

## 研究開発課題別中間評価結果

### 1. 研究開発課題名

シグナルオントロジーとバイオタームバンクの開発

### 2. 代表研究者名

高木 利久 (国立大学法人東京大学大学院新領域創成科学研究科 教授)

### 3. 研究開発概要

近年、大規模網羅的研究データの解釈、新たなパスウェイ発見の仮説生成、生命のシステムの理解（ネットワークの理解）に向けた解析に生体内パスウェイデータベースが必要とされている。そこで、本研究開発では、文献からの情報を用いて生物のシグナル伝達パスウェイを計算機上でモデル化するために、一貫性のある知識処理システムの枠組みを開発すること、および文献に登場する一連の概念を整理し体系化したオントロジーを構築することを目指している。これまでに世界で初めてグラフ（ネットワーク構造）を階層化した複合グラフによってパスウェイ構造を表現し、階層化された知識構造に対する生物学的に有用な検索機能の開発、ならびに階層化されたパスウェイ知識をキュレーションするための支援システム（パスウェイエディタ）の開発・公開を行った。パスウェイデータのキュレーションについてはヒト・マウス・ラットを中心に39のパスウェイを収集した。またオントロジー群の整備については、分子機能、細胞機能、生物種からフェノタイプ（表現型）に至るまで12余り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ている。

### 4. 中間評価

#### 4-1. 研究開発の進捗状況と今後の見込み

本研究開発は、文献中の細胞機能に関する生命知識を計算機に理解可能な形式で電子化するために、複合グラフ構造にもとづく（1）生命科学の視点からの新しい知識処理技術の開発、（2）複合グラフ構造をアノテーションするためのオントロジー群の構築、（3）背景知識をもらさずにアノテーションしたパスウェイデータコンテンツを新規にキュレートすることを目的としている。データ収集は、生物学者がパスウェイデータそのもののキュレーションとパスウェイデータ内に登場する各種の概念やオブジェクトを定義するためのオントロジー群の構築作業を並行して行っている。ここで構築するオントロジーは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。

また、当初は計算機に理解可能な形式で生物医学概念を定義したオントロジーを構築し、生命科学の知識の全容をカバーするために必要となる網羅的な用語辞書の構築を目標としていたが、生命科学全般は非常に広範の知識を含むため、現在は、生物学全般にわたる一般的な用語の収集と特にシグナル伝達パスウェイなどの細胞機能の分子機序にかかわりのある概念の収集をすすめている。

当初の予定より少し遅れ気味であるが、今後の展開に期待する。

#### 4-2. 研究開発成果の現状と今後の見込み

Web公開状況 INOH <http://www.inoh.org> FREX  
<http://www.ontology.jp:8083/FREX2/jsp/index.jsp>  
GEST <http://www.inoh.org:8083/GEST/> Ontology Viewer  
<http://www.inoh.org:8083/ontology-viewer/>

キュレーションされた複合グラフ構造のデータに対して生物学的に有用な検索機能を実現するためのウェブアプリケーションを開発している。このアプリケーションの特色は、パスウェイデータならびにオントロジーデータに対して、その構造を考慮した検索ならびにあいまい検索が実現されている点である。また、パスウェイデータのキュレーション支援のためにパスウェイエディタを開発している。このパスウェイエディタはパスウェイ構造に対してパスウェイ部分構造などの背景知識をアノテーションするために、複合グラフ構造に対応させている。上記のウェブアプリケーションおよびエディタはインターネットから公開されている。ここで開発されたパスウェイデータベースは、全てのオブジェクトがオントロジーに基づいてアノ

テーションされている点が他の同種のものにはない特徴である。この枠組みを採用することによって分子名だけでなく生物機能、反応、細胞内局在、細胞、組織といった様々な属性からパスウェイを検索でき、また検索条件の厳しさをオントロジーに基づいて緩めるなどの検索条件指定を拡張した知的な検索ができる。

一般的な語彙の収集(Bio Term Bank)についてウェブアプリケーションを開発し、特にシグナル伝達のオントロジーとデータベースに必要となる生命科学用語の整備に力を注いでいる。シグナル伝達パスウェイなどの細胞機能に関連する分子機序の知識を電子化するために必要となる用語の概念を、cell-line、Tissue、Species、Domain-family、Phenotype、DNA、RNA、Localizationとして体系化している。

本研究開発の成果はweb上での公開やBioinformatics等の国際雑誌への掲載、様々な学会での発表等、広範に発表されている。今までの検討をふまえた今後の成果に期待する。

#### 4-3. 今後の研究開発にむけて

本研究開発が達成すべき最終的な課題は、(1) 生命科学における知識処理技術の開発、(2) 文献に登場する概念を定義するためのオントロジー群の整備、(3) 生物学者によるデータのキュレーション、といった各課題の成果を統合し、文献に記載された分子機序に関する生命知識を検索することができる、全く新しいデータベースシステムを構築・公開することである。今後については上記の各課題において、それぞれ次の目標が掲げられており、その実現に期待したい。(1) においては、矛盾する知識の検出や制御・非制御情報の演繹的な導出などのより高度な検索機能を開発すること。(2) においては、データ形式を標準化・データ流通の観点からxmlベースのオントロジー記述言語のデファクトとして世界的に注目されるOWLフォーマットへ移行すること。(3) においては、主要なモデル生物種についてそれぞれ代表的な30~50のパスウェイを収集すること。本格的運用を行い、研究者に利用してもらうことで、さらに成果が得られると考える。

#### 5. 国内外のデータベース高度化・標準化の動向・状況と本課題の位置づけ

シグナル伝達パスウェイは種々のパスウェイデータベースの中でも扱うべき概念が最も多岐にわたっていて電子化が困難である。また手動による文献知識のキュレーション(マニュアルキュレーション)は、高品質なデータが得られその価値も高いが、非常にコストがかかるため国内外の動向として本手法を実施する研究機関は少数である。このような背景をもつシグナル伝達パスウェイデータベースプロジェクトの中で、本研究開発は背景知識をもらさずアノテーションした形式で高品質のマニュアルキュレーションを実施している。さらに、複雑で階層的な知識を格納でき、階層グラフによる視覚的に分かりやすい部分表示が可能であることから、特定のシグナル伝達にプロテアソーム(プロテイナーゼ複合体)が関与する制御機構をグラフ表示とその構成分子などの性質や分布などを含め、回答として提示でき、高度なデータベースとして優れている。また、用語の体系化、知識表現の統一的な枠組みとしてのオントロジーの整備がされており標準化の点で意義がある。

#### 6. 研究開発成果の社会への貢献

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るシグナル伝達系のような細胞機能全般にまつわる知識を扱ったパスウェイデータベースの開発は世界中から期待されている。この際に問題になるのはそれぞれの事項にどれだけの生物学的内容を独自に関連づけしているかである。本研究開発には、手作業によって得られた信頼性の高い情報やデータベースを発信していくことにより生命科学へ貢献をしていくことを期待する。本研究開発の進展により、研究者が論文の洪水を克服して、過去の知見の蓄積をすばやく自分の研究に位置付けられるようになる。

#### 7. 総合的評価

基盤的な情報と技術の開発という意味で根本から取り組んでおり、オリジナリティのある成果が期待でき、完成すれば、新しい言語体系による生物学が始まる可能性があり、基礎研究はもとより、予期せぬ生物機能の発見を通して、物質生産などの研究開発にも大きな貢献が期待でき、継続して進めていくべき事業である。これまでは技術的な試行錯誤に時間をとられ進捗が遅れ気味であるが、これからの展開に大きく期待したい。オントロジーの整備がされてお

り、標準化の点で大きな意義がある。今後は生物学情報を計算機にのせるだけでなく、生物学的意味論の抽出も必要であろう。また、国際的に使われるようになれば意義が深いと考えられるだけに、多くの生物学者と連携していく必要がある。テーマ自体が挑戦的なことから研究開発体制そのものについても進展に則した充実が望まれる。