シグナル伝達知識のデータベース化：オントロジーとパスウェイ表現

東京大学大学院新領域創成科学研究科
高木　利久

Pathway database: Ontologies and model

Toshihisa Takagi
Graduate School of Frontier Sciences, University of Tokyo, Japan

The target of biological knowledge acquisition has shifted from elucidating the features of genes and proteins to discovering combinations of their interactions that constitute biological functions, i.e. pathways. Pathway data are “processed” rather than “raw” knowledge or data, and are integrated from multiple knowledge sources. Their constituent biological entities are highly diverse and range from metal ions to proteins to biological processes in general. To provide machine-accessible pathway knowledge, we have developed a manually curated pathway database that focuses on biological processes at various levels and also a set of annotation ontologies.

1. はじめに

ゲノムにコードされた生命のメカニズムを解読するには、これまで蓄積されてきた膨大であるがバラバラな生物知識を整理統合し、計算機が理解し解析できるような形に整備すること、すなわち、生物知識の枠組みの体系化（オントロジー、辞書作成）とデータベース化が必要である。

しかしながら、シグナル伝達パスウエイを生体内分子のネットワークとして厳密に記述する（表現する）ことは簡単な作業ではない。分子間の相互作用から発展するパスウエイは、化学反応式を用いて普遍的な記述が可能なメタボリックパスウエイを除き、これまでデータベース化が遅れてきた。その理由として、パスウエイを構成する分子が、タンパク質、化合物、金属イオンなど様々であり、さらにこれら分子間の関係として、結合、酵素反応における酵素と基質、基質と生成物の関係、反応に対するアクチベーターやインヒビターの関係、conformation change などの分子の状態変化、核内輸送などの空間的移動など多種多様な概念が存在することがあげられる。

また、生物学者が文献を読んで知識を得る際には、文中に明示されないさまざまな背景知識（例えば、ERK1は MAPK の一つである、など）を用いて文脈を解釈している。このような背景知識を無視して単純に文献中の記述を電子化すると知識の断片化（ERK1と MAPK の間の関係を計算機が理解できない状況）がおきてしまう．文献中の分子機序に関する知識を電子化するためには、計算機に背景知識を持たせること、およびこれらの電子化された背景知識をパスウェイデータにきちんとアノテーションする新しい知識処理技術の枠組みを用意する必要がある。

本研究ではこのような動機のもとに、生命科学の中でもその要であるシグナル伝達系を主とする細胞機能の分子機序についての生命科学知識を扱う情報処理技術の研究開発を目指し、SPARK、 FREX、 INOH[1]の順に一連の知識ベースの開発に取り組んできた。

細胞機能の分子機序を説明した知識であるシグナル伝達パスウェイのオントロジー構築は、シグナル

伝達系研究全般の推進に貢献することが期待される。また、シグナル伝達系オントロジーの外延となるパスウェイデータベースの開発は、ゲノムからの機能予測、細胞機能のコンピュータシミュレーション、医薬品の開発に必要な情報源であり、その開発は世界中から期待されている。

2. 研究開発の成果

2.1 パスウェイ表現

シグナル伝達パスウェイは細胞の外界との応答を担う系であり、主にタンパク質間の相互作用が実現する生命現象の分子機序を記述している。しかし、伝達される“シグナル”に関する明確な定義はなく、むしろ様々なレベルでのタンパク質の機能を包含する用語として “シグナル”という用語が使用されている。つまり、実際には多種多様な役者、反応の集合から構成される知識を扱う必要がある. 具体的には金属イオン、低分子化合物、タンパク質などの物質にとどまらず、細胞応答などの現象の因果関係まで記述する必要がある. また、これらの事柄を結びつける関係(出来事)もタンパク質の“輸送”、“修飾”、出来事の“制御”、化合物の“生化学反応”など様々である。

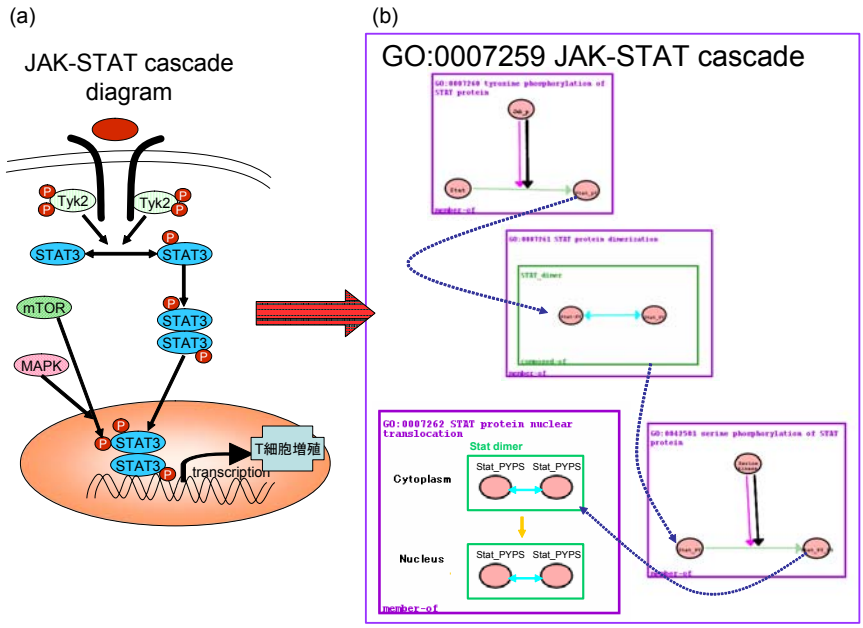

図1. (a)典型的なシグナル伝達パスウェイ知識の記述. (b)知識を明示化して記述したパスウェイ(細胞質内のみ). (a)細胞膜上の受容体に刺激が結合することでJAK, STATを主体とするメカニズムが応答を核内に伝達し、遺伝子の転写を引き起こしている.

図1(a)はダイアグラム図で記述された典型的なシグナル伝達パスウェイ知識である。細胞膜は二本の実線で表されており、大きな楕円は核、中くらいの楕円はタンパク質である。これらのオブジェクトを直感的に配置することでJAKタンパク質、STATタンパク質を主体とする細胞応答のメカニズムが表現されている。

このような知識表現を計算可能な形式に置き換えるには3つの問題がある. 第一は明示的でない部分構造やサブプロセスへの言及である。生物学者はタンパク質のリン酸化修飾が機能に重要な役割を果たしていることを知っているので、中くらいの大きさの各楕円に付随している小さな楕円がタンパク質の修飾

状態を表していることをすぐに理解する。そのため図1. の流れ図をみると即座に、リン酸化された STAT3 が二量体を形成して核内に移行し転写を活性化することを理解する。つまり、STAT3だけで4つのプロセスが発生していること、および二量体としての STAT3の存在が明示されていない。第二は構成要素の不均一な記述粒度である。前述の例では、リン酸基、タンパク質、細胞増殖などがそれぞれ独立のオブジェクトで表現されている。STAT3と修飾された STAT3の間の矢印はタンパク質間の状態遷移を表しているが、mTOR からの矢印はリン酸基を指している。また、核酸の二重螺旋が細胞応答という現象と矢印で結合している。このようにダイアグラム図では、様々な概念に属する役者が異種混交と登場するため異なるタイプ、異なる粒度の要素がお互いに相互作用しあう記述になっている。第三の問題は知識の不完全さである。ダイアグラム図では関連するオブジェクトを近くに配置することで、もしくはターゲットの曖昧な矢印を導入することで、実際には相互作用の詳細が不明なオブジェクトどうしを直感的に結びつけている。このような知識を単純にグラフ構造で表現することは困難である。

これらの問題点を解決するためには、様々な粒度の記述に対応できる階層的な記述が必要である。また、様々の異なる概念に属する要素を扱うためにはオントロジーによる意味づけが必要となる。

開発システムでは複合グラフに基づくパスウェイ表現を採用している[4]複合グラフはグラフを拡張した構造をもち以下で定義される。複合グラフ $CG = (G,T)$ はグラフ $G=(V,E^G)$ と根付き木 $T=(V,E^T,r)$ で定義される。r は木の根である。グラフ G を相互作用グラフ、木 T を分解木と呼ぶ。同様に、エッジ $e_I^G \in E^G$ を相互作用エッジと呼び、エッジ $e_I^T \in E^T$ を分解エッジと呼ぶ。CG の断片 $Frag(a)$ は分解木 Tの内点 aを根とする部分分解木 T'のノード集合から導かれる部分複合グラフである。図2は複合グラフの例である。複合グラフでは相互作用エッジの両端は分解木の葉に限定されず、内点も相互作用の始点もしくは終点となりうる。このため、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(図1. (b))。生体内プロセスの機械処理という観点からは以下の利点がある。分解木の各内点は複合グラフの部分構造を定義しており、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。また逆に、新しい根となるノードを導入することで、階層的な部分構造の情報を保持したまま複数の複合グラフを一つに結合することが可能である。

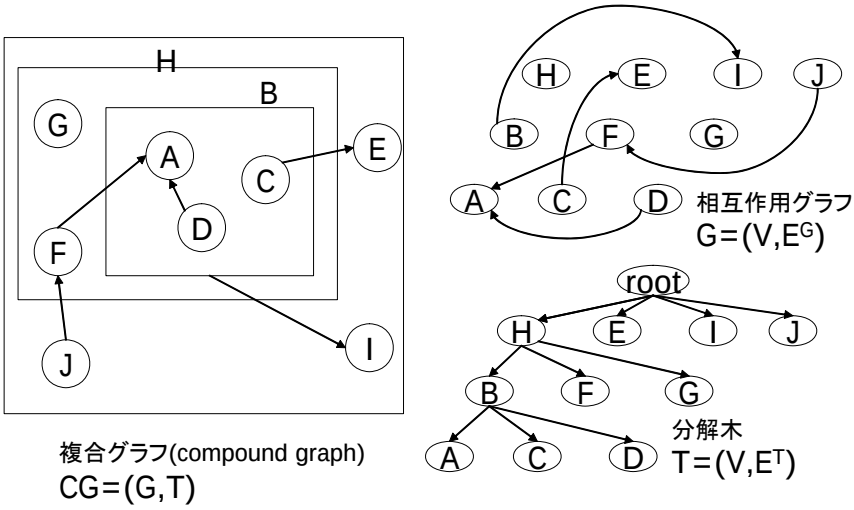

図2. 複合グラフに基づくパスウェイ表現

2.2 シグナル伝達オントロジー

シグナル伝達系は、多細胞生物における、発生、分化、成長、運動、日周期の調整、生体異物応答、ストレス応答といった、さまざまな生体作用を制御するシステムである。このように多くの構成要素が複雑に関わりあって構成されるパスウェイを表現するためには各構成要素の意味をきちんとアノテーションしなくてはならない。本研究開発では後述のとおり分子機能、細胞機能、生物種からフェノタイプに至るまでの概念体系について用語の収集ならびにオントロジーとしての体系化を進めた。特に、タンパク質オントロジー（MoleculeRole Ontology[2]）および生命現象オントロジー（Event Ontology[3]）については Gene Ontology Consortium が管理している Open Bio-medical Ontologies (OBO)として OBO サイトから公開している。また、相互作用オントロジー（Process Ontology）、細胞局在部位オントロジー（Location Ontology）を含めた4つのオントロジーを我々の開発した Ontology Viewer より公開している. 以下、タンパク質オントロジーおよび生命現象オントロジーについて説明する。

生命科学論文に登場する分子名は「ヒトの SMAD1」、「マウスの ERK1」、のようにただ一つの配列を特定できるようなものだけでなく、「R-SMAD」、「MAPK」、のように抽象的で、はっきりと一つの分子を指さないものも多数含まれる。特に総説などでは、このような分子名でパスウェイや相互作用が描かれる（Canonical pathway）。文献知識のデータベースでは、個々の分子について記述される相互作用・パスウェイと、このような Canonical なパスウェイの双方をデータベースに格納し、両者をまたがった柔軟な検索をできるようにすることが望ましい。

生物学者は、「SMAD1は R-SMAD の一つであり、ERK1は MAPK の一つである」、という関係を知っているが、論文中の分子名をそのまま計算機に移すだけでは ERK1は ERK1、MAPK は MAPK としてのみ扱われてしまう。このため、ERK1と MAPK の間の関係が失われ、知識が断片化してしまう。また、一方で文献知識のデータベースは、すでに広く利用されている配列データベースと連携を保ったものでなければならない。しかしながら、論文に登場する分子名ひとつひとつに対し、指す配列実体を正しく当てはめていくことは特に抽象的に書かれる分子については自動化が難しい作業となる。

これらの問題を解決するため、我々は Canonical な分子名（R-SMAD、 MAPK など）から具体的な分子名（SMAD1、 ERK1など）、配列データベースのエントリーまでをつなぐオントロジーMoleculeRole Ontology の構築にとりくんだ。具体的には、文献中におけるタンパク質の4種類の言及方法：1）機能のみの曖昧な言及（受容体、キナーゼなど）、2）通称名、ファミリー名など配列情報を特定できない程度に曖昧な言及（MAPK など）、3）配列情報まで特定できる具体的な言及、4）複合体の構成要素まで特定した言及を整理し階層的な分類体系を構築した。1）、2）、3）の知識を汎化－特殊化関係(is-a)、そして、これらと4）を部分－全体関係(part-of)で結んでいる。MoleculeRole Ontology 構築のために、分子生物学における代表的な総説記事・原著論文を読みながら、分子グループに関する知識を収集し、階層的な関係に整理した. 階層化に当たっては、文献内の記述、辞典、教科書、SwissProt などのアノテーションデータベースの記述などを参考にし、生物学者にとって一般的な、受け入れやすい階層になるよう十分注意をした。ここで我々のオントロジーが、配列に基づいたファミリーの分類とは必ずしも一致せず、分子間相互作用やシグナル伝達経路における分子の役割に基づいた階層となっていることに注意されたい。オントロジーの最下層にあるエントリーについて SwissProt/TrEMBL エントリーへのリンク付けを手作業で行った。その際、必要であればさらに細かい分類を行い、階層を整えた。分子複合体も階層に加え、複

合体を表す概念と、その要素を表す概念との間を相互リンクでつないだ。

生命現象オントロジーが扱うべき概念は、まず実験培地の栄養状況変化などの環境プロセスを扱うカテゴリーと生体内のプロセスを扱うカテゴリーに代別される。さらに、生体内プロセスは、1）分子レベル、2）細胞レベル、3）組織レベル、4）生理学レベルのプロセスに大別される。分子機序の説明となるパスウェイ、そして任意の部分パスウェイは1）の分子レベルのプロセスと汎化－特殊化関係で結ばれ、また、どのような生命現象に関わるのかによって、2）3）4）と部分－全体関係でつながっている。

シグナル伝達の構成要素をこのオントロジーでアノテーションすることによって、さまざまな粒度で記述された文献内の知識をもれなく電子化することができるようになる．各種のデータ･アノテーションを支援するためにこれらのオントロジーを我々の開発した Ontology Viewer（図3）から公開している。Ontology Viewer を用いることで各種の検索･データ閲覧が可能である。

図3．Ontology Viewer

これらの各オントロジーは SwissProt、KEGG、GeneOntology、PSI-MI などの既存データ、オントロジーへの参照リンクを持っている。

2.3 パスウェイデータベース

複雑に構造化された知識であるパスウェイデータを高い精度で文献から抽出するためには、専門家が実際に論文を読んで知識を抽出する必要がある。しかしながら、厳密に階層化されたオントロジーもしくは統制語句で構成要素の意味をアノテーションしている公共のパスウェイデータベースの事例は稀である。我々は前述の複合グラフによる表現にオントロジーによるアノテーションを組み合わせた高付加価値なデータを文献から抽出しデータベースを構築した。

パスウェイを構成する各ノードおよびエッジはタイプ（例えば、タンパク質ノード、プロセスノード、輸送エッジ、制御エッジ）に従って、それぞれ決められた属性集合を持っている。タンパク質ノードであれば表示名、分子情報、局在部位情報、組織情報などの属性を持っている．従来のデータベースではこれらの属性情報に専門家が自由形式で情報を入力し、結果的にデータの一貫性や統合化の際に大きな問題になっていた。我々のデータベースでは属性情報をオントロジーによってアノテーションすることでこの問題を解決している。例えば、分子情報には該当タンパク質の存在を規定している MoleculeRole Ontology への参照情報が、同様に局在部位情報や組織情報にはそれぞれの概念を規定するオントロジーへの参照情報が格納されている。

このようにオントロジーを強く意識したパスウェイ表現手法を採用することで、ユーザはパスウェイおよびパスウェイを構成する全てのオブジェクトに関して、様々な属性の組み合わせを条件に指定して検索を実行できる。とくに、本手法では複合グラフを用いることで部分パスウェイの検索も可能となっている（図4）。タンパク質 A と相互作用する分子は何か、という問い合わせは、A と相互作用グラフで結合されたノードを検索することに相当する(図2．ノード D、C)。タンパク質 A を含む生体内プロセスは何か、という問い合わせは A を含む上位構造のノードに対してアノテーションされたオントロジーのクラスを検索することに相当する(図2．ノードB)。

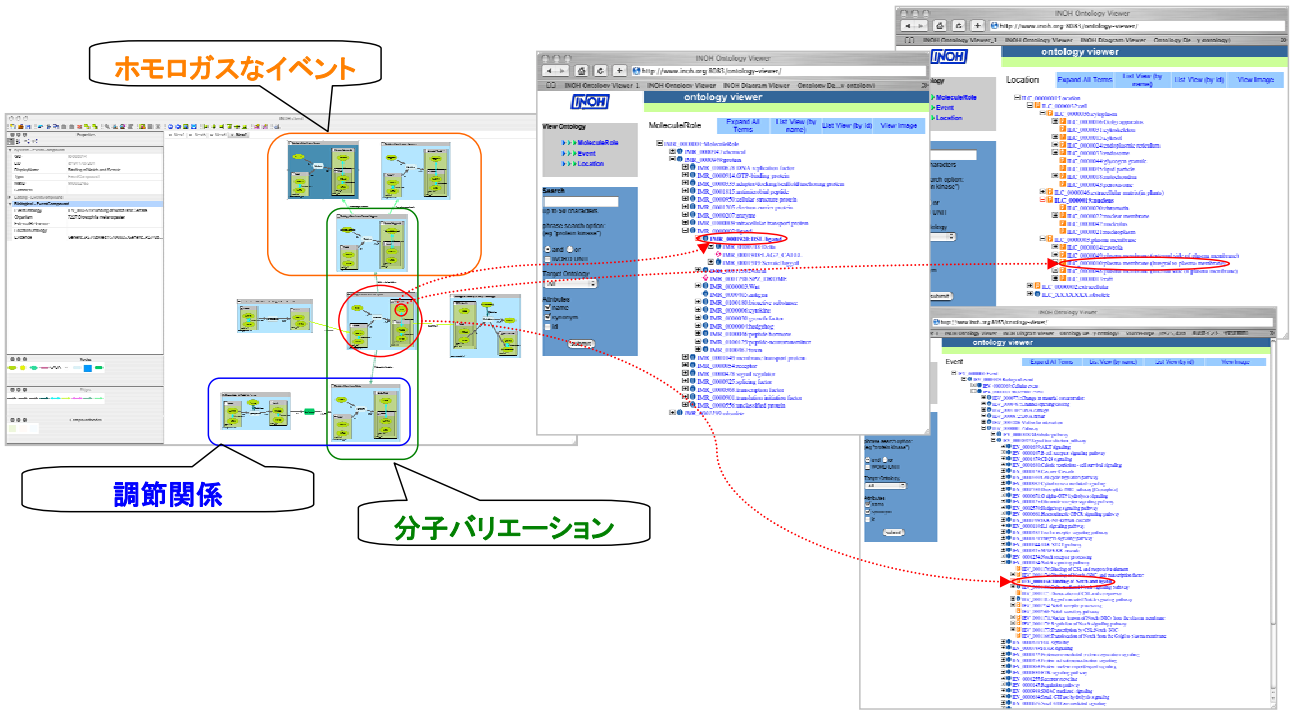

図4．オントロジーによるパスウェイデータのアノテーション

ダイアグラムもしくは自然言語で表現された複雑な知識を、強い構造の上に表現しかつオントロジーによるアノテーション付けを行うには、専門家が文献を読みながら自然に情報を入力できる GUI に基づく支援システムが不可欠である。パスウェイデータを扱う場合、特に属性付き複合グラフの編集に対応し、なおかつ数千のノードを持つパスウェイをストレスなく操作できる性能が要求される。このような背景から、我々は属性付き複合グラフ対応のパスウェイエディタ INOH Client を開発した。本エディタはデータ

ベースを検索するための検索機能も備えており、INOH データベースの検索インターフェイスの役割を果たす。新規のノードを追加する際に、通常のグラフエディタとは異なり、複合グラフに対応したエディタは相互作用グラフにノードが追加された情報だけでなく、追加されたノードが他のノードとどのような包含関係にあるかの情報、つまり前述の分解木の情報も管理される。随時編集をしながら、ノードを挿入・削除することで階層構造を変更できる。このため、ユーザは論文を読みながらパスウェイの部分構造を GUI で操作・管理できるようになっている。図5はパスウェイエディタで検索して取得したデータを編集表示しているスクリーンショットである。

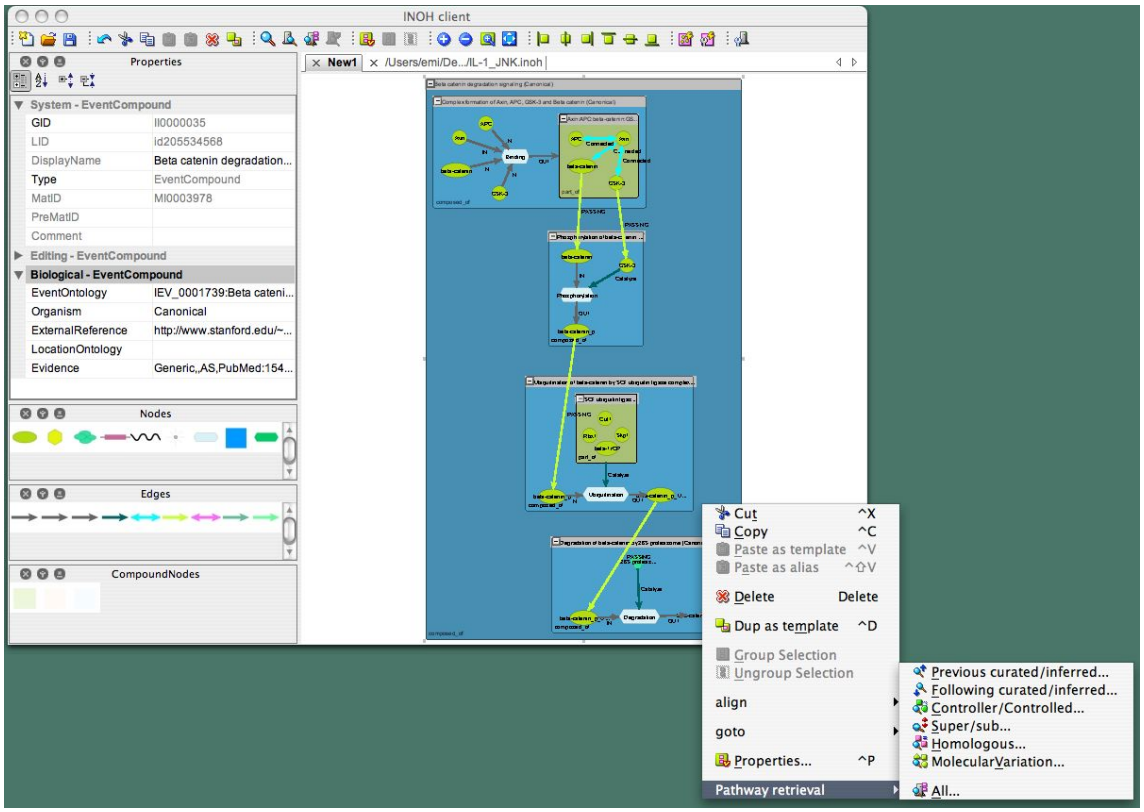

図5. パスウェイエディタのスクリーンショット

データ登録に際しては、1つの複合グラフが1つのパスウェイ登録単位となる。また、各ノード、エッジ、パスウェイはそれぞれ独立したオブジェクトとしてオブジェクト ID を持つ。前述のように、それぞれのオブジェクトはタイプ毎に生物学的情報を付加するための属性集合を持つ。専門家が背景知識を使って汲み取る高度な機能情報を獲得するために、生物学で修士号もしくは博士号をもつ専任のキュレーターが実際に文献を熟読し、シグナル伝達パスウェイデータベースに必要な生物知識の枠組みの体系化(オントロジー等)とパスウェイデータ入力を行っている。

現在、B cell receptor、TGF-beta、CD4 T cell receptor、EGF signaling pathway 、GPCR、Inter Leukin、Toll-like receptor をはじめ代表的なパスウェイが57の分類にわかれて INOH データ release 1.0として公開している。5654のタンパク質、2176の複合体、1619の部分パスウェイ、1379の相互作用が含まれるこれらのデータは http://www.inoh.org から入手可能である。

3. まとめ

本研究開発では生物学文献で共有されるパスウェイ知識をデータベース化する情報処理技術、知識基盤の構築を実施した。論文を介して共有されてきた知見を計算可能な形式でデータベース化するためには、知識を計算機が理解できる形式で蓄積しなければならない。特に、パスウェイのように多様な要素が複雑に関係しあう知識のデータベース化では様々な概念がオブジェクトとして登場するため、これらを自然に結合させるための表現手段とオントロジーの整備が必要となる。

本課題で開発したシステムは階層的で再帰的な表現モデルを採用し、形式化されたパスウェイを構成する全てのオブジェクトをオントロジーで意味づけている。オントロジーベースのパスウェイ検索システムは、ユーザが入力した検索文字列の意図をシステムが理解できる点でキーワードベースの検索と決定的に異なる。

パスウェイデータの知識基盤構築する上で標準化が重要となるが、本研究課題ではパスウェイデータの国際標準化を目指すBioPAXパスウェイデータフォーマットの仕様策定メンバーとして、パスウェイデータベースが表現できるべき情報を決めるデータ・モデル部分の設計に積極的に関与している。

4. 研究開発実施体制

代表研究者　高木　利久（東京大学大学院新領域創成科学研究科）

（1）シグナルオントロジー開発（平成13－4年度）
　　グループリーダー　高井貴子（東京大学）
（2）バイオタームバンク開発（平成13－4年度）
　　グループリーダー　高井貴子（東京大学）
（3）シグナルオントロジー開発（平成15年度）
　　グループリーダー　福田賢一郎（産業技術総合研究所）
（4）バイオタームバンク開発（平成15年度）
　　グループリーダー　高木利久（東京大学）
（5）知識処理設計とデータ入力（平成16年度）
　　グループリーダー　福田賢一郎（産業技術総合研究所）
（6）システム公開・運用（平成16年度）
　　グループリーダー　高木利久（東京大学）
（7）知識処理設計とデータ入力（平成17年度）
　　グループリーダー　福田賢一郎（産業技術総合研究所）
（8）知識処理インターフェイス研究（平成17年度）
　　グループリーダー　高木利久（東京大学）

5. 参考文献

[1] http://www.inoh.org/.

[2] Satoko Yamamoto, Takao Asanuma, Toshihisa Takagi and Ken Ichiro Fukuda, The Molecule Role Ontology: An Ontology for Annotation of Signal Transduction Pathway Molecules in the Scientific Literature, *Comparative and Functional Genomics 5*, 528-536, 2004.

[3] Tatsuya Kushida, Toshihisa Takagi and Ken Ichiro Fukuda, Event ontology: a pathway-centric ontology for biological processes*, Proc. Pacific Symposium on Biocomputing*, 2006. (in press)

[4] Ken Ichiro Fukuda and Toshihisa Takagi, Knowledge representation of signal transduction pathways, *Bioinformatics*, 17:829-837, 2001.