

Web上の情報を利用したタンパク質 相互作用ネットワークの構築

澁谷 翔吾^{†1} 廣 安 知 之^{†2} 三 木 光 範^{†3}
横 内 久 猛^{†2} 吉 見 真 聡^{†3}

本研究では、タンパク質とタンパク質の相互作用の関係を視覚的に確認できるシステムを提案する。本システムでは、生命科学分野の研究者があるタンパク質と相互作用の関係にあるタンパク質を調査する際、タンパク質名を入力することで、そのタンパク質と相互作用関係にあるタンパク質がネットワーク構造で表示される。相互作用の関係については、PubMed/MEDLINE データベースで公開されている論文を対象とし、テキストマイニングを行うことで抽出した。また、本システムにおいては、テキストマイニングによって得られていないタンパク質を検索した場合、システムは自動的に Web から情報を取得し、ネットワークを構築する。

Building a protein-protein Interaction Network using Information on the Web

SHOGO SHIBUTANI,^{†1} TOMOYUKI HIROYASU,^{†2}
MITSUNORI MIKI,^{†3} HISATAKE YOKOUCHI^{†2}
and MASATO YOSHIMI^{†2}

In this paper, we propose a protein-protein interaction network developing system. Using this system, researches in the life science field put the names of protein in which researches are interested and researches got the protein networks as the results. Protein-Protein interaction is extracted from papers in PubMed/ MEDLINE database using text mining. Even when researches put the proteins which does not existed in database, the protein network can be constructed using information on the Web automatically.

1. はじめに

近年、ゲノム情報を得るための実験機器の性能向上に伴い、塩基配列情報が短時間に大量に取得できるようになった。これにより、遺伝子の発現が及ぼす生物学的機能は何かといった知見を得る研究が活発に行われている。しかし、これらの作業は人手に頼っている部分が多く、手間がかかると言われている。加えて、多くの研究者が実験を行い、次々と論文で発表しているため、得られた知見は主に構造化されていない自然言語の形で集積されている。研究者にとって、特定の研究課題に関連する文献を効率よく見つけ出すこと、加えてそこに記述されているゲノム情報を把握することは重要である。そこで、近年、大量の文献を機械的に処理し、そこに記述されているタンパク質と他のタンパク質との相互作用の情報を抽出するテキストマイニングが行われている¹⁾。

本発表では、タンパク質の相互作用の関係をネットワーク構造で視覚的に確認できるシステムを提案する。多量のタンパク質相互作用情報が蓄積されている中、それら情報の理解、つまり、全体像を掴みながらも、注目すべき点を見つけること、データ全体の傾向や特徴を捉えるためにも可視化は必要である。そこで、研究者があるタンパク質について詳しく調査するといった状況において、タンパク質と他のタンパク質との相互作用の関係を視覚的に表示するシステムは有用であると考えられる。本研究では、PubMed/MEDLINE データベース²⁾で公開されている論文をテキストマイニングすることで、そこに記述されているタンパク質と他のタンパク質との相互作用（以下、タンパク質相互作用）情報を抽出し、タンパク質相互作用データベースを構築した。提案システムでは、タンパク質相互作用情報からタンパク質相互作用ネットワークを作成する。本システムでは、タンパク質相互作用データベースにないタンパク質、つまり、テキストマイニングの対象となった論文に現れていないタンパク質が検索された場合、自動的に Web から情報を取得し、既存のタンパク質相互作用データベースの情報と併せてタンパク質相互作用ネットワークを作成する。

2 章では関連研究について、3 章ではタンパク質相互作用情報の抽出について、4 章では提案システムについて、そして、最後に 5 章で結論を述べる。

^{†1} 同志社大学大学院工学研究科

Graduate School of Engineering, Doshisha University

^{†2} 同志社大学生命医科学部

Department of Life and Medical Sciences, Doshisha University

^{†3} 同志社大学理工学部

Department of Science and Engineering, Doshisha University

2. 関連研究

関連の研究を大別すると、遺伝子やタンパク質名などの領域固有語を同定するための研究と、タンパク質の相互作用の情報を抽出する研究に分類可能である。

遺伝子やタンパク質を示す名称は多くの同義語、多義語が存在する他、省略された表記や研究対象領域独自の表現方法があり、任意のテキストから高精度に固有語を同定することは困難である。このような課題に対して、固有語を同定するために、2つのアプローチがある。一つは、European Bioinformatics Institute (EBI)^{*1}が公開しているようなタンパク質の知識ベースを利用することでテキスト中のタンパク質を同定する³⁾⁴⁾⁵⁾。この方法では、高い確率でタンパク質名を同定することは可能であるが、知識ベースに存在しないタンパク質名を同定することはできないといった欠点がある。もう一つは、知識ベースなどは利用せず、タンパク質名の特徴からそれらを同定する方法である⁶⁾。タンパク質名は、複合語を形成している場合が多く、また、大文字や数字、記号文字が混在する特徴的な単語が多く存在するといった特徴がある。それらの特徴からタンパク質名を同定する。

タンパク質の相互作用の関係を抽出するには、上述したように、テキスト中のタンパク質を同定する必要がある。さらに、タンパク質の相互作用の関係を示したキーワードを特定する。キーワードには、予め生命科学の分野の研究者により定義されている場合が多い。タンパク質名とキーワードなどを同定し、文脈自由文法のルールからタンパク質の相互依存の関係を抽出する研究が行われている⁵⁾⁷⁾⁸⁾。

3. タンパク質相互作用情報の抽出

3.1 概要

本システムでは、PubMed/MEDLINE データベース²⁾で公開されている生命科学分野の論文を対象としてテキストマイニングを行う。タンパク質相互作用情報の抽出の流れを図.1に示す。

対象とする生命科学分野の論文からタンパク質を同定し、タンパク質とタンパク質の相互作用の関係を示すキーワードを同定し、文脈自由文法でルールに基づいてそれらを抽出する。

3.2 タンパク質名データベース

論文中のタンパク質を同定するためには、本研究では、タンパク質名データベースを利用

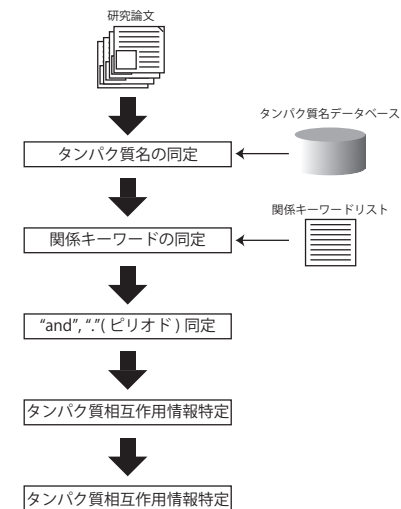


図1 タンパク質相互作用情報の抽出の流れ

している。本データベースは、European Bioinformatics Institute (EBI)で公開されているタンパク質名を保持したものであり、約18万語のタンパク質を保持している。

3.3 形態素解析とタグ付け

論文からタンパク質相互作用情報を抽出する最初のステップとして、入力文を形態素解析し、文脈自由文法でルールに基づいてタンパク質相互作用情報の抽出するために、各形態素にタグを付与する。タグの種類を表.1に示す。

表1 タグ一覧

タグ	説明	例
MOL	タンパク質名	kinase
KEY	関係キーワード(表.2)	activates
AND	並列	and
EOS	一文の終わり	.

タンパク質名データベースを利用して入力文内に存在するタンパク質名に”MOL”のタグ

*1 <http://www.ebi.ac.uk/>

を付与する．次に，タンパク質の相互作用を表すキーワードに”KEY”を付与する．タンパク質相互作用を表すキーワードを表.2 に示す．

表 2 関係キーワード一覧⁷⁾

accumulat(e,ed,es,ion)	cleav (e,ed,es)	inhibit (s,ed,ion)
activat(e,ed,es,or,ion)	demethylat (e,ed,es,ation)	reduc (e,ed,es,tion)
elevat(e,ed,es,ion)	Dephosphorylat (e,ed,es,ation)	repress (ed,es,ion)
hasten(ed,es)	sever (e,ed,es)	supress (ed,es,ion)
incit(ed,es)	influenc (e,ed,es)	modifi (ed,cation)
increas(ed,es)	contain (s,ed,es)	apoptosis
induct(e,ed,es,ion)	methyлат (e,ed,es,ation)	myogenesis
initiat(e,ed,es,ion)	phosphorylat (e,ed,es,ation)	interact (s,ed,ion)
promot(e,ed,es)	express (ed,es,ion)	react (s,ed,ion)
stimulat(e,ed,es,ion)	overexpress (ed,es,ion)	disassembl (e,es,ed)
transactivat(e,ed,es,ion)	produc (e,ed,es,tion)	discharg (e,es,ed)
up-regulat(e,ed,es,or)	block (s,ed)	mediat (e,ed,es)
upregulat(e,ed,es,or)	decreas (e,ed,es)	modulat (e,ed,es,ion)
associat(e,ed,es,ion)	deplet (e,ed,es,ion)	participat (e,ed,es,ion)
add(s,ition)	down-regulat (e,ed,es,ion)	regulat (e,es,ed,ion)
bind(s), bound	downregulat (e,ed,es,ion)	replac (e,ed,es)
catalyz(e,ed,es)	impair (s,ed)	substitut (e,ed,es,ion)
complex	inactivat (e,ed,es,ion)	

最後に，並列を示す”AND”と一文の終了を示す”EOS”をそれぞれ，形態素「and」と「.(ピリオド)」に付与する．

3.4 抽出ルール

前節では，タンパク質相互作用情報の抽出に必要な情報，つまり，タンパク質名，関係キーワード，「and」，「.(ピリオド)」にそれぞれ，”MOL”，”KEY”，”AND”，および”EOS”のタグを付与した．本節では，これらのタグから文脈自由文法において，与えたルール（文法）に基づいて，タンパク質相互作用情報を抽出する．ルールを表.3 に示し，抽出の流れを図.2 に示す．

図.2 に示すように，入力文に対してタグ付けを行い，文脈自由文法でルールを適用し，ルールに合うものだけをタンパク質相互作用情報として抽出する．

3.5 タンパク質相互作用情報

本節では，抽出したタンパク質相互作用情報について述べる．本研究では，PubMed/MEDLINE

表 3 タンパク質相互作用抽出ルール

S(start symbol) → Interaction
Interaction → Molecule Keyword Molecule . Interaction and Interaction
Molecule → MOL Molecule and Molecule
Keyword → KEY Keyword and Keyword
and → AND
. → EOS

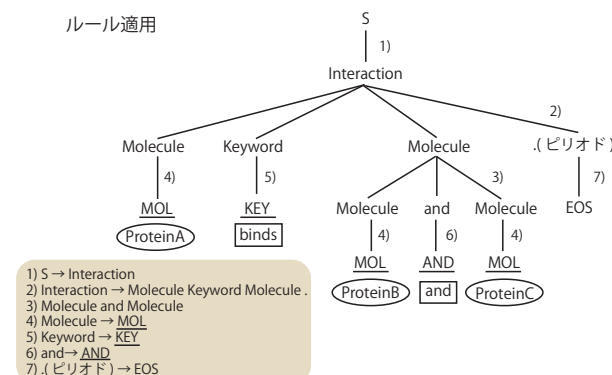
入力文

ProteinA binds to ProteinB and ProteinC .

タグ付け

ProteinA binds to ProteinB and ProteinC .
MOL KEY MOL AND MOL EOS

ルール適用



抽出結果

binds : ProteinA, ProteinB ProteinA と ProteinB は bind で関係している
binds : ProteinA, ProteinC ProteinA と ProteinC は bind で関係している

図 2 タンパク質相互作用情報の抽出

データベース²⁾において，タンパク質名「endothelin」で検索した結果，上位 100 件に現れた論文を対象としてテキストマイニングを行った．計 1250 個の文に対してテキストマイニングを行った結果，約 1100 個のタンパク質相互作用情報を抽出した．抽出例を以下に示す．

(a)
Input : We define a ASK1 sequence that binds to DJ-1.
Output : (binds : ASK1, DJ-1)
(b)
Input : We also demonstrate that a mouse strain lacking the dopamine signaling molecule DARPP-32 has significantly reduced levels of both Lrrk2 and alpha-synuclein.
Output : (reduced : DARPP-32, alpha-synuclein)

(a) の例では、タンパク質名”ASK1”と”DJ-1”を同定し、関係キーワード”binds”からそれらをタンパク質相互作用情報として抽出している。(b) の例では、タンパク質名”DARPP-32”と”alpha-synuclein”を同定し、関係キーワード”reduced”からそれらをタンパク質相互作用情報として抽出しているが、タンパク質”Lrrk2”をタンパク質名として特定できていない問題がある。これは、タンパク質”Lrrk2”がタンパク質名データベースに存在しないためである。

4. 提案システム

本章では、タンパク質の相互作用ネットワークを表示するシステムを提案する。

4.1 概要

本システムでは、検索したタンパク質を中心としたタンパク質相互作用ネットワークを作成する。タンパク質相互作用ネットワークの作成は、3章で抽出したタンパク質相互作用情報を基にしている。検索したタンパク質がタンパク質相互作用データベースにない場合は、システムが自動的に検索タンパク質に関する情報を Web 上から取得し、既存のタンパク質相互作用情報と併せてタンパク質の相互作用ネットワークを作成する。

4.2 タンパク質相互作用データベース

本システムでは、タンパク質相互作用情報を利用する。そこで、3章で解説したように、PubMed/MEDLINE データベース²⁾において、タンパク質名「endothelin」で検索した結果、上位 100 件に現れた論文を対象として、テキストマイニングを行い、タンパク質相互作用情報を抽出した。本データベースには、あるタンパク質とあるタンパク質が関係キーワードで関係しているという情報を保持している。

4.3 ユーザインタフェース

提案システムのユーザインタフェースを図.3 に示す。

「Load Protein」ボタンはタンパク質相互作用データベースに存在するタンパク質相互

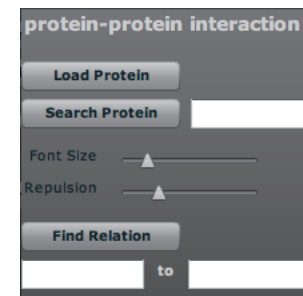


図 3 ユーザインタフェース

作用の全情報を基にネットワークを作成する。「Search Protein」ボタンはタンパク質名を指定し、そのタンパク質を中心としてタンパク質相互作用ネットワークを作成する。その下にある 2 つのスライダーは表示されたタンパク質相互作用ネットワークのノードの大きさ、およびエッジの長さを調整するためのものである。「Find Relation」ボタンはあるタンパク質とあるタンパク質がどのようなタンパク質を介して関係しているのかを表示するものである。

4.4 タンパク質検索

本節では、タンパク質検索について解説する。タンパク質検索の流れを図.4 に示す。

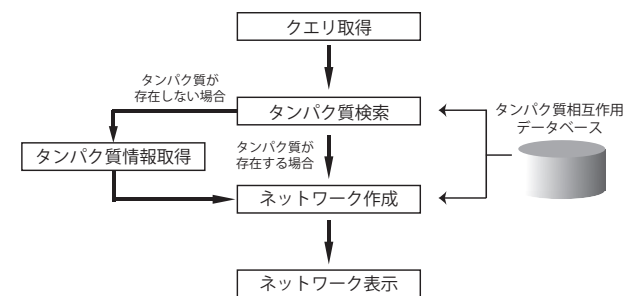
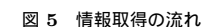


図 4 タンパク質検索の流れ

システムは検索クエリを取得し、そのタンパク質がタンパク質相互作用データベースに存在するか検索する。存在する場合は、そのタンパク質の相互作用情報を取得し、タンパク

CINII API を利用し、得た結果から文献の序論を取得し、タンパク質名、および関係キーワードを同定し、タンパク質相互作用情報として抽出し、タンパク質をノードとしエッジで繋ぐ。この場合、検索タンパク質と相互作用の関係にあるタンパク質は検索タンパク質からの距離 1 のタンパク質相互作用ネットワークとする。さらに、序論で同定したタンパク質がタンパク質相互作用データベースに存在するか検索し、存在する場合は、それらのタンパク質とノードで繋ぎタンパク質相互作用ネットワークとする。これを上述した通り、距離が N になるまで繰り返し、検索タンパク質からの距離 N のネットワークを作成する。



上述したように、タンパク質を検索すると、システムはそのタンパク質を中心としてタンパク質相互作用ネットワークを作成する。図.6 にタンパク質相互作用ネットワークの例を示す。

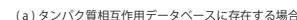


図 6 タンパク質ネットワーク表示例

★1 <http://www.nii.ac.jp/>
★2 http://ci.nii.ac.jp/info/ja/if_opensearch.html

がタンパク質相互作用データベースに存在しない状況におけるタンパク質相互作用ネットワークである。上述したように、タンパク質「kinase」がタンパク質相互作用データベースに存在しないので、CINII API を介し、Web 上から情報を取得し、タンパク質相互作用ネットワークを構築している。図.6 を見ることで、タンパク質間の相互作用の関係を一目で確認することができる。

4.6 雑音除去

本研究では、PubMed/MEDLINE データベース²⁾の論文をテキストマイニングすることで関係を抽出している。関係を抽出する上では、タンパク質名データベースを利用し、タンパク質名を同定している。そのため、少なからず単語をタンパク質名と誤って認識（以下、雑音という）している部分がある。そのため、本システムでは、ユーザがシステムを利用する中で、雑音を除去できる仕組みを考えた。

そこで、本システムでは、図.7 のように、表示されたタンパク質相互作用ネットワークのノード、つまりタンパク質名をダブルクリックすることで、削除することとした。

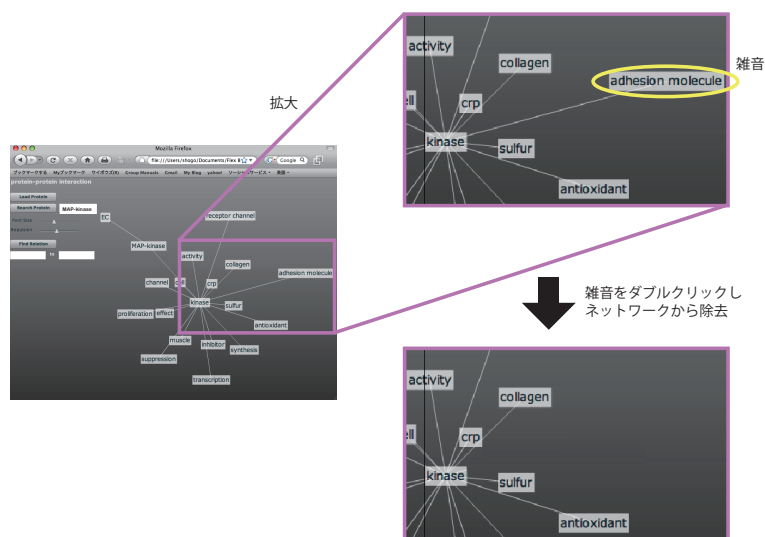


図 7 雑音の除去

図.7 に示したように、ユーザが雑音をダブルクリックすることで除去している。これに

より、ユーザがシステムを使えば使うほど、雑音の影響が少なくなっていく。

5. 結 論

本研究では、タンパク質とタンパク質の相互作用の関係を視覚的に確認できるシステムを提案した。本システムでは、生命科学分野の研究者があるタンパク質と相互作用の関係にあるタンパク質を調査する際、タンパク質名を入力することで、そのタンパク質と相互作用の関係にあるタンパク質がネットワーク構造で表示される。タンパク質相互作用データベースに存在しないタンパク質を検索すると、システムは Web から自動的にそのタンパク質に関する情報を取得し、ネットワークを作成した。今後の課題としては、タンパク質相互作用情報の妥当性の検証、および Web 上の情報からのタンパク質ネットワーク構築の妥当性の検証である。

参 考 文 献

- 1) 山本 泰智, 情報処理学会論文誌 Vol.50 No.9 Sep.2009, 生命科学分野におけるテキストマイニング
- 2) PubMed/MEDLINE
<http://www.ncbi.nlm.nih.gov/pubmed/>
- 3) Michael Krauthammer, Andrey Rzhetsky, Pavel Morozov and Carol Friedman, Vol. 259, Issues 1-2, 23 December 2000, Pages 245-252, Using BLAST for identifying gene and protein names in journal articles
- 4) L Tanabe and WJ Wilbur, Vol. 18 no. 8 2002, pages 1124-1132 Bioinformatics, 2002, Tagging gene and protein names in biomedical text
- 5) Joshua M. Temkin and Mark R. Gilder, Vol. 19 no. 16 2003, pages 2046-2053 DOI: 10.1093/bioinformatics/btg279, Extraction of protein interaction information from unstructured text using a context-free grammar
- 6) 福田賢一郎, 角田達彦, 田村あゆみ, 高木利久, 情報処理学会研究報告, NL-121 FI-47, p.103-110. 23, 医学生物学文献からの専門用語の抽出
- 7) Friedman, C., Krauthammer, M., Yu, H., Krauthammer, M. and Rzhetsky, A. (2001) GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. Bioinformatics, 17 (Suppl. 1), S74-S82.
- 8) Toshihide Ono, Haretsugu Hishigaki, Akira Tanigami and Toshihisa Takagi, Bioinformatics Vol. 17 no. 2 2001 Pages 155-161, Automated extraction of information on protein-protein interactions from the biological literature