修士論文

Webの情報を利用した
タンパク質相互作用の抽出

同志社大学大学院 工学研究科 情報工学専攻
博士前期課程 2008年度 753番

澁谷　翔吾

指導教授　三木 光範教授

2010年1月23日

Abstract

It is important for researchers to understand a protein-protein interaction when they try to figure out protein. Although researchers understand protein-protein interaction through reading many papers, it is difficult to figure out them. In recent years, text mining which extracts the information of protein-protein interaction is performed as the approach of information engineering.

In this paper, we propose a system that makes a protein-protein interaction network. Since a lot of information about protein-protein interaction is accumulated, we need to be visualized, in order to catch the tendency and the feature of the whole data. Then, the system that it displays the protein-protein interaction visually is useful in investigating in detail about protein for researchers. We extracted information of potein-protein interaction from papers in PubMed/MEDLINE database by text mining and built potein-protein interaction database. In a proposal system, it builds a protein-protein interaction network from information of protein-protein interaction. In this system, when researchers search the protein that system does not have in a protein-protein interaction database, a system acquires information from Web automatically and makes protein-protein interaction network from information. Next, system adds this network to the protein-protein interaction network that exist in the database and show the protein-protein interaction network which combines two network to researchers.

目 次

1　序論

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テキストマイニングにより抽出する試みがなされている[1)]．これにより，文献に書かれているタンパク質の相互作用を体系化することが可能である．しかし，それらのテキストマイニングは一定量の文献に対してであり，時間の経過に伴った新たな発見には対応できない．そういった状況においてもタンパク質の相互作用が得られることが望ましい．

本研究では，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．多量のタンパク質の相互作用情報が蓄積されている中，それらの情報の理解をしつつ，注目すべき点を見つけるため，また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．これまでのタンパク質の相互作用を調査するシステムでは，システムが保持しているタンパク質相互作用の情報のみでタンパク質相互作用ネットワークを構築していた．つまり，これはシステム側では大きなタンパク質相互作用ネットワークを保持しており，その一部を研究者の調査に応じて表示するものであった．よって，タンパク質相互作用ネットワークから外れるタンパク質の調査には対応できなかった．本システムでは，タンパク質相互作用ネットワークにないタンパク質が調査された場合，そのタンパク質に関する情報を Web から取得し，既存のタンパク質相互作用ネットワークに加えることで，ユーザの調査したタンパク質に関するタンパク質相互作用ネットワークを表示する．このようにタンパク質相互作用ネットワークにないタンパク質が検索されることで，タンパク質相互作用ネットワークは拡大されていく．

2 章では本研究のテキストマイニングの対象であるタンパク質について，3 章では関連研究について，4 章ではタンパク質相互作用情報の抽出について，5 章では提案システムについて，そして，最後に 6 章で結論を述べる．

2　タンパク質相互作用概説

本章では，本研究におけるテキストマイニングの対象であるタンパク質，およびタンパク質相互作用について解説する．

2.1　タンパク質

タンパク質とは，アミノ酸が多数連結してできた高分子化合物である．このタンパク質は生物固有の物質であり，その合成は生きた細胞の中で行われる．合成されたものは生物の構造そのものとなり，あるいは酵素などとして生命現象の発現に利用される．類似のタンパク質であっても，生物の種が異なれば一次構造，つまりアミノ酸の連結順序は異なるとされている．タンパク質はアミノ酸が多数結合してはいあるが，人工的な高分子のように単純な繰り返しではなく，順番が正確に決まっている．アミノ酸の種と順番は DNA に暗号で記述されている．

タンパク質の生体における機能としては，代謝などの化学反応を起こさせる触媒である酵素，生体構造の形成，生体内の情報のやりとり（シグナル伝達），運動への関与，抗原に対し特異的に結合することで免疫に重要な役割を果たす抗体，栄養の貯蔵，輸送，および蛍光など多種多様である．

これらのタンパク質が機能を果たす上で最も重要な過程は特異的な会合（結合）である．酵素および抗体はその基質，および抗原を特異的に結合することにより機能を発揮する．また構造形成，運動や情報のやりとりもタンパク質分子同士の特異的会合なしにはその機能は果たされない．よって，タンパク質を解明する上で，このタンパク質分子同士の特異的会合を解明することは必要不可欠である．本研究では，このタンパク質分子同士の特異的会合をタンパク質相互作用とよぶ．

2.2　タンパク質相互作用

タンパク質相互作用とは，複数の異なるタンパク質分子が状態に応じて特異的複合体を形成する現象である．全ての生命の基本現象は，生命を構成する分子と分子の特異的な相互作用によって行われている．最も顕著な例が，酵素である．ひとつひとつの酵素は特異的に気質を認識して別の物質に変換する．さらに，この物質を他の酵素が代謝するというように，特異的な認識に基づいて生体内の反応は行われている[3)]．

このような特異的な認識はシグナル伝達においても重要である．シグナルの受け渡しは全て特異的な分子間の認識によって行われている．そのため，あるシグナルタンパク質に結合してくる分子（タンパク質，核酸，低分子物質）を明らかにし，分子間の相互作用を調べることは，シグナルのルートや生理作用の調節メカニズムを明らかにする上で重要である[3)]．

2.3　タンパク質相互作用データベース

タンパク質相互作用データベースとは，タンパク質分子間の相互作用の情報を取り扱ったデータベースである．タンパク質相互作用データベースは研究者が実験よって得られた情報を集めたデータベース，それら研究者が発表した学術論文からタンパク質相互作用の情報を抽出したデータベース

に大別できる．前者としては，DIP(Database of Interacting)[1]，MIPS(muche information center for protein sequences)[2]，また後者としては，BIND(Biomolecular Interaction Network Databank)[3]などがある．研究者は，自身の専門とするタンパク質を研究する上で，上記などのデータベースを利用し，そのタンパク質相互作用を把握しながら研究を進めている．

3 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高分子化合物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高分子化合物を同定することは困難である．このような課題に対して，高分子化合物を同定するために，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[4)5)6)]．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[7)]．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，タンパク質の相互作用を抽出する研究が行われている[6)8)9)]．

4 タンパク質相互作用情報の抽出

4.1 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を抽出するために，着目すべきはタンパク質名とタンパク質相互作用を示す関係キーワードである．タンパク質相互作用情報の抽出の流れを Fig.4.1 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．

[1]http://dip.doe-mbi.ucla.edu/
[2]http://www.helmholtz-muenchen.de/
[3]http://bond.unleashedinformatics.com/
[1]http://www.ebi.ac.uk/

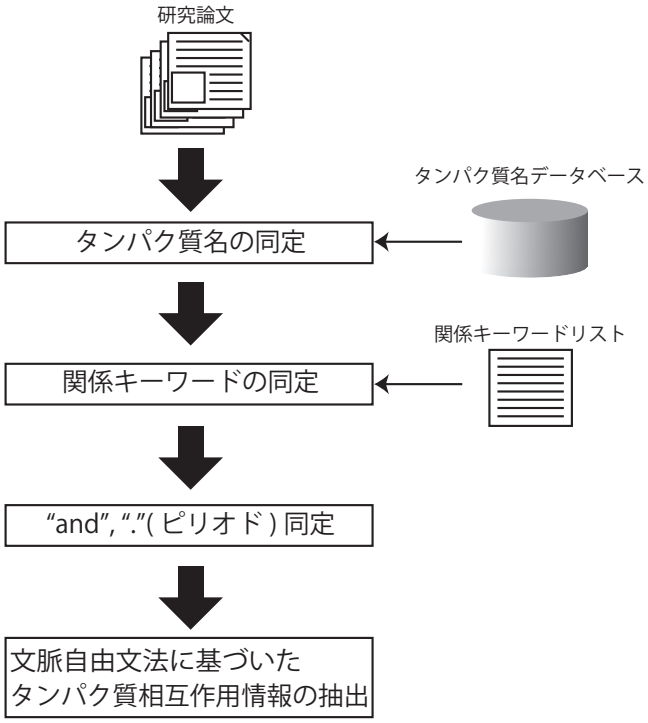

Fig. 4.1　タンパク質相互作用情報の抽出の流れ

4.2　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，1 つのタンパク質が複数の表記法を持つこともあり，約 25 万語のタンパク質を保持している．本データベースのデータを Table.4.1 に示す．

Table 4.1　タンパク質名データベース

ID	Protein Name
1	Alpha-amylase
2	Adenosine kinase
…	…
246038	YIL036W

4.3　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始めのステップとして，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 Table.4.2 に示す.

Table 4.2　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（Table.4.3）	activates
AND	並列	and
EOS	一文の終わり	.

4.3.1　タンパク質名

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”のタグを付与する．上述したように，タンパク質などの高分子化合物の多くは表記に多様性がある．この分野の専門用語には”正式な名称”というものが定着しておらず，既知のタンパク質について言及するときでさえ，著者間あるいは文献間で表記の対応がとれていない．タンパク質名の多様な表記の例を示す.

- 単語表記
  タンパク質名を表す単語は大文字，小文字，”-”，”/”などの記号文字，および数字からなる．しかし，小文字が大文字になったり，あるいは”-”，”/”などの記号文字の省略，追加または混合が起こる．タンパク質名，”c-Jun”，”c-jun”，”c jun”などがその例である.

- 複合語
  タンパク質名は複合語を形成しやすいといった特徴がある．例えば，”interleukin 1-responsive K protein kinase”がその例である．このタンパク質においても，”-”，”/”などの記号文字が混じったり，語順が入れ替わったりする.

- 省略形
  新規の用語はもともと，複合語の頭文字などをとった略語であるものもある．しかし，実際に文献などでは，分かりやすくするために，著者が意図して略語の一部または全部を復元して用いる場合もある．例えば，”epidermal growth factor receptor”，”EGF receptor”および”EGFR”は全て同じタンパク質である.

このようなタンパク質の表記の特徴から形態素とタンパク質名データベースとの完全一致による同定は困難である．そこで，タンパク質の同定においては，形態素とタンパク質名データベースとの完全一致ではなく，部分一致による同定を行う．また，上述したように，タンパク質名は複合語を形成しやすい特徴がある．タンパク質名同定においても複合語を考慮している．例を Fig.4.2 に示す.

Fig.4.2 のように，”MOL”が連続する場合，1 番目の単語と 2 番目の単語を結合し，その単語がタンパク質名データベースに存在するか否かを判断する．この時のタンパク質名データベースとの照合

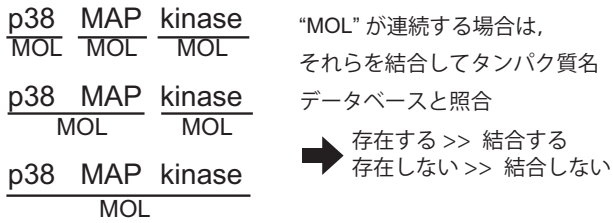

Fig. 4.2　複合語の形成

についても部分一致によるものである．存在する場合は結合し，存在しない場合は結合しない．これを繰り返すことにより複合語を形成する．以下の例では，”p38 MAPK”および”RGS2”をタンパク質名として同定している．

Input：The p38 MAPK inhibited RGS2.
Identification of Protein：p38 MAPK, RGS2

4.3.2　関係キーワード

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．キーワードは，bind(結合する), activate(活性化する) といったタンパク質の相互作用を示すものである．タンパク質相互作用を表すキーワードを Table.4.3 に示す.

以下の例では，”activates”を関係キーワードとして同定している．

Input：The ischemia activates Kupffer.
Identification of Relation Keyword：activates（活性化する）

4.3.3　その他

次節で文脈自由文法に基づいて，タンパク質相互作用情報を抽出する．そこで，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．

4.4　文脈自由文法に基づいたタンパク質相互作用情報の抽出

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を用いて，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 Table.4.4 に示し，抽出の流れを Fig.4.3 に示す.

入力文を Table.4.4 に示すルールに適用させることで，入力文を機械的に解析することができる．Fig.4.3 の中の”Interaction”に着目し,「前方のタンパク質”Molecule”が後方のタンパク質”Molecule”と関係キーワード”Keyword”で相互作用している」ということを示しており，その情報を抽出している．

Table 4.3　関係キーワード一覧[8)]

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

Table 4.4　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**

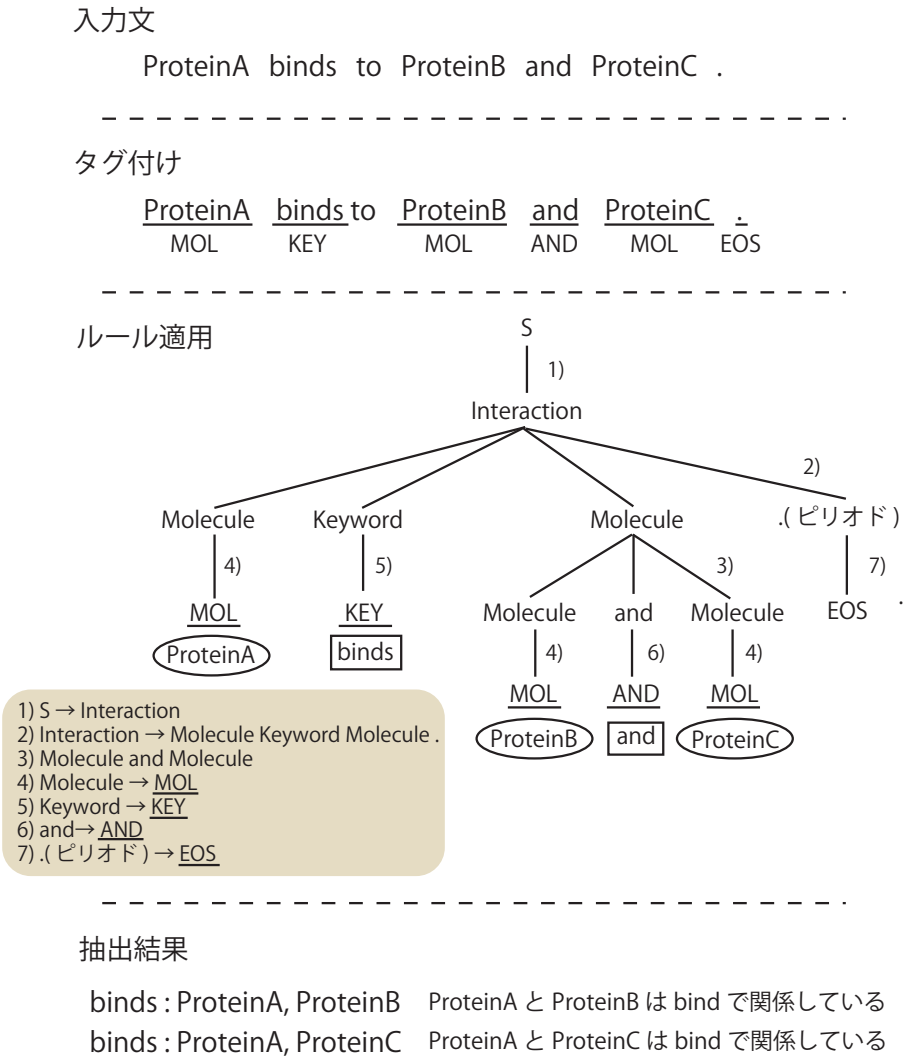

Fig. 4.3　タンパク質相互作用情報の抽出

4.5　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 データベース[2] において，検索語 x で検索した結果，検索の上位 n 件の論文序論を対象としてテキストマイニングを行った．検索語と論文の件数を Table.4.5 に示す．

合計 465 件の論文に対して，テキストマイニングを行った結果，約 4500 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．

Table 4.5　テキストマイニング対象

検索語 x	n 件
CXCR4	100
C10orf10	2
CYP51A1	22
LDLR	100
INSIG1	84
IDI1	11
SQLE	19
HMGCS1	17
HMGCR	100
SC4MOL	6

(a)
Input : The PNA zipper-GCN4 binds both the TRE and CRE DNA sites.
Output : (binds : PNA, TRE), (binds : PNA, CRE DNA)
(b)
Input : The ischemia activates Kupffer.
Output : (activates : ischemia, Kupffer)
(c)
Input : The Drosophila Hox gene produces isoforms.
Output : (produces : Drosophila, isoforms), (produces : Hox gene, isoforms)

(a) の例では，タンパク質名”PNA”，”TRE”および”CRE DNA”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．続いて，(b) の例ではタンパク質名”ischemia”および”Kupffer”を同定し，関係キーワード”activate”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．最後に，(c) の例では，タンパク質名として，”Drosophila”，”Hox gene”および”isoforms”を同定し，関係キーワード”produces”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．しかし，調べてみると単語”Drosophila”はタンパク質を示す単語ではなく，その意味は”《昆虫》ショウジョウバエ属”というものであった．つまり，タンパク質ではなく，タンパク質を修飾する単語（以下，タンパク質修飾語）であり，”Drosophila”と”isoforms”の相互作用というのは誤りである．抽出結果を見てみると，このようにタンパク質修飾語がタンパク質と同定され，タンパク質修飾語とタンパク質，もしくはタンパク質修飾語とタンパク質修飾語を相互作用として抽出しているケースが多く見られた．これは形態素がタンパク質か否かの判断する際，部分一致によるタンパク質の同定を行っているから生じている．

そこで，タンパク質修飾語は部分一致によってタンパク質の候補となるが，タンパク質修飾語がタンパク質を修飾することなく，その語のみで相互作用となる場合には，抽出しないようにした．つま

り，上の例では，以下のように，タンパク質修飾語”Drosophila”がタンパク質を修飾することなく，単独で相互作用となっているので抽出しない．

Input : The Drosophila Hox gene produces isoforms. Output : (produces : Hox gene, isoforms)

補足であるが，”Drosophila Hox gene”のように複合語を形成しない理由は，タンパク質名データベースに”Drosophila PROTEIN”(PROTEIN にはタンパク質名が入る) というタンパク質が存在するが，”Drosophila Hox gene”というタンパク質が存在しないからである．

5　提案システム

本章では，タンパク質の相互作用のネットワークを表示するシステムを提案する．

5.1　概要

本システムでは，タンパク質の相互作用をネットワーク構造で可視化する．研究者があるタンパク質に着目するということを想定し，タンパク質の検索機能を有する．このタンパク質検索では，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表示する．しかしながら，必ずしもタンパク質相互作用情報が得られているとは限らない．また，タンパク質，タンパク質相互作用は新たに発見されうるもので，常にテキストマイニングの結果を得ている状態を保つことは難しい．そのような場合，つまりテキストマイニングにより，タンパク質相互作用が得られていない場合，システムは自動的にタンパク質に関する情報を Web から取得し，その情報からタンパク質相互作用ネットワークを作成，Fig. 5.1 のように既存のタンパク質相互作用ネットワークに追加し，結合したタンパク質相互作用ネットワークから結果を表示する．

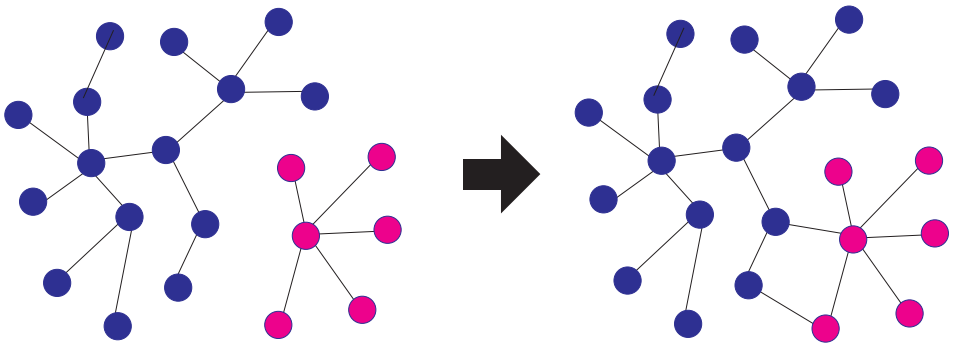

Fig. 5.1　ネットワークに追加

5.2　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，4.5 節で解説したように，合計 465 件の論文序論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

タベースのデータを Table.5.1 に示す．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．

Table 5.1　タンパク質相互作用データベース

ID	Protein Name	Protein Name	Interaction
1	PNA	TRE	binds
2	PNA	CRE DNA	binds
3	CENP-V	SH3	associates
…	…	…	…

Table.5.1 の例では，タンパク質「PNA」とタンパク質「TRE」が関係キーワード「bind」つまり，結合の関係を示している．

5.3　タンパク質相互作用ネットワーク

タンパク質相互作用ネットワークはタンパク質相互作用情報を基に構築する．タンパク質をノード，相互作用をエッジで表現している．Fig. 5.2 にタンパク質相互作用ネットワークを示す．

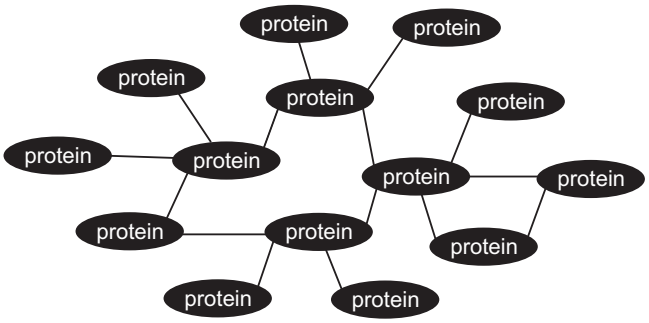

Fig. 5.2　タンパク質相互作用ネットワーク

このタンパク質相互作用ネットワークを見ることで，研究者は一目でタンパク質の相互作用を確認できる．

5.4　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 Fig.5.3 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 つのスライダーバ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

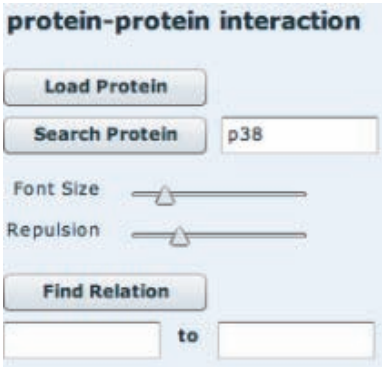

Fig. 5.3　ユーザインタフェース

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.

5.5　タンパク質相互作用ネットワーク表示

5.5.1　タンパク質検索

本項では，タンパク質検索について解説する．生命科学分野の研究者はあるタンパク質に着目した際，そのタンパク質相互作用を把握する必要がある．そこで，本システムでは，ある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見ることができる．このことを本システムでは，タンパク質検索という．タンパク質検索の流れを Fig.5.4 に示す.

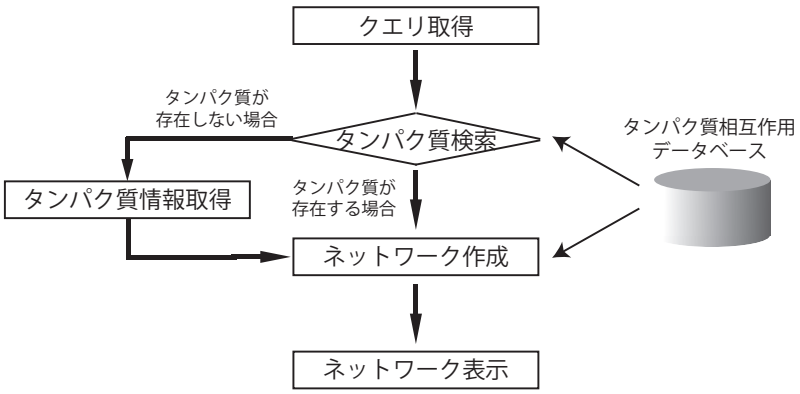

Fig. 5.4　タンパク質検索の流れ

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．つまり，これはテキストマイニングによって対象とするタンパク質のタンパク質相互作用が得られているかどうかを調べる．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，タンパク質相互作用ネットワーク作成，表示する．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存在しない場合，システムは自動的に Web から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，表示する．以下では，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存在

する場合としない場合に分けて解説する．

- データベースに存在する場合
  データベースに存在する場合とは，つまり，テキストマイニングによってタンパク質相互作用情報が得られている場合である．この場合，検索タンパク質と相互作用の関係があるタンパク質をタンパク質相互作用データベースから取得し，それらをエッジで繋ぐ．これらのタンパク質を検索タンパク質から距離 1 のネットワークとする．続いて，タンパク質相互作用データベースから取得したタンパク質に対して，同様に相互作用の関係にあるタンパク質を取得し，エッジで繋ぐ．このネットワークは検索タンパク質からの距離が 2 のタンパク質相互作用ネットワークである．これを距離が N になるまで繰り返すことで，検索タンパク質からの距離が N のタンパク質相互作用ネットワークを作成する．

- データベースに存在しない場合
  タンパク質相互作用情報は必ずしもタンパク質相互作用データベースに存在するとは限らない．タンパク質，およびタンパク質相互作用は新たに発見させる可能性もあり，それらを全てカバーすることは困難である．そこで，本システムでは，タンパク質相互作用データベースに検索タンパク質が存在しない場合，Web の情報を利用してタンパク質相互作用ネットワークを作成する．このとき，Web から取得した情報からタンパク質相互作用ネットワークを作成し，既存のタンパク質相互作用ネットワークと併せて表示する．今回，Web の情報を取得するにあたり，国立情報学研究所[1]が運営する学術文献のデータベースから同研究所の OpenSearch[2]といわれる API(以下，CiNii API) を利用し，検索タンパク質に関する文献の序論を取得している．情報取得の流れを Fig.5.5 に示す．

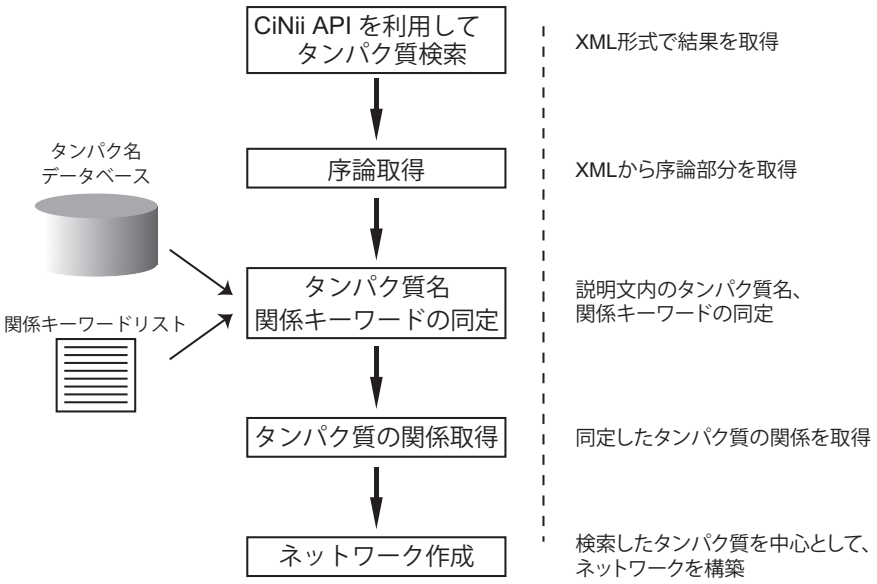

Fig. 5.5　情報取得の流れ

[1]http://www.nii.ac.jp/
[2]http://ci.nii.ac.jp/info/ja/if_opensearch.html

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．これをタンパク質相互作用データベースに存在する場合と同様，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．ここで，Web から取得した情報は次からの検索にも生かされるように情報を蓄積している．つまり，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ことで，Web から情報を取得し，タンパク質相互作用ネットワークを作成し，既存のタンパク質相互作用ネットワークに順次追加される．

5.5.2 相互作用パス検索

あるタンパク質とあるタンパク質に着目したとき，それらのタンパク質がどのようなタンパク質との相互作用を介して，関係しているかを検索することが可能である．例えば，ProteinA と ProteinB の関係を調べたいとする．そういった場合，本システムを利用することで，Fig.5.6 のように,「ProteinA >> ProteinX >> ProteinY >> ProteinZ >> ProteinB」(”>>”は相互作用「interaction」を示す) という形で相互作用のパスを得ることが可能である．

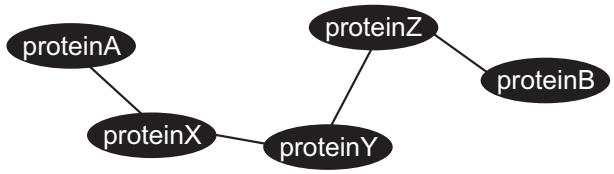

Fig. 5.6 相互作用パス

5.6 誤タンパク質除去

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，タンパク質ではない単語をタンパク質名と誤って認識（以下，誤タンパク質という）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，誤タンパク質を除去できる仕組みを考えた．

そこで，本システムでは，Fig.5.7 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．

Fig.5.7 に示したように，ユーザが誤タンパク質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，誤タンパク質の影響が少なくなっていく．

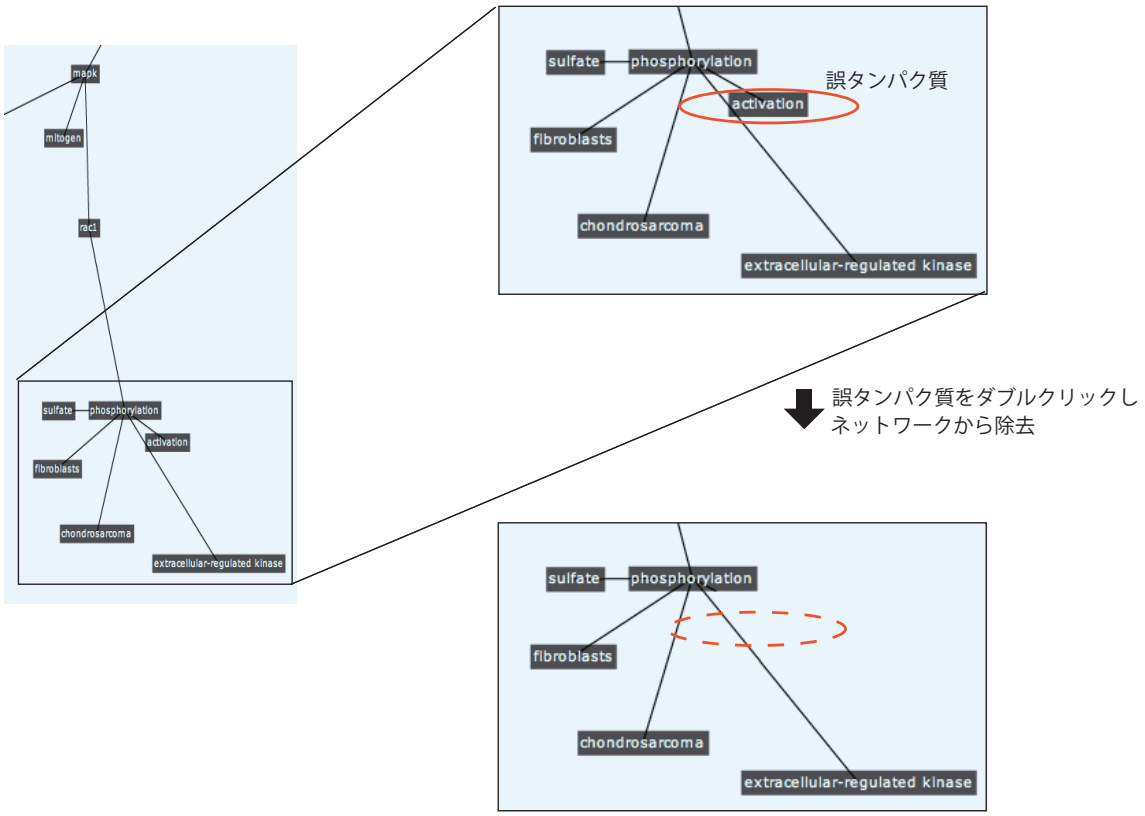

Fig. 5.7　誤タンパク質の除去

5.7　表示例

本節では，本システムを利用することで得られたタンパク質相互作用ネットワークの例を示す．

- 検索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場合のタンパク質相互作用ネットワーク

  ユーザがあるタンパク質を検索した際，そのタンパク質の相互作用が得られている場合，その情報を利用してタンパク質相互作用ネットワークを作成する．ここでは，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「SREBP」を検索した結果を Fig.5.8 に示す．

  Fig.5.8 から分かるように，タンパク質「SREBP」と相互作用のあるタンパク質がネットワーク構造で表示された．ここで検索したタンパク質「SREBP」は，転写因子といい，DNA に特異的に結合するタンパク質群の一つである．よって，Fig.5.8 のように多くのタンパク質（遺伝子）と相互作用の関係にある．

- 検索タンパク質がタンパク質相互作用データベースに存在しない場合のタンパク質相互作用ネットワーク

  ユーザがあるタンパク質を検索した際，そのタンパク質の相互作用が得られていない場合，本システムでは，Web から検索タンパク質に関する情報を取得し，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データベースに存在しないタンパク質「cysteine」を検索した結果を Fig.5.9 に示す．

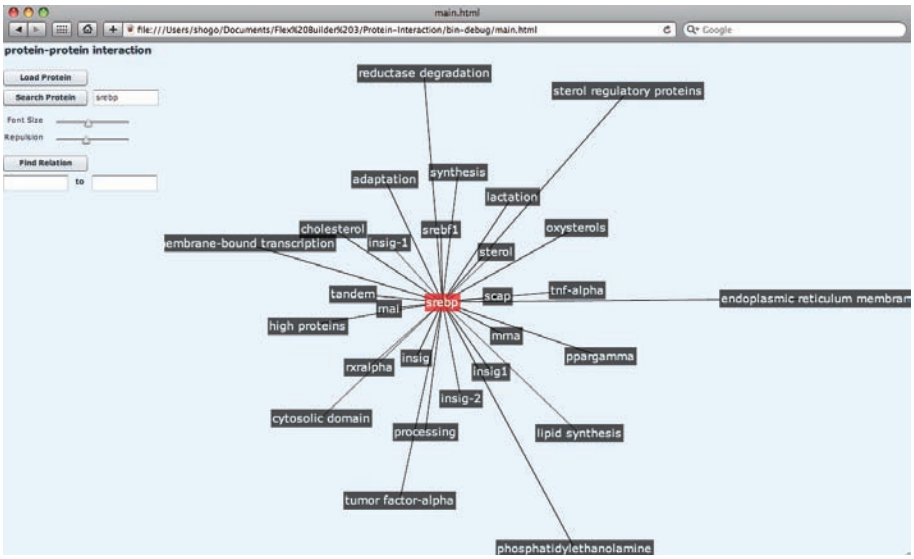

Fig. 5.8　タンパク質相互作用データベースからのタンパク質相互作用ネットワーク

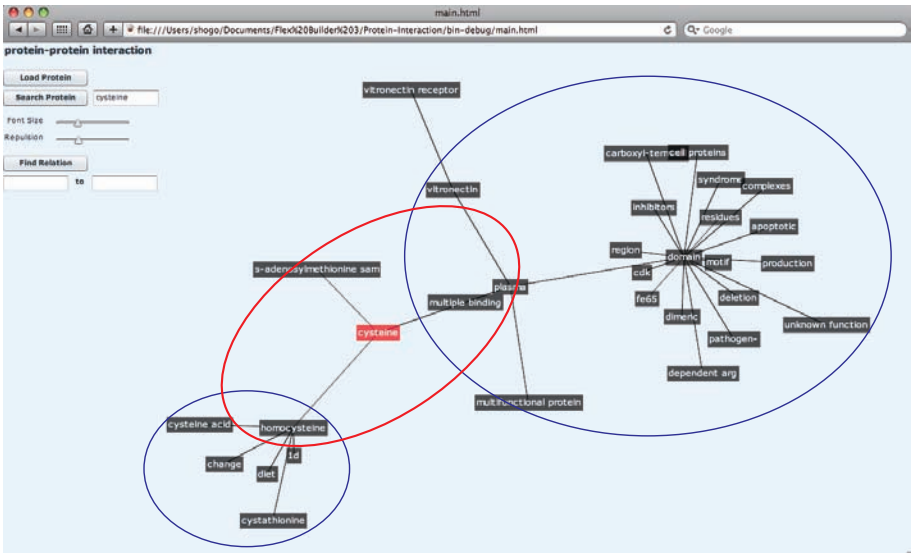

Fig. 5.9　Web の情報を利用したタンパク質相互作用ネットワーク

タンパク質「cysteine」が検索され，システムはタンパク質相互作用データベースにタンパク質「cysteine」の相互作用が存在するか調べる．しかし，タンパク質「cysteine」の相互作用はタンパク質相互作用データベースに存在しないので，Web からタンパク質「cysteine」に関する情報を取得し，その情報を基にタンパク質相互作用ネットワークを作成した．Fig.5.9 における，赤で囲ったネットワークが Web の情報から得たタンパク質相互作用ネットワーク，そして青で囲ったネットワークはシステムが既に保持しているタンパク質相互作用ネットワークであった．Fig.5.9 を見れば分かるように，Web の情報から得られたタンパク質相互作用ネットワークがシステムが保持する既存のタンパク質相互作用ネットワークに追加されている．今回の結果では，タンパク質「homocysteine」およびタンパク質「plasma」が既存のタンパク質相互作用ネット

ワークと Web から取得した情報からの共通ノードであったために，既存のタンパク質相互作用ネットワークへ追加された．

- 相互作用パス
  相互作用パスとは，ある 2 つのタンパク質に着目した際，それらのタンパク質が他のどのようなタンパク質との相互作用を介して，関係しているか表すものである．相互作用パスの例を Fig.5.10 に示す．

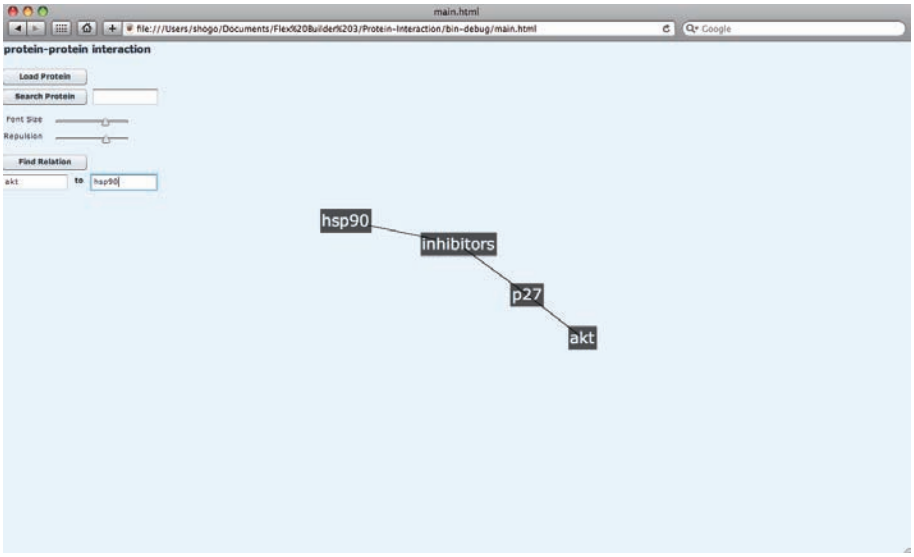

Fig. 5.10　相互作用パス

Fig.5.10 の例では，タンパク質「akt」とタンパク質「hsp90」の相互作用パスを示している．タンパク質「akt」とタンパク質「p27」，タンパク質「p27」とタンパク質「inhibitors」，およびタンパク質「inhibitors」とタンパク質「hsp90」の相互作用からタンパク質「akt」とタンパク質「hsp90」の相互作用パスは「ark >> p27 >> inhibitors >> hsp90」（”>>”は相互作用「interaction」を示す）という結果が得られた．

6　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，タンパク質名を検索することで，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．この際，タンパク質相互作用ネットワークの構築は，タンパク質相互作用から作成した．しかし，必ずしも全てのタンパク質に関する相互作用を保持することは困難であった．そういった場合，つまり，検索したタンパク質に関する相互作用を保持していない場合，システムは Web から検索タンパク質に関する情報を自動的に取得し，タンパク質相互作用ネットワークを作成した．そして，そのタンパク質相互作用ネットワークを既存のタンパク質相互作用ネットワークに追加し，そのタンパク質相互作用ネットワークの中からユーザの求めるものを表示した．この Web

の情報からのタンパク質相互作用ネットワークを既存のタンパク質相互作用ネットワークに追加していくことで，タンパク質相互作用ネットワークは拡大された．

謝辞

本研究を遂行するにあたり，多大なる御指導そして御協力を頂きました，同志社大学生命医科学部の廣安知之教授に心より感謝いたします．日ごろのご指導の他，国内学会，Super Computing 2008 (SC08) など貴重な機会を多く与えていただきました．また，様々な指摘，助言をして下さいました，同志社大学理工学部の三木光範教授，吉見真聡先生に心より感謝いたします．三木先生，吉見先生とは研究グループこそ違いましたが，研究や生活などにおいて多くの知的刺激を受けました．

本論文の校正にあたっては，協力していただいた，研究室の芝野功一郎君，戸松祐太君に感謝します．また，日ごろから研究に関するご意見，ご協力をいただきました宮地正大君に心より厚く御礼申し上げます．最後に，私の大学院への進学を快く了承してくださった母，私を支えてくださったすべての人に感謝したいと思います．

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) 竹縄 忠臣, 渡邉 俊樹, タンパク質の分子間相互作用実験法, 羊土社, 1996

4) 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

5) L Tanabe and WJ Wilbur , Vol. 18 no. 8 2002, pages 1124?1132 Bioinformatics, 2002, Tagging gene and protein names in biomedical text

6) 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

7) 福田賢一郎, 角田達彦, 田村あゆち, 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

8) Friedman,C., Kra,P. 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.

9) 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

付 図

付 表

付録：発表論文リスト

1. **澁谷 翔吾**，廣安 知之，三木 光範，横内 久猛：
   対話的なキーワード抽出によるブログ推薦システム，
   第 72 回数理モデル化と問題解決研究会（MPS 研究会），2008.12.

2. **澁谷 翔吾**，廣安 知之，三木 光範，横内 久猛，吉見 真聡：
   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，
   第 76 回数理モデル化と問題解決研究会（MPS 研究会），2009.12.