計画研究：2000～2004年度

文献からの生物知識の抽出と体系化

●高木　利久[1)]　◆辻井　潤一[2)]　◆高井　貴子[3)]　◆福田　賢一郎[4)]　◆小池　麻子[5)]

1）東京大学大学院新領域創成科学研究科　2）東京大学大学院情報学環　3）東京大学大学院情報理工学系研究科
4）産業技術総合研究所生命情報科学研究センター　5）㈱日立製作所中央研究所ライフサイエンスセンター

〈研究の目的と進め方〉

ゲノム配列はもとより遺伝子発現や分子間相互作用などの大量データを解釈し、生物学的・医学的な意味を付与するためには、これまでにおもに論文の形で蓄えられた、遺伝子やタンパク質の相互作用情報や機能情報を取り出し、データベース化することが不可欠である。生命機能の形式化およびそれに基づくデータベース化は生命のシステム的理解に向けた解析を進める上でも欠かせない。このような観点から、我々は文献に書かれた生命機能に関わる知識をいかに自動的に抽出するか、そして、それをいかに計算機の中に表現し利用するかという課題に取り組んできた。より具体的には、以下の4つの研究課題を設定し、その解決に向けた取り組みを行ってきた。

a）知識抽出システムの開発

文献等のテキストから遺伝子やタンパク質の相互作用に関する知識や生命機能に関する用語を自動的に抽出するための手法を考案し、それに基づいた知識抽出システムを開発すること。

b）生命機能に関する辞書・データベースの構築と公開

上述の知識抽出に使用する専門用語（遺伝子名やファミリー名など）の辞書を整備する。また、知識抽出システムを用いて種々の辞書や相互作用データベースや遺伝子機能データベースなどを構築し公開すること。

c）知識抽出のためのコーパスやオントロジーの整備

上記の知識抽出システム開発のために必要なコーパスやオントロジー等を整備すること。これらは、実際に知識抽出を行ったり、その結果を評価したりする際にも使用する。

d）パスウェイ等の知識の表現法と利用法の開発

生命機能に関する複雑な知識をその本質を損なうことなく計算機上に表現し、利用する技術を開発すること。具体的にはおもにシグナル伝達を対象として、そのためのパスウェイやオントロジーの表現方法、利用方法等を開発すること。

〈研究開始時の研究計画〉

a）知識抽出システムの開発

1. テキスト形式で記述された文献データベースとファクトの形のゲノムデータベースとから、生物知識が記載された文献や文章を漏れなく、かつ、検索者の意図した通りに正確に捜し出すための知的情報検索技術の研究を行なう。
2. テキスト中に出現する重要な専門用語（遺伝子名やタンパク質名あるいは機能用語など）とその意味クラスを認識する技術を開発する。
3. テキスト中で、専門用語で表される対象（概念）同士がどのような関係性を持つ（「結合する」、「活性化する」など）と記述されているかを認識する技術を開発する。
4. 上記2および3などの技術を実装したシステム、すなわち、文献からの知識抽出を支援するシステムを構築する。

b）生命機能に関する辞書・データベースの構築と公開

より高度な知識抽出が可能か否かは、生命科学の専門用語に関するシソーラス（語彙集）や辞書がどの程度整備されているかに大きく依存する。本研究開始当初は、Gene Ontologyなど各種オントロジーも作成され始めたばかりで、活用可能なシソーラスや辞書がほとんどなかったため、自前でそれらの整備を始めることとした。ゲノム研究では当然のことながらほとんどの研究は、遺伝子を中心に行われていることが多いことから、遺伝子名やその同義語、略語など辞書の整備を、また、遺伝子の機能面に着目していることが多いことから、生物学機能用語の整備から着手することとした。これらの準備が整い次第、それらと上記a）で開発の知識抽出システムを用いて、タンパク質間相互作用に関するデータベースや遺伝子とその機能を収集したデータベースなどを構築し公開する。

c）知識抽出のためのコーパスやオントロジーの整備

1. 知識抽出システム開発に向けて、それに必要となるオントロジーの知識表現とその自動構築に関する基礎研究を行なう。生物学オントロジーの基礎研究を基盤にして、オントロジー構築を試みる。
2. 上記オントロジーを基にして専門用語を意味的に分類するためのタグセットを設計し、そのセットに基づいたタグをゲノム科学分野の論文中の専門用語に付加したコーパスを作成する。また、コーパスを作成、管理するシステムの研究開発を行う。

d）パスウェイ等の知識の表現法と利用法の開発

1. おもに真核細胞のシグナル伝達系を対象分野として、それを計算機上に記述するためのオントロジー開発を行う。より具体的には、「シグナル」が伝播するという従来のモデルの限界を打破するために、生物種間で共通な反応のグループを単位としたモデルを提案する。すなわち、(1) 共通なグループと表現型との関係、(2) グループのユニットである化学反応と生化学的性質との関係、(3) 化学反応のユニットである分子とゲノムとの関係、の3層からなる、解析技術とシグナル伝達系の知識ベースをつなぐリファレンスを構築する。
2. タンパク質相互作用データなどは、2つのタンパク質の間の2項関係として表現できる。しかし、シグナル伝達系などのパスウェイは登場する役者（分子）も多くのその関係も種々雑多で複雑である。そこで、これらの複雑な関係を計算機にとっても研究者にとっても分かりやすい形で形式化する必要がある。そのための知識表現法や検索法を開発する。

〈研究期間の成果〉

上記の4つの研究課題の成果のおもなものを以下に示す。

a）知識抽出システムの開発

1. 文書検索によく用いられるブーリアン検索では、単

語の出現関係をAND、OR、NOT、前後関係で指定することができるが、XMLに代表される半構造化文書に対して、構造の関係を指定することは困難である。領域代数は、ブーリアン検索の機能に加え、半構造化テキストに対する構造の指定を可能とする検索のための代数である。既存の領域代数では構造間の関係を指定した質問をシステムに与えた場合、質問と完全にマッチする文書領域のみ返すため、ほとんど解が得られないか、もしくは順序付けされない大量の解が得られるかのどちらかになることが多く、柔軟な検索を行うことが難しい。我々は、質問を部分質問に分解し、各部分質問に対しTFIDFスコア(情報検索でよく使われる用語重み付け手法）を与えることにより、ランク付きXML検索を実現することに成功した〔17〕。

これらの技術と以下の2)、3）で開発した技術とを統合し、2つの物質とその関係の3つ組みを検索するシステムのプロトタイプを開発した。このプロトタイプでは、フル・パーザから出力される統語・依存構造解析結果に対して領域代数による検索を行うことによって、例えば、動詞が activate、主語がproteinA、 目的語がproteinBである文書を探す、という検索ができる。このプロトタイプシステムでは統語解析を行っているため、簡単な依存構造の指定により多様な統語構造を吸収した検索をすることが可能になった。例えばa）受身（"activated"）や動名詞（"activating"）の表現に対してもマッチさせることができるb）より遠い並列構文になっている主語目的語関係や、関係節を介する主語目的語関係の検索ができる、c)関係が否定されている記述を検索対象としないことが可能になっている。

2. テキストから専門用語である部分を切り出すとともに、切り出した用語を後述のGENIAオントロジーと呼ばれるオントロジーに基づいてタンパク質名・遺伝子名・生物種名・組織名・細胞名などの粗い意味クラスに分ける、固有名認識システムを開発した〔10、11、15、16、26、29〕。手法としてはHMM・SVM等の機械学習と用語辞書マッチングを用いた。結果は再現率71.5%、適合率70.2%、両者の幾何平均であるFスコアでは70.8%となった。また、用語とその略称を対応付けるシステム〔2、14、45〕、異なる表記をされている同一の用語(NF kappa B、NF-kBなど）を認識するシステムを作成した〔18、48〕。

また、後述のGENIAコーパスを用いて品詞タガー(品詞を認識するプログラム）の訓練を行い、ゲノム科学分野に適合したタガーを作成した。従来のPennTreebankのみで訓練したタガーでは、PennTreebank（に含まれる新聞記事）に対しては97.05%の精度が得られるが、GENIAコーパス上では85.19%に下がる。これに対して訓練コーパスにGENIAコーパスとPennTreebankの両方を用いると、PennTreebank上で96.89%、GENIAコーパス上で98.20%と両者で高い精度を得られることが分かった。なお、本研究で開発した品詞タガーはWebページ上で公開している。

3.タンパク質間相互作用など生体内でのイベントに関する情報を抽出するためには、動詞の格フレームを自動学習する必要があり、このためにフル･パーザを用いて効率よく構文解析を行う技術を開発した。この技術を用いて、動詞とその主語、述語などの格フレームを抽出するとともに、「主語・目的語などに注目する意味クラスの名詞が出現する動詞は情報抽出の際に重要な役割を持つ（たとえば、主語・目的語にタンパク質名が出てくる動詞は、タンパク質相互反応関係を抽出するために重要な動詞である)」という経験則を用いて、情報抽出の鍵となる動詞を抽出する手法〔25〕、情報抽出パターンを自動学習する手法〔46〕を開発した。これにより、1文で完結している（イベント自身とその中で役割を果たす物質などがすべて1文中に現れる)イベントに関しては、その情報をテキストから抽出することが可能になった。

また、このフル･パーザを用いて、MEDLINE上のアブストラクトの一部（100,000件）を構文解析し、その統語構造をXML形式で自動付与したデータを作成しWebページ上で公開した。今後、MEDLINE全体に規模を広げる予定である〔5、23、34〕。

4.上述の要素技術などを用いて、タンパク質、遺伝子、化合物の間の相互作用抽出システムを開発した。そのシステムにおける処理の流れは以下の通りである。1）遺伝子名を認識しIDに変換（遺伝子はIDで管理)、2）浅い構文解析、3）名詞句を認識するとともに、従属接続節、等位接続節、挿入句などの解析をし、ACTOR(doer of action、動作主）とOBJECT（receiver of action、被動作主）の関係を抽出する。4）ACTORとOBJECTが特定の関係で記述されているときは、2項関係として取り出す。5）使用した関係(特定の動詞や名詞句)により、相互作用の物理的な種類：直接・間接、生物学的な種類：活性化、抑制、輸送、制御、その他の何らかの関係性、などに分類する。なお、このシステムでは、前述のフル・パーザではなく浅い構文解析（シャロウ・パーザ）を使用している。相互作用抽出でのフル・パーザの利用は今後の課題であり、現在それに向けて開発を進めている。

文献中では、相互作用は様々な方法で記述されている。ACTOR とOBJECTは主語と目的語の関係だけに留まらない。"activation of protein-A by protein-B"や"protein-B-induced protein-A"のような名詞句の中で係り受け関係で記述されるものもあれば、"protein-A is a ligand of protein-B"のような、activate、inhibitなどの言葉とは無縁の形で記述されるものもある。また"the expression of protein-A causes the activation of protein-B"のように動詞と目的語中のキーワード（expressionやactivation等）とのペアで相互作用として認識できるものもある。本研究においては、1）特定の動詞の主語、目的語で取り出すもの、2）動詞は特定せず、目的語の中にキーワードを認識して主語と目的語中の遺伝子名を取り出すもの（この場合は、遺伝子名とキーワードの相対位置に制限を課す)、3）名詞句の中で特定のフレーズを構成するものなど、様々なタイプで取り出している。なお、情報抽出にはグレーゾーンが存在する。相互作用といっても、利用する研究者によって、何を取り出したいかは大きく異なるためである。"Protein-A activates protein-B under the expression of protein-C."からprotein-A とprotein-Cの何らかの関係を取り出したいと思うか否かは、その研究目的に依存する。我々のシステムにおいては、幅広い検索要求に答えるために、この類の関係は低い信頼度のマークを付けて抽出している。相互作用は、コーパスにより違いがあるが、50%台の再現率と90%台の精度（正解基準は相互作用の向きと遺伝子を特定することであり、不正解例には遺伝子名などの認識エラーを含む）で抽出している。複数の文にまたがる、もしくは単一文中での照応関係はかなり多く、再現率を大きく下げる要因となって

いる。照応関係の解消は試みているが、精度を下げない効果的な方法は他の分野と同様に見つかっていない。また、浅い構文処理での動詞/過去形と形容詞/過去分詞との間違い、名詞と動詞との間違いも精度・再現率を下げる要因となる。

b）生命機能に関する辞書やデータベースの構築と公開

1. 遺伝子名辞書、ファミリ名辞書、機能用語辞書の構築と公開

文献から遺伝子/タンパク質名を自動抽出する手法は、大まかにはルールを使うもの、確率・機械学習を使うもの、辞書/シソーラスを使う方法とその混合方式がある。遺伝子の命名法は生物種ごとに大きく異なるため、精度を上げるためには、種毎のチューニングが必要である。また、遺伝子は沢山の同義語（シノニム）を有しているため、どの遺伝子かを特定したい、もしくは文献情報と配列情報にリンクさせたい場合、辞書を利用する必要がある。遺伝子名については、各生物種の主要研究機関のデータベースが収集・整理しているが、多くの場合不十分である。従って、主要な真核生物に対し遺伝子名辞書GENAを開発した（図1）。そのカバー率は生物種により違いがあるが90-95%である。GENAは各種データベースからの収集、略語抽出プログラムによる収集、主語、目的語の意味クラスを遺伝子/タンパク質/化合物に限定する動詞（phosphorylate、 methylate等）を利用した収集、生物学者の知識による収集、など様々な手法を使って収集している。残念ながら、収集先のデータベースには相当数の誤った名前を含むため半自動的にそれらを削除しているが、様々な原因から自動的に除去することが困難な名前も存在する。GENAにはその他、様々なサイトから収集した化合物情報も蓄積している。こちらは、MEDLINEのアブストラクトに出現する化合物の約85%程度が収集されている〔44〕。

一方、とかく遺伝子/タンパク質が着目されがちだが、文献中には遺伝子まで特定できない上位の概念（例えば、RasはH-ras、K-rasの上位概念）が記されていることが多い。本研究では、これに対応するために、半自動的に階層構造を持つファミリ名辞書を構築している。その他、遺伝子の機能を自動抽出するために機能用語を収集している〔35〕。

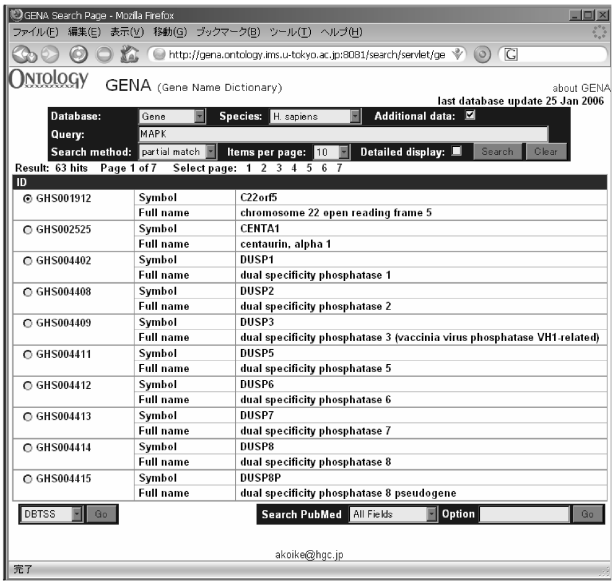


図 1: GENA 検索結果例

2. タンパク質/遺伝子/化合物間相互作用データベースの構築と公開

タンパク質は生体内で他の生体分子と相互作用することによりその機能を発現している。従って、生命現象を解明するためには、タンパク質/遺伝子/化合物のネットワークを解明することが必要である。ここでいうネットワークとは、"ERKがELK1を活性化し、ELK1がc-fosのプロモータ領域に結合して発現させることにより増殖・分化が誘導される"などの一連の作用のことである。これらの相互作用情報は、大規模実験データの場合は各研究室のデータベースで管理されていることが多いが、小規模実験データは文献中に埋もれてしまっているのが現状である。現在では、文献から人手で相互作用情報を抽出し、代謝系データベースやシグナル伝達系データベースが構築されている。これらのデータは詳細な情報まで精度高く整理されている反面、全体としてのコンテンツ量が不十分で最新のデータも不足しがちという短所がある。また、各研究者が発見した相互作用を自ら登録できるデータベースもあるが、なかなか広く認知されない上、相互作用の属性に関しては各研究者が判断するには複雑であることから、提出されるデータの質のばらつきも問題である。そこで我々は、文献（実際には文献のアブストラクト）からタンパク質/遺伝子/化合物間の相互作用データを自動的に収集するシステムを開発した。

2-1. 遺伝子名/タンパク質名/ファミリ名/化合物名の認識

相互作用を見つけるためには、まず遺伝子名やファミリ名を認識する必要がある。この過程の精度が、相互作用抽出に限らず情報抽出の精度を大きく左右するといっても過言ではないが、見かけよりも難しい課題である。遺伝子名認識の問題として、1）遺伝子名の多数のシノニムの存在、2）遺伝子名の表記の揺れ、3）一般動詞、名詞と同じ綴りを持つ遺伝子名の存在、4）多数の遺伝子で共通のシノニムを持つ曖昧性の問題（例えばNIKはMAP3K14とMAP4K4のシノニムである）、が挙げられる。本研究では1）を辞書の構築により、2）を名前の認識方法の工夫により、3)、4）を、名前を認識した後の処理によりできる限り回避している〔38〕。

遺伝子の表記にはかなり揺らぎがあるので、辞書に登録されている名前を完全一致で検索するとカバー率が大きく下がる。ERK-1をERK1と記述するハイフンの有無、mitogen activated ser/thr kinase 1 を mitogen activated kinase 1と記述する単語の挿入/欠如、STE11/STE20をSTE11/20と記述する省略形など様々である。しかしながら、その揺らぎはある程度規格化（特殊文字はスペースに入れ替え等）すれば数個のルールに帰着可能である。本研究では、GENAのエントリをベースにバリエーションを自動生成した後にトライ構造（辞書引きに適した木構造形式の探索アルゴリズム)にして高速に遺伝子名を認識している。このトライ構造もSTE11/20をSTE11/STE20と認識できるように工夫した構造となっている。遺伝子を認識した後、1）フルネームと略語のペアの利用、2）各遺伝子と特徴的に共起する用語(keyword)のチェック、3）浅い構文解析（品詞と係り受けの関係のみ付与)後の名詞のチェックなど多数の後処理を行い上述の3)、4)の問題を解決する。この一連の処理により、生物種によるばらつきがあるが90%程度の精度・再現率で遺伝子名/タンパク質名まで特定できる。また、ファミリ名や化合物名においても同様の処理で認識している。

2-2. 相互作用情報の抽出

文献に記述されている遺伝子/タンパク質/化合物間の相互作用情報の抽出を考えた場合、相互作用には、物理的な相互作用と共に、遺伝子間相互作用がある。遺伝子間相互作用とは、"Protein-A induced the expression of gene-B mRNA"や"Gene-A is synthetically lethal in combination with a deletion of the gene-B"などの、間接的な相互作用である。自動抽出においては、物理的な相互作用もしくはactivate、inhibit（他のタンパク質を介す可能性があるので直接的な相互作用とは限らない）などの明らかな制御関係に着目する傾向にあるが、本研究での抽出は両者を対象にしている。また、タンパク質のプロモータ領域への結合、タンパク質-化合物間の相互作用なども収集対象としている。

相互作用抽出の流れは、前述したように、以下の通りである。1）遺伝子名を認識しIDに変換（遺伝子はIDで管理）、2）浅い構文解析、3）名詞句を認識するとともに、従属接続節、等位接続節、挿入句などの解析をし、ACTOR（doer of action、動作主）とOBJECT（receiver of action、被動作主）の関係を抽出する。4）ACTORとOBJECTが特定の関係で記述されているときは、2項関係として取り出す。5）使用した関係（特定の動詞や名詞句）により、相互作用の物理的な種類：直接・間接、生物学的な種類：活性化、抑制、輸送、制御、その他の何らかの関係性、などに分類する。

現在、主な真核生物に関して約300万の相互作用情報を抽出しPRIME databaseとして公開している〔35〕（図2）。このデータベースでは、配列情報と組み合わせ、種間でのパスウェイ比較や他生物のデータを使用して、ドメイン構成（配列保存領域）や配列類似性などの条件下で対応するパスウェイを描画することが可能である。Viewer上では、カーソルを動かすと対応するオルソログスの遺伝子の色が変化するなど生物種間の対応関係がわかる。また、発現情報と組み合わせて、ある臓器におけるパスウェイの表示など条件付のパスウェイ描画も可能である。エッジとノードは、相互作用の情報抽出に使用した文章と配列情報（辞書情報を含む）にそれぞれリンクしており、エビデンスとなる文章、文献にもユーザーは簡単に辿り着くことができる。

3. 遺伝子/タンパク質の機能データベースの構築と公開

従来の分子生物学では研究者が対象とする遺伝子数は数十個であり、その機能の殆どは研究者の頭の中に整理され得る量であった。しかし、大規模実験法が導入されてから、扱うべき遺伝子数が数百以上となり機能をよく知らない遺伝子を解析する必要性が出てきた。そこで、個々の遺伝子の機能情報を文献から抽出・整理するという技術が注目されつつある。Gene ontology（GO）がよく整理されていることから、その多くはGOベースである。Gene ontologyは、おもに遺伝子配列のアノテーションを目的として作成されたオントロジーで、階層構造（有向非循環グラフ）を持つ。主な真核生物の研究機関において、人手で各遺伝子に該当するGO-IDを関連付けるアノテーションを行っているが、現時点では十分なアノテーションには至っておらず、自動的な機能情報の抽出が着目されている。全アブストラクト/単一アブストラクトを用いた機械学習でも予測は可能であるが、多くの生物学者は証拠となるセンテンスの抽出を望むことから、本研究においてはセンテンスレベルで各遺伝子にGO-IDを自動的に付与する方法を開発した〔35、44〕。

3-1. 機能用語の収集

センテンスレベルでGOベースの遺伝子機能の認識を行うためには、機能を表す用語を認識する必要がある。しかし、GOの語彙（GO-term）は、機能の分類を主な目的として構築された統制語であり、文中で使用される機能用語のカバー率は決して高くない。従って、GOの語彙を基に予め機能用語を収集しておく方法がある。我々は、おもに次の5種類の手法で機能用語を収集している。

1）GO-termとの共起の利用、2）GO-termとのcollocation（連語構成語）の類似度の利用、3）ルールベースでの統語的/意味的バリエーションの生成、4）パターンマッチでの酵素名の収集、5）動詞と専門用語のコンビネーションの作成。その他、UMLS、MeSH、WordNetなどのシソーラスを使い、下位語も追加するなど、様々な工夫を施している。共起ではおもに関連語が収集され、局所文脈の類似度ではおもに類似語が収集される。類似語は、常に同じ意味クラスに属するが、関連語は必ずしも同じ意味クラスには属さない。例えば、"metamorphosis" と "metabolism"とは類似語であり、"chaperon"と"protein folding"とは関連語である。"chaperon"は他のタンパク質の折り畳み（folding）を助けるタンパク質であり、"protein folding"は作用である。"Protein-A helps the protein folding of protein-B."でも"Protein-A is a chaperon of protein-B."と表現してもprotein-Aには同様に、"GO:0006457:protein folding"が付与される。機能は、関連語でも類似語を使っても表現可

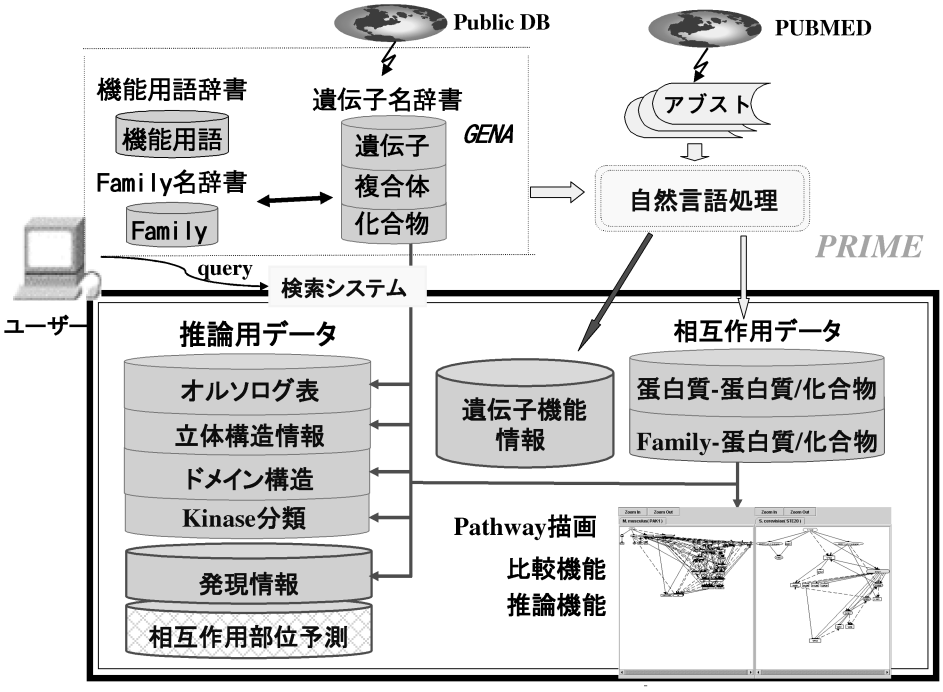


図2: PRIME Database の概要

能であり、この多様性を考慮しないと情報抽出の再現率は向上しない。collocation 類似度の高い名詞句は、あらかじめ、各名詞句のcollocationをベクトル化し、与えられた名詞句のcollocationベクトルとの類似度が高いベクトルを持つ名詞句として収集している。1)、2) の完全自動化はかなり難しいので、候補用語が正しいか否かは最終的に博士課程レベルの生物専攻の学生が判定している。3) においては、"apoptosis<-->apoptotic"、"transport<-->transporter"などの派生語、逆成語との半自動生成と共に、"regulation of osteoblast differentiation" <--> "osteoblast differentiation regulation"などの統語的バリエーションの自動生成を行う。また、単語ごとに収集した関連語/類似語を基に(例えば"metabolism" -> "metabolic、metastasis、 metamorphosis、 reducer、 reduction")その単語を含む用語のバリエーションの自動生成を行うなど、様々な方法で用語を生成/収集している。

3-2. 遺伝子機能の抽出

各遺伝子にGO-ID（Gene ontologyのID）を自動的に付与する過程を示す。機能の抽出の流れは、相互作用と同様以下の通りである。1）遺伝子名、機能用語の認識 2）浅い構文解析 3）文構造解析を行いACTORとOBJECTの関係を抽出 4）ACTORとOBJECT（または動詞のみ、もしくはOBJECTと動詞のペア）が特定の関係で記述されているとき、遺伝子はその機能を有すると判定する。機能の表現は実に様々であり、相互作用の抽出に比べて数段に難しい。基本的は、動詞には殆ど制限を掛けずにACTORが遺伝子、OBJECTが機能になるときに遺伝子機能として抽出しているが、requireなど幾つかの動詞に関しては、その逆を抽出対象としている。相互作用と同様、ACTORとOBJECTの抽出は、主語と目的語だけに留まらず様々である。また、"GO:0006846:acetate transport"（アセテートの輸送、GO:0006846はGO-ID）などの場合は、動詞が、"transport"、 "locate"、 "localize"、"translocate"、 "import"、 "export"のいずれかであり、目的語にacetateもしくはその下位語が存在し、かつ主語に遺伝子名が存在するとき、その遺伝子にこのGO-IDを付与する。動詞とキーワードとのコンビネーションはある程度までは自動的に行っている。その他、palmitoylateなどの動詞の場合は、目的語が何であれ、主語になる遺伝子の機能は自動的に決定する（この場合は"GO:0018318:protein amino acid palmitoylation"）。

ある機能に関与するか否かのグレーゾーンはかなり広い。"Protein A is highly expressed during mitogenesis"（タンパク質Aは有糸分裂中に多く発現している）からprotein Aにmitogenesisを付与するか否かは前後の文脈にも依存する。また、人手によるアノテーションにおいても、キュレータによるアノテーション基準のばらつきは国際的な情報抽出に関する会議などにおいても指摘されている。

相互作用に比べると、照応関係の率も増加し、表現の多様性も広がるので遺伝子機能の自動抽出は非常に難しい課題である。特に、詳細な機能のGO-IDを付与することはかなり困難である。上位の概念のID付与を可とすると精度は90%レベルになるが、この性能評価はかなり複雑になるので、詳細は発表論文を参照して頂きたい。現在では、おもな真核生物に関して36万件（非冗長）程度の遺伝子-機能のアノテーションを行っており、相互作用と同様に公開している。

c）知識抽出のためのコーパスやオントロジーの整備

1. MEDLINEデータベースに登録された論文アブストラクト上に現れる専門用語を意味的に分類し、タンパク質、DNA、 RNA、 細胞種などの32種のカテゴリーに分けた。またそれらのカテゴリーを、Substance（物質）・Source（物質の所在）・Other（その他）をトップノードとする階層構造に整理した〔8、19、47〕。これをGENIAオントロジーとして公開している（図3）。GENIAオントロジーは、次項のGENIAコーパス上での専門用語タグ中で、専門用語を分類するクラスとして用いた。また、教科書等からオントロジーを自動学習する手法の研究も行った〔20〕。

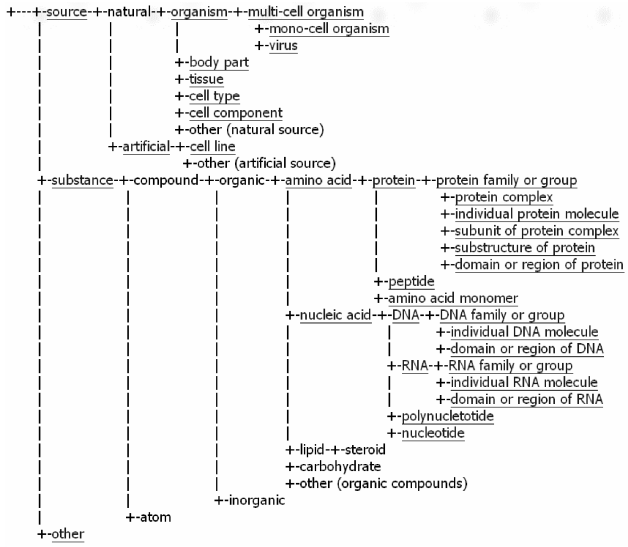


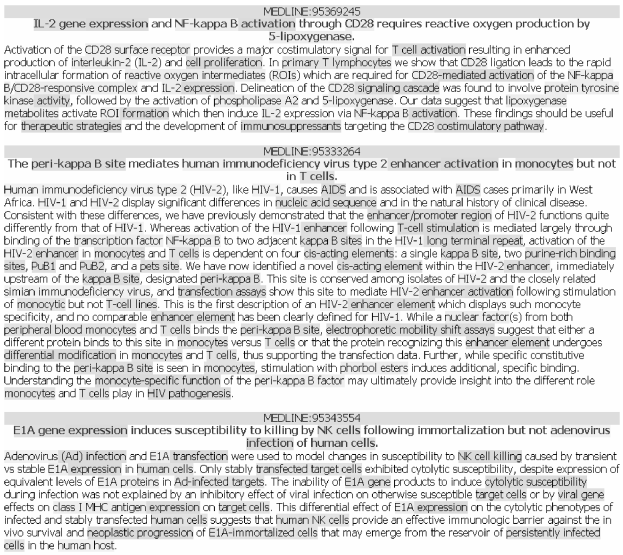
図3: GENIA オントロジー

2. 機械学習ベースの自然言語処理の手法を応用するための学習および検証データとして、MEDLINEデータベース上のアブストラクト2000件に専門用語・品詞をXML形式でタグ付けしたコーパス（GENIAコーパス）を開発した（図4）。対象とするテキストは、MEDLINEデータベースからhuman（ヒト）、transcription factors（転写因子）、blood cells（血球細胞）の3つをキーワードとして検索された結果のアブストラクト群である。GENIA専門用語コーパスは、物質とその所在（ソース）の名の位置を同定するとともに、各々の用語についてタンパク質名・細胞名などのクラス分けを、GENIAオントロジーで定義した用語の意味クラスに基づいて与えた。専門用語には当然多義語も存在し、同一の表現で複数の意味クラスに属する可能性があるが、それぞれの出現箇所では、その文脈に依存して一つの意味クラスを割り当てることが可能であるので、文脈に依存した唯一のクラスを割り当てた。GENIA品詞コーパスでは、自然言語処理分野で広く使われているPennTreebankコーパスのスキーマを用い、既存のシステムの評価・訓練に利用しやすいようにした。作成したコーパスは辻井研究室のWebページ上で一般に公開している〔3、9、18、27〕。

さらに、同一のテキストに付与された複数のXMLタグを管理するシステムTIMSを開発し、公開した〔8〕。このシステムでは、複数のDTDによるXMLタグの範囲の重なり、差分を検索することができ、例えば動名詞句（統語構造タグとして「動名詞句」が割り当てられている）の中にタンパク質名(専門用語タグとして「タンパク質名」が割り当てられている)が含まれる部分を検索するなどの作業が可能になる。

d）パスウェイ等の知識の表現法と利用法の開発

1. 細胞内の機能である、シグナル伝達、遺伝子発現、代謝、輸送、分解、を対象として、これらの機能は、リン酸化等の物理化学的現象のレベルと、それを生物の活動として解釈するレベルの二層に分離できることを明らかにした。各機能の本質的な違いは、解釈のレベルにおいて注目される物理化学的属性の違いにあるので、これを体系化した〔30、31、42、43〕。この体系をオントロジーと呼ぶ。


MEDLINE:95369245
IL-2 gene expression and NF-kappa B activation through CD28 requires reactive oxygen production by 5-lipoxygenase.
Activation of the CD28 surface receptor provides a major costimulatory signal for T cell activation resulting in enhanced production of interleukin-2 (IL-2) and cell proliferation. In primary T lymphocytes we show that CD28 ligation leads to the rapid intracellular formation of reactive oxygen intermediates (ROIs) which are required for CD28-mediated activation of the NF-kappa B/CD28-responsive complex and IL-2 expression. Delineation of the CD28 signaling cascade was found to involve protein tyrosine kinase activity, followed by the activation of phospholipase A2 and 5-lipoxygenase. Our data suggest that lipoxygenase metabolites activate ROI formation which then induce IL-2 expression via NF-kappa B activation. These findings should be useful for therapeutic strategies and the development of immunosuppressants targeting the CD28 costimulatory pathway.

MEDLINE:95333264
The peri-kappa B site mediates human immunodeficiency virus type 2 enhancer activation in monocytes but not in T cells.
Human immunodeficiency virus type 2 (HIV-2), like HIV-1, causes AIDS and is associated with AIDS cases primarily in West Africa. HIV-1 and HIV-2 display significant differences in nucleic acid sequence and in the natural history of clinical disease. Consistent with these differences, we have previously demonstrated that the enhancer/promoter region of HIV-2 functions quite differently from that of HIV-1. Whereas activation of the HIV-1 enhancer following T-cell stimulation is mediated largely through binding of the transcription factor NF-kappa B to two adjacent kappa B sites in the HIV-1 long terminal repeat, activation of the HIV-2 enhancer in monocytes and T cells is dependent on four cis-acting elements: a single kappa B site, two purine-rich binding sites, PuB1 and PuB2, and a pets site. We have now identified a novel cis-acting element within the HIV-2 enhancer, immediately upstream of the kappa B site, designated peri-kappa B. This site is conserved among isolates of HIV-2 and the closely related simian immunodeficiency virus, and transfection assays show this site to mediate HIV-2 enhancer activation following stimulation of monocytic but not T-cell lines. This is the first description of an HIV-2 enhancer element which displays such monocyte specificity, and no comparable enhancer element has been clearly defined for HIV-1. While a nuclear factor(s) from both peripheral blood monocytes and T cells binds the peri-kappa B site, electrophoretic mobility shift assays suggest that either a different protein binds to this site in monocytes versus T cells or that the protein recognizing this enhancer element undergoes differential modification in monocytes and T cells, thus supporting the transfection data. Further, while specific constitutive binding to the peri-kappa B site is seen in monocytes, stimulation with phorbol esters induces additional, specific binding. Understanding the monocyte-specific function of the peri-kappa B factor may ultimately provide insight into the different role monocytes and T cells play in HIV pathogenesis.

MEDLINE:95343554
E1A gene expression induces susceptibility to killing by NK cells following immortalization but not adenovirus infection of human cells.
Adenovirus (Ad) infection and E1A transfection were used to model changes in susceptibility to NK cell killing caused by transient vs stable E1A expression in human cells. Only stably transfected target cells exhibited cytolytic susceptibility, despite expression of equivalent levels of E1A proteins in Ad-infected targets. The inability of E1A gene products to induce cytolytic susceptibility during infection was not explained by an inhibitory effect of viral infection on otherwise susceptible target cells or by viral gene effects on class I MHC antigen expression on target cells. This differential effect of E1A expression on the cytolytic phenotypes of infected and stably transfected human cells suggests that human NK cells provide an effective immunologic barrier against the in vivo survival and neoplastic progression of E1A-immortalized cells that may emerge from the reservoir of persistently infected cells in the human host.

図 4: GENIA コーパス

本オントロジーの特徴は、すべての機能を共通に捉える物理化学的レベルと、生物の活動として別々に解釈するレベルの関係を演繹するルールを与えることである。このルールとは、例えばGene Ontologyのように概念の階層構造を提供するオントロジーと比較した場合、その階層構造が定義される理由に相当する。したがって暗黙的知識がより深いレベルで洗い出されていると言えるものである。

本オントロジーを具体的に表現する媒体としてペトリネットを選び、細胞機能を達成する細胞内ネットワークを対象としたペトリネットモデルの部品化を行った〔41〕。機能を合目的に操作するためには、目的とする機能が部品の組み合わせで再構築できることが必要である。ペトリネットモデルについて、これまで物理化学的レベルにおける部品化は行われていたが、それと機能レベルとの組織的関連付けがなかったため、機能デザインを目的としてモデルを共有し再利用することが困難であった。細胞機能は、複数のネットワークの相互作用により達成されており、かつその相互作用は細胞の環境に応じて動的に変化することが明かとなってきている。その完全な姿をモデル化するためのモデル記述量は、莫大となることが予想される。モデルを部品化し、部品の組立と機能の関係を組織化して、モデル構築作業を計算機支援する技術の開発が必要である。本研究では、その技術開発の基盤となるものである。

2. 各種のグラフ構造で表現された生体内ネットワーク情報、相互作用情報から新しい生物学的洞察を導く計算機手法の開発は大きな期待が寄せられており、さまざまなデータベースが公開されている。特にパスウェイデータベースという観点からはデータベースへの問い合わせ手続きとして、パス検索やグラフの共通部分構造の発見など、構造に基づく研究が試みられてきた。しかし、生物学的に意味のある妥当な構造を推論から導くためには、様々な生物学的な意味づけを制約として推論手続きに与える事が必要であると考えられる。このような機構を実現するために、生命科学のパスウェイデータベース構築に必要となる生命情報の高度な格納、検索、演繹方式、および表現方法の研究に取り組んだ。

パスウェイ表記を実現するためには、パスウェイを階層的に表現できる必要がある。このために複合グラフによるパスウェイ表現方法を開発した〔4, 7〕。複合グラフは入れ子グラフやクラスターグラフと比較して生物学文献中のあいまいな知識を構造化するモデルとしてより適している特徴があり、例えば、パスウェイを構成する部分プロセスを表現しやすい構造になっている（図5, 6）。

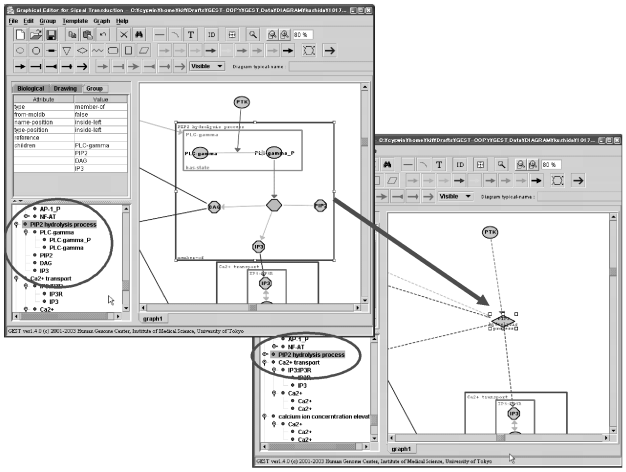


図 5: パスウェイ情報電子化支援ツール GEST。具体的に構造が記述されているプロセスを縮退して、単一物としても扱えるようになっている。

この複合グラフによるパスウェイ表記法を基盤として、(1) 複雑なデータの電子化支援ツール、および (2) 電子化された複雑な構造のデータを検索するためのインターフェイスの研究・開発を行った。

(1) 複雑なデータの電子化支援ツール

生物医学文献に記述されるパスウェイなどの複雑な知識は、抽出作業に非常にコストがかかる。また、従来データベース化されていなかったデータであるため、本質的に表などのデータにしづらいという性質がある。例えば、生体内プロセスへの言及については、生体内プロセスを実現する生化学反応が具体的に記述されている場合と、抽象化して生体内プロセス自体が名詞として使われるケースがある。電子化支援ツールは、高度なGUI (Graphical User Interface) を備えた操作性に優れたツールが必要なだけでなく、内部構造まで記述したプロセスと内部構造を省略・縮退して単体として言及したプロセスをシームレスに繋げてユーザーに提示できる必要がある。また、文献知識を記述するのに過不足ないオブジェクトを用意する必要がある。これらの機能を備えたツールを設計した〔32〕。

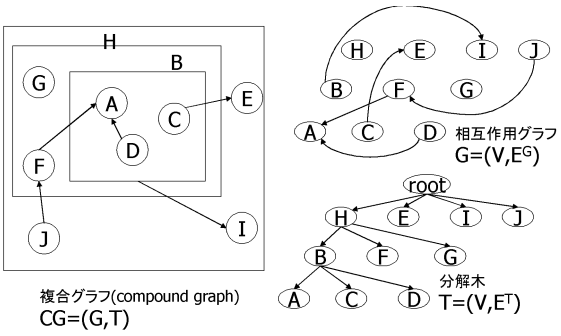


図6：複合グラフによるパスウェイ表現

(2) 複雑な構造のデータを検索インターフェイス

複雑に構造化されたデータは、住所録などの単純な表で表現されたデータと違い、ユーザーへの提示方法、検索手法などが複雑になりやすい。そこで、ユーザーにとって簡便であり、なおかつ、電子化された複雑なデータの構造情報から柔軟に生物学的に意味深い情報を検索できる検索インターフェイスおよび検索サーバーを開発する必要がある。このような観点からFREXを設計した〔21〕。FREXはさまざまなパスウェイを指定された情報に基づき検索・表示し、全体を俯瞰することが可能である。また、タンパク質間相互作用などの情報を統合して表示することも可能である。タンパク質間相互作用のような巨大なデータについても、各タンパク質の性質に基づいてレイアウト位置、表示方法を調整する高速な自動レイアウトシステムを複数提供することでユーザーにわかりやすく情報を提示している。

〈国内外での成果の位置づけ〉

a) 知識抽出システムの開発

生命分野のテキスト処理と知的情報検索の研究を行なうグループが世界的に急増しており、BioCreative、BioLink、BioNLPという生命分野のテキスト処理と情報検索を対象とするワークショップが開かれている。このうち専門用語認識について、2004年度のBioNLPワークショップ（2004年7月、スイス）で、我々の作成したGENIAコーパスを使って開発されたプログラム性能を比較するタスクが設定され、海外から15のグループが参加した。我々のグループはコーパスの作成者であるため、参加者間の条件の公平さという観点から比較自体には参加しなかったが、主催者の一部として評価に加わった。もし参加していればわれわれのグループのシステムは2位の成績を収めることができた。また、国内でも、同年の領域合同班会議においてGENIAコーパスを使った専門用語抽出のコンテストが行われた。

b) 生命機能に関する辞書・データベースの構築と公開

正確な比較はできないが、公開データベースでは、我々の開発したPRIMEは世界で最も多く、文献からの情報抽出データを蓄積していると思われる。欧米の民間企業を含め多くの研究機関から情報抽出したデータの利用申し出があった。データベースへのアクセス件数も増加しており、知名度は着実に上がっている。

c) 知識抽出のためのコーパスやオントロジーの整備

1. GENIAコーパスは、現在、国内外の240以上のグループがダウンロードし、専門用語抽出システムの訓練および評価コーパスとしてデファクトスタンダードの1つとなっている。コロンビア大学のコーエンらは、2005年のBioLinkワークショップにおいて生命科学分野でのコーパスの使用状況を比較し、GENIAコーパスが他のコーパスに比べて有意に多く使用されていることを示し、その理由として生物学的情報（専門用語）と言語学的情報（品詞等）が同一テキストに付与されている統合コーパスであることが大きいと推測している。また、同一のテキストに多種の情報を付与することが有用であることから、世界の他のグループでGENIAコーパスを補完する試みもなされている。Institute for Infocomm Research（シンガポール）におけるMedCoプロジェクトではGENIAコーパスの一部（670アブストラクト）に照応関係（代名詞などの参照関係）をタグ付けし、リクエストベースで公開している。また、スイスのチューリヒ大学では、GENIAコーパスに対して依存文法による文法構造を自動的に付与し、コーパスとしてWeb上に公開している。
2. 2002年2月18日～20日に医学・生物学分野におけるオントロジー構築と自然言語処理に関するワークショップを開催し、8カ国38名が参加した。これが契機となり、この後継ワークショップが2003年に日本、2004年にドイツで開かれた。（参照URL: http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/WS.html）

d) パスウェイ等の知識の表現法と利用法の開発

1. 本研究は、知識工学における、工学機能デザイン支援技術の研究成果を、生物機能の解析に世界で初めて応用したものである。知識工学の成果が、工学知識ドメインにとどまるのではなく、他の知識ドメインにも応用可能であり、かつそのドメインの問題解決に有用であることを示したことから、日本人工知能学会より高い評価を受けた。
2. 複合グラフによるシグナル伝達表現は、世界で始めて階層的なパスウェイ表現を数学的に定式化したものである。その後の多くのパスウェイデータベースは同様な階層構造をもっている。GESTは階層化されたテータ構造を持つパスウェイ・エディタのさきがけとなった。

〈達成できなかったこと、予想外の困難、その理由〉

a) 知識抽出システムの開発

知識抽出の精度について、対象とした分野に関する大規模な正解コーパスが存在しないことから、正確な検証ができなかった。

b) 生命機能に関する辞書・データベースの構築と公開

遺伝子名、ファミリ名については、かなりカバー率が高いが、化合物については、更なる名称収集と、同義語同定が今後必要である。

我々の遺伝子名辞書を作る際に参照した各種データベースのエントリに誤りが多く見受けられ、かつ、それらの情報源が明示されていないことから、正しい遺伝子名か明らかでないものも残った。

c) 知識抽出のためのコーパスやオントロジーの整備

特になし。

d) パスウェイ等の知識の表現法と利用法の開発

特になし。

〈今後の課題〉

a) 知識抽出システムの開発

1. 1つの文では完結しない、前後の文を参照してはじめて抽出できる情報についての情報抽出の研究を行う。そ

のために代名詞の参照、名詞の共参照の関係を抽出する研究を行う。
2. 現在までに開発した、フル・パーザを用いた情報抽出システムの基礎技術を統合し、ユーザーインターフェースを整え、イベント抽出システムとして実用可能なものを作成する。
3. 作成した情報抽出システムと、外部データベース検索、辞書検索との統合を行う。
4. イベント情報抽出について、現在GENIAコーパスが対象としているヒトの情報のみならず他の生物種についても対象を拡張する。また、相互反応以外のイベントについても対象とする範囲を拡張する。
5. 情報抽出の対象となるテキストをアブストラクトからフルペーパーに拡張する。
6. 情報抽出システムについての定量的な検証を行う。

b）生命機能に関する辞書・データベースの構築と公開

疾患名をはじめとする、他の意味クラスの用語の整備、名詞/形容詞の派生語の自動生成技術の拡張など、用語周りの整備はまだ、かなり開発の余地がある。

c）知識抽出のためのコーパスやオントロジーの整備

1. 現在のGENIAオントロジーは物質名と所在を対象としているが、反応、実験などのオントロジーの整備を、現在のOtherカテゴリーの分類と拡張により行う。また、現在のGENIAオントロジーでは、物質の化学的な性質のみに基づいて分類しており、機能に関する分類を行っていない。今後、生体内での役割などの物質の機能に関する概念を定義し、現在タグ付けされている用語に対してさらに属性を追加していく。
2. GENIAオントロジーをGene Ontologyなどの外部オントロジーと連結し、GENIAコーパスに外部オントロジーの情報も付加してGENIAコーパスの情報を豊かにする。また、外部オントロジーのクラスの自動付与についても研究する。
3. GENIAコーパスに文中で生体内の相互作用イベントがどのように表現されているのかを明示的に記述する相互作用イベントコーパスを作成する。また、構文木、述語項構造のような、動詞を中心とした言語構造についてもタグを付与したコーパスを作成する。作成したコーパスは研究用に公開する。
4. XMLで記述されたコーパスのメンテナンスツールを整備し、公開する。開発はEclipse上で行う。Eclipseは広く大学、研究機関、産業界で用いられているオープンアーキテクチャーを採用しており、多数のプラグインが存在しているので、それらを利用することによって開発コストを削減する。我々のグループではXMLタグに対する検索モジュールを開発し、XMLエディタに組み込んで統合エディタを開発する。また、コーパスの変更の履歴と共同作業を管理するヴァージョン管理システムを整備する。

d）パスウェイ等の知識の表現法と利用法の開発

1. 本研究では、細胞内機能を対象とした解析を行い、その統一的表現法の開発を行った。今後は、細胞間および組織間の機能について同様に解析を進め、将来には生体内の機能を統一的に表現する方法の開発までを達成したい。本研究で提案したペトリネットモデルの部品を用いて、実際にペトリネットモデル構築を支援する計算機システムの開発を行う。開発するシステムでは、自然言語で書かれているテキストを入力として、それをペトリネットモデルに変換して出力する機能の開発を目指す。この開発においては、自然言語で書かれた生物機能とリン酸化等の物理化学的現象との、より具体的な関係の事例を組織的に蓄積することを目指す。
2. 複合グラフ表現によるシグナル伝達パスウェイのデータベースのコンテンツの充実を図り、その普及に努める。これは現在ＪＳＴ ＢＩＲＤプロジェクトの一環として進行中である。これと並行してパスウェイの高度な検索機能を開発することを目指す。

〈研究期間の全成果公表リスト〉

1）論文／プロシーディング

1. 201231454
Kato, M., Tsunoda, T. and Takagi, T. "Inferring genetic networks from DNA microarray data by multiple regression analysis" GIW2000, pp. 118-128, 2000.
2. 602211522
Yoshida, M., Fukuda, K. and Takagi T. "PNAD-CSS: a workbench for constructing a protein name abbreviation dictionary" Bioinformatics, 169（2）, 169-175, 2000.
3. 0202261553
Kim, J-D., Ohta T., Tateisi Y., Mima H., and Tsujii J. "XML-Based Linguistic Annotation of Corpus" The First NLP and XML Workshop, pp. 47-53, 2001.
4. 0202261553
Fukuda, K., and Takagi, T. "Knowledge representation of signal transduction pathways" Bioinformatics, 17, pp. 829-837, 2001.
5. 201231146
Ono, T., Hishigaki, H., Tanigami, A., and Takagi, T. "Automated extraction of information on protein interactions from the biological Literature" Bioinformatics, 17, pp. 155-161, 2001.
6. 201231157
Hishigaki, H., Nakai, K., Ono, T., Tanigami, A., and Takagi, T. "Assessment of prediction accuracy of protein function from protein-protein interaction data" Yeast, 18（6）, pp. 523-531, 2001.
7. 201231439
Fukuda, K., and Takagi, T. "Signal transduction pathways and logical inferences" METMBS2001, pp. 297-303, 2001.
8. 0202261603
Mima, H., Ananiadou, S., Nenadic, G., and Tsujii, J. "XML Tag Information Management System -- A Workbench for Ontology-based Knowledge Acquisition and Integration" Proc Human Language Technology Conference, March 2002.
9. 0202261606
Ohta, T., Tateisi, Y., Kim, J-D., Mima, H., and Tsujii, J. "GENIA Corpus: an Annotated Research Abstract Corpus in Molecular Biology Domain" Proc Human Language Technology Conference, March 2002.
10. 0304251813
Kazama, J., Makino, T., Ohta, Y., and Tsujii, J. "Tuning Support Vector Machines for Biomedical Named Entity Recognition" Proc Workshop on Natural Language Processing in the Biomedical Domain（ACL 2002）, pp. 1-8, 2002.
11. 0304251818

Nenadic, G., Mima, H., Spasic, I., Ananiadou, S., and Tsujii, J. "Terminology-driven literature mining and knowledge acquisition in biomedicine" International Journal of Medical Informatics Vol. 67 (1-3), pp. 33-48, 2002.

12. 602211516
Koike, A., Nakai, K., and Takagi, T. "The origin and evolution of eukaryotic protein kinases" Genome Lett, 1 (2), 83-104, 2002.

13. 602211518
Hirakawa, M., Tanaka, T., Hashimoto, Y., Kuroda, M., Takagi, T., and Nakamura, Y. "JSNP: a database of common gene variations in the Japanese population" Nucleic Acids Res, 30, 158-162, 2002.

14. 0404081418
Yu, Z., Tsuruoka, Y. and Tsujii, J. "Automatic Resolution of Ambiguous Abbreviations in Biomedical Texts using Support Vector Machines and One Sense Per Discourse Hypothesis" Proc SIGIR'03 Workshop on Text Analysis and Search for Bioinformatics pp. 57-62, 2003.

15. 0404081433
Tsuruoka, Y. and Tsujii, J. "Boosting Precision and Recall of Dictionary-Based Protein Name Recognition" Proc ACL-03 Workshop on Natural Language Processing in Biomedicine, pp. 41-48, 2003.

16. 0404081439
Kim, J-D., Rim, H-C., and Tsujii, J. "Self-Organizing Markov Models and Their Application to Part-of-Speech Tagging" Proc ACL 2003, pp. 296-302, 2003.

17. 0404081452
Masuda, K., Ninomiya, T., Ohta, T., and Tsujii, J. "A Robust Retrieval Engine for Proximal and Structural Search" Proc HLT-NAACL 2003, pp.50-57, 2003.

18. 0404081459
Tsuruoka, Y. and Tsujii, J. "Probabilistic Term Variant Generator for Biomedical Terms" Proc 26th Annual International ACM SIGIR Conference pp.167-173, 2003.

19. 0404081531
Kim, J-D., Ohta, T., Tateisi, Y., and Tsujii, J. "GENIA corpus - a semantically annotated corpus for bio-textmining" Bioinformatics, Vol 19 (suppl. 1). pp. i180-i182, 2003.

20. 0404081625
Kawasaki, Y., Kazama, J., and Tsujii, J. "Extracting Biomedical Ontology from Textbooks and Article Abstracts" Proc SIGIR'03 Workshop on Text Analysis and Search for Bioinformatics, pp. 44-50, 2003.

21. 0404161517
Fukuda, K-I., Yamagata, Y., and Takagi, T. "FREX: a query interface for biological processes with a hierarchical and recursive structures" In Silico Biology, 4, pp. 63-79, 2003.

22. 0303302237
Poluliakh, N., Takagi, T., and Nakai, K. "Melina: a web server for motif extraction from promoter regions of potentially co-regulated genes" Bioinformatics, 19 (3), pp. 423-424, 2003.

23. 404091851
Koike, A., Kobayashi, Y., and Takagi, T. "Kinase pathway database: an integrated protein-kinase and NLP-based protein-interaction resource" Genome Res., Jun;13 (6A), pp. 1231-43, 2003.

24. 602211514
Yada, T., Totoki, Y., Takaeda, Y., Sakaki, Y. and Takagi, T. "DIGIT: a novel gene finding program by combining gene-finders" Proc. Pacific Sympo. on Biocomputing '03, 8, 375-387, 2003.

25. 0602031138
Yakushiji, A., Tateisi, Y., Miyano, Y., and Tsujii, J. "Finding Anchor Verbs for Biomedical IE Using Predicate-Argument Structures" Companion Volume to the Proceedings of 42st Annual Meeting of the Association for Computational Linguistics, pp. 157-160, 2004.

26. 0602031155
Kim, J-D. and Tsujii, J. "Word Folding: Taking the Snapshot of Words Instead of the Whole" Proc of the IJCNLP 2004, pp. 61-68, 2004.

27. 0602031243
Tateisi, Y. and Tsujii, J. "Part-of-Speech Annotation of Biology Research Abstracts" Proc LREC2004, pp. 1267-1270, 2004.

28. 0602031651
Tsujii, J. "Thesaurus or logical ontology, which do we need for mining text?" Proc LREC2004, IX-XVI, 2004.

29. 0602031712
Tsuruoka, Y. and Tsujii, J. "Improving the Performance of Dictionary-based Approaches in Protein Name Recognition" Journal of Biomedical Informatics 37 (6), pp. 461-470, 2004.

30. 0602061540
Takai-Igarashi, T. and Mizoguchi, R. "Cell signaling networks ontology" In Silico Biol. 4, pp. 81-87, 2004.

31. 0602061543
Takai-Igarashi, T. and Mizoguchi, R. "Ontological integration of data models for cell signaling pathways by defining a factor of causality called 'signal'" Genome Informatics, 15, pp. 255-265, 2004.

32. 0404161458
Fukuda, K-I. and Takagi, T. "A Pathway Editor for Literature-based Knowledge Curation" Conferences in Research and Practice in Information Technology, 29, pp. 339-344, 2004.

33. 404091845
Koike, A. and Takagi, T. "Prediction of protein-protein interaction sites using support vector machines" Protein Engineering Design and Selection, 17 (2), pp. 165-173, 2004.

34. 602211455
Koike, A. and Takagi, T. "PRIME: automatically extracted PRotein Interactions and Molecular Information databasE" In Silico Biology, 5, pp. 9-20, 2004.

35. 602211456
Koike, A., Niwa, Y., and Takagi, T. "Automatic extraction of biological functions using semi-automatically gathered biological terms, Development and Applications of Ontology on OMICS Rearch" The Third Wordshop on Ontology and Genome, 2004.

36. 602211501
Dohkan, S., Koike, A., and Takagi, T. "Prediction of protein-protein interactions using Support Vector Machines" In Proceedings of the Fourth IEEE

Symposium on BioInformatics and BioEngineering（BIBE2004）, 576-584, 2004.
37. 602211502
Imanishi, T., ..., Takagi, T. et al. "Integrative annotation of 21,037 human genes validated by full-length cDNA clones" PLoS BIOLOGY, 2（6）, 0001-0020, 2004.
38. 602211507
Koike, A. and Takagi, T. "Gene/protein/family name recognition in biomedical literature, Linking Biological Literature, Ontologies and Databases: Tools for Users" Workshop in conjunction with NAACL/ HLT 2004, 9-16, 2004.
39. 602211509
Terai, G. and Takagi, T. "Predicting rules on organization of cis-regulatory elements, taking the order of elements into account" Bioinformatics, 20（7）, 1119-1128, 2004.
40. 0602031240
Miyao, Y. and Tsujii, J. "Deep Linguistic Analysis for the Accurate Identification of Predicate-Argument Relations" Proceedings of COLING 2004, pp. 1392-1397, 2004.
41. 0602061533
Takai-Igarashi, T. "Ontology based standardization of Petri net modeling for biological pathways" In Silico Biol. 5, 0047, 2005.
42. 0602061546
Takai-Igarashi, T. and Mizoguchi, R. "Development of an ontology for systems functions of signal transduction" VIIIth European Conference on Artificial Life. UK, 2005.
43. 0602061536
高井貴子, 溝口理一郎. デバイスオントロジーに基づくシグナル伝達パスウェイの統一的記述枠組みの開発. 人工知能学会論文誌, 20, pp. 406-416, 2005.
44. 602211441
Koike, A., Niwa, Y., and Takagi, T. "Automatic extraction of gene/protein biological functions from biomedical text" Bioinformatics, 21（7）, pp. 1227-1236, 2005.
45. 0602082114
Ao, H. and Takagi, T. "ALICE: An Algorithm to Extract Abbreviations from MEDLINE" J. Am. Med. Inform. Assoc., 12, pp. 576-586, 2005.
46. 0602031813
Yakushiji, A., Miyao, Y., Tateisi, Y. and Tsujii, J. "Biomedical Information Extraction with Predicate-Argument Structure Patterns" Proc. 1st Int. Symp. on Semantic Mining in Biomedicine, pp. 60-69, 2005.
47. 0602031816
Tateisi, Y., Yakushiji, A., Ohta, T. and Tsujii, J. "Syntax Annotation for the GENIA corpus" Proc. IJCNLP 2005, pp. 222-227, 2005.
48. 0602031821
Tsuruoka, Y., Ananiadou, S. and Tsujii, J. "A Machine Learning Approach to Acronym Generation" Proc. ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics, pp. 25-31, 2005.
49. 0602061736
Tsujii, J. and Ananiadou, S. "Thesaurus or logical onotology, which do we need for mining text?" Language Resources and Evaluation, 60（1-3）, pp. 77-90, 2005.

2）データベース/ソフトウェア

1. 109
GENIA専門用語コーパス
http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/topics/Corpus/
2. 110
GENIA品詞コーパス
http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/topics/Corpus/pos3.02p.html
3. 112
TIMS（XMLタグ管理システム）
http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/topics/TIMS/
4. 113
GENIA　Tagger（品詞タガー）
http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/
5. 111
MEDLINE自動構文解析サンプルデータ
http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/topics/Parser/parsedMEDLINE.html
6. 304301105
SPARK,シグナル伝達ネットワークデータベース
http://www.ontology.jp/SPARK/
7. 202131625
シグナルオントロジー
http://ontology.ims.u-tokyo.ac.jp
8. 304301124
Kinase Pathway Database
http://kinasedb.ontology.ims.u-tokyo.ac.jp/
9. 304301140
Tissue DB, ヒトの組織データベース
http://tissuedb.ontology.ims.u-tokyo.ac.jp/
10. 0602081602
収集した化合物名称、疾患名、その他の用語データベース
GENA http://gena.ontology.ims.u-tokyo.ac.jp:8081/search
11. 0602141641
MEDLINE全体のテキストをEnjuで解析した結果。リクエスト先メールアドレス：genia@is.s.i-tokyo.ac.jp
12. 0602141645
GENIAコーパスの一部にPenn Treebankタグセットに基づいて構文木構造を付与したもの。500アブストラクト分について公開
http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/topics/Corpus/GTB.html
13. 0602101205
略字とその展開形の自動的抽出システム
ALICE http://uvdb3.hgc.jp/ALICE/ALICE_index.html