

課題名「シグナルオントロジーとバイオタームバンクの開発」

代表研究者 東京大学大学院新領域創成科学研究科 教授 高木 利久

1. 代表研究者による成果概要報告

1-1. 研究開発のねらい

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ェ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ェイは、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ェイを除き、これまでデータベース化が遅れてきた。その理由として、パスウェ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。

本研究開発ではこのようなねらいのもとに、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 の順に一連の知識ベースの開発に取り組んできた。

1-2. 研究開発の成果

1-2-1. 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。

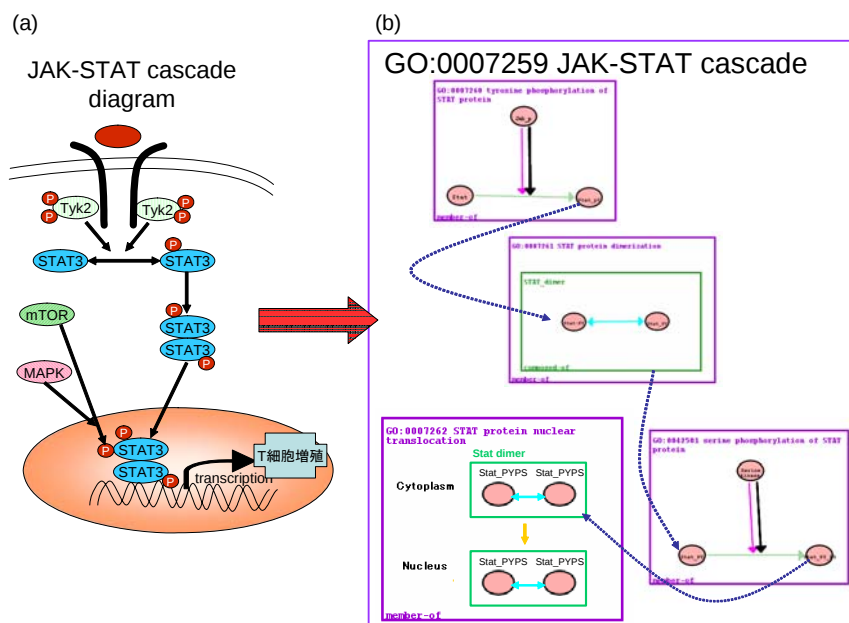


図 1.(a)典型的なシグナル伝達パスウェイ知識の記述
(b)知識を明示化して記述したパスウェイ（細胞質内のみ）

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、様々な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。

開発システムでは図 1(b)にあるような複合グラフに基づくパスウェイ表現を採用している。複合グラフはグラフを拡張した階層的なグラフ構造で、ノードの内側にもグラフが存在する特徴がある。

1-2-2. 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)および生命現象オントロジー(Event Ontology)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。MoleculeRole Ontology はこのような生物学者の知識を体系化し計算機に利用可能にしたオントロジーである。

パスウェイデータベースに関係する生命現象概念は、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－特殊化関係にあり、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-全体関係になっている。Event Ontology はこのような生命現象に関する知識を体系化しオントロジーとして提供している。

1-2-3. 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し、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する必要がある。

INOH では、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(図 2.)。

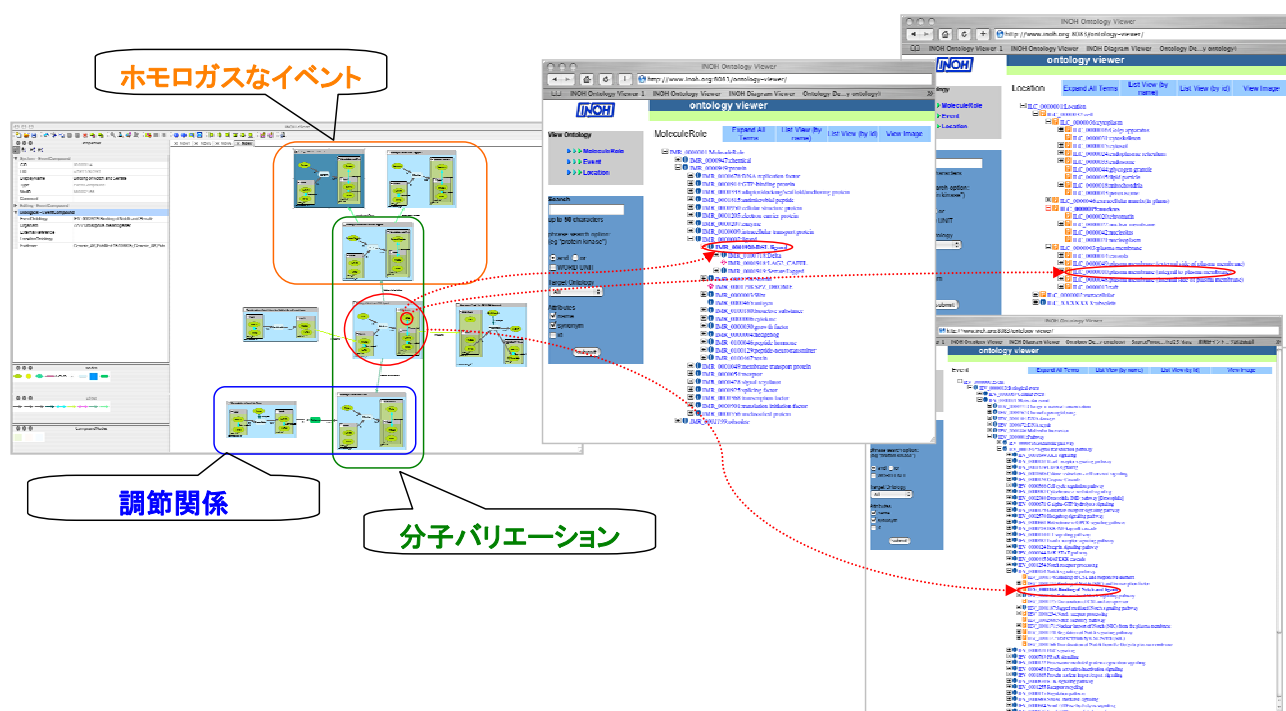


図 2. オントロジーによるパスウェイデータのアノテーション

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を 57 の分類にわけて INOH データ release 1.0 として公開している。5,654

のタンパク質、2,176 の複合体、1,619 の部分パスウェイ、1,379 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。

2. 事後評価結果

2-1. 当初計画の達成度

数年先の研究動向や需要をにらんだ、先進的かつ、高度で、標準化を指向したデータベースを開発する、という挑戦的な計画に基づいた研究開発である。その結果、既存のデータベースにはない高付加価値のデータを新たに電子化したデータベースが構築された。現在の公開までに変遷を要したのは、枠組み構築は研究的側面が非常に強く、また人材育成をしながらの研究開発であったことから考えれば、相応であったと考えられる。

2-2. 知的財産権、外部発表(論文等)等研究開発成果の状況

INOH データ : <http://www.inoh.org/download.html>

INOH クライアント : http://www.inoh.org/Web_Installers/install.htm

Ontology Viewer : <http://www.inoh.org/ontology-viewer/>

オントロジーデータ : <http://www.inoh.org/download.html>

現在のデータベース公開から 1 年余りであるが、確実にアクセス数を増やしている。原著論文発表、学会発表なども、国内・海外ともに積極的に行っている。

2-3. 研究開発成果の公開による波及効果

ライフサイエンス研究において、オントロジーと階層的なデータ構造により生物学者の背景知識を計算機の中に構造化する必要性の認識が広まった。

INOH データベースは高度で精度の高いパスウェイ情報をマニュアル・キュレーションしている。国内外の動向と比較して、パスウェイデータベースに登録されている各種の生体高分子、反応をすべて厳密に階層化されたオントロジーでアノテーションしている公共データベースは INOH 以外にはほとんど存在せず、意義が大きい。アノテーションのためのオントロジーは Open Biomedical Ontologies (OBO) から MoleculeRole Ontology および Event Ontology が公開されている。また、パスウェイデータフォーマットの国際標準化をすすめる BioPAX では仕様策定のワーキンググループメンバーとして積極的に関与し、第一回 BioPAX 国際シンポジウムおよび仕様策定作業部会を日本において主催した。公共データベースの高度化・標準化を目指して開発を進めると同時にパスウェイデータベースやオントロジーの国際標準化動向の急速な展開にも対応し、イニシアティブをとりつつ連携を深めることで、当初想定していた以上に国際社会に対して貢献していることは、高く評価できる。

2-4. 成果の実用化の可能性及び成果から予想される波及効果

生物学者の背景知識をうまく電子化し、検索処理などの諸々の知識処

理に利用可能な形式で提供できる独自色が強みとなっており、敬遠されがちなコストのかかるキュレーションを地道に継続する意義は非常に大きい。高価かつ有用な電子情報の所有率を高めることで、データベースの国際的な価値を強めていき、長期的にデータベースの存在意義を維持することが可能となるであろう。従来のパスウェイデータ解析は利用者がパスウェイデータの構造など、技術的な内容を把握した上で特定の応用事例のために設計されたツールに過ぎなかったが、本研究開発成果を継続してより発展させていくことで、厳密なオントロジーで記述されたパスウェイデータベースが標準化される可能性が開ける。この結果、初めて遺伝子配列から文献知識までの情報が計算機に理解可能な形式でシームレスに結合されることになると考えられる。パスウェイデータの高度化・国際標準化を積極的にすすめることにより、パスウェイ解析技術の新たなシーズが生まれ、技術に駆動される形で新しい生命科学研究が発展すると期待される。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を提供できることが期待できる。

2-5.総合評価

挑戦的研究開発課題であり、我が国に欠けていたオントロジーに挑戦し、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ことは高く評価できる。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り独自性が高い。なお、マニュアルキュレーションは精度は高いが時間がかかる。この分野は進展が速いので、今後は体制を充実し、スピードアップをはかるべきであろう。本研究開発課題が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日本から参加し、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ことは意義深い。本研究開発で新しい公共の知識基盤を創出することに成功したが、今後はこれらの高付加価値データを用いて、演繹的な推論、仮説生成に代表される高度な推論を実現することを期待する。

3.主な論文発表

- 1)Fukuda, K. and Takagi, T., “Knowledge Representation of Signal Transduction Pathways”, BIOINFORMATICS, Vol.17, pp.829-837,(2001)
- 2)Fukuda, K , Y Yamagata and Takagi, T., “FREX: a query interface for biological processes with a hierarchical and recursive structures”, In Silico Biology,4(1),pp63-79,(2004)
- 3)福田賢一郎、山縣友紀、高木利久「パスウェイデータベース：バイオロジーにおける高次知識のデータベース化」情報処理学会シンポジウムシリーズ、情報処理学会:2003 巻、8 号、ISSN 1344 – 0640、pp.267-272, 2003 年 11 月

- 4)Yamamoto, S., Asanuma, T., Takagi, T. and Fukuda, K., “The Molecule Role Ontology: an ontology for annotation of signal transduction pathway molecules in the scientific literature.”, Comparative and Functional Genomics, Vol.2004, 5, 528-536.(2004.12)
- 5)Kushida, T., Takagi, T. and Fukuda, K., “Event Ontology : A Pathway-Centric Ontology for biological processes.”, Pacific Symposium on Biocomputing 2006, Vol.11, pp.152-163, 2006.1

参考

論文発表 国内 11 件、海外 12 件

口頭発表 国内 11 件、海外 9 件

ポスター発表 国内 3 件、海外 8 件

特許出願 なし