オントロジーによるパスウェイの高度化および国際標準化（INOH パスウェイデータベース）

東京大学大学院新領域創成科学研究科
高木　利久

INOH: an ontology-based highly structured database of signal transduction pathways

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

INOH is a highly structured, manually curated database of signal transduction pathways including *Mammalia*, *X. laevis*, *D.melanogaster*, *C.elegans* and canonical. As most part of pathway knowledge resides in scientific articles, the database focuses on curating and encoding textual knowledge into a machine-processable form. We use a hierarchical pathway representation model with compound graph and every pathway component in INOH is annotated by a set of uniquely-developed ontologies. Therefore we can provide rich semantics and a powerful querying facility. It is difficult to achieve such functions in a typical keyword-search-based database. And we also cooperate in the effort to establish a pathway description standard format, BioPAX format in an affirmative way.

1. はじめに

現在、生物のパスウェイに関する情報を提供するデータベースが世界に 300 以上あると言われている。しかしその情報の多くはイラストや文章で書かれたものや、単純な分子同士の相互作用関係によるものが多く、複雑な知識を体系化したものは少ない。INOH データベースでは、従来のように人間による利用を前提として教科書／論文の知見を電子化するのみではなく、応用プログラムが直接利用することを想定して計算機が処理可能な形式でパスウェイ情報を統合することを目的に、開発を行った。様々なパスウェイデータの中でも特に扱うべき概念が非常に多岐にわたるシグナル伝達パスウェイを体系化するにあたって、記述された様々な概念の違いや相対的な関係を計算機に識別させるために[1]、オントロジーに基づいた新たな知識処理の枠組みを開発した。

また同時に、オントロジーによる詳細なアノテーションが付与された高精度の一次データ公開機関としての独自性をいかし、よりいっそうの成果の普及と知識基盤としての浸透を目的に BioPAX パスウェイデータ交換フォーマットなどの国際的な標準化活動に積極的に取り組んだ。

2. 研究開発の成果

2.1 INOH データベースの表現力と機能の拡張

応用プログラムが直接利用することを想定し計算機が処理可能な形式でパスウェイ情報を統合するため、シグナル伝達パスウェイ・データベースシステムを開発した（INOH パスウェイデータベースシステム）。インターフェースとなる INOH client は、キュレーション機能をもち論文中で自然言語や図で表現されていた情報を柔軟に表現することができる。また、INOH パスウェイデータベースへの検索機能を持ち、キーワード検索、および、前後パスウェイ、ホモロガスパスウェイ、分子バリエーションなどのパスウェイ検索が可能で、取得したパスウェイデータに対してユーザが自由にパスウェイを追加する機能をもつ。また、2.3 で述べる類似パスウェイ検索機能をもっている。

INOH client は、Java Web Start を使った起動ができ Web ブラウザからの検索・表示が可能である。また INOH client から外部データベース（PubMed、GO、UniProt、KEGG、EC など）や INOH Ontology Viewer（http://www.inoh.org/ontology-viewer/：現在公開中であるオントロジー検索のための Web アプリケーション）へのリンク機能により、オントロジー（Ontology Viewer）とパスウェイデータ（INOH client）の両方向からの利用ができる。

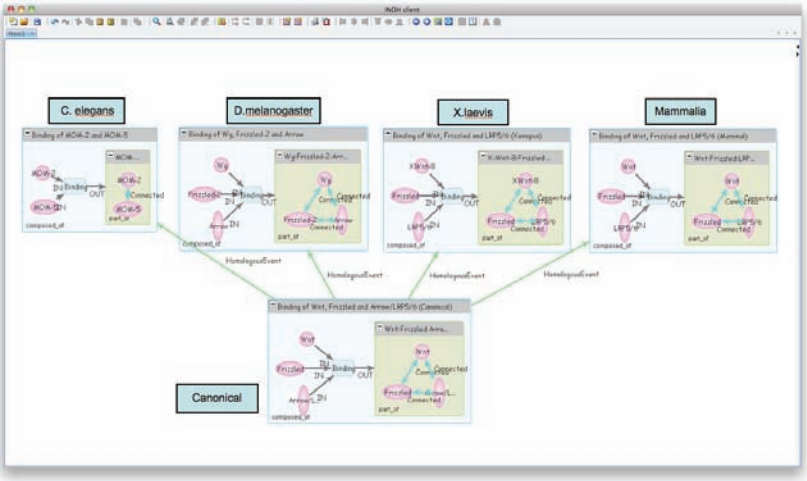
更に、パスウェイ全体を視覚的に把握し易くする、分子主体の表示機能をもち、パスウェイを階層的に表示する Normal View と、分子間相互作用をわかり易く表示する Reduced View 機能を提供している。

またキュレーション作業の効率化を図る機能をもち、データのエラーチェック機能も自動で行うことができる。

INOH サーバについては、速度や移植性の向上を考慮し、新型データベース eXist（オープンソースネイティブ XML データベースシステム）を用いている。

2.2 知識の抽出およびオントロジー構築

INOH パスウェイデータは、専門家が科学文献中に記述されている生体内分子機序に関する知識を抽出し、計算機による様々な処理ができるような形式で電子化したものである。本研究では、カノニカルな（生物種を区別しない）59 のシグナル伝達パスウェイに加え、生物種別のシグナル伝達パスウェイ（*Mammalia* 4、*M.musculus* 1、*X.laevis* 2、*D.melanogaster* 9、*C.elegans* 5）、代謝パスウェイ（*H.sapiens* 29）を新たにキュレーションした。INOH では、生物種別のパスウェイ内の各プロセスをカノニカルなパスウェイと関連づけて記述しているため、プロセスの生物種間での比較が可能である。このように、マニュアルキュレーションした生物種別のパスウェイを、カノニカルなパスウェイと関連づけているデータベースは他にない（下図：Wnt pathway のホモロガスイベントの例）。その他、各パスウェイについて、未知の情報を含むパスウェイへの新しい情報の追加、パスウェイ本流を調節するプロセスの追加、パスウェイ同士の関連付けなど、常に最新の情報を反映させている。パスウェイデータは、INOH 形式、BioPAX 形式ともにプロジェクトの Web サイト（http://www.inoh.org/）からフリーで提供している。また、パスウェイ統合データベース（ConsensusPathDB、InnateDB など）においてもデータ利用されている。



INOH オントロジーは、パスウェイデータをアノテーションする標準オントロジーが存在しないため、INOH 独自で開発したオントロジーである。オントロジーによってパスウェイに関する生物学者の知識を体系化し、パスウェイデータをオントロジーでアノテーションすることにより、パスウェイデータが一貫性を持ち、用語の階層をたどることにより用語の背景知識を得ることができる。さらに、パスウェイデータの高度な推論検索も可能である。

パスウェイに登場する Protein ノードや ChemicalSubstance ノードのアノテーションには MoleculeRoleOntology を[2]、Event ノードや EventCompound ノードのアノテーションには EventOntology を使用し[3]、各オントロジー用語には様々な外部データベースへのリンクや定義およびその出典を記述した（UniProt、KEGG COMPOUND、GO、KEGG REACTION、PSI-MI、PubMed など）。

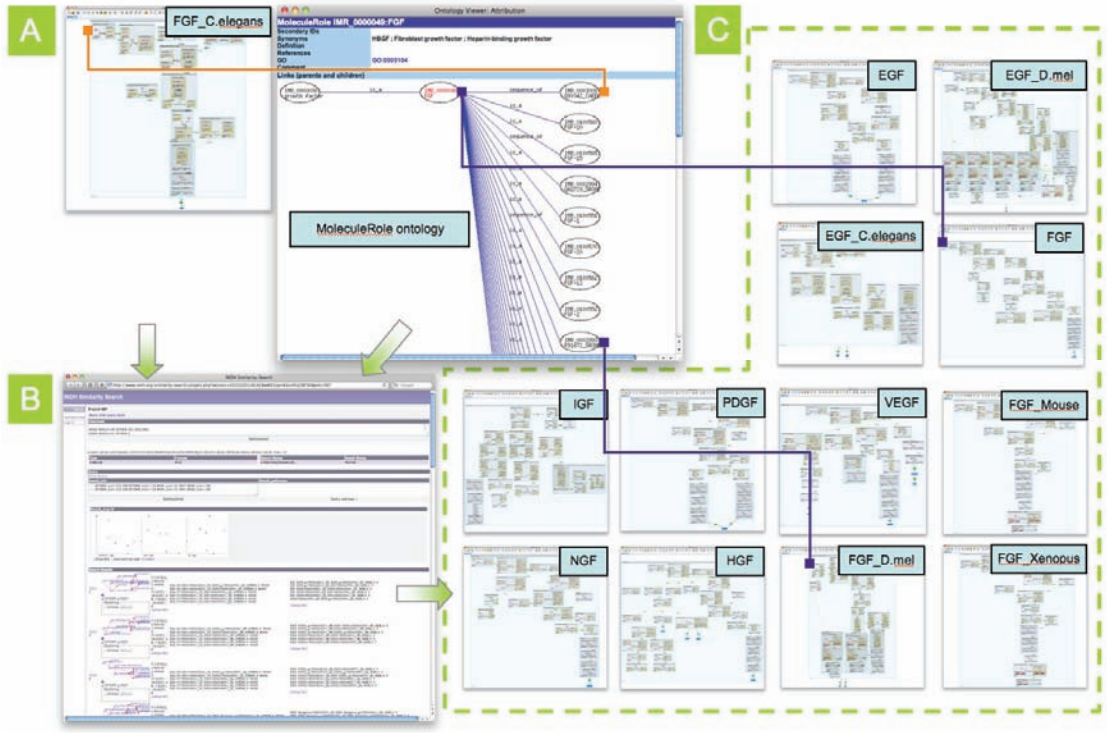
MoleculeRoleOntology、EventOntology はプロジェクトの Web サイト、OBO Foundary、NCBO BioPortal、BRENDA Ontology Explorer などから OBO 形式で提供している。

2.3 高度なパスウェイ推論検索システムの構築

オントロジーの階層とイベントの包含関係を利用した、類似パスウェイを検索・提示するシステムとして、INOH client から利用できる SimilaritySearch を開発した。利用者が検索したい部分パスウェイをクエリとし、与えられたクエリに対して、パスウェイデータと MoleculeRoleOntology や EventOntology から類推される、類似パスウェイを検索し出力することができる。これによりキーワード検索等では出ないパスウェイ同士の類似性を推測することが可能である。

類似パスウェイを検索するにあたっては、各クエリノードと類似のノードを求めることで算出する手法を用いた。類似ノードは、ノードに付与された MoleculeRoleOntology や EventOntology と、その階層から、分子間やイベント間の意味的距離を算出し、ユーザの許容する閾値以内のものを類推される検索対象として抽出した。抽出された類似ノードおよびエッジは、一連のつながりを持つ部分（連結成分）ごとに分け、類似ノードの意味的距離と対応エッジの構造的距離から算出する 3 種類の類似係数を算出し、利用者の利用目的に応じて順序付けられるようにした。

（下図：*C.elegans* の FGF pathway からの類似検索の例）



2.4 ウェブサービス API の開発

INOHデータベースをユーザがWebブラウザやプログラムから利用できるようINOH Web API（http://www.inoh.org/inoh_api_manual.html）を公開した。INOHデータベースに対するキーワード検索、および、前後パスウェイ、ホモロガスパスウェイ、分子バリエーションなどのパスウェイ検索やパスウェイデータの取得ができる。このAPIを組み合わせることで、実験データから得られたタンパク質を複数指定し、前後に起こるイベントを再帰的に検索することで、大規模なパスウェイのネットワークを仮説生成することが可能である。

2.5 国際標準化への取り組み

最新の国際動向として、パスウェイデータやシミュレーションデータなど、生命現象の分子メカニズムに関連したデータの標準化はパスウェイデータベースのデータ交換フォーマットである BioPAX、シミュレーションデータなどのシステムズバイオロジー研究を指向した SBML/CellML などが様々な活動を展開している。このような世界的な標準化の動向を考慮し、BioPAX の一員として標準化の議論に参加し、各種の仕様の提案および検証を行った。

論文中で自然言語や図で表現されていた情報を柔軟に表現できる高度なデータ表現の INOH 形式について、BioPAX フォーマットで適切に表現できないケースがあり、BioPAX の仕様提案としてフィードバックを行い、仕様が反映された[4]。

更に、INOH から BioPAX フォーマットへのエクスポートツールを開発し、BioPAX 形式も公開した。これにより INOH パスウェイデータを外部参照として利用する Web サイトが増え、INOH パスウェイデータのよりいっそうの普及に結びついた。

3. まとめ

本研究開発では、INOH データベースの知識処理技術の拡張を行い、高精度のパスウェイデータの公開、様々な分子生物学概念を定義したオントロジーの公開、オントロジーの階層とパスウェイの包含関係を利用した高度なパスウェイ推論検索システムの公開、およびパスウェイデータをプログラムから利用するための API の公開を行った。これらは従来の人間による利用を前提として教科書／論文の知見を電子化したパスウェイデータベースとは異なり、応用プログラムが直接利用することを想定して計算機が処理可能な形式でパスウェイ情報を統合した、未来指向型のデータベースであると言える。これからの生命科学分野において重要な、大規模実験データの自動解釈や仮説生成などによる知識発見に貢献するものと確信している。

また同時に、パスウェイデータの世界的な標準化の動向を考慮し、INOH パスウェイデータのよりいっそうの普及を目的として、特に BioPAX パスウェイデータ交換フォーマットにおいて各種の仕様の提案および検証など標準化にむけた独自の貢献を行うことが出来た。

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)　統括グループ

　　グループリーダー　高木　利久（東京大学大学院新領域創成科学研究科）

(2)　データベース設計・開発、オントロジー構築・データ抽出グループ

　　グループリーダー　福田　賢一郎（産業技術総合研究所）

5. 参考文献

[1] Fukuda, K. and Takagi, T. (2001) Knowledge representation of signal transduction pathways. Bioinformatics, 17, 829-37.

[2] Yamamoto, S., Asanuma, T., Takagi, T. and Fukuda, K.I. (2004) The molecule role ontology: an ontology for annotation of signal transduction pathway molecules in the scientific literature. Comp. Funct. Genomics, 5, 528-536.

[3] Kushida, T., Takagi, T. and Fukuda, K.I. (2006) Event ontology: a pathway-centric ontology for biological processes. Pac Symp Biocomput, 152-63.

[4] Demir, E., Cary, M.P., Paley, S., Fukuda, K, et al. (2010) The BioPAX community standard for pathway data sharing. Nat. Biotechnol., 28, 935-942.