` などに相当するような，シグナルを下流に

1) ソースとして KEGG が記載されている (1) 方向性がある（有向エッジで表される） (2) 全てのソースで情報が一致している (3) など．

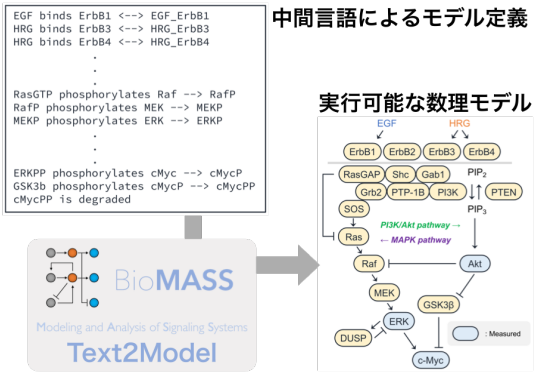

図 1　BioMASS の Text2Model の概要図．[11] より改変．

伝達する因子が明示されておらず，また無向と有向のエッジが混在し，かつ閉路が存在するような PPI ネットワークのみから上流・下流の情報を得ることが困難であるためである．

そこで，次に，このような反応の方向性が既にわかっているシグナル伝達経路に対してシステマティックな処理を行うことで，このネットワークを実行可能な数理モデルに変換可能かを試行した．

ここでは，シグナル伝達経路の知識データベースである KEGG PATHWAY[12] に登録されている Human ErbB Pathway を元に作成したネットワークを用いた．KEGG から得られたネットワークに対する事前処理として，まず明示されていない逆反応や，リン酸化や活性化状態を表す中間的なノードを自動的に追加し，その後一部のタンパク質の分解反応を手動で加えることを行った．動的なモデル生成には，公開されている時系列の実験データ [11] を用いてパラメータ推定とシミュレーション解析を行った．パラメータ推定とは，数理モデルの中の反応定数などのパラメータに関して，再現したい現象とのずれが最も小さくなもの遺伝的アルゴリズムなどを用いて推定する手法である．また，実験データには乳がんの異なるサブタイプ由来の細胞株（MCF-7, MDA-MB-231）に対して 2 種類の成長因子（EGF, HRG）で処理したデータが含まれている．最終的に，生成したモデルで細胞内のタンパク質の時空間動態をある程度再現可能であることを実証した．

3　結果と考察

構築した PPI ネットワークから抽出されたサブネットワークの例を図 2 に示す．このネットワークを抽出する際に用いたキーワードは “Breast Cancer” である．この結果から，本手法により，与えられたキーワードの文脈において強調されたノードやエッジを中心的に抽出することができ，なおかつ少数の重みの小さいものも抽出できていることがわかった．また，抽出されてくるネットワークの大きさは重みの変換のパラメータを変えることで大まかに調節することできた．

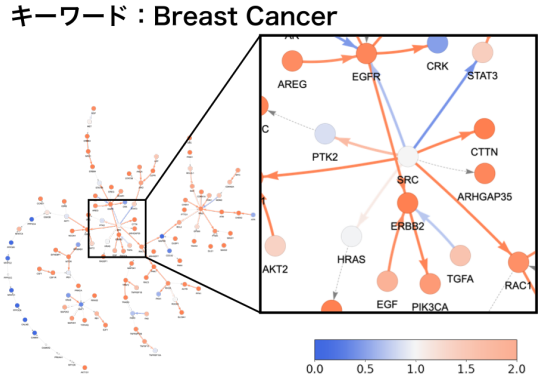

図 2　PPI ネットワークから抽出されたサブネットワークの例．各ノードやエッジに紐づけられている重みが色で表現されている．

また，抽出された遺伝子からそれぞれのコンテキストにおいて生物学的に意味のあるネットワークが抽出できていることが確認できた．例えば図 2 で示した “Breast Cancer” のサブネットワークにおいては，がん化と関わりがあることが知られている ERBB2[13] やそれと同一のタンパク質ファミリーに属する遺伝子 EGFR が含まれていた．また “Inflammation” をキーワードとして抽出した場合では，細胞の炎症反応において中心的役割を果たすと考えられている NF-κB シグナル伝達経路 [14] の遺伝子が連なって抽出できていることが確認された．

また，与えられた文脈において強調されていなかったノードも注目に値する．すなわち，キーワードに関連するものとして抽出できた遺伝子は，換言すればその働きもよく研究されている有名な遺伝子だと言える．一方で，この手法で抽出されてきた関連性の低いと思われるノードやエッジは，この後の解析で注目する現象を再現するために重要な働きをする可能性がある．本稿で紹介した手法では，このように細胞内の遺伝子のシステムの全体像を抽出できる利点があると考えている．

しかしながら，今回の結果として得られた PPI のサブネットワークは，この後に想定している数理モデリングによるシミュレーション解析に用いることはできない．これは，2.4 節で述べたような理由の他に，今回用いた PCST アルゴリズムが無向グラフ

を解くものであったことがより大きな要因として挙げられる．

シグナル伝達経路のモデリング解析では，しばしば入力と出力ノードが定義される．しかし，無向グラフを解いた場合，想定される入力と出力ノードの間に路が存在するネットワークが得られるとは限らない．その上，生物ネットワークにおいては閉路のような特殊な構造が重要な働きをする場合が多い．そのような特徴を有したネットワークは PCST 問題を解くことでは得られない．そのため，本研究の目的を達成するためには異なる手法が必要だと考えられた．

次に，KEGG PATHWAY 上の Human ErbB pathway を元に作成したネットワークをシステマティックに数理モデルに変換し，それを用いて実際にシミュレーション解析を行った結果を述べる．

KEGG から得られたネットワークに対しては，2.4 で述べたような処理を通して BioMASS において Text2Model で対応する形式に変換することができた．

このようにして得られた数理モデルに対し，過去に得られた 2 種類の細胞株の 2 条件の実験データ [11] を用いてシミュレーションを行った結果，本手法で得られたモデルはこれらの細胞の応答をある程度表現できることがわかった（図 3）．KEGG から得られるような公共のネットワーク情報から，ほぼ自動的に実行可能な数理モデルを生成できた点は成果として大きい．同モデルが異なる細胞株由来のデータを説明することができたという結果は，自動的な ODE モデル生成が実現可能であることと，今後論文データベース等から抽出した情報をどのように構造化すれば良いかを示唆するものである．

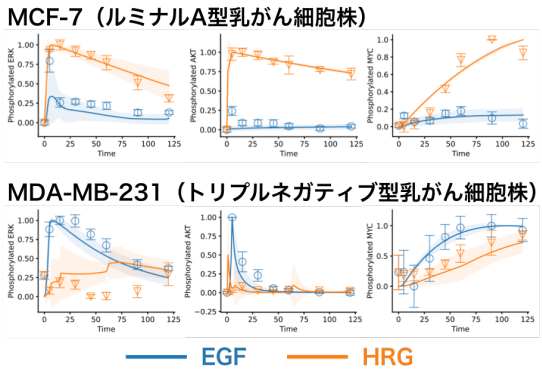

図 3　ErbB pathway のシミュレーション結果．実線がシミュレーション結果，エラーバー付きの点が実験データを表す．

4　おわりに

本稿では，(1) 論文中の固有表現の出現頻度や共起頻度情報を用いて特定の生命現象に関連するネットワークを疾患名などのコンテキスト依存的に抽出する手法と，(2) 知識データベース KEGG PATHWAY に登録されている情報を元に実行可能なモデルをシステマティックに構築する手法を紹介した．

今回示した (1) の結果では論文情報からデータ駆動的にシミュレーション解析を行えるようなシグナル伝達経路を抽出することは困難であった．これは抽出されたネットワークとシミュレーション解析が行える数理モデルとの間にまだ隔たりが存在するためであった．

(2) の手法に関しては，公共データベースにある有向ネットワーク情報から，人の手による調整をほとんど介さずに実行可能な数理モデルに変換できた．また，同モデルは実験データをある程度再現することが可能であった．まだ完全に自動的にシステマティックにモデルを生成するには至っていないが，今回の結果は自動的なモデル生成に向かうための足がかりとなると考えている．

今後の課題として，(1) と (2) の間をつなぐような手法を開発する必要がある．今回は，(1) でグラフ理論で発展した PCST 問題の応用を扱ったが，この問題の解を求めるだけでは十分ではない可能性があることを述べた．ここで，より目的に沿うネットワークを抽出するために，PCST 問題を解くアルゴリズムの代わりに深層学習モデルを導入することも考えられる．グラフを扱うことに長けたグラフニューラルネットワークモデルと強化学習モデルを組み合わせた手法などが適用できると考えている．

またそのほかの改善点としては，よりデータ駆動型の研究に近づくために，PPI ネットワークを得るために OmniPath のようなデータベースを参照するのではなく，直接 PubMed 等の論文データベースから抽出することが考えられる．これは自然言語処理の分野では関係性抽出に該当するタスクであり，深層学習モデルを導入することで達成できると思われる．

このように，今回導入することに能わなかった手法を今後取り入れることで，当初の目的であったシグナル伝達経路のモデルの自動抽出を実現できるのではないかと考えている．

謝辞

本研究は，JST，CREST バイオ DX（JPMJCR21N3）（研究代表者　岡田眞里子）の支援を受けた．本研究の議論に関して，京都大学　下平英寿研究室，理化学研究所　泰地真弘人研究室に感謝いたします．

参考文献

[1] Uri Alon. **An Introduction to Systems Biology: Design Principles of Biological Circuits**. CRC Press, July 2006.

[2] Chih-Hsuan Wei, Alexis Allot, Robert Leaman, and Zhiyong Lu. PubTator central: Automated concept annotation for biomedical full text articles. **Nucleic Acids Research**, 47(W1):W587–W593, July 2019.

[3] Dénes Türei, Tamás Korcsmáros, and Julio Saez-Rodriguez. OmniPath: Guidelines and gateway for literature-curated signaling pathway resources. **Nature Methods**, 13(12):966–967, December 2016.

[4] Maria Minkoff. **The Prize Collecting Steiner Tree Problem**. PhD thesis, Massachusetts Institute of Technology, 2000.

[5] Shao Shan Carol Huang and Ernest Fraenkel. Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks. **Science Signaling**, 2(81):1–11, 2009.

[6] M. Bailly-Bechet, C. Borgs, A. Braunstein, J. Chayes, A. Dagkessamanskaia, J. M. François, and R. Zecchina. Finding undetected protein associations in cell signaling by belief propagation. **Proceedings of the National Academy of Sciences of the United States of America**, 108(2):882–887, 2011.

[7] Nurcan Tuncbag, Pamela Milani, Jenny L. Pokorny, Hannah Johnson, Terence T. Sio, Simona Dalin, Dennis O. Iyekegbe, Forest M. White, Jann N. Sarkaria, and Ernest Fraenkel. Network Modeling Identifies Patient-specific Pathways in Glioblastoma. **Scientific Reports**, 6:1–12, 2016.

[8] Jiajie Peng, Linjiao Zhu, Yadong Wang, and Jin Chen. Mining Relationships among Multiple Entities in Biological Networks. **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, 17(3):769–776, May 2020.

[9] Yahui Sun, Chenkai Ma, and Saman Halgamuge. The node-weighted Steiner tree approach to identify elements of cancer-related signaling pathways. **BMC Bioinformatics**, 18(Suppl 16), 2017.

[10] Hiroaki Imoto, Suxiang Zhang, and Mariko Okada. A Computational Framework for Prediction and Analysis of Cancer Signaling Dynamics from RNA Sequencing Data—Application to the ErbB Receptor Signaling Pathway. **Cancers**, 12(10):2878, October 2020.

[11] Hiroaki Imoto, Sawa Yamashiro, and Mariko Okada. A text-based computational framework for patient - specific modeling for classification of cancers. **iScience**, 25(3):103944, March 2022.

[12] Minoru Kanehisa and Susumu Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. **Nucleic Acids Research**, 28(1):27–30, January 2000.

[13] Dihua Yu and Mien-Chie Hung. Overexpression of ErbB2 in cancer and ErbB2-targeting strategies. **Oncogene**, 19(53):6115–6121, December 2000.

[14] Ting Liu, Lingyun Zhang, Donghyun Joo, and Shao-Cong Sun. NF-κB signaling in inflammation. **Signal Transduction and Targeted Therapy**, 2(1):1–9, July 2017.