修　士　論　文

オントロジーを用いた生物医学文献からの知識抽出手法に関する研究

指導教官　　佐藤賢二　助教授

北陸先端科学技術大学院大学
知識科学研究科知識システム基礎学専攻

150024　亀谷 聡

審査委員：　佐藤　賢二　助教授（主査）
小長谷　明彦　教授
本多　卓也　教授
Ho Tu Bao　教授

2003 年 2 月

目　次

図　目　次

表　目　次

第　1　章

はじめに

1.1　研究の背景と目的

ゲノムプロジェクトにより，各種モデル生物の全 DNA 配列が決定されつつある．現在では，遺伝子およびその生産物であるタンパク質の機能解明とそれらがいつ発現し，他の物質とどのように関係しているかを解明することが求められている．

生物医学に関する文献には，遺伝子やタンパク質の機能や相互作用に関する情報など，ゲノム解析の研究を進めるための重要な情報が記述されており，その数は膨大なものとなっている．そのため，このような有用な情報を抽出するための手法の確立が必要とされている．こうした背景を受け，生物医学に関する専門用語を整理したオントロジーが盛んに構築される一方で，文献データベースからタンパク質間相互作用などの情報を抽出する研究も行われている．しかし，前者に関しては専門家が人手で構築していることもあり，最大でも数万程度の用語しか網羅されていないという問題があり，後者に関しては相互作用を記述する際によく用いられるいくつかの限られた動詞に着目したテンプレートマッチングが主流であるため，抽出できる情報の量が少ないという問題がある．そこで，本研究では，研究室で既に構築されている大量の専門用語リストを利用して，既存のオントロジーを拡張することを試みる．そして，従来は人手で行っていた動詞やテンプレートの発見を，大規模なオントロジーを用いて機械的に抽出することを目指す．これにより，単純なタンパク質間相互作用に限らず，多様な情報を抽出するための複雑なテンプレートを多数発見することが期待できる．

1.2　本論文の構成

本論の構成は以下のようになっている.

第 1 章において，本研究の背景と目的について述べる.

第 2 章において，生物医学における情報抽出の現状に関して述べる.

第 3 章において，オントロジーの拡張を試みた結果と考察について述べる.

第 4 章において，生物医学文献から動詞やテンプレートの抽出を行った結果と考察を行う.

第 5 章において，抽出したテンプレートの評価を行う.

第 6 章において，本研究のまとめについて述べる.

第 7 章において，今後の展望について述べる.

第　2　章

生物医学における情報抽出の現状

生物医学分野における文献には，タンパク質や遺伝子などの物質名に関する情報やそれらの機能や相互作用に関する情報などが自然言語の形で記述されている．ゲノム解析研究を進めるための重要な情報源として，このような自然言語情報をいかに活用していくかという研究が近年活発におこなわれている．ここでは，専門用語とそれらの関係の情報を抽出する研究の現状に関して述べる．なお，本論文では以下のような定義を使用する．

[単語]：空白を含まない，1 文字以上の連続した文字列
[用語]：1 つ以上の単語を空白で接続したもの
[専門用語]：生物学や医学などの分野で特によく使われる用語

2.1　専門用語の抽出

物質に関する機能や相互作用情報を文献から抽出するためには，まず専門用語を認識しなければならない．そこで，タンパク質名などの生物医学における専門用語を抽出する手法を紹介する．

2.1.1　NLP（自然言語処理）による手法

関連する研究として，固有名詞の特徴を利用した PROPER(PROtein Proper-noun phrase Extracting Rules)という手法を用いた研究がある[1]．この研究では，kinase, receptor, ligand, enzyme, compound などの単語を含んでいるタンパク質名や，より

狭い領域であるタンパク質ドメイン*[1]やモチーフ*[2]などの名前の抽出を行っている．この研究の特徴は，タンパク質名に見られる特徴を利用した抽出を行っている点である．以下にその研究の概要を述べる．まずタンパク質名の特徴として，例えば下線部のように大文字や数字，特殊記号を含む単語が頻繁に見られる．

Src homology (SH) 2 domains

p54 SAP kinase

このような単語を，“core-term”と呼ぶ．また，以下の下線部のように専門用語の機能や特徴を表す単語は重要な単語とみなすことができる．

EGF receptor

Ras GTPase-activating protein (GAP)

このような単語を，“f-term”(feature-term)と呼ぶ．これら２つの特色を利用して，タンパク質名の抽出を行う．

タンパク質名の抽出は，core-term の抽出，core-term と f-term の連結という 2 つの段階からなる．core-term の抽出では，まず大文字や数字，特殊記号を持つ単語を core-term の候補として抽出する．それらの候補から core-term にふさわしくないものを以下の処理で取り除く．

- ９文字以上の単語や“-”と小文字からなる単語(“full-length”や“dual-specificity”など)は除外する．
- 特殊記号が単語の半分以上を占める単語を取り除く．これにより，“+/-”のような単語を除去する．
- 単位のような，数値に関係する単語（aa，AA，fold，bp，nM，microM，%，UV の単位が語尾についた単語）は除去する．
- あらかじめ用意したテンプレートとマッチする単語は除去する．これにより参考文献の人名やジャーナル名などを除去することができる．

---

*[1] 構造の面から見ると，分子量の大きなタンパク質，特に球状タンパク質にはその立体構造がいくつかの構造単位に分かれていることがあり，これらの構造をドメインという．脱水素酵素では補酵素を結合するドメインと触媒作用に関係したドメインとに分かれる．

*[2] ドメインよりも小さな部分構造であり，機能的な最小単位となる部分構造をさす．

こうして抽出した core-term に関して，テキストの文中で注釈を行い core-block の構築を行う．注釈とは，core-term を隣接した単語へ拡張することや他の注釈を用いて連結することである．この時，core-term と f-term の連結には以下のようなルールを用いる．矢印の右辺の下線部分は，左辺の下線部分の単語を連結した結果である．

・core-term と f-term が隣接している時は連結する．

Src SH3 domain　→　Src SH3 domain

・括弧に挟まれる場合は以下のように処理する．

ⅰ．(SH3)　→　(SH3)

ⅱ．(SH2 (and|or) SH3)　→　(SH2 (and|or) SH3)

・POS tagger*[3]の利用

ⅰ.注釈を与えられた部分が離れていて，その間に名詞や形容詞，数字しかないような場合は連結する．

Ras guanine nucleotide exchange factor Sos　→

Ras guanine nucleotide exchange factor Sos

ⅱ.冠詞，指示詞，所有格代名詞，前置詞が左側にある時は注釈を延ばす．

the focal adhesion kinase (FAK)　→　the focal adhesion kinase (FAK)

ⅲ.単体の大文字やギリシア文字の単語が右側にある場合は注釈を延ばす．

p85 alpha　→　p85 alpha

こうして構築された core-block を以下のルールでさらに連結する．A，B，C，D，E は，core-block を表す．

・“A, B, …C and D f-term”

Src, Fyn, Lyn, Yes, and PI3K SH3 domains

→　Src, Fyn, Lyn, Yes, and PI3K SH3 domains

---

*[3]　与えられた特定分野の文献集合のみの情報から専門用語を特定，抽出するシステム

・“A, B, …, C and D of E”

Src homology 2 (SH2) and 3 (SH3) domains of Vav

→ Src homology 2 (SH2) and 3 (SH3) domains of Vav

・“A of B, C and E”

SH2 domains of Abl, Lck, Fyn, and p85

→ SH2 domains of Abl, Lck, Fyn, and p85

・“A f-term core-term and core-term”

GTP-binding proteins Rac1 and Cdc42

→ GTP-binding proteins Rac1 and Cdc42

・“A of B”

p85 alpha subunit of PI 3-kinase

→ p85 alpha subunit of PI 3-kinase

・“A, B”

the Src-related tyrosine kinase, Hck

→ the Src-related tyrosine kinase, Hck

これらの他にも誤った注釈を除外するためのルールなどを用いてタンパク質名の抽出を行っている.

一般的にこのような NLP による手法の長所として，未知語の認識ができる可能性があることがあげられる．一方，短所としては，構文解析など複雑な処理が必要となり，処理が重くなることがあげられる.

2.1.2 辞書・オントロジーベースによる手法

Thomas らは，遺伝子や細胞，薬品名の認識に UMLS(Unified Medical Language System) Metathesaurus というバイオ医学に関する概念及び用語に関する辞書を利用している[2].

このような辞書・オントロジーベースによる手法の長所として，専門用語の認識が単純なマッチングですむために，NLP による手法に比べて処理が少なくてすむとい

うことがあげられる．一方，短所として，辞書やオントロジーが十分な専門用語を網羅していない場合や，最新版でなければ，必要な専門用語の認識が出来ないということがあげられる．

2.2　専門用語間の関係の抽出

Sekimizu らは, Medline のアブストラクトから約 100 万の単語を抽出し, Lingsoft が提供している EngCG というシステムを用いて簡単な構文解析を行っている．そして，activate，bind，interact，regulate，encode，signal，function といった，遺伝子や遺伝子産物の関係を表す際に頻繁に出現する動詞に対して，文献中での主語と述語を抽出することにより，遺伝子や遺伝子産物の間の相互作用情報の抽出を試みている[3].

Thomas らは，200 のアブストラクトから人手によって相互作用を表す共通した記述を抽出した. 約 30 種類の動詞(including, activate, inhibit, modulate, suppress, isolate，promote，characterise など)について調査し，タンパク質間の相互作用を表す動詞（およびそれに続く前置詞）として，interact (with)，associate (with)， bind (to)の 3 種類に限定した．そして，これらを用いたテンプレートにより，タンパク質間相互作用情報の抽出を試みている[4].

第　3　章

オントロジーの拡張

3.1　オントロジーとは

これまで生命科学分野では，生物種や研究対象ごとに専門性の高い研究が行われ，細分化された各分野に特有な概念や用語が多く生み出されてきた．このことが，分野の垣根を超えた知識の共有を行う際に障害となっている．例えば，遺伝子(gene)という生命科学分野における最も基本的な用語でさえ，２つの異なる意味で定義されている．GDB(Genome database)では，「遺伝子とは転写翻訳されてタンパク質となるDNA 上の領域(コード領域)」と定義している．しかし，GenBank と GSDB(Genome Sequence Database)では，「遺伝子とは遺伝的な特徴(表現型)の制御に関する DNA 上の領域」と定義している．後者は，前者の定義に加えてイントロン，プロモーター，エンハンサーといった，DNA から転写・翻訳を行う際に関与する非コード領域も含むことになる．生命科学分野では，こうした用語の多義性が多く見られる．こうした問題を解決するために生命科学におけるオントロジーが構築されるようになった．生命科学におけるオントロジーとは，生命科学における概念を抽出し，概念の属性と関係について，人間と計算機が理解できる形で記述したものである．そして，これまで蓄積された知識を洗い直し，生物種間に共通の性質を取り出して，その性質を形式的に記述することである[5].

近年，多くのゲノムデータベースがインターネット上に公開されるようになり，知識を共有するための手段として標準化された概念体系が必要であるとの認識が高まっている．現在では，生命科学に関するオントロジーの構築が活発に行われており，用途によって種々のものがある．以下に構築されているオントロジーをいくつか紹介する．

・TaO，MBO

TaO(Transparent Access to Multiple Biological Information Sources Ontology)と MBO(Molecular Biology Ontology)は, 統合データベースを目的として最初に開発されたオントロジーである．これらのオントロジーは，生命科学全体という広い領域を対象としているので, 必然的に抽象度の高い概念のみが記述されているという特徴がある．例えばルートは，“Being”から始まり，末端は“Disease”や“Reaction”といったレベルで終わっており，個々の分子の機能までには達していない．TaO や MBO はむしろ，生命科学の普遍的な概念について，それらの関係の種類と構造を詳細に考察し，生命科学の概念の性質を明らかにしている．

・Gene Ontology

実際に比較ゲノム生物学に利用できる大規模なオントロジーを作ろうとする研究は，Gene Ontology コンソーシアムによって始められた．Gene Ontology は遺伝子の生体機能に対象を限定したオントロジーである．プロジェクトが始まってからわずか 2 年あまりで 10 万個を超える概念を格納しており，その勢いは米国におけるゲノムプロジェクト推進の勢いをそのまま受け継いでいる. Gene Ontology は TaO や MBO と異なり，分子の詳細な機能を対象としている．

・Interaction Ontology

Interaction Ontology は，分子機能を分子間相互作用の側面から定義したオントロジーである．Gene Ontology が機能全般を対象としているのに対し，Interaction Ontology は複数分子の間の相互作用による機能に対象を限定している．生物の機能は分子，分子間相互作用，パスウェイやネットワーク，細胞間の相互作用，器官，臓器，生態，個体発生，系統発生と階層的に分類できるとする考えに基づいており，その二段階目である分子間相互作用を対象としたオントロジーを開発している．Interaction Ontology は特に概念(機能)の属性について詳細な考察を行っているが，概念の階層化は行っていない．

3.2 Gene Ontology

この章では，生命科学に関するオントロジーの中でも本研究に利用した Gene Ontology について述べる．

3.2.1 Gene Ontology の歴史

Gene Ontology コンソーシアムは，3 種のモデル生物(ショウジョウバエ，酵母，マウス)の共同プロジェクトとして，真核生物遺伝子の機能アノテーションに使用する用語とその意味，概念体系，アノテーションのルールなどを標準化する作業を 1998 年の夏にスタートした[6]．その後，2000 年にシロイヌナズナと線虫のチームが加わり，現在では 13 チームにまで増え，真核生物に限らずすべての生物に対して適用可能な語彙体系を構築することを目指している．Gene Ontology の有用性はバイオインフォマティクスやゲノムデータベース分野で認知され，2001 年には真核生物遺伝子のアノテーション語彙体系として，事実上の標準といえる地位を確立している．

3.2.1 Gene Ontology の構成

Gene Ontology は，３つのオントロジー(molecular function, biological process, cellular component で構成されている．それぞれのオントロジーは，以下のような特徴を持つ．

・molecular function：遺伝子産物の生化学レベルでの挙動に関する記述
・biological process：主要な生物学の現象についての記述
・cellular component：遺伝子産物の場所や細胞内と高分子複合体内の構造に関する記述

これらのオントロジーは独立している．つまり，2 つ以上のオントロジーにまたがって存在する用語はない[7]．2002 年 1 月 4 日の時点で，それぞれ 947 語，4819 語，4541 語，合計 10307 語を網羅している．Gene Ontology の各用語には識別のための

ID 番号がつけられており，1 つの用語は複数の親用語と複数の子用語を持つことができる．各親用語との関係は，“is-a”または“part-of”のどちらかであり，この関係は Resource Description Framework(RDF)[8]の文法によって表現されている(図 3.1)．一方，子用語との関係を記述する項目は設けられておらず，その用語を親用語として指し示している用語を見つけることで知ることができる．

図 3.2 は，go.xml のデータ構造の概略を示している．go.xml における各 term 要素は Gene Ontology の各用語に相当する. go.xml は, 浅く平らな木構造をしているが, RDF 構文を利用することにより概念上では複雑な包含関係を表現している．図 3.3 は AmiGO[9]のスクリーンショットで，階層化された Gene Ontology の 3 階層までを記述している．図 3.4 は，Gene Ontology の一部のデータ構造を表している GeneAround[10]のスクリーンショットで，single-stranded RNA binding という専門用語に着目した時のデータ構造を描写している．親用語として RNA binding という用語を持ち，子用語として poly(A) binding，poly(B) binding，poly(C) binding という用語を持つことを示している．そして，Gene Ontology の上位の階層から single-stranded RNA binding の子用語に至る経路を示している．

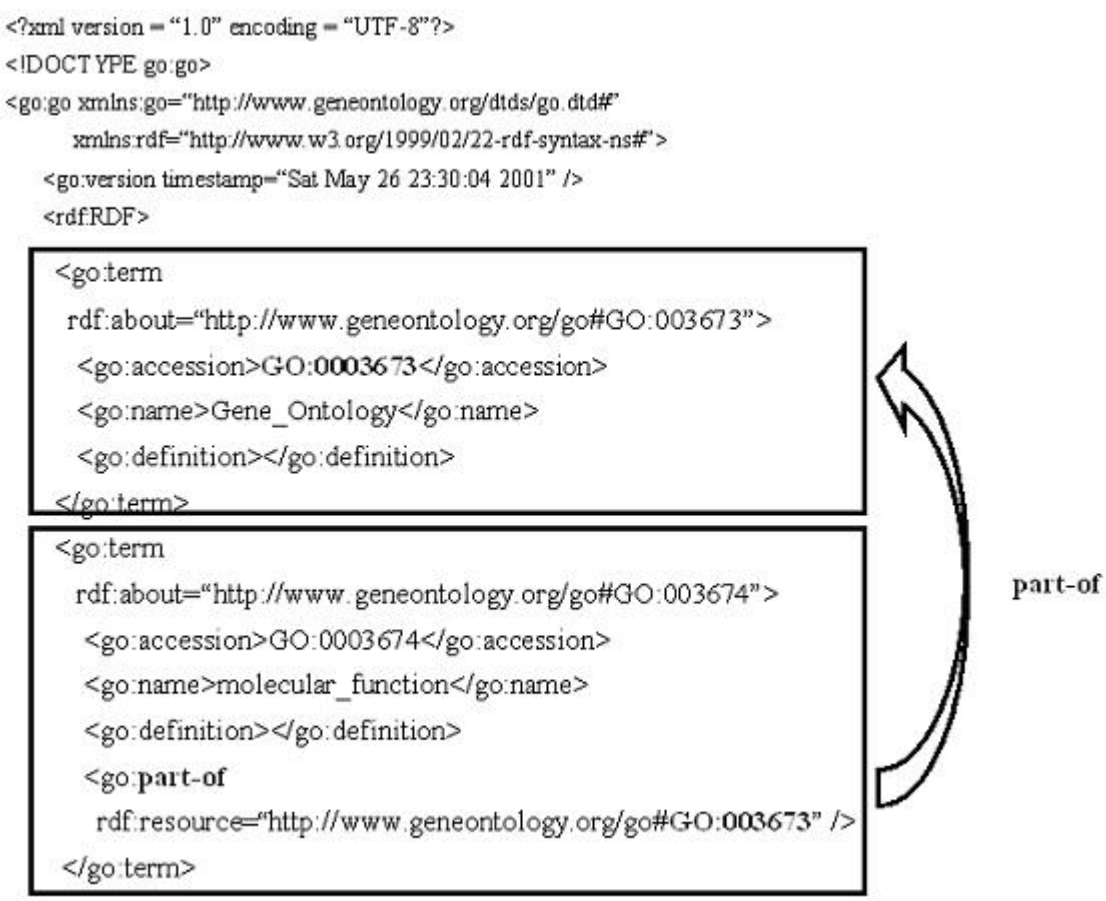

図 3.1　RDF による用語間の関係の記述

```
<!ELEMENT rdf:rdf (go:term*)>
<!ELEMENT go:term(
   go:accession          →     ID
   go:name               →     名称
   go:synonym*,          →     同義語
   go:definition?.       →     定義
   (go:part-of | go:isa)*,  →  親用語へのポインタ
   go:dbxref*,           →     他 DB 参照
   go:association*,      →     遺伝子との対応
   go:history*           →     更新履歴
   )>
```

図 3.2　go.xml のデータ構造

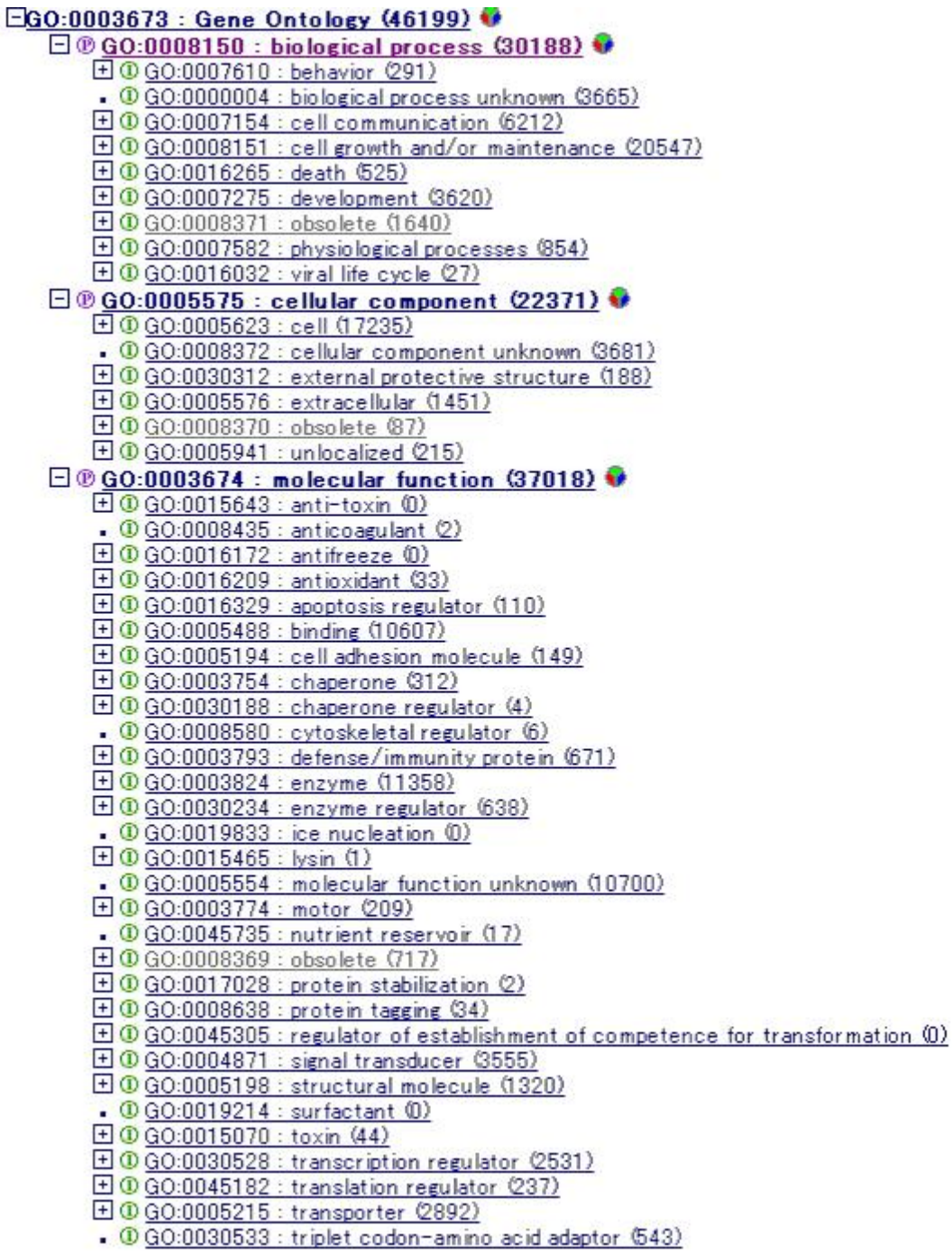

図 3.3　AmiGO のスクリーンショット：Gene Ontology は３つのオントロジーから構成され，用語の下に子用語がある場合は＋が表示されている

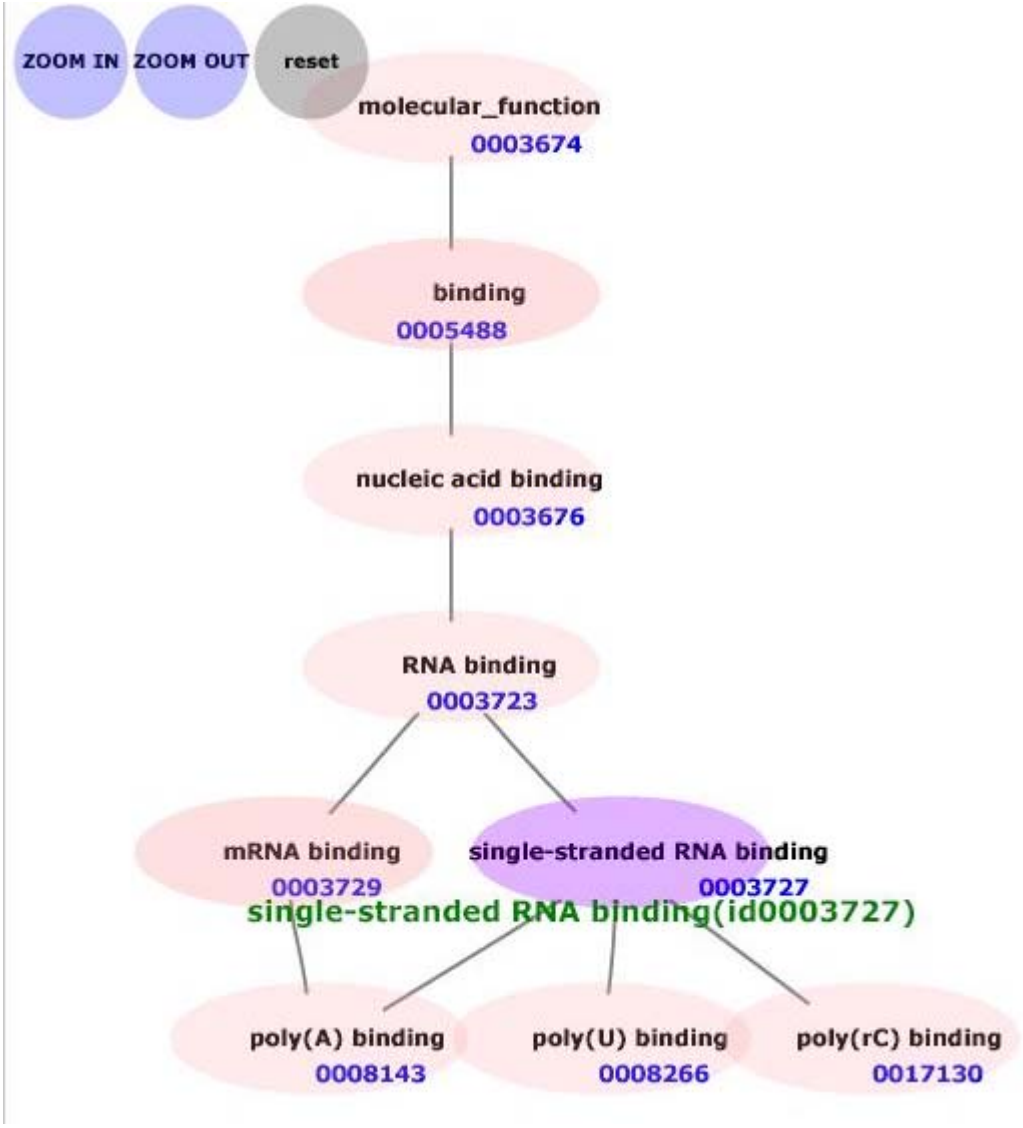

図 3.4　GeneAround のスクリーンショット

3.3　外延的オントロジー

今まで説明してきたオントロジーは，用語の意味を厳密に定義し，矛盾がないように概念同士を関係づけていく必要があるため，その構築には大変な労力を必要とする．さらに，その作業は，領域の専門家による合意を得ながら作成するため，多くの時間がかかり，大規模なオントロジーを構築することが非常に難しい．

これに対し，本研究室で開発している外延的オントロジーでは，専門用語の外延的な意味をその用語のあらゆる出現情報を基に，グループ化や階層化を行い，自動的にオントロジーを構築することを目標にしている．このオントロジーの特徴は，大量の専門用語を網羅することが可能で，オントロジーの更新も容易に行うことができる．

この章では，柳生ら[11]により構築された外延的オントロジーについて解説する．この外延的オントロジーは，ゲノム分野の知識の集合ともいえるゲノムネットで利用できるデータベースから専門用語の収集を行っている．これらのデータベースは世界各地で運営・管理されており，生命科学における文献情報の他に，ゲノムの地図と塩基配列，タンパク質のアミノ酸配列と立体構造，代謝系や制御系の分子ネットワーク，神経系や免疫系における細胞ネットワーク，そして発生・分化・老化や疾病に関する個体レベルのデータなど，多種多様なデータが含まれている．ゲノムネットでは，これらのデータベースを利用することができる[12]．ゲノムネットで利用できるデータベースに関する詳しい解説は，「ゲノムネットのデータベース利用法」[12]を参考にして欲しい．

これらのデータベースはエントリーと呼ばれる単位の情報がたくさん集まったものである．エントリーはいくつかのフィールドに分類されており，フィールドにはそれぞれ記述すべき内容や書式が定められている．この特徴を利用して，ある専門用語がどのデータベースのどのフィールドに出現しているかという情報とともに，外延的オントロジーを構築している．

3.4　オントロジーの拡張

3.4.1　オントロジーを拡張する目的

Gene Ontology は約 1 万語の専門用語を網羅しているのに対し, 外延的オントロジーは約 200 万語の専門用語を網羅している．そこで，まず本研究では，網羅している専門用語の数が少ないという既存のオントロジーの欠点を外延的オントロジーで拡張することにより解消し，効果的に専門用語を抽出することを試みる．最終的には，拡張したオントロジーを利用して，生物医学に関する文献から様々な機能や相互作用に関する情報を抽出するための動詞やテンプレートの発見を目指す．また，階層化されたオントロジーの特性を利用することにより，新たに機能や相互作用に関する情報を抽出できる可能性がある[図 3.5].

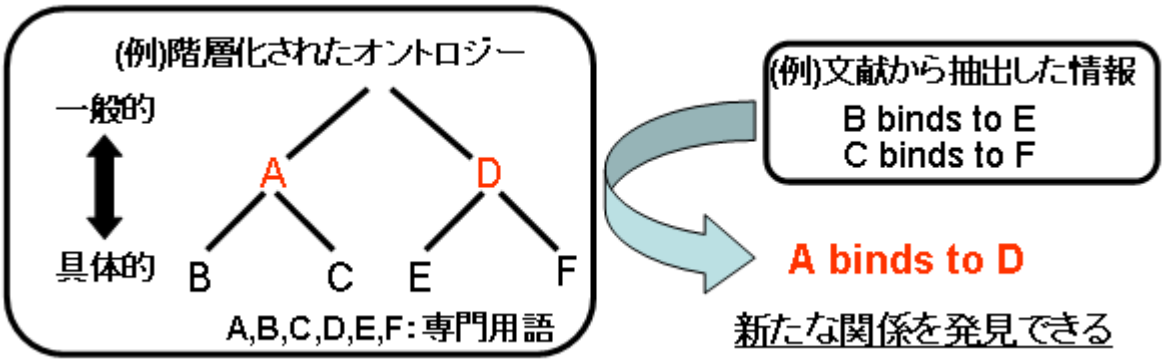

図 3.5　オントロジーの階層構造を利用した情報抽出

まず予備実験として，オントロジーを拡張することで多くの専門用語を認識できるようになる可能性はあるか，また，拡張するためにオントロジーの間で対応づけられる専門用語がどの程度あるかを調査した．これらの結果を参考にして，オントロジーの拡張を試みることにした.

3.4.2　準備

予備実験をする前の処理として以下のことを行った.

(ⅰ)　オントロジーからの専門用語の切り出し

Gene Ontology からは，name と synonym のタグに囲まれた専門用語をそれぞれ抽出し，外延的オントロジーからも同様に専門用語を抽出した(図 3.6).

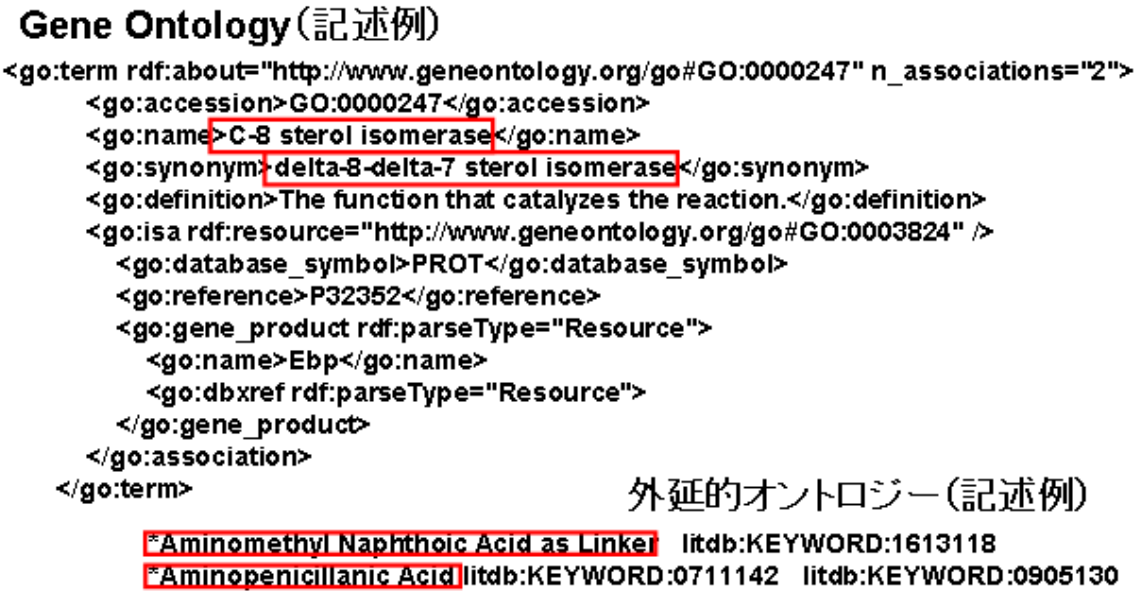

図 3.6　専門用語の切り出し例

(ⅱ)　用語の正規化

本来は同じ物を指しているが，表記のされ方が違うために予備実験の際に別の用語として扱われてしまう場合がある．この問題を回避するために以下のような処理を行った．この処理を今後は正規化と呼ぶ（図 3.7）.

・大文字はすべて小文字に変換

・特殊記号の除去

特殊記号(!”#$%&’()*+,-/;:<=>?@[¥]^_`{|}~ の 31 種)を全て空白に変換した後，2 つ以上連続している空白を 1 つの空白に置き換える.

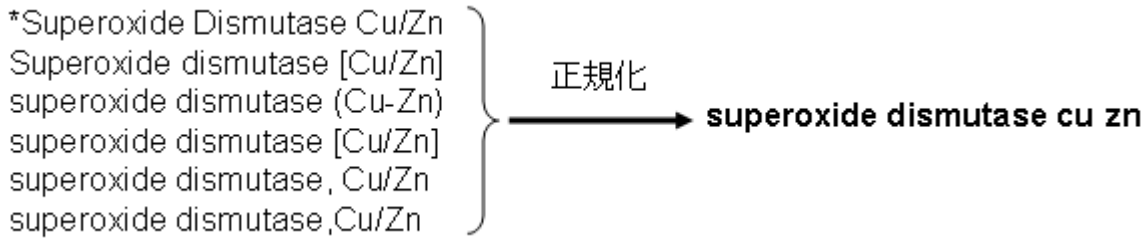

図 3.7　専門用語の正規化の例

これらの用語の正規化を行った結果，Gene Ontology からは 11,494 個の専門用語を抽出し，外延的オントロジーからは 2,138,347 個の専門用語を抽出した．これらの専門用語を用いて以下の 2 つの予備実験を行った.

・Gene Ontology と外延的オントロジーが，テキストに記述されている専門用語をそれぞれどの程度認識できるか.
・Gene Ontology と外延的オントロジーの間で，どの程度対応づけられる専門用語があるか.

3.4.3　予備実験（1）

この予備実験では，Gene Ontology と外延的オントロジーが，テキストに記述されている専門用語をどの程度認識できるか調査した．予備実験に用いた生物医学に関する文献データは，ジャーナル「Cell」の 1998~2002 年の文献のテキスト部分(総単語数：937,589 個)と OMIM(遺伝病に関する DB)のテキスト部分(総単語数：6,261,479 個)である．どの程度専門用語を認識できたか(認識率)の計算には，式〔3.1〕を用いた.

$$\frac{\text{マッチした専門用語が被覆している単語数}}{\text{テキストに含まれる単語数}} \times 100 \qquad \text{式〔3.1〕}$$

計算結果は, 表 3.1 のようになった. Cell のテキスト部分に対して, Gene Ontology は 4.0%の認識率を示し，外延的オントロジーは 34.6%の認識率を示した．OMIM のテキスト部分に対して，Gene Ontology は 2.6%の認識率を示し，外延的オントロジーは 35.5%の認識率を示した．いずれも外延的オントロジーの方が，Gene Ontology より専門用語の認識率が非常に高いことがわかった.

文献データ：Cell のテキスト部分(単語数 937,589 個)

オントロジー	被覆している単語数	専門用語の認識率
Gene Ontology	37,340 個	4.0%
外延的オントロジー	324,637 個	34.6%

文献データ：OMIM のテキスト部分(単語数 6,261,479 個)に対する

オントロジー	被覆している単語数	専門用語の認識率
Gene Ontology	159,972 個	2.6%
外延的オントロジー	2,221,119 個	35.5%

表 3.1　専門用語の認識率

3.4.4　予備実験（２）

この予備実験では，Gene Ontology と外延的オントロジーの間で，どの程度対応づけられる専門用語があるかを調査した．方法としては，専門用語の包含関係を見ることにより対応づけられる専門用語がどの程度あるかを調べた．例えば，genome と human genome では human genome の方がより狭い概念と考えられる．このように専門用語の包含関係を調べることにより，単純に概念の大きさを比べることができる．そのため，オントロジーを拡張する時の手掛かりにすることができる．包含関係を用いた専門用語の対応づけには，以下のような場合が考えられる.

(ⅰ) 完全に一致する場合

(Gene Ontology) cell death = (外延的オントロジー) cell death

(ⅱ) 部分的に連続かつ順序保存で含む場合

(Gene Ontology) blood coagulation factor

⊃ (外延的オントロジー) coagulation factor

(ⅲ) 部分的に不連続かつ順序無視で含む場合

(Gene Ontology) alcohol dehydrogenase nadp

⊃ (外延的オントロジー) nadp dehydrogenase

(ⅰ)～(ⅲ)のそれぞれの場合について，対応づけられる専門用語の数を，それぞれのオントロジーについて計算した結果を表 3.2 に示す．またそれぞれのオントロジーの中で，対応づけが可能な専門用語の数の合計を計算した結果を表 3.3 に示す．

(ⅰ)完全一致する時

専門用語の包含関係	対応づけられる専門用語数 Gene Ontology		外延的オントロジー
Gene Ontology ＝ 外延的オントロジ	4,120 個	＝	4120 個

(ⅱ)部分的に連続かつ順序保存で含む場合

専門用語の包含関係	対応づけられる専門用語数 Gene Ontology		外延的オントロジー
Gene Ontology ⊃ 外延的オントロジ	10,267 個	⊃	5,554 個
Gene Ontology ⊂ 外延的オントロジ	4,301 個	⊂	310,705 個

(ⅲ)部分的に不連続かつ順序無視で含む場合

専門用語の包含関係	対応づけられる専門用語数 Gene Ontology		外延的オントロジー
Gene Ontology ⊃ 外延的オントロジ	3,728 個	⊃	3,530 個
Gene Ontology ⊂ 外延的オントロジ	2,882 個	⊂	48,970 個

表 3.2　包含関係から対応づけられる専門用語の数

オントロジー	対応づけ可能な専門用語数	全専門用語数
Gene Ontology	11,378 個	11,494 個
外延的オントロジー	330,097 個	2,138,347 個

表 3.3　対応づけが可能な専門用語数の合計

Gene Ontology のほとんどの専門用語は, 外延的オントロジーの専門用語と対応づけ可能なことがわかった．一方，外延的オントロジーの専門用語は，この方法だけでは Gene Ontology の専門用語と約 15%しか対応づけられないことがわかった.

3.4.5　考察

予備実験(１)の結果から，網羅している専門用語の量が少ないという既存のオントロジーの問題を, 外延的オントロジーで拡張することにより解消できる可能性が示唆された．予備実験(２)の結果から専門用語の包含関係を利用するだけでは，外延的オントロジーの約 33 万語の専門用語しか Gene Ontology に対応づけられないことがわかった．そのため専門用語の包含関係以外に対応づける方法を模索する必要がある．しかし，専門用語の包含関係以外に対応づける方法を模索したところ，Gene Ontology には，機能や現象に関する階層化が多くなされており，遺伝子名や酵素名といった具体的な物質名に関するカテゴリーが無いため, 単純には対応づけられないことがわかった．そのためこれ以降は，外延的オントロジーのみから抽出した専門用語リストを用いて，人手で行っていた動詞やテンプレートの発見を，機械的に抽出することを目指した.

第　4　章

専門用語間の関係の抽出

この章では，本研究での情報抽出手法に関して述べる．

4.1　テンプレート抽出の流れ

専門用語の関係を表す動詞およびテンプレートの抽出は以下のような方法で行った．

（1）専門用語リストの作成(4.2 節)

専門用語を認識するための専門用語リストとして，外延的オントロジーを利用する．既存のオントロジーが最大数万程度の専門用語しか網羅していないのに対して，　柳生らより構築された外延的オントロジーは約 200 万の専門用語を網羅しているため，文献中の専門用語を高い確立で認識できることが期待される．

（2）インターバルの抽出(4.3 節)

作成した専門用語リストを生物医学の文献に適用する(図 4.1)．専門用語間に挟まれている部分(以後インターバルと呼ぶ)に着目すると，(図 4.2)に見られるように前後の専門用語の関係を表すような記述が多く見られる．そこで，インターバルを抽出して着目すべき動詞やテンプレートの発見を目指す．

finally it was discovered recently that ***apc*** binds to ***asef*** an exchange factor that apparently activates the ***small g protein rac*** which in turn controls the ***actin*** cytoskeleton kawasaki et al. 2000

図 4.1　専門用語リストの文献への適用例：赤字は専門用語を表す．

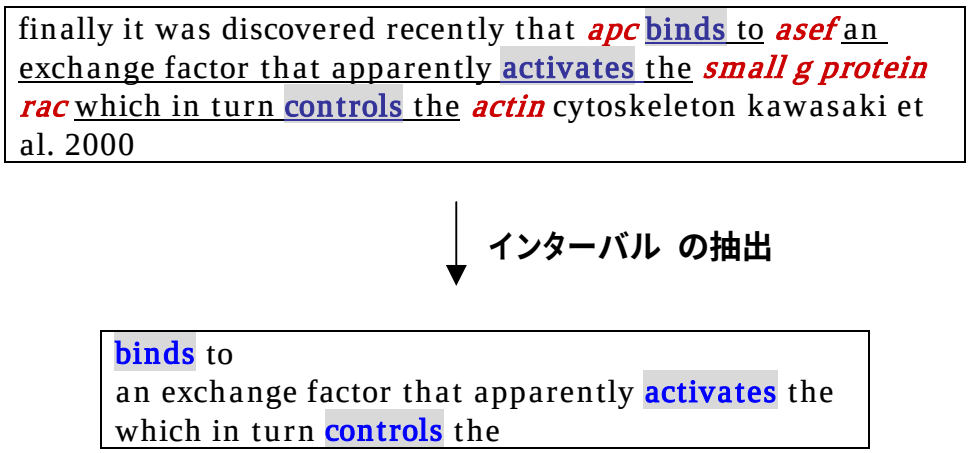

図 4.2　インターバルの抽出：赤字は専門用語を表し，下線部分が専門用語の間に挟まれた部分(インターバル)を表す．青字は，インターバルの前後の専門用語の関係を表すと考えられる動詞を表す．

（3）インターバルに特異的な単語の抽出(4.4 節)

テキスト中で特にインターバルに出現する単語は，インターバルに特異的な単語であるといえる．このようなインターバルに特異的に出現している単語は，インターバルの前後の専門用語と何らかの関係性を持っていると考えられる．そこで，インターバルに特異的な単語の抽出を行った．

（4）インターバルに特異的なテンプレート抽出(4.5 節)

4.4 節で抽出したインターバルに特異的な単語を含むテンプレートの抽出を行った．

4.2　専門用語リストの作成

柳生らにより構築された外延的オントロジーから専門用語を抽出した．3.3 節でも述べたが，外延的オントロジーは，ゲノムネットで利用できるデータベースから専門用語の収集を行っている．これらのデータベースはエントリーと呼ばれる単位の情報がたくさん集まり構築されている．エントリーはいくつかのフィールドに分類されており，フィールドにはそれぞれ記述すべき内容や書式が定められている．このようなデータベースのどのフィールドに出現しているかという情報を基に外延的オントロジーを構築している．オントロジーを構築する際に，柳生らはフィールドのカテゴリ

ー化を行っている. 例えば genbank というデータベースの gene というフィールドには遺伝子に関する用語が記述されており，brite というデータベースの ORGANISM というフィールドには生物種に関する用語が記述されている．フィールドのカテゴリー化とは，それぞれのフィールドには記述すべき内容や書式が定められているという特徴を利用して，同じ内容を表すフィールドが複数ある場合にそれらのフィールドを 1 つのカテゴリーに統一することである．クラスタリングによる単語の分類をする際には，同じ内容が記述されているフィールドは統一して処理する方が良い結果が得られるためカテゴリー化を行っている．表 4.1 は，柳生らによって行われたフィールドのカテゴリー化の一覧である．本研究では，専門用語リストを作成する際に，これらのカテゴリー化されたフィールドを利用した．これらのカテゴリーの中で，物質名を多く含んでいるカテゴリーである organism, organism_class, protein, compound, gene に属するフィールドから専門用語を抽出した．これら５つのカテゴリーのフィールドから抽出できた専門用語の数は 1,082,830 個であった.

本来は同じ物を指しているが，表記のされ方が違うために別の用語として扱われてしまう場合を回避するために，3.4 節と同様に正規化を行った．５つのカテゴリーのフィールドから抽出できた専門用語は 1,082,830 個であったが，正規化を行うことにより 887,574 個になった.

カテゴリー	データベース：フィールド
organism	brite:ORGANISM
	epd:OS
	genbank:organism
	refseq:organism
	genome:NAME
	genome:DEFINITION
	pmd:EXPRESSION-SYSTEM
	pmd:SOURCE
	swissprot:OS
	transfac:OS
	genbank:specific_host
	refseq:specific_host
	genbank:lab_host
	refseq:lab_host
	genbank:sub_species
	refseq:sub_species
	genbank:variety
	refseq:variety
organism_class	genome:LINEAGE
	swissprot:OC
	transfac:OC
protein	pmd:PROTEIN
	prf:NAME
	transfac:DE
	genbank:product
	refseq:product
	enzyme:NAME
compound	compound:NAME
morphorogy	genome:MORPHOLOGY
physiology	genome:PHYSIOLOGY
sex	genbank:sex
	refseq:sex
gene	genbank:gene
	refseq:gene
mnemonic	brite:MNEMONIC

カテゴリー	データベース:フィールド
strain	genbank:strain
	refseq:strain
phenotype	genbank:phenotype
	refseq:phenotype
plasmid	genbank:plasmid
	refseq:plasmid
organelle	genbank:organelle
	refseq:organelle
tissue_type	genbank:tissue_type
	refseq:tissue_type
cell_type	genbank:cell_type
	refseq:cell_type
cell_line	genbank:cell_line
	refseq:cell_line
enzyme_product	enzyme:PRODUCT
enzyme_cofactor	enzyme:COFACTOR
enzyme_effector	enzyme:EFFECTOR
enzyme_inhibitor	enzyme:INHIBITOR
enzyme_substrate	enzyme:SUBSTRATE
transposon	genbank:transposon
	refseq:transposon
function	brite:FUNCTION
	genbank:function
	refseq:function
dev_stage	genbank:dev_stage
	refseq:dev_stage
bound_moiety	genbank:bound_moiety
	refseq:bound_moiety
clone_lib	genbank:clone_lib
	refseq:clone_lib
rpt_family	genbank:rpt_family
	refseq:rpt_family
sub_clone	genbank:sub_clone
	refseq:sub_clone

表 4.1　カテゴリー化したフィールド

4.3　インターバルの抽出

4.3.1　本研究で用いた文献データ

動詞やテンプレートの抽出に用いた文献データは,

・ジャーナル「Cell」の 1998~2002 年の文献のテキスト部分(文献数：299 種類)
・OMIM(遺伝病に関するデータベース)のテキスト部分
(エントリー数 : 13,818 エントリー)
・Medline に 2002 年登録された文献のアブストラクト
(アブストラクト数 : 33,622 種類)

である．それぞれの文献のテキスト部分は，文単位に変形した．そして，専門用語リストの場合と同様に正規化を行い, さらに一般用語と機能語の抽象化を行った(図 4.3). 表 4.2 は抽象化した一般用語と機能語を示す.

Grier et al. (1983) reported father **and** 2 sons **with** typical Aarskog syndrome, including short stature, hypertelorism, **and** shawl scrotum. **They** tabulated **the** findings **in** 82 previous cases. X-linked recessive inheritance **has been** repeatedly suggested (see 305400). **The** family reported **by** Welch (1974) **had** affected males **in** 3 consecutive generations.

grier et al. 1983 reported father **CONJUNCTION** 2 sons **PREPOSITION** typical aarskog syndrome including short stature hypertelorism **CONJUNCTION** shawl scrotum

PRONOUN tabulated **ARTICLE** findings **PREPOSITION** 82 previous cases

x linked recessive inheritance **HAVE BE** repeatedly suggested see 305400

ARTICLE family reported **PREPOSITION** welch 1974 **HAVE** affected males **PREPOSITION** 3 consecutive generations

図 4.3　テキスト処理の例

抽象化後	抽象化前
ARTICLE	a an the
RELATIVE	who whose whom which that what when where why how whoever whomever whichever whatever wherever whenever
PRONOUN	i my me you your he his him she her it its we our us they their them mine yours theirs myself ourselves yourself yourselves himself herself itself some any all both each every either neither no none everybody everyone everything somebody someone something anybody anyone anything nobody nothing
AUXILLARY-VERB	can could must may might will shall need ought to did does
CONJUNCTION	and or but for neither also besides then however nevertheless still yet else so therefore consequently hence namely whether if while as before until till since directly immediately because unless although while whereas than both
PREPOSITION	after at by for from in of on upon onto till to under with without up aboard about above across along alongside around before behind below besides between beyond by down less near off opposite over outside past round since through throughout under underneath within without against during toward towards across among into
BE	is are was were am been being be
HAVE	have has had having

表 4.2　抽象化を行った一般用語と機能語

4.3.2　テキストからのインターバル抽出

外延的オントロジーから作成した専門用語リストを用いて，文献のテキスト部分に記述されている専門用語の認識を行った(図 4.4).

finally PRONOUN BE discovered recently RELATIVE ***apc*** binds PREPOSITION ***asef*** ARTICLE exchange factor RELATIVE apparently activates ARTICLE ***small g protein rac*** RELATIVE PREPOSITION turn controls ARTICLE ***actin*** cytoskeleton kawasaki et al. 2000

図 4.4　専門用語リストの文献への適用例：赤字は専門用語を表す

次に，　インターバル(専門用語間に挟まれている部分)を抽出した(図 4.5).

finally PRONOUN BE discovered recently RELATIVE ***apc*** binds PREPOSITION ***asef*** ARTICLE exchange factor RELATIVE apparently activates ARTICLE ***small g protein rac*** RELATIVE PREPOSITION turn controls ARTICLE ***actin*** cytoskeleton kawasaki et al. 2000

↓ インターバルの抽出

binds PREPOSITION ARTICLE exchange factor RELATIVE apparently activates ARTICLE RELATIVE PREPOSITION turn controls ARTICLE

図 4.5　インターバルの抽出例

Cell の文献のテキスト部分と OMIM のテキスト部分から抽出したインターバルの集合と Medline の文献のアブストラクトから抽出したインターバルの集合から，それぞれ専門用語の間の関係を表す動詞やテンプレートの抽出を試みた.

4.4　インターバルに特異的な単語の抽出

４.4.1　単語の抽出と評価

4.3 節で抽出したインターバルの集合から，インターバルに特異的な単語の抽出を試みた.

まず，テキスト中で特にインターバル中に多く出現する単語は，インターバルに特異的な単語であると考えることができる．このようなインターバルに特異的な単語は，インターバルの前後の専門用語と何らかの関係性を持っていると考えられる．インターバルに特異的に出現して一定の出現数をもつ単語は，専門用語の関係を表す動詞やテンプレートを抽出すために着目すべき単語であるといえる．そこで，インターバルに出現したすべての単語に対し，全テキスト中では何回出現し，インターバル中では何回出現したかを計算した．そして，式〔4.1〕により，着目するべき単語としてふさわしいかどうかを評価した．なお，着目するには明らかに不適当な単語は評価の際に除去した．例えば不適当な単語として，j ，b，az，bv，xd，80s，n187，73.0 などがある．そこで，評価する単語は 3 文字以上で数字を含まない単語に限定した.

$$\frac{a \times \log(a)}{b} \times 100 \qquad \text{式〔4.1〕}$$

a：ある単語がインターバル中に出現した回数

b：ある単語が全テキスト中に出現した回数

4.4.2　結果

インターバルに出現した単語に評価式を適用した結果を示す．Cell と OMIM のテキストのインターバルに出現した単語(32,952 個)の評価値と単語数の分布図を図 4.6 に示した．また，評価値が上位から 100 番目までの単語を表 4.3 に示し，表 4.4 の(a)は評価値が上位から 1,000～1,050 番目の単語を示し，(b)は 32,902～32,952 番目の単語を示している．

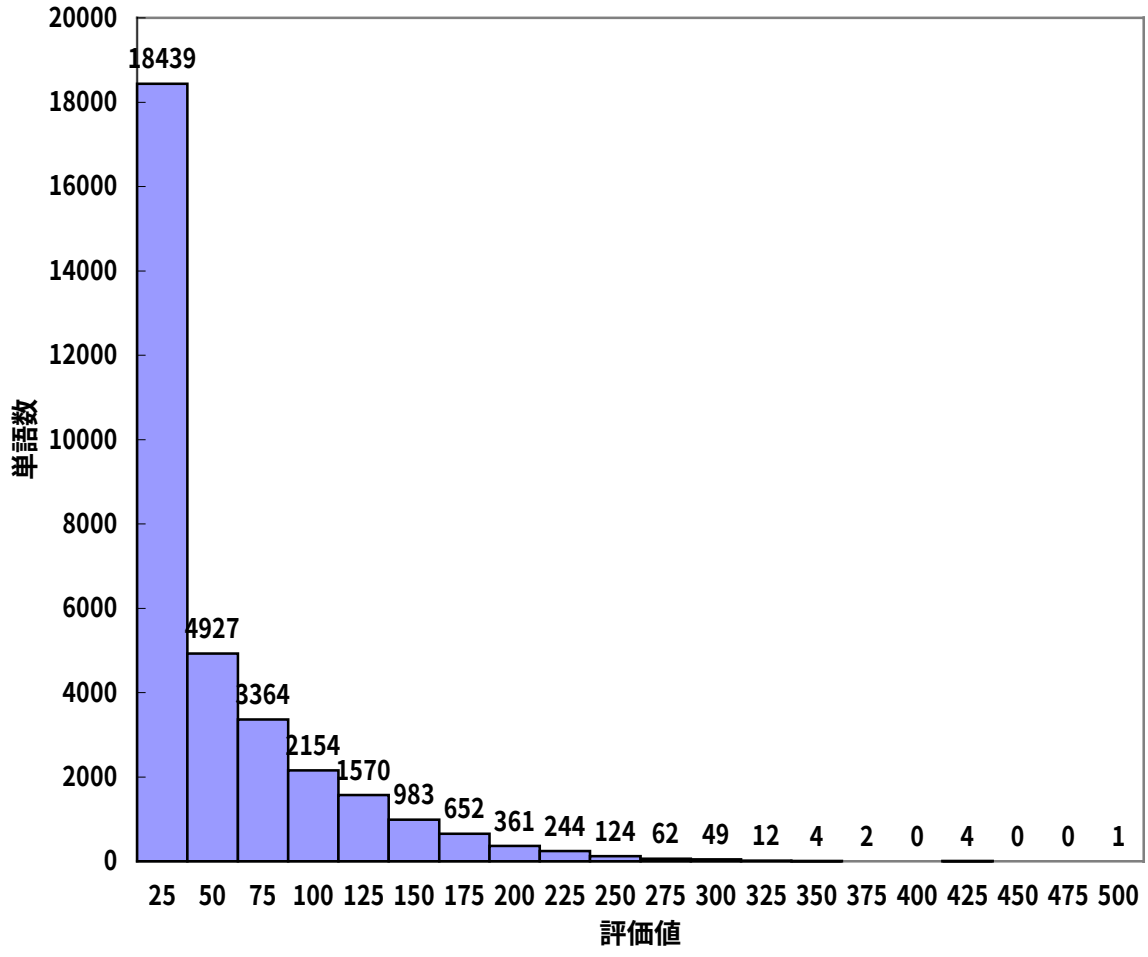

図 4.6　Cell と OMIM のインターバルの単語分布図(抽象化した単語は除外)

評価値	単語数
477	roychoudhury
413	CONJUNCTION
411	hybrids
403	PREPOSITION
402	deficient
359	symbolized
351	binds
349	ARTICLE
345	located
333	RELATIVE
333	ceacam
324	encoded
324	chromosome
320	bind
319	homolog
318	produced
316	lacking
314	electrophoresis
311	carrying
308	mediated
307	homologous
305	activates
301	containing
299	activate
299	expressing
298	bound
298	product
296	required
295	via
294	telomeric
294	site
293	codes
293	coded
293	inhibits
292	catalyzes
292	induced
291	stimulates
291	express
290	chains
289	increase
288	tabulated
288	produce
288	resulting
288	prime
288	deficiency
287	hybridize
287	substitution
287	encodes
286	ternary
286	promotes

評価値	単語数
286	synthesized
286	activity
286	not
284	stimulated
283	amplify
283	dnas
282	recognizes
281	mutant
280	plays
280	interacts
279	regulate
279	release
278	haploid
278	leads
278	activation
278	genes
277	driven
277	cultured
277	due
276	carries
276	stimulate
276	conversion
275	amplification
275	panel
274	regulates
274	synthesis
273	specific
272	mediates
272	resulted
272	shown
271	converts
271	induces
271	suggesting
270	equivalent
270	concentrations
269	resistant
269	assigned
269	including
269	levels
266	lies
266	induce
266	production
266	intron
266	single
265	inhibit
265	products
264	upstream
263	potentiation
263	promoter
262	tropic

表 4.3　Cell と OMIM のインターバルで評価値が高い単語

評価値	単語
169	pushing
169	homogenate
169	dyl
169	arylsulfatases
169	aiv
169	werdnig
169	nonspherocytic
169	guard
169	abolishes
169	coreceptor
169	rendered
169	correspond
169	upregulation
169	confirming
169	regulators
169	link
169	exclusively
169	density
169	double
168	suppressive
168	retrovirus
168	nonfunctional
168	establishing
168	readily
168	specialized
168	microscopic
168	oxidative
168	effective
168	switch
168	indicates
168	determining
168	event
168	insertion
168	allelic
168	defined
168	primary
168	nuclear
168	common
168	most
167	hairpins
167	augment
167	responds
167	coordinates
167	colored
167	vegfr
167	phosphorylating
167	competing
167	mossy
167	imported
167	bacterium

(a)

評価値	単語
0	borsani
0	bondeson
0	boggs
0	biggs
0	barak
0	bader
0	assortment
0	asparaginyl
0	appel
0	antioxidants
0	ankyloglossia
0	albuminuria
0	albipunctatus
0	alanyl
0	ailment
0	acheiropody
0	gaucher
0	friedreich
0	maroteaux
0	guanylate
0	johnston
0	znfs
0	protocadherins
0	january
0	pcdhg
0	october
0	allen
0	sonic
0	baraitser
0	niece
0	senile
0	pcdhb
0	cisternal
0	teebi
0	dehydrocholesterol
0	fetoprotein
0	endoplasmic
0	retinoic
0	mice
0	alpha
0	complex
0	small
0	yeast
0	epsilon
0	gamma
0	exclusion
0	bcl
0	heparan
0	humans
0	hippel

(b)

表 4.4　Cell と OMIM のインターバルで評価値が低い単語：(a)は上位から 1,000～1,050 番目の単語を示し，(b)は 32,902～32,952 番目の単語を示す

Medline のテキスト部分のインターバルに出現した単語(108,888 個)の評価値と単語数の分布図を図 4.7 に示した．また，評価値が上位から 100 番目までの単語は表 4.5 に示し，表 4.6 の(a)は評価値が上位から 1,000～1,050 番目の単語を示し，(b)は 108,838～108,888 番目の単語を示している．

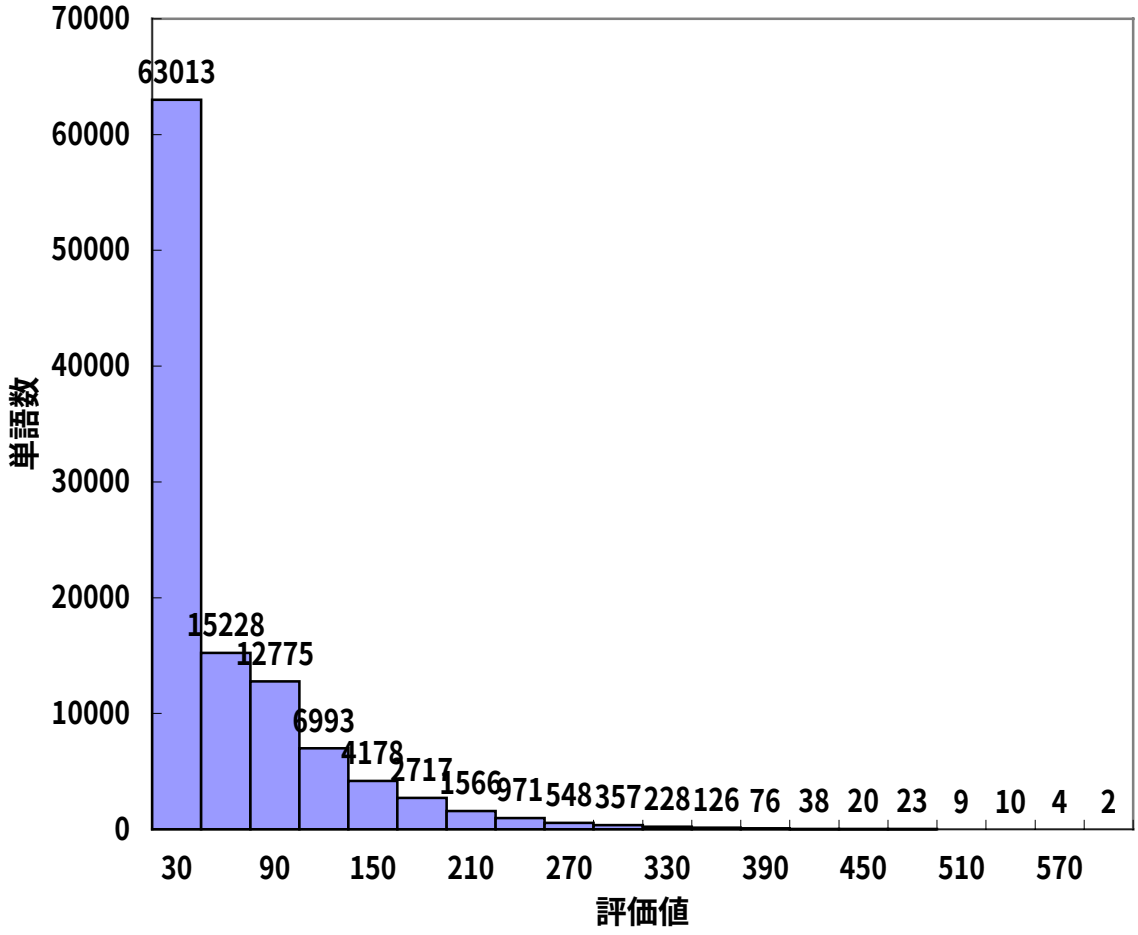

図 4.7　Medline のインターバルの単語分布図(抽象化した単語は除外)

評価値	単語
598	encodes
587	deficient
569	mediated
556	binds
552	induced
545	expressing
538	encoding
531	phosphorylation
530	catalyzes
528	stimulated
523	suggesting
522	regulates
521	inhibited
518	nick
515	interacts
515	expression
509	inhibits
506	containing
501	CONJUNCTION
498	expressed
496	kinases
493	activates
493	bind
491	activation
489	dodecyl
488	express
477	blocked
477	induces
474	encoded
471	plays
470	via
469	mediates
469	stimulates
469	regulate
467	reporter
467	signaling
467	PREPOSITION
465	bound
464	phosphorylated
463	homologue
459	member
459	inhibit
458	activate
458	inhibitors
454	transfected
454	regulated
454	production
452	antagonist
452	indicating
452	promoter

評価値	単語
446	overexpressing
443	activity
442	methyl
442	plus
441	receptors
438	regulating
437	downstream
437	synthesis
435	induce
433	release
432	regulation
431	subunits
431	contains
430	proteins
428	secretion
428	dependent
428	levels
426	phosphorylates
425	pathway
422	homolog
420	intracellular
420	mrna
420	cells
418	transduction
417	polyacrylamide
417	homology
416	glucopyranosyl
416	mutant
414	stably
414	abolished
414	anti
413	peroxidation
413	interact
408	suppresses
407	expresses
407	residues
406	agonist
406	infected
405	homologous
405	prevented
404	catalyzed
403	promotes
403	suppressed
403	ARTICLE
402	phosphorylate
402	potent
401	stimulate
401	cultured
399	tagged
399	constitutively

表 4.5　Medline のインターバルで評価値が高い単語

評価値	単語
265	only
264	dephosphorylates
264	arabinofuranosyl
264	pregenomic
264	opener
264	hyperphosphorylation
264	hyperphosphorylated
264	nfat
264	acetylated
264	synthesizing
264	polycyclic
264	suppressing
264	monomeric
264	stimulatory
264	junctions
264	immature
264	drinking
264	blockade
264	affects
264	tolerance
264	primary
264	related
263	downregulates
263	upregulate
263	rnas
263	altering
263	genus
263	initiate
263	conditioned
263	exclusively
263	hydrophobic
263	normally
263	primarily
263	mainly
263	female
263	single
263	methylethyl
263	khalid
262	hydrochloric
262	competed
262	ionomycin
262	metabolizing
262	implicating
262	constitute
262	generating
262	precursors
262	pair
262	allele
262	alone
262	reduction

(a)

評価値	単語
0	shigella
0	piperacillin
0	king
0	hai
0	camptothecin
0	adhesin
0	var
0	pneumophila
0	nep
0	lyn
0	angiotensinogen
0	push
0	dichloromethane
0	birch
0	ankyrin
0	reticulocyte
0	proteus
0	groel
0	cart
0	pkg
0	fnr
0	imp
0	dithiothreitol
0	sps
0	recombinase
0	synechocystis
0	syk
0	gpa
0	arginase
0	avidin
0	natal
0	abr
0	som
0	nucleocapsid
0	caa
0	aas
0	tce
0	spider
0	dioxygenase
0	chs
0	scopolamine
0	lpc
0	prs
0	aeromonas
0	mad
0	aquaporin
0	drosophila
0	tcr
0	mitogen
0	knockout

(b)

表 4.6　Medline のインターバルで評価値が高い単語：(a)は上位から 1,000～1,050 番目の単語を示し，(b)は 108,838～108,888 番目の単語を示す

4.4.3　結果の考察

評価値が上位から 100 番目までの単語に着目する．Cell と OMIM のインターバルから抽出した評価値が上位の単語(表 4.3)には，インターバルの前後の専門用語の間の関係を表すような単語が多く見られた．また，100 個中の約 6 割の単語がインターバルの前後の専門用語の間の関係を表すような動詞で占められていた．しかし，評価値が一番高い roychoudhury は人名であり，インターバルの前後の専門用語の間の関係を表す単語としては不適当で好ましくない結果であった．Medline のインターバルから抽出した評価値が上位の単語(表 4.5)も同様に，インターバルの前後の専門用語の間の関係を表すような単語が多く見られた．また，こちらも約 6 割の単語がインターバルの前後の専門用語の間の関係を表すような動詞で占められていた．

表 4.4 や表 4.6 からは，評価値が低くなるにつれてインターバルの前後の専門用語の間の関係を表すような単語が減少していくことがわかった．どちらの場合も，特に評価値が下位から 50 番目までの結果(表 4.4.b や表 4.6.b)を見ると，専門用語の間の関係を表すような単語は見られず，人名や記号のような単語が多く見られた．

インターバルから抽出した全単語の評価値と数の分布に着目すると, Cell と OMIM のインターバルから抽出した単語(図 4.6)は，最大評価値を 500 とした場合，最大評価値の 25%以下の値(評価値 0～125)である単語が，全単語(32,952 個)の 92.4%を占めていた．また，最大評価値の 75%以上の値を持つ単語(評価値 375～500)は 5 個存在した．Medline のインターバルから抽出した単語(図 4.7)は，最大評価値を 600 とした場合, 最大評価値の 25%以下の値(評価値 0～150) である単語が, 全単語(108,888 個)の 93.9%を占めていた．また，最大評価値の 75%以上の値を持つ単語(評価値 450～600)は 48 個存在した．これら単語の分布の結果から，Medline のインターバルから抽出した単語の方が，Cell と OMIM のインターバルから抽出した単語よりもインターバルに特異的な単語とそうでない単語の差別化がされていると考えられる．よって，Medline から抽出した単語(動詞)やテンプレートの方が，より効果的な情報抽出ができる可能性が高いのではないかと思われる．

4.5　インターバルに特異的なテンプレート抽出

4.5.1　テンプレートの抽出と評価

インターバルに出現した単語で, 評価値が上位から 100 番目までの単語を含むテンプレートの抽出を行った．テンプレートを抽出するにあたり，図 4.8 のアルゴリズムに従った．式〔4.2〕は，テンプレートの抽出に用いた評価式である．

１．インターバルに特異的な単語を入力する．
２．その単語を含んでいるすべてのインターバルを抽出する．
３．その単語を長さ１のテンプレートとして，式〔4.2〕を用いて評価する．
４．そのテンプレートに他の単語を加えて１つ長くしたテンプレートを作成する．
(例)　bind　→　bind PREPOSITION
５．そのテンプレートを式〔4.2〕で評価する．
６．(a) １ つ長くしたテンプレートの評価値の方が，長くする前のテンプレートよりも高い場合は４の操作に戻る．
(b) １つ長くしたテンプレートの評価値の方が, 長くする前のテンプレートよりも低い場合は，長くする前のテンプレートを結果として出力する

図 4.8　テンプレート抽出のアルゴリズム

$$\frac{a \times \log(a)}{b} \times 100 + \log(c) \times 100 \qquad \text{式〔4.2〕}$$

a：ある単語がインターバル中に出現した回数

b：ある単語が全テキスト中に出現した回数

c：テンプレートの長さ

4.5.2 結果

それぞれのインターバルに出現した単語の中で, 評価値が上位から 100 番目までの単語を含むテンプレートの評価を行った結果を示す．Cell と OMIM の場合のテンプレートの評価結果を表 4.7，表 4.8，表 4.9，表 4.10 に示した．表 4.7 は着目した単語を含む全テンプレートの中で評価値が上位から 50 番目までのテンプレートを示しており，表 4.8 は評価値が下位から 50 番目までのテンプレートを示している．また，表 4.9 と表 4.10 は，着目した各単語を含むテンプレートの中で一番評価値が高かったテンプレートを示している．Medline の場合のテンプレートの評価結果を表 4.11，表 4.12，表 4.13，表 4.14 に示した．表 4.11 は着目した単語を含む全テンプレートの中で評価値が上位から 50 番目までのテンプレートを示しており，表 4.12 は評価値が下位から 50 番目までのテンプレートを示している．また，表 4.13 と表 4.14 は，着目した各単語を含むテンプレートの中で一番評価値が高かったテンプレートを示している．全テンプレート中からのテンプレート抽出と各単語に着目した場合のテンプレート抽出を行った．

評価値	テンプレート
701	frequencies PREPOSITION allelic variants BE tabulated PREPOSITION roychoudhury CONJUNCTION
678	allelic variants BE tabulated PREPOSITION roychoudhury CONJUNCTION
646	BE tabulated PREPOSITION roychoudhury CONJUNCTION
524	CONJUNCTION ARTICLE site PREPOSITION ARTICLE
509	genes BE tandemly oriented PREPOSITION ARTICLE 5 prime PREPOSITION 3 prime direction PREPOSITION telomere PREPOSITION
502	ARTICLE site PREPOSITION ARTICLE
490	BE present PREPOSITION single
489	PREPOSITION ARTICLE panel PREPOSITION
484	CONJUNCTION ARTICLE site PREPOSITION
465	chromosome 1.
462	site PREPOSITION ARTICLE
461	homolog PREPOSITION ARTICLE
458	subfamily ARTICLE ceacam subfamily see 109770 CONJUNCTION ARTICLE ceacam
458	specific probes indicated PRONOUN ARTICLE order PREPOSITION ARTICLE 11
454	BE ARTICLE site PREPOSITION ARTICLE
454	chromosome 7.
453	candidate CONJUNCTION ARTICLE site PREPOSITION ARTICLE
450	present PREPOSITION single
450	BE located PREPOSITION
450	hybrids PREPOSITION
447	PREPOSITION chromosome 1.
446	cluster CONJUNCTION 6 genes belonging PREPOSITION ARTICLE third
443	PREPOSITION chromosome 7.
441	ARTICLE site PREPOSITION
440	PREPOSITION ARTICLE region PREPOSITION homology PREPOSITION synteny PREPOSITION
439	BE closely linked PREPOSITION
439	chromosome 3.
434	BE encoded PREPOSITION ARTICLE single
434	PREPOSITION homology PREPOSITION synteny PREPOSITION
434	PREPOSITION chromosome 1. ARTICLE
434	chromosome 5.
433	CONJUNCTION ARTICLE candidate
432	chromosome 2.
428	substitution PREPOSITION ARTICLE
428	hybrids PRONOUN
425	encoded PREPOSITION ARTICLE single
423	chromosome 1. ARTICLE
422	homolog PREPOSITION
421	BE located PREPOSITION ARTICLE
420	BE located PREPOSITION ARTICLE telomeric
420	encoded PREPOSITION ARTICLE
420	BE ARTICLE site PREPOSITION
418	site PREPOSITION
418	hybrids containing
415	PREPOSITION ARTICLE telomeric
414	located PREPOSITION
413	BE closely linked PREPOSITION ARTICLE
412	hybrids CONJUNCTION
411	encoded PREPOSITION
410	closely linked PREPOSITION

表 4.7　評価値が上位から 50 番目までのテンプレート(Cell と OMIM の場合)

評価値	テンプレート
107	b locus
103	PREPOSITION carrying
103	induced growth
103	ii activity
102	encoding genes
101	closely resemble
101	activation pathway
101	ii deficiency
100	PRONOUN cultured
99	genes cause
98	cdna encoded
98	activation requires
98	receptor deficiency
98	PREPOSITION overexpressing
98	homolog designated
98	HAVE deficient
94	action potential
94	expression increased
94	cdnas containing
94	locus contains
94	inducing activity
94	completely lacking
92	deficiency see
92	induced angiogenesis
92	bacterially expressed
90	irradiation induced
89	sequence containing
88	data suggesting
88	containing 3
88	homology regions
85	activation induced
84	domain containing
82	oxidation pathway
81	2 chains
81	stimulated peripheral
81	homolog ARTICLE
80	domains containing
79	2000 expressed
78	sequence encodes
78	thereby promoting
78	motor activity
78	containing 8
77	PRONOUN generate
74	receptor deficient
69	region genes
69	receptor locus
69	1 deficient
69	chromosome number
69	type locus
69	termination site

表 4.8　評価値が下位から 50 番目までのテンプレート(Cell と OMIM の場合)

単語	テンプレート
roychoudhury	frequencies PREPOSITION allelic variants BE tabulated PREPOSITION roychoudhury CONJUNCTION
hybrids	hybrids PREPOSITION
deficient	deficient PREPOSITION ARTICLE
symbolized	BE CONJUNCTION symbolized
binds	binds PREPOSITION ARTICLE
located	BE located PREPOSITION
ceacam	subfamily ARTICLE ceacam subfamily see 109770 CONJUNCTION ARTICLE ceacam
encoded	BE encoded PREPOSITION ARTICLE single
chromosome	chromosome 1.
bind	PREPOSITION bind PREPOSITION
homolog	homolog PREPOSITION ARTICLE
produced	produced PREPOSITION
lacking	lacking ARTICLE
electrophoresis	electrophoresis CONJUNCTION
carrying	carrying ARTICLE
mediated	mediated cleavage PREPOSITION
homologous	homologous PREPOSITION PRONOUN PREPOSITION ARTICLE
activates	activates ARTICLE
containing	hybrids containing
activate	PREPOSITION activate
expressing	expressing ARTICLE
bound	BE bound PREPOSITION
product	PREPOSITION ARTICLE product PREPOSITION ARTICLE
required	BE required CONJUNCTION
via	603258 marks PRONOUN CONJUNCTION destruction via ARTICLE ubiquitination pathway thereby allowing activation PREPOSITION ARTICLE
telomeric	BE located PREPOSITION ARTICLE telomeric
site	CONJUNCTION ARTICLE site PREPOSITION ARTICLE
codes	codes CONJUNCTION
coded	BE coded PREPOSITION
inhibits	inhibits ARTICLE
catalyzes	catalyzes ARTICLE
induced	induced PREPOSITION
stimulates	CONJUNCTION stimulates
express	not express
chains	chains PREPOSITION
increase	PREPOSITION ARTICLE increase PREPOSITION
tabulated	frequencies PREPOSITION allelic variants BE tabulated PREPOSITION roychoudhury CONJUNCTION
produce	PREPOSITION produce
resulting	resulting PREPOSITION ARTICLE
prime	genes BE tandemly oriented PREPOSITION ARTICLE 5 prime PREPOSITION 3 prime direction PREPOSITION telomere PREPOSITION
deficiency	deficiency CONJUNCTION
hybridize	hybridize PREPOSITION ARTICLE
substitution	substitution PREPOSITION ARTICLE
encodes	RELATIVE encodes
ternary	PREPOSITION ARTICLE ternary
promotes	PRONOUN promotes
synthesized	synthesized PREPOSITION
activity	activity PREPOSITION
not	not respond PREPOSITION
stimulated	BE stimulated PREPOSITION

表 4.9　着目した各単語を含むテンプレートで最高評価値のテンプレート(1)

(Cell と OMIM の場合)

単語	テンプレート
dnas	PREPOSITION dnas PREPOSITION
recognizes	recognizes ARTICLE
mutant	BE mutant PREPOSITION ARTICLE
plays	plays ARTICLE important role PREPOSITION
interacts	CONJUNCTION interacts PREPOSITION ARTICLE
regulate	approximately 8 PREPOSITION 14 kd mostly basic structurally related molecules PRONOUN regulate
release	CONJUNCTION ARTICLE release PREPOSITION
haploid	PREPOSITION ARTICLE haploid
leads	leads PREPOSITION activation PREPOSITION
activation	PREPOSITION activation PREPOSITION
genes	genes BE tandemly oriented PREPOSITION ARTICLE 5 prime PREPOSITION 3 prime direction PREPOSITION telomere PREPOSITION
driven	driven PREPOSITION ARTICLE
cultured	CONJUNCTION cultured skin
due	due PREPOSITION ARTICLE
carries	PRONOUN carries
stimulate	PREPOSITION stimulate
conversion	conversion PREPOSITION
amplification	amplification PREPOSITION genomic
panel	PREPOSITION ARTICLE panel PREPOSITION
regulates	regulates ARTICLE
synthesis	synthesis PREPOSITION
specific	specific probes indicated PRONOUN ARTICLE order PREPOSITION ARTICLE 11
mediates	mediates ARTICLE
resulted	resulted PREPOSITION
shown	BE shown PREPOSITION
converts	PRONOUN converts
induces	induces ARTICLE
suggesting	suggesting PRONOUN ARTICLE
equivalent	equivalent PREPOSITION
concentrations	concentrations CONJUNCTION
resistant	BE resistant PREPOSITION
assigned	hybrids PRONOUN assigned ARTICLE
including	including ARTICLE
levels	levels CONJUNCTION
lies	PRONOUN lies PREPOSITION ARTICLE
induce	BE sufficient PREPOSITION induce
production	PREPOSITION ARTICLE production PREPOSITION
intron	PREPOSITION ARTICLE first intron PREPOSITION ARTICLE
single	BE present PREPOSITION single
inhibit	PREPOSITION inhibit
products	products PREPOSITION
upstream	upstream PREPOSITION ARTICLE
potentiation	potentiation PREPOSITION
promoter	promoter PREPOSITION
converted	AUXILIARY-VERB BE converted PREPOSITION
carry	PRONOUN carry
proximal	PREPOSITION ARTICLE proximal long
phosphorylates	PRONOUN phosphorylates
closely	BE closely linked PREPOSITION
leading	leading PREPOSITION ARTICLE

表 4.10　着目した各単語を含むテンプレートで最高評価値のテンプレート(2)

(Cell と OMIM の場合)

評価値	テンプレート
627	ARTICLE member PREPOSITION ARTICLE
624	BE ARTICLE member PREPOSITION ARTICLE
624	BE ARTICLE member PREPOSITION
616	member PREPOSITION ARTICLE
608	phosphorylation PREPOSITION
607	ARTICLE member PREPOSITION
602	suggesting PRONOUN
593	expression CONJUNCTION
589	expressed PREPOSITION
586	catalyzes ARTICLE
586	BE expressed PREPOSITION
580	encoding ARTICLE
579	amino 3 hydroxy 5 methyl 4
576	binds PREPOSITION
573	interacts PREPOSITION
571	member PREPOSITION
565	activation PREPOSITION
562	PREPOSITION ARTICLE regulation PREPOSITION
557	phosphorylation CONJUNCTION
555	indicating PRONOUN
554	suggesting PRONOUN ARTICLE
554	PREPOSITION ARTICLE activation PREPOSITION
553	expression PREPOSITION
551	transfected PREPOSITION
551	expressed PREPOSITION ARTICLE
549	signaling PREPOSITION
547	BE inhibited PREPOSITION
545	encodes ARTICLE
544	activity CONJUNCTION
544	activation CONJUNCTION
541	expression PREPOSITION ARTICLE
541	BE expressed PREPOSITION ARTICLE
540	cells CONJUNCTION
538	containing ARTICLE
537	inhibited PREPOSITION
537	binds PREPOSITION ARTICLE
536	plays ARTICLE
533	promoter CONJUNCTION
533	phosphorylation PREPOSITION ARTICLE
533	RELATIVE encodes ARTICLE
533	BE blocked PREPOSITION
532	induced activation PREPOSITION
532	induced activation PREPOSITION
531	interacts PREPOSITION ARTICLE
531	expression BE
530	homologue PREPOSITION
528	production PREPOSITION
528	homologue PREPOSITION ARTICLE
528	encoded PREPOSITION
526	Indicating PRONOUN ARTICLE

表 4.11　評価値が上位から 50 番目までのテンプレート(Medline の場合)

評価値	テンプレート
109	enos expression CONJUNCTION
109	cytokine release PREPOSITION
109	cortisol secretion CONJUNCTION
109	corticosterone levels CONJUNCTION
109	collagen induced arthritis
109	cholesterol synthesis PREPOSITION
109	cellular proliferation PREPOSITION
109	cells pbmc CONJUNCTION
109	caspase activity PREPOSITION
109	cardiac expression PREPOSITION
109	calcium levels CONJUNCTION
109	cadherin expression CONJUNCTION
109	c expression PREPOSITION
109	atp release PREPOSITION
109	antigen expression CONJUNCTION
109	anti inflammatory medication
109	alpha levels CONJUNCTION
109	activated receptors ppars
109	CONJUNCTION anti il
109	8 mrna expression
109	2 signaling CONJUNCTION
109	2 secretion CONJUNCTION
109	2 promoter PREPOSITION
108	mutant shows
108	levels normalized
108	anti e
105	dependent efflux
105	anti c.
105	activity pa
101	production remained
101	mutant type
101	induced dilatation
101	induced chemical
101	induce similar
101	cells include
98	residues form
96	plus two
96	containing oligonucleotides
95	homology ph
92	expressing transgenic
82	proteins included
82	proliferation indices
82	activity coefficient
80	stimulated cyclic
80	dependent disease
80	deficient b
80	cultured islets
76	anti tuberculosis
69	mediated il
69	cells nk

表 4.12　評価値が下位から 50 番目までのテンプレート(Medline の場合)

単語	テンプレート
encodes	encodes ARTICLE
deficient	deficient PREPOSITION
mediated	mediated PREPOSITION
binds	binds PREPOSITION
induced	induced activation PREPOSITION
expressing	expressing ARTICLE
encoding	encoding ARTICLE
phosphorylation	phosphorylation PREPOSITION
catalyzes	catalyzes ARTICLE
stimulated	stimulated PREPOSITION
suggesting	suggesting PRONOUN
regulates	regulates ARTICLE
inhibited	BE inhibited PREPOSITION
nick	nick
interacts	interacts PREPOSITION
expression	expression CONJUNCTION
inhibits	inhibits ARTICLE
containing	containing ARTICLE
expressed	expressed PREPOSITION
kinases	kinases CONJUNCTION
activates	activates ARTICLE
bind	bind PREPOSITION
activation	activation PREPOSITION
dodecyl	dodecyl sulphate
express	express ARTICLE
blocked	BE blocked PREPOSITION
induces	induces ARTICLE expression PREPOSITION
encoded	encoded PREPOSITION
plays	plays ARTICLE
via	via activation PREPOSITION
mediates	mediates ARTICLE
stimulates	stimulates ARTICLE
regulate	PREPOSITION regulate ARTICLE
reporter	reporter PREPOSITION
signaling	signaling PREPOSITION
bound	bound PREPOSITION
phosphorylated	phosphorylated PREPOSITION
homologue	homologue PREPOSITION
member	ARTICLE member PREPOSITION ARTICLE
inhibit	inhibit ARTICLE
activate	activate ARTICLE
inhibitors	inhibitors CONJUNCTION
transfected	transfected PREPOSITION
regulated	regulated PREPOSITION
production	production PREPOSITION
antagonist	antagonist CONJUNCTION
indicating	indicating PRONOUN
promoter	promoter CONJUNCTION
overexpressing	overexpressing ARTICLE
activity	activity CONJUNCTION

表 4.13　着目した各単語を含むテンプレートで最高評価値のテンプレート(1)

(Medline の場合)

単語	テンプレート
methyl	amino 3 hydroxy 5 methyl 4
plus	plus anti
receptors	receptors PREPOSITION
regulating	PREPOSITION regulating ARTICLE
downstream	downstream PREPOSITION
synthesis	synthesis CONJUNCTION
induce	induce ARTICLE
release	release PREPOSITION
regulation	PREPOSITION ARTICLE regulation PREPOSITION
subunits	subunits PREPOSITION
contains	contains ARTICLE
proteins	proteins PREPOSITION
secretion	secretion PREPOSITION
dependent	dependent manner CONJUNCTION
levels	levels CONJUNCTION
phosphorylates	phosphorylates ARTICLE
pathway	pathway CONJUNCTION
homolog	homolog PREPOSITION
intracellular	PREPOSITION intracellular stores CONJUNCTION
mrna	mrna CONJUNCTION
cells	cells CONJUNCTION
transduction	transduction pathway PREPOSITION
polyacrylamide	polyacrylamide gels
homology	homology PREPOSITION
glucopyranosyl	d glucopyranosyl 1 6
mutant	mutant CONJUNCTION
stably	stably transfected PREPOSITION
abolished	BE abolished PREPOSITION
anti	ARTICLE anti apoptotic
peroxidation	peroxidation CONJUNCTION
interact	PREPOSITION interact PREPOSITION
suppresses	suppresses ARTICLE
expresses	expresses ARTICLE
residues	residues PREPOSITION
agonist	agonist PREPOSITION
infected	infected PREPOSITION
homologous	homologous PREPOSITION
prevented	prevented ARTICLE
catalyzed	catalyzed PREPOSITION
promotes	promotes ARTICLE
suppressed	suppressed ARTICLE
phosphorylate	phosphorylate ARTICLE
potent	more potent CONJUNCTION
stimulate	stimulate ARTICLE
cultured	cultured PREPOSITION ARTICLE
tagged	tagged PREPOSITION
constitutively	BE constitutively expressed PREPOSITION
proliferation	proliferation CONJUNCTION
lacking	lacking ARTICLE
microg	microg x kg 1

表 4.14　着目した各単語を含むテンプレートで最高評価値のテンプレート(2)

(Medline の場合)

4.5.3 考察

まず，Cell と OMIM のインターバルから抽出した単語で，評価値が上位から 100 番目までの単語を含むテンプレートを評価した結果に関して述べる．表 4.8 は着目した単語を含むテンプレートの中で，評価値が上位から 50 番目までのテンプレートを示している．これらのテンプレートの一部には，専門用語の間の関係を抽出するのに効果的と考えられるものもあった．しかし，適切ではないと考えられるテンプレートも多く見られた．例えば，“frequencies PREPOSITION allelic variants BE tabulated PREPOSITION roychoudhury CONJUNCTION”というテンプレートがあげられる．このテンプレートは，全テキスト中に 125 回出現し，インターバルに 125 回出現していた．同様の例として，“PREPOSITION ARTICLE region PREPOSITION homology PREPOSITION synteny PREPOSITION”や“genes BE tandemly oriented PREPOSITION ARTICLE 5 prime PREPOSITION 3 prime direction PREPOSITION telomere PREPOSITION“などがあげられる．これらに共通する特徴は，テキスト中にペーストされたと思われる記述が多く見られた点である．このような場合には，テンプレートの単語を 1 つずつ増やして評価を行っていくという本研究で用いたアルゴリズムが有効に働かなかったと考えられる．そのため，適切な長さのテンプレートを抽出することができず，ペーストされたと思われる共通した記述部分がテンプレートとして抽出されてしまった．この問題を解決する手段として，インターバルを抽出するテキストの量を増やすことが考えられる．テキストの量を増やすことにより，統計的にペーストされた部分の影響の緩和が期待できるからである．しかし，単純にテキストを増やせばいいとは限らない．なぜなら，生物医学分野であっても領域が異なれば文献に記述される内容が異なる．また，同じ領域であっても年代が異なれば研究対象や焦点が異なることもあり，テンプレートの抽出結果に影響する可能性がある．そのため，データ量を持つ領域がある場合には，その領域に限定したテンプレート抽出の方が効果的なものが発見できる可能性がある．表 4.9 と表 4.10 は，各単語(評価値が上位から 100 番目までの単語)を含むテンプレートの中で一番評価値が高かったテンプレートを示している．こちらにもペースト部分と思われる部分の影響が見られるが，一部を除けば効果的なテンプレートとして利用できる可能性が十分期待できる．

次に，Medline のインターバルから抽出した単語で，評価値が上位から 100 番目までの単語を含むテンプレートを評価した結果に関して述べる．表 4.11 は着目した単語を含むテンプレートの中で，評価値が上位から 50 番目までのテンプレートを示している．これらのテンプレートには，専門用語の間の関係を抽出するのに効果的と考えられるものが多く含まれていた．また，表 4.13 と表 4.14 は，各単語(評価値が上位から 100 番目までの単語)を含むテンプレートの中で一番評価値が高かったテンプレートを示しているが，こちらにも効果的と思われるテンプレートが非常に多く見られた．

これらの結果を見ると，Cell と OMIM の場合に比べて，Medline の方が好ましい結果が出ていると思われる．これは，単語やテンプレート抽出に用いた文献データが影響している．OMIM は遺伝病に関するデータを扱っており，特殊性の強い内容が蓄積されているデータベースである．一方，Medline は生物医学全般の広い分野に関する文献データを大量に扱っている．Medline から抽出したテキストデータの方が，Cell と OMIM から抽出したテキストデータより約 10 倍は量が多い．そのため Medline の方が様々な内容が記述された文章が数多くあり，ペーストなどによる影響は見られず，好ましいテンプレートの抽出結果がでたと考えられる．

4.6 抽出したテンプレートの特異性

この章では，抽出したテンプレートと専門用語の５つカテゴリー(organism，organism_class，protein，compound，gene)を利用して，どのカテゴリーとどのカテゴリーの間に頻繁に出現しているテンプレートかどうかを調べた．例えば，遺伝子名と遺伝子名の間に頻繁に出現しているテンプレートならば，遺伝子名と遺伝子名の間に特異的なテンプレートであるということがわかる．どのカテゴリーの間に特異的なテンプレートであるかを知ることにより，より効果的にテンプレートを利用することができる．

4.6.1　専門用語のカテゴリー化

4.2 節で述べたが，外延的オントロジーから専門用語リストを作成する際に，柳生らにより行われたフィールドのカテゴリー化を参考にした．専門用語は，organism，organism_class，protein，compound，gene のカテゴリーに属するフィールドから抽出した．実際に抽出したフィールドを表 4.15 に示す．

カテゴリー	データベース：フィールド	カテゴリー	データベース：フィールド
organism	brite:ORGANISM		refseq:sub_species
	epd:OS		genbank:variety
	genbank:organism		refseq:variety
	refseq:organism	organism_class	genome:LINEAGE
	genome:NAME		swissprot:OC
	genome:DEFINITION		transfac:OC
	pmd:EXPRESSION-SYSTEM	protein	pmd:PROTEIN
	pmd:SOURCE		prf:NAME
	swissprot:OS		transfac:DE
	transfac:OS		genbank:product
	genbank:specific_host		refseq:product
	refseq:specific_host		enzyme:NAME
	genbank:lab_host	compound	compound:NAME
	refseq:lab_host	gene	genbank:gene
	genbank:sub_species		refseq:gene

表 4.15　専門用語リストに用いたカテゴリーとそのフィールド

例えば，以下のような文章があったとする．

ARTICLE c *met proto oncogene* encodes ARTICLE *receptor tyrosine kinase* met CONJUNCTION HAVE BE shown PREPOSITION play ARTICLE role PREPOSITION oncogenesis.

赤い単語は，専門用語リストにより専門用語と認識している単語である．下線部分はインターバルを表す．この時，専門用語のカテゴリーを利用して以下のような処理をする．矢印の左辺は処理前，右辺は処理後を表す．

met proto oncogene encodes ARTICLE *receptor tyrosine kinase*

→ *gene* encodes ARTICLE *protein*

この処理により，“encodes ARTICLE”というインターバルは遺伝子名とタンパク質名の間の関係を表していることがわかる．このような処理をすることにより，抽出したテンプレートが，どのカテゴリーとどのカテゴリーの間に頻繁に出現しているかを知ることができる．着目した各単語を含む一番評価値が高かったテンプレートに対してこのような処理を行った．

4.6.2　結果

着目した各単語を含む一番評価値が高かったテンプレートに対して，そのテンプレートが，どのカテゴリーとどのカテゴリーの間の関係を表しているかを調べた．この結果，あるカテゴリーの間の関係を特異的に表すテンプレートやいろんなカテゴリーの間の関係を表すテンプレートがあることがわかった．あるカテゴリーの間の関係を特異的に表すテンプレートの例を図 4.9 に，いろんなカテゴリーの間の関係を表すテンプレートの例を図 4.10 に示した．

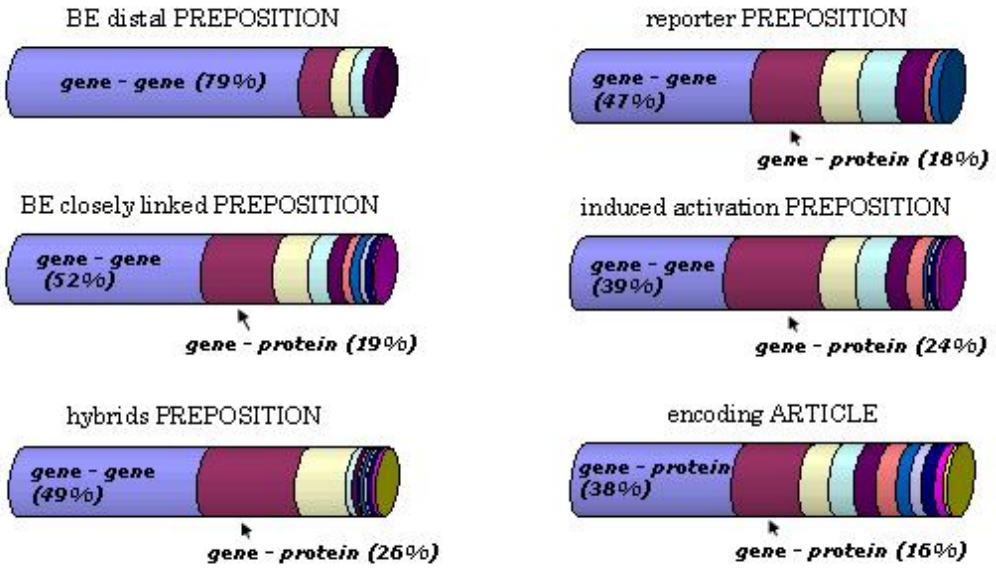

図 4.9　あるカテゴリーの間の関係を特異的に表すテンプレートの例

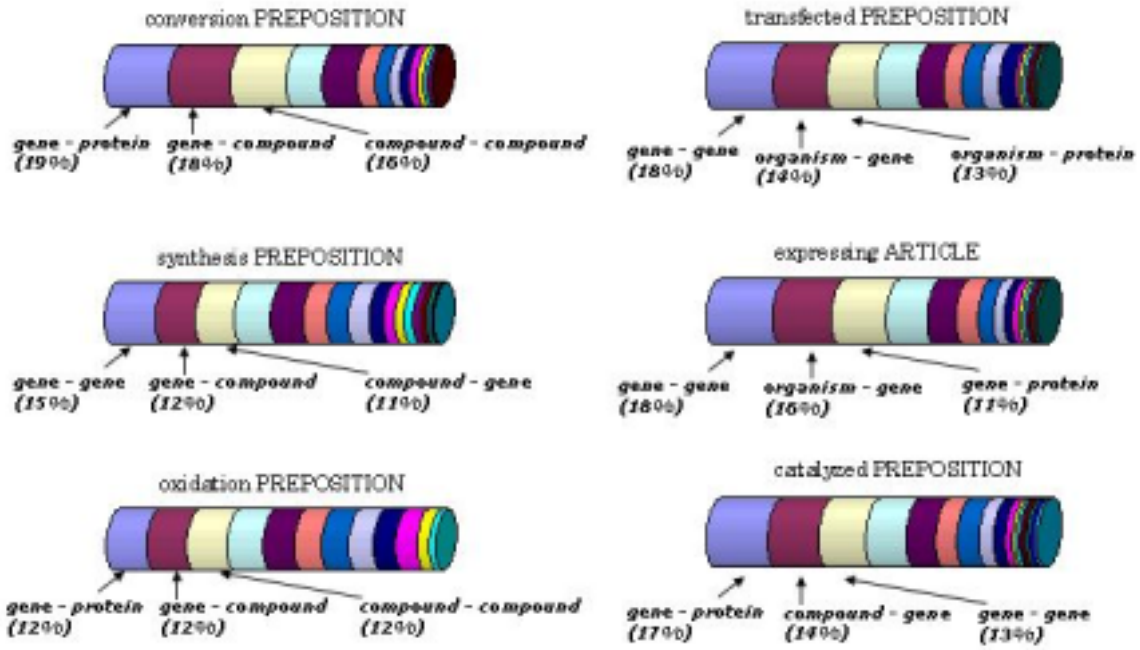

図 4.10　いろんなカテゴリーの間の関係を表すテンプレートの例

第５章

本研究の評価

この章では，関連研究を参考にして，インターバルから抽出した単語やテンプレートの評価を行う．

5.1　関連研究との比較(1)

2.2 節で紹介した Thomas らによる研究と本研究の比較を行った．Thomas らの研究では, タンパク質間相互作用を表す動詞（およびそれに続く前置詞）として, interact (with)，associate (with)，bind (to)の 3 種類のテンプレートを発見している．そして, これらのテンプレートを用いて, タンパク質間相互作用情報の抽出を試みている．

そこで，まずテンプレートの基になっている interact，associate，bind の３つの動詞に着目した．これらの動詞が，本研究でインターバルに出現した単語としてどのような評価結果になったかを調べた(表 5.1)．表 5.1 は，Cell と OMIM のインターバルに出現した単語(32,946 個)と, Medline のインターバルに出現した単語(108,882 個)の中で，評価値が上位から何番目であったかを示す．なお，それぞれの動詞に関係する単語は，上位から 1,000 番以内を示す．

Thomas らが着目した動詞	Cell と OMIM のインターバルから抽出した単語	Medline のインターバルから抽出した単語
interact	interacts (56) interaction (329) interact (589) interactions (603)	interacts (15) interact (81) interaction (294) interacted (378) interacting (460) interactions (561)
associate	associated (167) associates (355) associate (500)	associates (137) associated (372) associate (435)
bind	binds (5) bind (10) binding (719)	binds (4) bind (22) binding (132)

表 5.1　Thomas らが用いた動詞の評価結果 : ( )内の数字は，それぞれのインターバルから抽出した単語の中で評価値が何番目であったかを示す

表 5.1 から，いずれの動詞も関係する単語の評価値が上位に位置することがわかった．また，これらの単語の順位を，インターバルから抽出した単語の評価結果と比べると，Medline を用いた場合の方が単語の評価値の順位が上位に位置することがわかった．したがって，着目すべき単語の抽出については，Medline を用いた場合の方がより成功したといえる．

また本研究では, 評価値が上位から 100 番目までの単語に対してテンプレート抽出を行った．その中に，bind および interact の動詞と関係する単語が含まれていたため, それぞれどのようなテンプレートが抽出できたかを調べた(表 5.2). なお, Thomas らは，これらの動詞が，interact with，associate with，bind to という形で文献に頻繁に出現していることを発見して，これらをテンプレートとしている．

Thomas らの発見したテンプレート	Cell と OMIM のインターバルから抽出したテンプレート	Medline のインターバルから抽出したテンプレート
interact with	CONJUNCTION interacts PREPOSITOIN ARTICLE	interacts PREPOSITION PREPOSITON interact PREPOSITION
bind to	PREPOSITION bind PREPOSITION binds PREPOSITION ARTICLE	bind PREPOSITION binds PREPOSITION

表 5.2　Thomas らの用いたテンプレートとの比較

表 5.2 から，本研究で interact，bind に関係する単語を基に抽出したテンプレートは，いずれもこれらの単語に PREPOSITION が続く形となっており，Thomas らが発見したテンプレートを支持する形であることがわかった．さらに，これらの各テンプレートの動詞に関係する単語に続く PREPOSITION の内訳(抽象化する前の単語)を調べた(図 5.1)．それぞれのテンプレートがインターバルに出現した回数は，Cell と OMIM から抽出したテンプレートの場合は，CONJUNCTION interacts PREPOSITOIN ARTICLE（211 回），PREPOSITION bind PREPOSITION(102 回)，binds PREPOSITION ARTICLE(23 回)であり，Medline から抽出したテンプレートの場合は，interacts PREPOSITION（742 回），PREPOSITON interact PREPOSITION(278 回)，bind PREPOSITION(845 回)，binds PREPOSITION(912 回)であった．

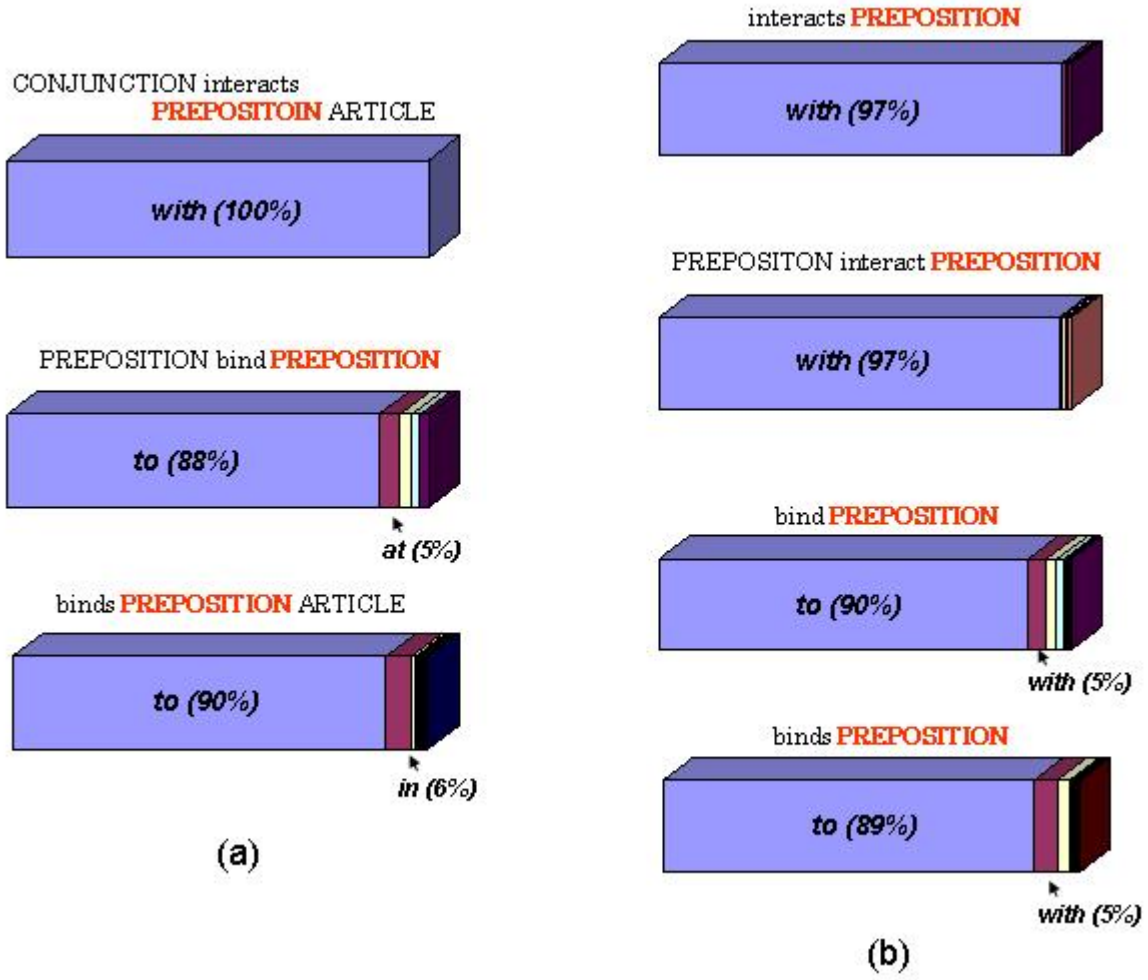

図 5.1　interact と bind に続く PREPOSITION の内訳

PREPOSITION の内訳を調べた結果, 圧倒的に interact(s)に続く単語は with であり，bind(s)に続く単語は to であることが明らかになった．これは Thomas らが発見したテンプレートと一致する結果となった．そして，本研究で抽出したテンプレートは，Thomas らが発見したテンプレートと一致するか，または含んだ形になっていることがわかった．

Thomas らは，タンパク質間相互作用情報を抽出するためにこれらのテンプレートの発見を行っている．そこで，本研究で抽出したテンプレートは，どの専門用語とどの専門用語の関係を表していたかを調べた(図 5.2)

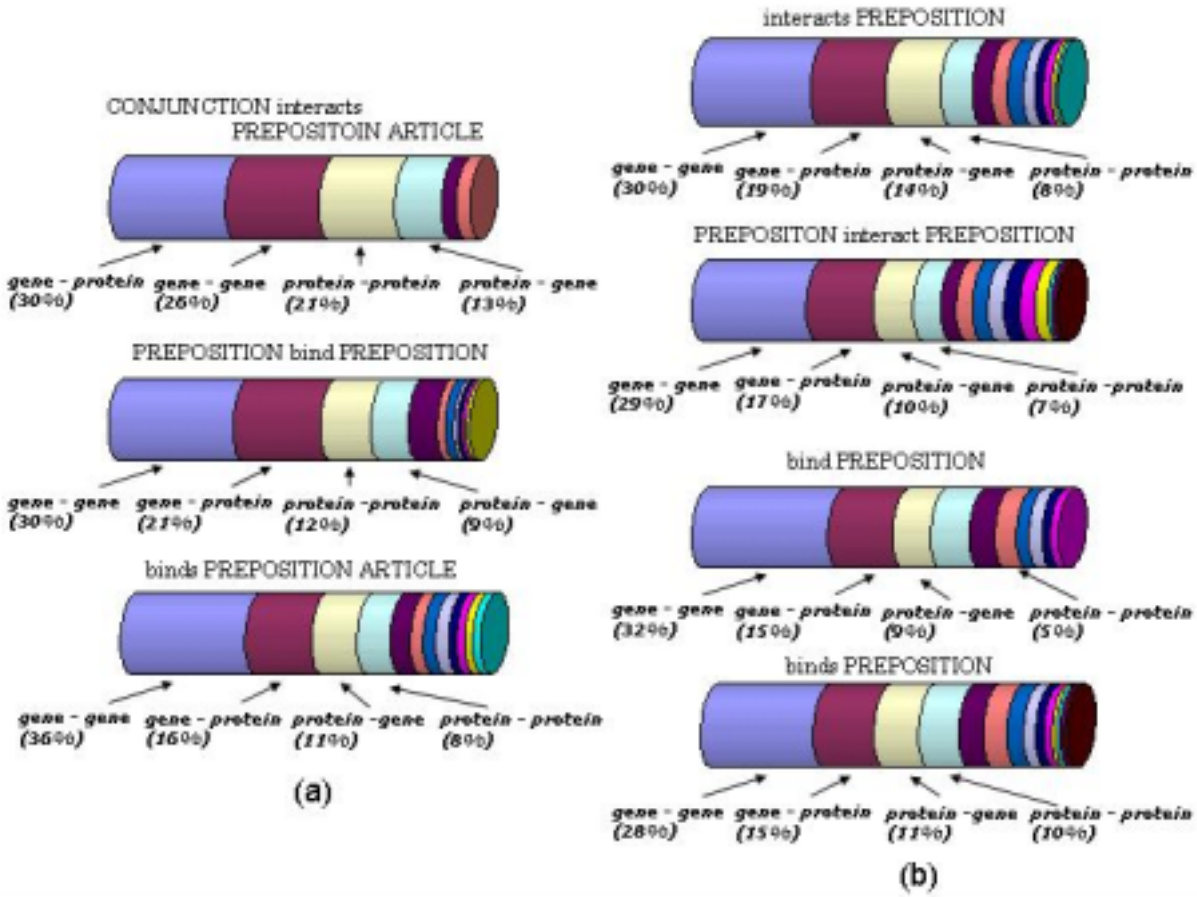

図 5.2　interact と bind を含むテンプレートの特異性：テンプレートが，テキスト中でどのカテゴリーの専門用語の間に出現したか比率を求めた

図 5.2 から，interact と bind を含むテンプレートが際立ってタンパク質とタンパク質の間の関係を表すとは限らないことがわかった．しかし，タンパク質とタンパク質，またはタンパク質と遺伝子の間の関係を含めた比率を見ると，Cell と OMIM から抽出したテンプレートの場合，CONJUNCTION interacts PREPOSITOIN ARTICLE （ 60% ）， PREPOSITION bind PREPOSITION(42%) ， binds PREPOSITION ARTICLE(35%)であり，Medline から抽出したテンプレートの場合は, interacts PREPOSITION（41%）, PREPOSITON interact PREPOSITION(34%), bind PREPOSITION(29%)，binds PREPOSITION(36%)であった．したがって，何らかの形でタンパク質との関係を表す場合が比較的多いといえる．しかし，どのテンプレートにも共通しているのは，遺伝子と遺伝子の間の関係を表す比率が高いということであった．そして，それはこれらのテンプレート以外のテンプレートにも見られ

る特徴であった．これは，そもそも文献の中に記述されている相互作用に関する情報が，遺伝子に関することが多いということを意味している．ある程度の傾向のあるテンプレートは見られたものの，例えば遺伝子やタンパク質というあるカテゴリーに特異的な間の関係を表すテンプレートというものは，ほとんど見られないということがわかった．

5.2　関連研究との比較（２）

2.2 節で紹介した Sekimizu らよる研究との比較を行った. この研究は, activate, bind, interact，regulate，encode，signal，function という動詞に着目して，文献中でのそれら動詞の主語と述語を抽出することにより，遺伝子や遺伝子産物間の相互作用情報の抽出を試みている．そこで，この研究で用いられた動詞が，本研究のインターバルに出現した単語の評価ではどのような結果になったかを調べた(表 5.3)．表 5.3 は，表 5.1 と同様にインターバルに出現した単語の中で評価値が上位から何番目であったかを示す．なお，それぞれの動詞に関係する単語は，上位から 1,000 番以内を示す．

Sekimizu らが着目した動詞	Cell と OMIM のインターバルから抽出した単語	Medline のインターバルから抽出した単語
activate	activates (18) activate (20) activation (61) activators (703) activating (972)	activates (21) activation (23) activates (41) activating (114) activators (483)
bind	表 6.1 参照	
interact	表 6.1 参照	
regulate	regulate (57) regulates (71) regulated (212) regulation (278) regulating (502) regulatory (975)	regulates (12) regulate (33) regulated (44) regulating (54) regulation (59) regulators (222) regulatory (426)
encode	encoded (8) encodes (44) encoding (215) encode (383)	encodes (1) encoding (7) encoded (28) encode (154)
signal	signaling (165) signals (953)	signaling (35) signals (849)
function	function (214) functions (538) functional (694) functionally (834)	function (434) functions (519) functionally(783)

表 5.3　Sekimizu らが用いた動詞の評価結果：( )内の数字は，それぞれのインターバルから抽出した単語の中で評価値が何番目であったかを示す

Sekimizu らが用いた動詞のいずれも，それらに関係する単語の評価値が上位に位置することがわかった．唯一，function の評価値の順位が若干低いのは，動詞として使われると共に名詞としての使われ方も多く見られるため，インターバル以外にも多く記述があり，その結果評価値が低くなってしまったことが理由である．また，Thomas らの研究と同様に，Sekimizu らの関連研究との比較においても，これらの動詞と関係する単語の順位を，インターバルから抽出した単語の評価結果と比べると，Medline を用いた場合の方が単語の評価値の順位がより上位に位置することがわかった．

Medline のテキストは，Cell と OMIM のテキストよりも約 10 倍の量を持っている．しかし，5.1 節と 5.2 節の関連研究との比較結果を見ると，Cell と OMIM のインターバルから抽出した単語の評価結果よりも Medline のインターバルから抽出した単語の評価結果の方が好ましい結果となった．このことから本研究の手法は，日々増加の一途をたどる生物医学分野の文献データに十分に対応できるといえる．したがって，生物医学分野の文献からの動詞やテンプレートの抽出に適切な手法だといえるであろう．

5.3 新たな動詞の発見

5.1 節と 5.2 節の関連研究との比較により，本研究における動詞やテンプレートの抽出手法が妥当であることがわかった．次は，関連研究では使われなかったが，着目するのにふさわしい動詞があるかどうか調べた．関連研究では，人手によって物質間の相互作用を表す動詞の発見が行われている．そこで，本研究では関連研究で用いられた動詞に関係する単語の評価値の結果を参考にして，関連研究では用いられなかったが，評価値の高かった動詞の発見を試みた．これにより，文献データベースからタンパク質間相互作用などの知識を抽出する研究は，いくつかの限られた動詞に着目したテンプレートマッチングが主流であるため，抽出できる情報の量が少ないという問題を解決することができる可能性がある．表 5.4 は，関連研究では用いられなかったが，Cell と OMIM の場合と Medline の場合で共に評価の高かった動詞の例を示す．なお，それぞれの動詞に関係する単語の中で上位から 100 番以内を示す．表 5.5，表 5.6 では，関連研究では用いられなかったが，Cell と OMIM の場合か Medline の場合のどちらか一方で特に評価が高かった動詞の例を示す．Cell と OMIM の場合と Medline の場合の両方で評価の高い動詞の方が，どちらか片方のみで評価が高い動詞よりもふさわしいと考えられるが，必ずしもそうとは言い切れないため，これらの結果も示す．

新たに発見した動詞	Cell と OMIM のインターバルから抽出した単語	Medline のインターバルから抽出した単語
mediate	mediated (16) mediates (74)	mediated (3) mediates (31)
express	expressing (21) express (34)	expressing (6) express (25)
contain	containing (19)	containing (18)
induce	induced (32) induces (78) induce (87)	induced (5) induces (27) induce (57)
catalyze	catalyzes (31)	catalyzes (9)
inhibit	inhibits (32) inhibit (91)	inhibits (17) inhibit (40)
stimulate	stimulated (50) stimulates (33) stimulate (67)	stimulated (10) stimulates (32) stimulate (94)
lack	lacking (13)	lacking (99)
release	release (58)	release(58)
promote	promotes (46) promoter (95)	promotes (90) promoter (48)
culture	cultured (64)	cultured (95)

表 5.4　関連研究にはない評価値の高い動詞の例：( )内の数字は，それぞれのインターバルから抽出した単語の中で評価値が何番目であったかを示す

単語	Cell と OMIM のインターバルから抽出した単語	Medline のインターバルから抽出した単語
carry	carrying (15)	carrying (119)
produce	produced (12) product (23)	produced (159) product (441)
increase	increase (36)	increase (171)
synthesize	synthesized (47)	synthesized (273)
require	required (24)	required (409)

表 5.5　Cell と OMIM の場合に特に評価値の高い動詞の例

単語	Cell と OMIM のインターバルから抽出した単語	Medline のインターバルから抽出した単語
block	blocked (223)	blocked (26)
phosphorylate	phosphorylated (229)	phosphorylated (37)
transfect	transfected (356)	transfected (43)

表 5.6　Medline の場合に特に評価値の高い動詞の例

第６章

本研究のまとめ

6.1　まとめ

本研究では，網羅している専門用語の数が少ないという既存のオントロジーの問題(Gene Ontology で約 1 万語)を本研究室で開発している外延的オントロジー(約 200 万語)の専門用語を対応づけて拡張することを試みた．しかし，専門用語の包含関係を利用するだけでは，外延的オントロジーの約 33 万語の専門用語しか Gene Ontology に対応づけられないことがわかった．また，この方法以外に対応づける方法を模索したが，例えば Gene Ontology には遺伝子やタンパク質名といった具体的な物質名に関するカテゴリーが無いために，単純には対応づけられないことがわかった．したがって，不完全に拡張したオントロジーを利用するよりも，外延的オントロジーから大規模な専門用語リストを作成し，従来は人手で行っていた動詞やテンプレートの発見を，機械的に抽出することを試みた．まず，外延的オントロジーから作成した専門用語リストを，生物医学に関する文献に適用した．そして，同じ文中においてそれらの専門用語に挟まれている部分(インターバル)を抽出した．このようなインターバルには, その前後の専門用語の関係を表すような記述が多く見られる. そこで, インターバルから相互作用を表す動詞やテンプレートの抽出を行った．こうして機械的に抽出した動詞やテンプレートは，専門家によりそれらの発見が行なわれている関連研究を参考に評価したところ良い結果が得られた．したがって，本研究の動詞やテンプレートの抽出手法が妥当な手法であることが証明できた．また，関連研究では用いられていないが，相互作用を表す動詞として期待できる動詞もいくつか発見することができた．さらに，２つの文献(Cell と OMIM，Medline)の文献に対して動詞やテンプレートの抽出を行ったところ，データ量の増加による弊害は見られなかった．よ

って，本研究の手法は，日々増加している生物医学分野の文献データに十分対応できる適切な手法であると考えられる．

6.2　改善すべき問題点

本研究を進めるにあたり，以下のような改良すべき点が見つかった．

6.2.1　専門用語リストのフィルタリング

外延的オントロジーから作成した専門用語リストには，数字のみの用語や 001r や 0.6a といった専門用語として不適切な用語が含まれていた．どのようにして，どこまでのフィルタリングを行うかは難しい問題であるが，１つの方法として Gene Ontology に対応づけられた用語から専門用語リストを作成する方法が考えられる．Gene Ontology は生命科学に関するオントロジーであるから，それに対応づけられた用語は専門用語と見なして問題ないであろう．

また，本研究では，文献から専門用語を認識する際に，専門用語の表記のされ方の違いにより別の用語として扱われる場合を除くための正規化を行った．この処理により回避できた問題は，大文字小文字の違いと特殊記号の違いであった．もう 1 つの問題として語順による違いがあげられる．例えば，

Cu/Zn Superoxide dismutase
Superoxide dismutase (Cu/Zn)
dismutase,Cu/Zn superoxide

という場合があげられる．この語順の問題も解決する必要がある．この問題に対しては，専門用語を単語の集合と考えることにより解決することができる．こうすることにより，上記のような 3 種類の専門用語を 1 つの専門用語として認識することができる．

6.2.2　単語とテンプレートの評価方法

本研究では，インターバルから抽出した単語やテンプレートの評価式に，出現率と出現回数を用いた．これらの評価式は，“インターバルに特異的に出現する単語は，前後の専門用語と何らかの関係をもっている単語である”という仮定の基に成り立っている．そして，関連研究を参考にした評価結果から，この評価式の妥当性をある程度証明できた．しかし，出現率が高くテンプレートとしては適しているが，出現回数が少ないために評価値が低い場合や，テンプレートとして不適当であるが，出現率が高いため評価値が高くなってしまう場合があった．したがって，これらの問題を解決するために評価式を改善する余地がある．そこで，評価式を検討した経緯について解説し，問題点について考察する．

本研究の評価式には，頻度の概念が使われている．一般的に，あらゆる文書の単語の頻度を計算してみると，高頻度語には冠詞や前置詞といった機能語が多く含まれ，その文書の内容を表すような内容語は，機能語の次に頻度が高い単語のグループ(中頻度語)に含まれるという傾向がある．そして，文書の主題が異なれば中頻度語に出現する単語が異なるという傾向がある．この傾向は，本研究で抽出したインターバル内の単語にも見られた．つまり，高頻度語には冠詞や前置詞が多く見られ，その次の中頻度語に専門用語の関係を表すような単語が多く見られた．よって，単語の評価に頻度は重要な要素だと考えられる．

これから紹介するのは情報検索における自動索引語づけを行う際に用いられる方法である．自動索引語づけの基本的な考え方は，索引語づけの対象となる文章に出現する単語の中から，その文書の主題に関連する単語を選択するというものである．中頻度語にその内容を表す内容語が含まれると述べたが，自動索引づけとはこの中頻度を索引語として抽出する方法である．中程度語を索引語として抽出するために，以下のような方法が用いられる．

1．機能語の辞書を用意し，その文書に現れる機能語を削除する．このように，索引語として選択しない単語のことをストップワードと呼ぶ．
2．ストップワード以外のすべての単語に対して，単語T_jの文書D_iにおける頻度tf_j^iを計算する．

３．ある閾値 N を選び，$tf_j^i > N$ を満たす単語 T_j のすべてを文書 D_i の索引語として割り当てる．

このような単純な方法でも，対象となる文書集合に含まれる文書の内容が十分にばらついていれば，検索がある程度うまく働くような索引語を付与することが可能である．しかし，この方法には問題がある．例えば，遺伝子に関する多数の文献からなる文書集合が与えられたとする．これらの文書には，当然のことながら「gene」という単語が多数含まれるため，この方法ではすべての文書において「gene」が索引語として採用されることとなる．しかし，この「gene」という単語で検索したとしても，それは検索結果として出力すべき文書集合をまったく絞り込む力を持っていない．したがって，その文書集合の中で限られた文書にのみ高い頻度で現れる単語を採用する必要がある．つまり，ある単語がその文書集合においてまんべんなく出現する単語なのか，それとも一部の文書のみに出現する単語なのかを調べ，それを索引語の採用基準に反映させる必要がある．

これを実現する１つの方法は文書頻度の逆数を用いる方法である．いま，N 個の文書からなる文書集合が与えられたとする．このとき，ある語単 T_i の文書頻度 df_j とは，その単語が出現する文書数として定義される．

$$\text{文書頻度 } df_j \ = \ \text{単語 } T_i \text{ を含む文書数} \qquad \text{式〔6.1〕}$$

この頻度が小さければ，その単語が検索質問に用いられた場合に該当文書を小さな集合に絞り込むことができるわけであるから，索引語としての望ましさの指標としてはこの文書頻度の逆数を用いればよい．標準的に用いられる指標は $\log(N / df_j)$ である．

この指標と先に用いた指標である tf_j^i を組み合わせると，文書 D_i において単語 T_i を索引語として採用するかどうかを決定する評価値は以下のようになる．

$$w_j^i = tf_j^i \times \log \frac{N}{df_j} \qquad \text{式〔6.2〕}$$

この値は文書 D_i における単語 T_i の頻度 tf_j^i が高く(その文書に何度も出現し)，かつそ

の単語の文書頻度 df_j が低い(特定の文書にしか出現しない)場合に大きな値をとる．この値は，そのままその文書におけるその索引語の重要度を表す重みとして利用することができる．

この索引語を採用する評価式は，抽出した単語やテンプレートの評価式に応用することができる．なぜなら，自動索引づけは，中頻度語を索引語として抽出する方法であり，インターバルに出現する単語で相互作用を表す動詞は，機能語の次の中頻度語に多く見られたからである．本研究で用いた評価式は，自動索引づけを抽出するための方法を参考にして，以下のように導いた．

1．機能語を抽象化することにより，機能語を評価から削除する．
2．ストップワード以外の inteval に出現したすべての単語に対して，単語 T_j のインターバル D_i における頻度 tf_j^i を計算する．
3．索引語の場合は，低頻度の語を省くために閾値を設定している．しかし，この場合はテンプレートとしてふさわしい単語の抽出が目的である．そこで，インターバルに特異的に出現しているか(出現率)を評価式に加えた． N 個の文献集合が与えられたとすると，この時の出現率は， $tf_f^i / N \times 100$ となる．
4．2 の頻度と 3 の出現率を考慮して，実際に単語の評価に以下の式を用いた．

$$w_j^i = \log tf_j^i \times (\frac{tf_j^i}{N} \times 100) \qquad \text{式〔6.3〕}$$

自動索引づけでは，文書頻度の逆数を用いることにより，文書集合の中で限られた文書にのみ高い頻度で出現する単語を抽出している．本研究で考えると，文献中のインターバルに特異的に高い頻度で出現する単語を抽出することにあたる．そのために本研究ではインターバルに特異的に出現しているか(出現率)を評価に加えた．したがって，この評価式により高い評価を得た単語は，インターバルに特異的であり，テンプレートとするのに適した頻度をもつ単語であると考えた．関連研究との比較結果により，この評価式の妥当性を証明することができた．しかし，本研究で用いた評価式では，インターバルに特異的に出現しているためテンプレートとしての効果は期待できる単語であるが，頻度が少ないために評価が低くなってしまうという問題が生じた．そこで，特異的に出現している単語ほど高い重みづけを行う必要があると考えられる．

テンプレートの評価式も，単語の評価式の場合と基本的に同じ考え方である．すなわち，インターバルに特異的に出現して，一定以上の頻度を持つテンプレートの評価が高くなるように設定した．異なる点は，テンプレートの長さをどのように考慮するかという点である．テンプレートとしては，ある程度の長さを持つであることが望ましいであろう．今回は専門用語が n 個の単語からなる場合，単語の評価式に $\log n \times 100$ を加えることにより，長さの重みづけを行っている．テンプレートの長さをいかに評価式に反映するか，また反映しなくてもテンプレートの頻度と出現率で評価できるかどうかも含めて，今後の検討課題としたい．

第 7 章

今後の展望

前章で述べた研究内容の改良点の他に，本研究で取り組めなかった新たな課題を述べる．

7.1　オントロジーの拡張

本研究では，最初に Gene Ontology と外延的オントロジーの対応づけを試みた．しかし，研究を進める上で，単純には対応づけられないことがわかった．今後は，専門用語の包含関係を利用して，外延的オントロジーの約 33 万語の専門用語を Gene Ontology に実際に対応づけることを試みる．また包含関係以外にも専門用語を対応づける方法を模索する必要がある．オントロジーを拡張する理由は，オントロジーを利用したテンプレートの抽出が，非常に魅力的だからである．拡張したオントロジーには，外延的オントロジーの大量の専門用語と Gene Ontology の詳細に階層化された構造の 2 つの性質を利用したテンプレート抽出を行うことができる．したがって，インターバルから抽出したテンプレートを挟んでいる専門用語を，Gene Ontology の階層構造を利用して上位の概念に変換することができるため，専門用語を含めたテンプレートが抽出できるようになる(図 7.1)．既存のテンプレートを用いた情報抽出の再現率は 20%～60%，適合率は 60%～80%ぐらいの性能であるが，拡張したオントロジーから抽出したテンプレートを用いた場合には，再現率と適合率の向上が期待できる．なぜならば，既存の研究より多くのテンプレートが抽出でき，専門用語も考慮したテンプレートであるため確実に物質間の相互作用情報を抽出できるからである．

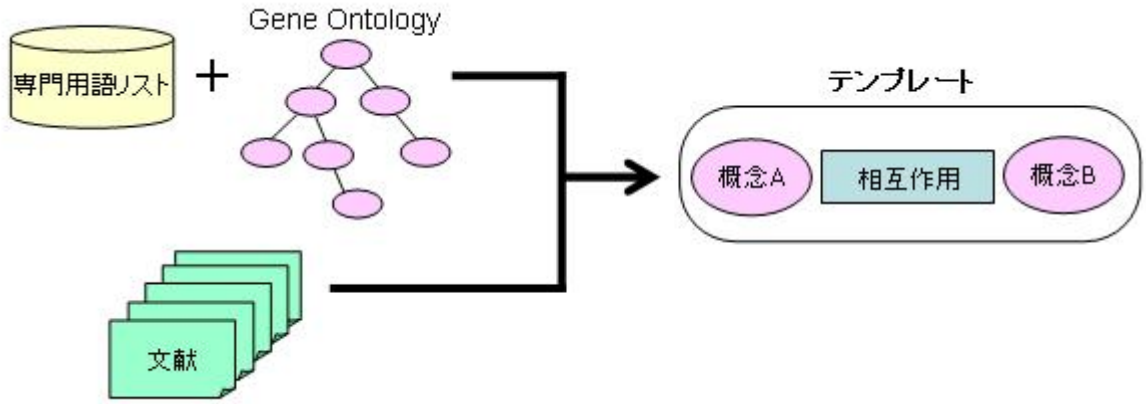

図 7.1　Gene Ontology の階層化を利用したテンプレートの抽出

7.2　ステミング

英語の場合，単語*4 は空白で区切られた文字列と考えられる．形態素*5 は単語を構成するもので，大きく語基と接辞に分類され，接辞はさらに接頭辞と接尾辞に分類される．bind，play，kind などは 1 形態素(語基)で 1 単語であり，playing(play-ing)，smaller(small-er)，unkind(un-kind)，kindly(kind-ly)などはそれぞれ 2 形態素(語基と接辞)で 1 単語である．ステミングとは，この接辞を取り除くことで，同じ概念を指す索引語を作成する技術のことである．例えば，formula，formulate，formulation，reformulate という単語は，formula という元の語義を表す形に変換される．どこまで厳密に変換するかということが問題となるが，本研究の場合は，インターバルに出現した単語の評価をする際に例えば bind，binds，bound，binding のような動詞の変化形は bind に統一して評価する方が望ましいであろう．したがって，実際に発見したテンプレートを用いて情報抽出をする際だけでなく，インターバルから抽出した単語やテンプレートの評価にもステミングは重要である．最もよく知られており，最

---

*4　第 2 章の冒頭に定義したが，ここではさらに詳しく定義する．第 2 章の定義に加えて，1 つの意味のまとまりをなし，文法上 1 つの機能をもつ最小の言語単位．1 つ以上の形態素からなる．

*5　形態素とは，意味をもつ最小の言語単位であり，1 つ以上の音素からなる．音素とは，人間の意味(意志)伝達において音声をどのように使っているかを基に考えた音の単位．

も広く用されているステミングアルゴリズムは, Lovins[13], Paice[14], Porter[15], simple S-removal[16]などがある[17]．しかしながら，今回は，Paice と simple S-removal のアルゴリズムを適用してみたものの，誤った例が多く見られたため採用しなかった．これら多くのステミングアルゴリズムでは言葉の意味を無視して，ルールに基づいた字面処理を行うことでステミングを行っているため，誤りを完全に無くすことができないというのが現状である．字面だけの処理には限界があるため，単語の形態情報に関する辞書を用いた解析(形態素解析)が必要だと考えられる．単語の形態情報とは，発音や品詞，語形変化が含まれる．例えば，EngCG[18]という構文解析を行うシステムがある．このシステムは，Sekimizu らの研究において，遺伝子や遺伝子産物間の相互作用情報を抽出する際に用いられた．このシステムが形態素解析をする際に分解した単語の語基を利用することにより接辞処理を行うことができる．“This suggests that both hemispheres are capable of encoding and retrieval.”という文章をこのシステムに適用した例を図 7.2 に示す.

解析した単語	解析結果
"<*this>"	"this" <*> PRON DEM SG @SUBJ
"<suggests>"	"suggest" <Vcog> <SVO> V PRES SG3 VFIN @+FMAINV
"<that>"	"that" <**CLB> CS @CS
"<both>"	"both" CC @CC
	"both" <Quant> DET PRE PL @QN>
"<hemispheres>"	"hemisphere" N NOM PL @SUBJ
"<are>"	"be" <SV> <SVC/N> <SVC/A> V PRES -SG1,3 VFIN @+FMAINV
"<capable>"	"capable" <DER:ble> A ABS @PCOMPL-S
"<of>"	"of" PREP @<NOM-OF
"<encoding>"	"encode" <SVO> PCP1 @<P-FMAINV
"<and>"	"and" CC @CC
"<retrieval>"	"retrieval" <-Indef> N NOM SG @OBJ
"<$.>"	

図 7.2　EngCG の適用結果：“　”は語基，＜　＞は形態素タグ，@は構文タグを表す

図 7.2 を見ると，suggests は suggest，hemispheres は hemisphere，are は be，encoding は encode の語基からなっていることがわかる．このシステムにより名詞や代名詞の数や格による屈折と動詞の数や人称や時制による活用の変化を語基に変換することができる．生物医学の文献に対してこのシステムを用いて接辞処理を行い，改めて単語やテンプレートの抽出を試みたい．このようにして抽出したテンプレート

を用いて，生物医学文献からの情報抽出を行いその性能を評価したい．

7.3　新たに発見した動詞やテンプレートの評価

関連研究を用いた評価により，既に発見されている動詞やテンプレートが本研究の手法で抽出できることは証明できた．次は，本研究手法により新たに発見した動詞やテンプレートの評価を行う必要がある．一般的には，再現率と適合率という２つの値で評価する手法がある．この手法は，索引語を利用した検索システムの良し悪しの評価をする際に用いられる．再現率とはユーザーの求める文書(検索意図に該当する文書)がどの程度検索されるかということを表す指標で，以下のように定義される．

$$\text{再現率 } R = \frac{\text{検索された文書中の該当文書の数}}{\text{全文書中の該当文書の数}} \qquad \text{式〔7.1〕}$$

一方，適合率とは検索された文書中にユーザーが求める文書がどの程度の割合で存在するかということを表す指標で，以下のように定義される．

$$\text{適合率 } P = \frac{\text{検索された文書中の該当文書の数}}{\text{検索された文書数}} \qquad \text{式〔7.2〕}$$

索引語を利用した検索システムの評価の場合は，2 つの値を両者とも 1 に近づけることが望ましいとされている[19]．本研究で抽出したテンプレートを評価する場合の再現率と適合率を求める式は以下のようになる．

$$\text{再現率 } R = \frac{\text{マッチしたテンプレートが物質間の相互作用を表す数}}{\text{全文献中の物質間の相互作用情報の数}} \qquad \text{式〔7.3〕}$$

$$\text{適合率 } P = \frac{\text{マッチしたテンプレートが物質間の相互作用を表す数}}{\text{マッチしたテンプレート数}} \qquad \text{式〔7.4〕}$$

これらの評価を行うには，事前に文献に物質間の相互作用情報がどれだけ出現しているかを人手で数える必要があるため非常に困難であり，どの文献を対象にするかという選択も難しい問題である．しかし，今後は，本研究の手法により抽出したテンプレートの評価を行いたい．本研究の手法は，動詞やテンプレートの抽出に専門家の知識を必要とすることなく機械的に行えるため，物質間の相互作用情報が飛躍的に抽出できるようになる可能性を秘めている．こうして抽出された相互作用情報は，文献などの情報検索への応用や KEGG*[6] への応用，マイクロアレイ*[7] 実験や実際の実験現場での活用など様々効果が期待できる．また，このオントロジーに基づくテンプレート抽出手法は，生物医学分野だけではなく他の分野への応用も期待できる．情報化が進み，あらゆる膨大なデータが蓄積されつつある現代において，本研究の人手を必要としないテンプレート抽出手法は，有益であり各種の応用が期待される手法であるといえる．

---

*[6] KEGG(遺伝子ゲノム百科辞典)は，遺伝子の調節ネットワークや代謝経路に関するデータベースである．分子生物学や生化学の教科書には広範な知識がまとめられており，それらの中でも代謝系および膜輸送やシグナル伝達や細胞周期などの経路に関することを中心に電子化を行い，コンピュータを使った解析に直接利用することを目指している．

*[7] 遺伝子発現の解析手法．発現量から遺伝子と遺伝子の関係を解明し，遺伝子機能の解明を目指している．

謝辞

本研究を進めるにあたり，終始熱心な御指導を賜りました佐藤 賢二助教授に心から御礼申し上げます．パソコンでは，インターネットやメールができるという認識しかなかった私が，こうして修論を書き上げることができたのは先生のおかげでした．先生から受けた御恩を忘れることなく，将来に出世払いで御返しできるように今後も努力を続けていきたいと思います．

また，終始貴重な御助言を賜りました小長谷 明彦教授に心から御礼申し上げます．また，Xavier Defago 先生や山本 知幸先生，そして，研究室の皆様には，様々な側面から御助力戴き心から御礼申し上げます．

今の私があるのは，すべて両親のおかげでした．これからもすべての事に対して最善を尽くして，両親に恩返しができるように真面目に生きていくことを誓って，結びの言葉とさせていただきます．

参 考 文 献

[1] Fukuda, K., Takuma, T., Tsunoda, T., Takagi, T. "*Toward Information Extraction: Identifying protein names from biological papers*", BIOCOMPUTING, pp.707-718, 1998.

[2] Thomas, C. Rindflesch. "*EDGAR: Extraction of Drugs, Genes And Relations from the Biomedical Literature*" , BIOCOMPUTING, pp.517-528, 2000.

[3] Sekimizu, T., Hyuu S. Park, Tsujii, J. "*Identifying the Interaction between Genes and Gene Products Based on Frequently Seen Verbs in Medline Abstracts*" , Genome Informatics, pp.62-71, 1998.

[4] Thomas, J., Milward, D., Ouzounis, C., Pulman, S., Carroll, M. "*Automatic Extraction of Protein Interactions from Scientific Abstracts*" , BIOCOMPUTING, pp.541-551, 2000.

[5] 高井貴子,高木利久. "*生命科学のためのオントロジー*", 実験医学 Vol.19 No.11, pp.47-53, 2001.

[6] The Gene Ontology Consortium. "*Gene Ontology: tool for the unification of biology*" , nature genegics Vol.25, pp25-29, 2000.

[7] The Gene Ontology Consortium. "*Creating the Gene Ontology Resource: Design and Implementation*", genome research Vol.11, pp.1425-1433, 2001.

[8] The World Wide Web Consortium. "*Resource Description Framework*", http://www.w3.org/RDF/.

[9] "*AmiGO*", http://www.godatabase.org/cgi-bin/go.cgi.

[10] 田上純子,吉川正俊,植村俊克: "*Gene Ontology グラフィカルブラウズシステムの設計と開発* ",情報処理学会論文誌 Vol.0 No.0, pp1-23.

[11] Yagyuu, T., Satou, K. "*Toward Automatic Construction of Extensional Ontology from Genome Databases*", Genome Informatics , pp442-443, 2000.

[12] 高木利久,金子實. "*グノムネットのデータベース利用方法[第 2 版]*", 共立出版, 1998.

[13] Lovins, J.B. "*Development of a stemming algorithm*", Mechanical Translation and Computational Linguistics Vol.11 No.12, pp22-31, 1968.

[14] Paice, Chris D. "*Another Stemmer*", SIGIR Forum Vol.24 No.3, pp56-61, 1990.

[15] Porter. M.F. "*An algorithm for suffix stripping*", Program Vol.14 No.3, pp130-137, 1980.

[16] Harman, D. "*How Effective is Suffixing?*", Journal of the American Society for Information Science Vol.42 No.1, pp7-15, 1991.

[17] Brian, F., Christopher, J.Fox. "*Efficient Stemmer Generation*", Information Processing and Management vol.38 No.4, 547-548, 2002.

[18] "*EngCG*", http://www.lingsoft.fi/cgi-bin/engcg.

[19] 長尾真. "自然言語処理", 岩波出版 ,1996.

研究業績

[1] Satoshi Kamegai, Kenji Satou, Akihiko Konagaya. "*Toward Ontology-based Knowledge Extraction from Biomedical Literature*", Genome Informatics 2003, UNIVERSAL ACADEMY PRESS,INC. TOKYO,JAPAN