人工知能学会研究会資料
SIG-SW&ONT-A302-06

バイオロジーにおける高次知識の体系化：パスウェイデータベースとオントロジー

福田　賢一郎
産業技術総合研究所　生命情報科学研究センター
〒 135-0064 東京都江東区青海 2-43 青海フロンティアビル 17F
Tel: 03-3599-8049 Fax: 03-3599-8081
fukuda-cbrc@aist.go.jp

高木　利久
東京大学大学院新領域創成科学研究科情報生命科学専攻

概 要

生物学分野では生命現象の分子機序を説明する “事柄間の関係の組み合わせ” などの高次知識がますます重要になってきている．多様で不均一な粒度の要素が複雑に組み合わさって構成されるこれらの知識のデータベース化にはオントロジーを強く意識した知識システムが不可欠となる．本発表では我々の構築しているシグナル伝達パスウェイデータベースについて報告する．

1　はじめに

ヒト，マウスのゲノム配列の解読が完了し，タンパク質が互いにどのように関わり合って生命現象の分子機序を成り立たせているかを表す “関係の組み合わせ”，すなわちパスウェイと呼ばれる高次知識がますます重要になってきている．このため，パスウェイ知識を可能な限りデータベース化する意義は大きい．ここで問題になるのが，これらの知見をどのように計算機上で処理するかである．というのは，このような意味的な情報 (背景知識) は通常論文中に自然言語や図によって表現されているからである．

文献に記載されているパスウェイ知識は，化合物，タンパク質，金属イオン，さらには “ 細胞死” のような生命現象など様々な概念を構成要素とし，また要素間の関係も “輸送”，“修飾”，“制御”，“生化学反応” など様々である．このように多様で不均一な粒度の要素が複雑に組み合わさって構成される知識のデータベース化にはオントロジーを強く意識した知識システムが不可欠となる．

本発表では我々の構築しているシグナル伝達パスウェイデータベースに関するプロジェクトを紹介する．このプロジェクトでは，(1) 分子生物学分野に関わる高次知識の計算機処理技術の開発，(2) 専門家による文献からの高次知識の収集，(3) 必要なオントロジー群の整備，(4) パスウェイデータベースシステムの開発までを含めて総合的にパスウェイデータベース構築という課題に取り組んでいる．

以下の各節では，文献に記述された生体機能の知識表現の問題を議論し，生物学オントロジについて触れる．また，知識表現とオントロジーの対によって実現される検索機能を述べる．その後，文献情報から適切に効率よく情報を抽出するためのグラフィカルユーザーインターフェイスの開発および実際のシステムの構成を論じる．

2　生体内パスウェイ

細胞は生き延びるために必要な機能を，遺伝子やタンパク質がお互いを複雑に制御する機構で実現している．この制御のネットワーク的な構造をパスウェイと呼ぶ．

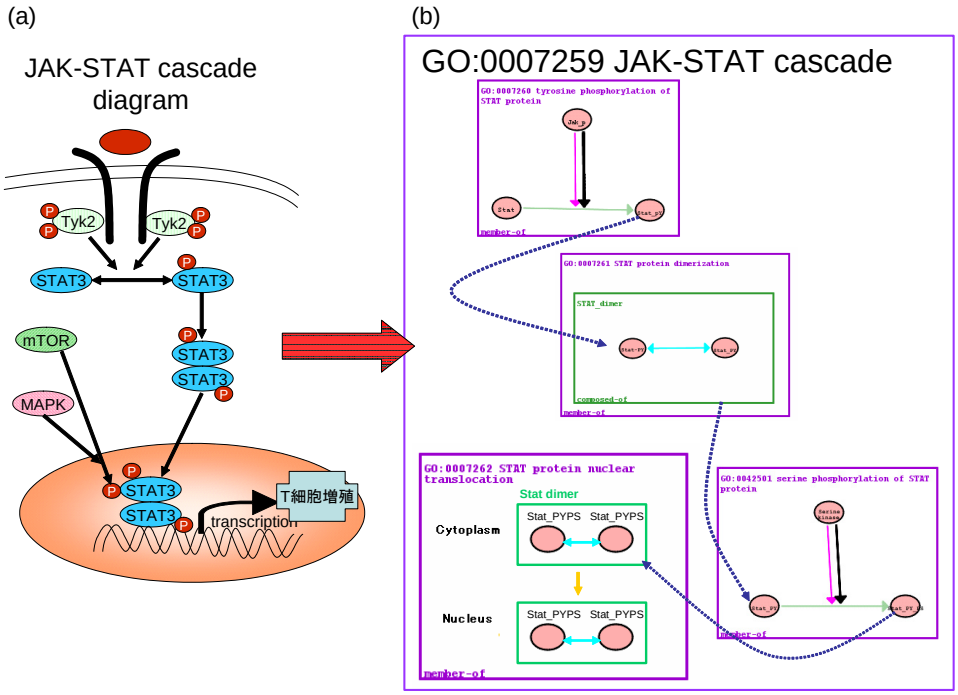

図 1: (a) 典型的なシグナル伝達パスウェイ知識の記述．(b) 知識を明示化して記述したパスウェイ（細胞質内のみ）．(a) 細胞膜上の受容体に刺激が結合することで JAK，STAT を主体とするメカニズムが応答を核内に伝達し，遺伝子の転写を引き起こしている．事柄間の関係づけに黒い矢印が使用されているが，様々な意味用いられることに注意されたい．

本報告では，生体内パスウェイに関する高次知識を，細胞内の物質・現象の間の関係の情報と定義する．このように定義される生体内パスウェイ情報の中で我々のデータベースが蓄積していくのは主に科学論文を介して生物学者の間で共有されている，自然言語およびダイアグラム図を用いて記述されたパスウェイ情報，特にシグナル伝達パスウェイに関する知識である．

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 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(a) 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

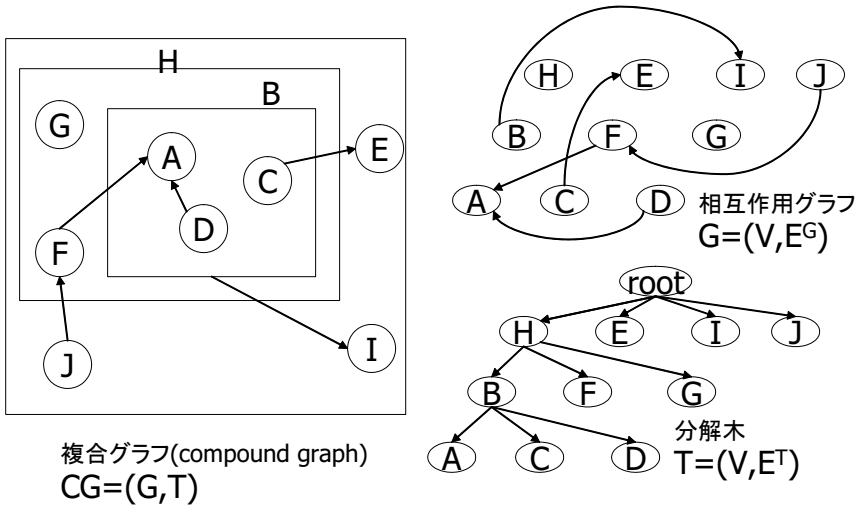

図 2: 複合グラフに基づくパスウェイ表現

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．

3 パスウェイの複合グラフ表現

提案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 [1]．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．本論文ではグラフ 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))．

生体内プロセスの検索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．

- [1] Localization
- [2] Signal passing process (bio-process) Ontology
  - Biological functions & molecular interactions
- [3] Signaling molecule Family (signal-family) Ontology
  - Families of proteins classified by functional similarities
    - Smad2,3[regulatory], Smad6,7[inhibitory]
    - ligand, antigen, scaffold protein, transporter
- [4] Species
- [5] Phenotype [MeSH/disease]
- [6] Tissue and Organs [TissueDB, http://tissuedb.ontology.ims.u-tokyo.ac.jp/]
- [7] Molecule DB : protein [SWISS-PROT], protein complex
- [8] Chemical DB [KEGG]
- [9] DNA DB [ATCC]
- [10] Reaction DB
- [11] Cell-line DB
- [12] Reference DB [PubMed]

図 3: オブジェクトのアノテーションに使用されるオントロジーの一覧.

3.1 オントロジーによる構造のアノテーション

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．分子情報は該当タンパク質の存在を規定している生体高分子データベースへのリンク情報を格納している．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リンクが格納される．

本報告で構築しているパスウェイデータベースはアノテーションのために以下を含む６つのオントロジーをシステム内に持ち，さらに外部データベースへのリンクを持っている[1]：細胞内局在部位，生体プロセス (bio-process)，シグナル伝達機能分類 (sig-family)，生物種，表現型，組織・器官，など (図 3)．

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．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，D と 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．

3.2 グラフエディタ GEST

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．

このような背景から，我々は属性付き複合グラフ対応のグラフエディタ GEST (Graphical Editor for Signal Transduction) を開発している [2]．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，つまり分解木の情報も管理しなければならない．

GEST では随時編集をしながら，ノードを挿入・削除することで分解木の階層を変更できる．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 を介して操作・管理できるようになっている．図 4 は GEST のスクリーンショットである．分解木の内点に相当するノードを “展開”，“縮退” させる操作が示されている．オブジェクトの属性を表示・編集する左上の表の下にあるツリーが分解木情報を表示しており，ツ

[1] 本報告では，オントロジーは DAG 構造を持った概念の階層分類もしくは概念階層と各概念階層に属するインスタンスの集合を指す

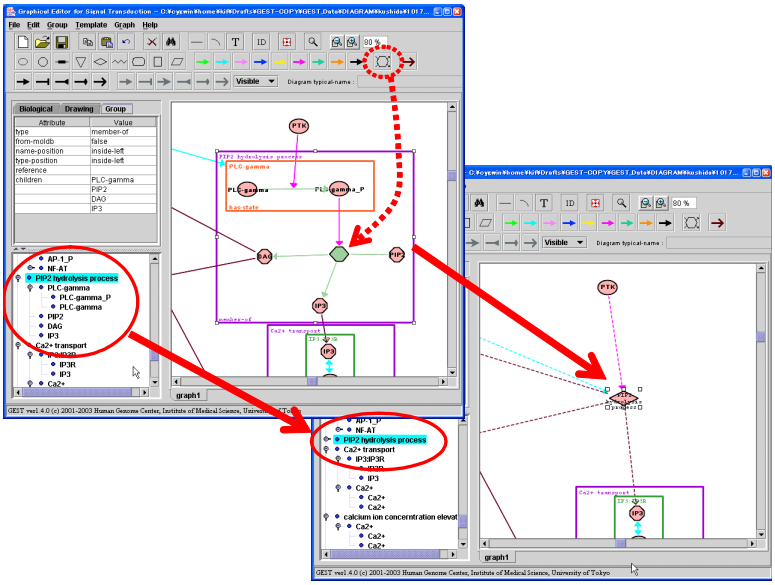

図 4: 複合グラフエディタのスクリーンショット．ノードの “展開”，“縮退” 動作を行っている．

リー表示の各ノードの開閉動作に連動して，該当するノードがグラフ編集画面上で展開・縮対している．なお，画面内の PIP2 の加水分解によって IP3 と DAG が生じる例に示されるように，GEST では多対多のエッジならびにエッジを指すエッジの存在を許している．

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．

4 FREX システム

蓄積されたパスウェイ情報は FREX (Functional Relation EXplorer) と名付けられた Web アプリケーションベースのインターフェイスを介して検索可能である．

パスウェイデータベースシステム全体は FREX サーバー，XML ミドルウェアとリレーショナル DBMS の三層からなる．XML データベース上にはパスウェイデータベース，およびパスウェイの各構成要素を規定する各種のオントロジーならびにタンパク質データベースなどの分子生物学データベースが格納されている．

ブラウザからの検索問い合わせを受け取った FREX サーバーは JAVA で実装された XML ミドルウェアの提供する API を介して XML データベースにアクセスする．実際の XML データは PostgreSQL 上に用意したテーブルに分解されて格納されている．

4.1 FREX インターフェイス

ウェブブラウザでアクセスするとユーザは生体高分子などの事柄に関する情報，事柄間の関係にまつわる情報，またはそれらの関係情報の集合であるパスウェイに関する情報を指定することで，科学論文に蓄積された知見を検索することができる．検索は指定した項目の論理積として扱われる．ここで注意したいのは，FREX の検索機能が属性と属性値の対によるキーワード検索を実現しているのではなく，オントロジー情報に基づいた検索を実現していることである．

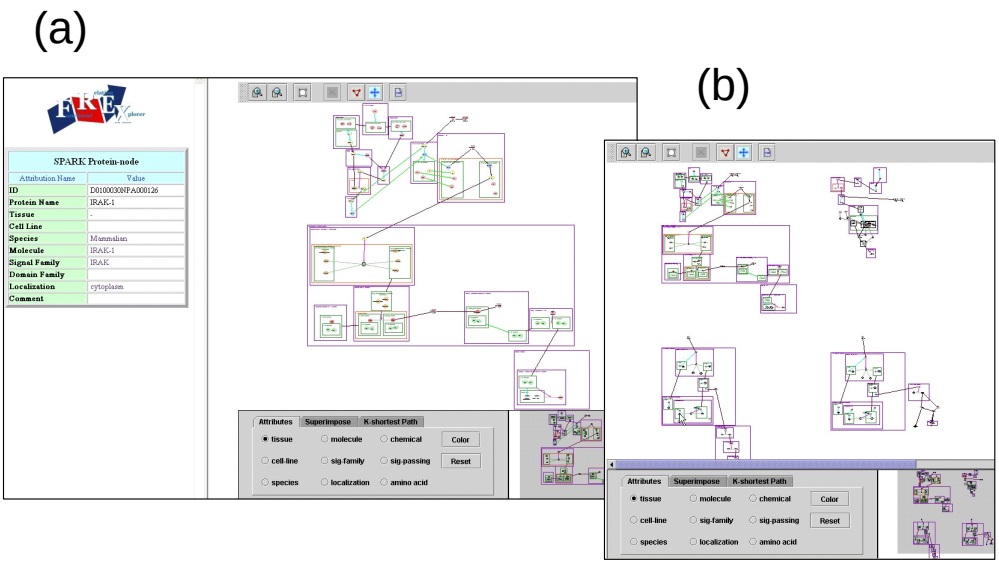

図 5: パスウェイ検索結果の表示 (a) 単一パスウェイ (b) ４つのパスウェイの同時表示

4.2 FREX による高次知識の検索

FREX では検索対象として複合グラフ表現のノード，エッジ，パスウェイを指定できるが，ここではパスウェイ検索とノード検索について述べる．

トップ画面でプルダウンメニューから検索項目を指定し，指定した項目について値を入力する．検索ボタンを押した結果，条件に適合したオブジェクトのリストが表示される．リスト内から１つ以上，複数の結果を選択し表示することが可能である (図 5)．表示画面にはサムネイル表示や拡大縮小，属性値によるオブジェクトの色の塗り分け，k-最短経路検索 [3] などの機能が実現されている．

ノード検索の場合には結果は二通りに表示できる．一つは該当ノードを含むパスウェイの表示である．もう一方は，検索条件に適合したノードが分解木の内点である場合（例えば，プロセスノード），ノード自身の下位構造だけを表示することができる（図は割愛）．

5 まとめ

本報告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取り組みについて論じた．

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．

提案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．

6　今後の課題

現在のシステムは独自開発の XML ミドルウェア，オントロジー記述フォーマットによって構築されている．また，ミドルウェアとサーバー間の通信には RMI 技術が採用されている．このため，オープンな環境での利用およびデータ配布に関して問題がある．我々は現在，Web サービスやセマンティック Web で標準とされる技術に基づいたシステムの改修に取り組んでいる．

参考文献

[1] Ken ichiro Fukuda and Toshihisa Takagi. Knowledge representation of signal transduction pathways. *Bioinformatics*, 17:829–837, 2001.

[2] K. Fukuda and T. Takagi. A pathway editor for literature-based knowledge curation. *Conferences in Research and Practice in Information Technology (APBC2004)*, 29:to appear, 2004.

[3] D. Eppstein. Finding the k shortest paths. *SIAM J. Computing*, 28:652–673, 1998.

[4] Ken ichiro Fukuda and Toshihisa Takagi. Signal transduction pathways and logical inferences. In *Proc. of The 2001 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences, METMBS '2001*, pages 297–303, 2001.