複合グラフ構造による生命現象メカニズムのデータベース化

福田 賢一郎[†] 山縣 友紀[††] 高木 利久[†††]

ヒト，マウスのゲノム配列の解読が完了し遺伝子のカタログ化がすすんだことで，生体高分子単体の機能とあわせて，多数の生体高分子で構成される生命現象のメカニズム（パスウェイ）に関する知識がますます重要になってきている．これらの知識は文献を通して共有されており，そのデータベース化が知識基盤として果たす役割は大きい．文献で共有された知識は，化合物，タンパク質，金属イオン，さらには “細胞死” などの生命現象を構成要素とし，また，要素間の関係も “輸送”，“制御”，“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を強く意識した知識システムが不可欠となる．本稿はこれらの課題を解決するために我々が開発したパスウェイデータベースについて報告する．まず，文献に記述された高次知識であるパスウェイの特徴を概観しその知識表現の問題を論じる．その後，知識表現とオントロジの対によって実現される検索機能と実際のシステムの構成を論じている．

Pathway Database: Higher Order Knowledge in Biology

Ken-Ichiro Fukuda,[†] Yuki Yamagata[††] and Toshihisa Takagi[†††]

Molecular mechanisms of biological processes are typically represented as a diagram called “pathways”, which have a graph-analogical network structure. However, due to the diversity of topics that pathways covers, the kind of biological entities that constitutes them are highly diverse and ranges from metal ion, protein to biological processes in general. In addition, the kinds of interactions that relate biological entities are likewise diverse. As a consequence, the current knowledge about pathways is very heterogeneous both in the sense of the types of constituents and the granularity of descriptions. To cope with this problem, a strongly structured ontology-aware knowledge system is required. This paper describes a biological pathway database system for higher order knowledge. The proposed database adopted a recursive and hierarchical representation model which enables to annotate and query pathways or sub-pathways of arbitral granularities.

1. 背　　景

様々な実験技術の革新によって，生物学者は今まで以上に様々なタンパク質の機能を把握する必要に迫られている．特に近年重要視されているのは，生体高分子単体が担っている機能の情報ではなく，それらが互いにどのように関わりあって生命現象の分子機序を成り立たせているかを表す “関係の組合せ” から成り立つ高次知識（パスウェイ）である．これらの知識は文献を通して共有されており，そのデータベース化が知識基盤として果たす役割は大きい．文献で共有されたパスウェイ知識は，化合物，タンパク質，金属イオン，などの物理的実体をともなう事柄だけでなく，“細胞死” などの抽象的な生命現象や概念を構成要素とし，また，要素間の関係も “輸送”，“制御”，“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，事柄の抽象度に柔軟に対応できる記述形式と記述された各事柄を意味づけるオントロジを強く意識した知識システムが不可欠となる．

本稿では我々の開発した高次知識データベースシステムについて報告する．我々は上述の問題を解決するために，階層的な知識表現に一群の生物学オントロジを組み合わせパスウェイデータベースを開発した．格納される情報は生物学者が実際に論文を精読して抽出し，ユーザはこの情報を Web アプリケーションで実装された検索インタフェースを介して検索することができる．

本報告の構成は以下のとおりである．2 章で生体内パスウェイ知識の特徴を概観し，計算可能な形式に変

† 産業技術総合研究所生命情報科学研究センター
CBRC, AIST
†† 科学技術振興機構バイオインフォマティクス推進センター
BIRD, JST
††† 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
Graduate School of Frontier Sciences, The University of Tokyo.

換する際の課題点を指摘する．3 章では我々の採用した知識表現手法および文献からの知識獲得について紹介する．4 章では知識表現とオントロジの対によって実現される検索機能を説明し，我々の開発したシステムについて述べる．5 章で関連研究に言及し，6 章，7 章で本報告のまとめと今後の課題を論じる．

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．

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．この定義は非常に広範であり，シミュレーションのための生化学反応レベルの解像度を持つシステム記述からより抽象的で定性的な分子機序に関する知識構造，さらには文献内の単語の共起頻度解析に基づく近似的な知識構造の記述までが含まれる．

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る．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，細胞応答などの生命現象の因果関係まで記述する必要がある．各生命現象はそれ自身のメカニズムを記述したパスウェイを持つことから，シグナル伝達パスウェイではパスウェイの中に再びパスウェイが現れる再帰的な構造が形成される．

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 2 本の実線で表されており，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酵母のメイティング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．

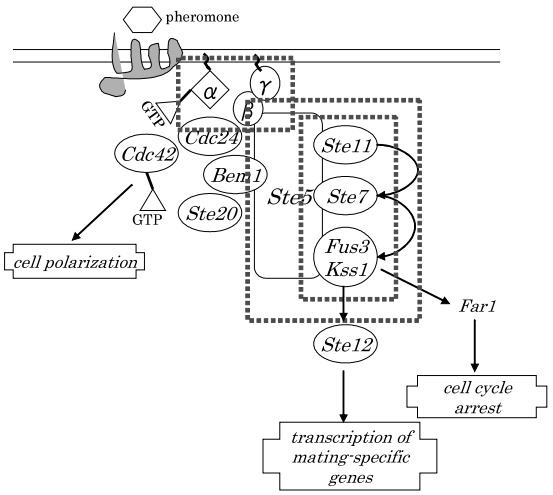

図 1 酵母のメイティングシグナルの伝達例．フェロモンが 7 回膜貫通型受容体（G-タンパク質共役受容体，GPCR）に結合した際の刺激が下流に伝達されている
Fig. 1 Mating signal transduction in yeast.

第 1 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．生物学者は菱形 α と 2 つの小さな円 β，γ が G-タンパク質のサブユニットであることを背景知識から知っている．そのため，図 1 をみると即座に，フェロモンが G-タンパク質共役受容体に結合した後に G-タンパク質の $\beta\gamma$ サブユニットが α サブユニットからリリースされてタンパク質 *ste5* に結合することを理解する．*Ste11*，*Ste7*，*Fus3/Kss1* は “MAPK cascade” と呼ばれる有名なシグナルのカスケードを形成しており，MAP キナーゼ *Fus3/Kss1* が MAP キナーゼ・キナーゼ *Ste7* によってリン酸化されている．*Ste7* は MAP キナーゼ・キナーゼ・キナーゼ *Ste11* によってリン酸化されている．シグナルはこの一連のリン酸化反応によって *Ste11* から *Fus3* に伝達される．また，*Ste5* はこのリレーシステムを実現するための物理的な “足場” 構造を形成することが知られている．このように明示されない部分構造には，タンパク質の複合体とその構成要素のように物理的な実体に言及する際の抽象度の違いを表すものと，パスウェイを構成するプロセス–サブプロセスを表すものがある．点線による矩形はこれらの暗にしか言及されない部分構造を示している．

第 2 は構成要素の不均一な記述粒度である．G-タンパク質の表現には各サブユニットごとに独立のオブジェクトを用意している一方で他のタンパク質はそれぞれ 1 つのオブジェクトで表現されている．*Ste11* からの矢印は触媒反応（リン酸化）を表しているが *Ste12* からの矢印は活性型のタンパク質と抽象化された現象

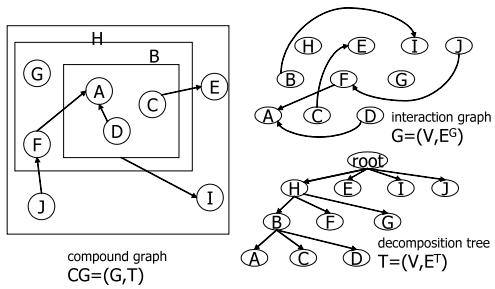

図 2 複合グラフに基づくパスウェイ表現
Fig. 2 Compound graph based pathway representation.

の間の因果関係を表現している．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し，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．このため，同列に記述されるダイアグラム図内の個々のオブジェクトがまったく異種の概念を表していることを判別できる仕組みが必要となる．

第 3 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，実際には相互作用の結合相手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による意味づけが必要となる．

3. パスウェイ知識の形式化

3.1 複合グラフ

提案システムでは複合グラフ（compound graph）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持ち，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．本稿ではグラフ G を相互作用グラフ（interaction graph），木 T を分解木（decomposition tree）と呼ぶ．同様に，エッジ $e_I^G \in E^G$ を相互作用エッジ（interaction edge）と呼び，エッジ $e_I^T \in E^T$ を分解エッジ（decomposition edge）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．

3.2 パスウェイの複合グラフ表現

複合グラフのすべてのオブジェクト（ノード，相互作用エッジ，分解木エッジ）には，それぞれ何を表すかによってタイプが存在する．たとえば，ノードにはタンパク質ノード，DNA ノード，プロセスノードなどがあり，相互作用エッジには輸送エッジ（Transport），制御エッジ（Modulate）などがある．分解エッジも同様に状態への分解（has_state），プロセスのメンバへの分解（has_member），分子のサブユニットへの分解（has_component）などのタイプを持つ．各オブジェクトが何種類のタイプを持つべきかは，データ入力の煩雑さが軽減されなおかつ必要な知識を記述するのに十分かどうかを指標にして，生物学者との議論を通して決定した．

階層的なグラフ表現を採用することにより，パスウェイを部分構造に従って階層的に “分かち書き” して暗に埋め込まれている部分構造に関する知識を明示的に記述することが可能になる．複合グラフでは，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，各部分構造の階層を下るごとに記述の粒度が細かくなっていくことで多段階の抽象化レベルを明示的に表記の中に埋め込むことが可能になる．たとえば，G-タンパク質を表すタンパク質ノードが分解木エッジ has_component によってアルファ，ベータ，ガンマを構成要素に持つことになる．同様に，プロセスノードの下位階層は has_member 分解木エッジによって該当プロセスのメンバを規定する．該当プロセスのメカニズムはプロセスのメンバ間に張られる相互作用エッジによって定義される．逆に，プロセスのメンバのみが明らかで，実際の相互作用のメカニズムが不明な不完全な知識を記述する場合には，分解木エッジによってメンバに関する情報だけを記述すればよい．

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 3**）．また，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の 1 つに結合することが可能である．

3.3 オントロジによる構造のアノテーション

各ノードはノードタイプに従って，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

☆ 入れ子グラフでは分解木の兄弟関係のノードでしか相互作用エッジが定義できない．クラスタグラフでは分解木の葉間でしか相互作用エッジが定義されない．

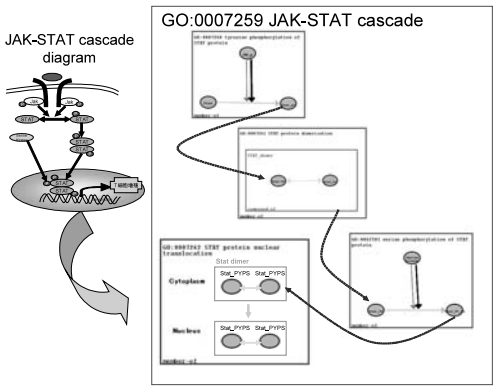

図 3 プロセスのダイアグラム図と複合グラフへの変換．各サブプロセス（矩形）は生体内プロセスオントロジで意味づけされている．それぞれの矩形は異なる粒度のプロセスを表している
Fig. 3 A process diagram and its compound graph representation.

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へのリンクが格納される．ノードと同様に各エッジについてもエッジタイプごとにオントロジで規定された情報を付加する．

提案システムはパスウェイ情報のアノテーションのために以下を含む 12 のオントロジもしくはデータベースをシステム内に持っている☆：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，生体高分子 DB，化合物 DB，リアクション，参考文献，など．パスウェイ表現を構成するすべてのオブジェクトをこれらのオントロジで規定することでオブジェクトの表現する内容が区別可能になる．

さらに，オントロジを使用することで検索対象概念の表記の揺れの吸収および検索対象概念の緩和検索が可能になる．緩和検索については 4.4 節で述べる．バイオロジの分野では概念名称の表記の揺れの問題は古くから知られ大きな問題である[2]．特にタンパク質名の表記のばらつき，シノニムの多様性はバイオロジにおける自然言語処理技術の応用では大きな課題となっている．パスウェイデータとは別にオントロジを用意しオントロジ内でシノニム情報を管理することで，データ登録時の表記のばらつきへの対応と検索時のシノニム情報による検索意図の拡張が可能になる．これらの機能はパスウェイの中で直接各オブジェクトの意味を規定した場合には実現困難である．

☆ FREX では，オントロジは 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である．

3.4 パスウェイデータの取得

我々は高次知識処理システムの開発だけでなく，パスウェイに関する知識の収集も行っている．特に，専門家が実際に論文を読んで理解しなければ抽出できない高次知識の蓄積に重点を置いており，物質間の単純な二項関係情報を収集する試みとは異なる．

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．知識の蓄積プロセスは 2 段階に分けており，第 1 段階目ではレビュー記事からパスウェイ情報を抽出する．通常，レビュー記事は複数の論文を著者が再構成しているので，ある程度まとまった知識が記述されている．典型的には，70 程度のノード数と 5 階層程度の分解木から構成される．これらの知識の更新頻度は比較的低い．第 2 段階目では通常の論文から知識を収集する．これらの論文は典型的には，レビューで紹介されたパスウェイ構造を数ステップ伸張させたことを報告している．これらの小規模のデータの更新頻度は比較的高いことが予想される．現在は，レビューに基づく知識の収集に取り組んでいるところである．

4. FREX システム

我々の構築しているシステム FREX（Functional Relation EXplorer）は分子生物学データベースであり，パスウェイとして計算可能な形式で蓄積された文献中の高次知識を検索するためのシステムである．

システムは XML データベースとして実装されている．検索処理サーバ，XML ミドルウェアとリレーショナル DBMS の 3 層からなり，ブラウザからの検索問合せを受け取った検索処理サーバは 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（または MySQL，Oracle）上に用意したテーブルに分解されて格納されている．

XML データベース上にはパスウェイ情報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．

4.1 FREX の検索機構

前述のようにオントロジ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すべてのオブジェクトに関する様々な項目を条件として指定しながら検索をすることができる．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** (a)）．たとえば，“IL-13 と IL-4 がともに関わっ

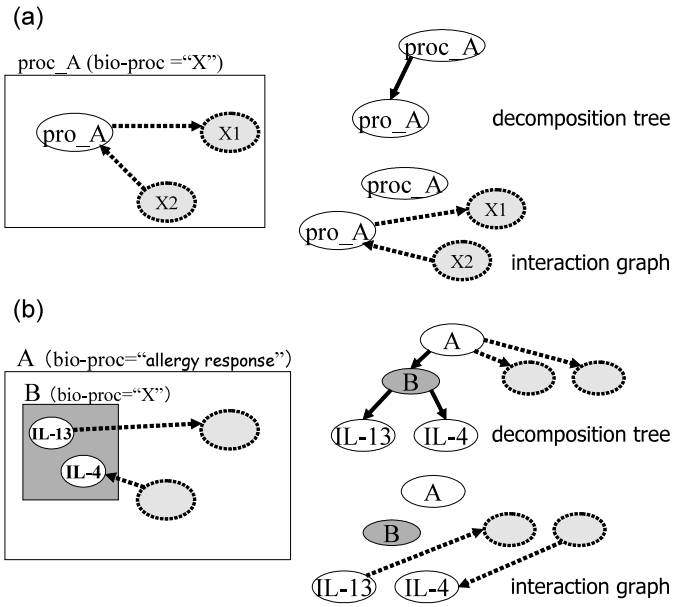

図 4 複合グラフ上での検索要求の処理方法．(a) タンパク質 A と相互作用する分子は何か？（X1 と X2）タンパク質 A を含む生体内プロセスは何か？（オントロジ ID bio-proc=X を持つ proc_A）；(b) アレルギー反応のパスウェイで IL-13 と IL-4 が関わる部分パスウェイは何か？（オントロジ ID bio-proc=X を持つノード B）

Fig. 4 Biological queries on compound graph structures.

表 1 検索要求の例．ノード，相互作用エッジ，分解木エッジの属性値を組み合わせることで，例示された自然言語による概念指定と同等の検索が可能になる

Table 1 Examples of biological queries.

- IL-1とNFkB(molecule)の関わるパスウェイはあるか？
- コレラ毒素の下痢(phenotype)にいたるパスウェイは？
- RTKからアダプタータンパク(sig-family)を介するパスウェイは？
- SRE配列(DNA)に結合(reaction)することにより転写を調節するプロセスを示せ
- 2002年の Li らによる Nfkb のreview(PMID)に関するパスウェイを示せ
- シロイヌナズナ(species) でサーカディアンリズム(sig-passing)の関わるパスウェイはあるか？
- MEKファミリー(sig-family)でSer218のリン酸化により活性化される分子とその分子の関わるパスウェイを示せ
- Perタンパク質(source node)と二量体(reaction)になる分子をピックアップせよ
- 視交叉上核(tissue)で薬剤カルバコール(chemical)はどの分子を活性化することが知られているか？
- 視交叉上核(tissue)において、サーカディアンリズムにおけるコリン作用薬(chemicals)の作用機序を示せ

ているアレルギー反応を担うサブプロセス（部分パスウェイ）があるか” という検索は，“IL-13 という名称のタンパク質ノードと IL-14 という名称のタンパク質ノードを持ち，かつプロセスノードが生体プロセス反応としてアレルギー反応もしくはその下位の概念を指しているパスウェイを返せ” という処理に変換される（図 4 (b)）．**表 1** は各オブジェクトの属性を組み合わせることで実現される検索内容を自然言語で記述した例である．

4.2 FREX インタフェース

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．最大で 15 の属性とそれぞれの値の対を指定できる．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，オントロジ情報に基づいた検索を実現していることである．

4.3 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

☆ プレリミナリーなバージョンは http://www.ontology.jp/FREX/index.html

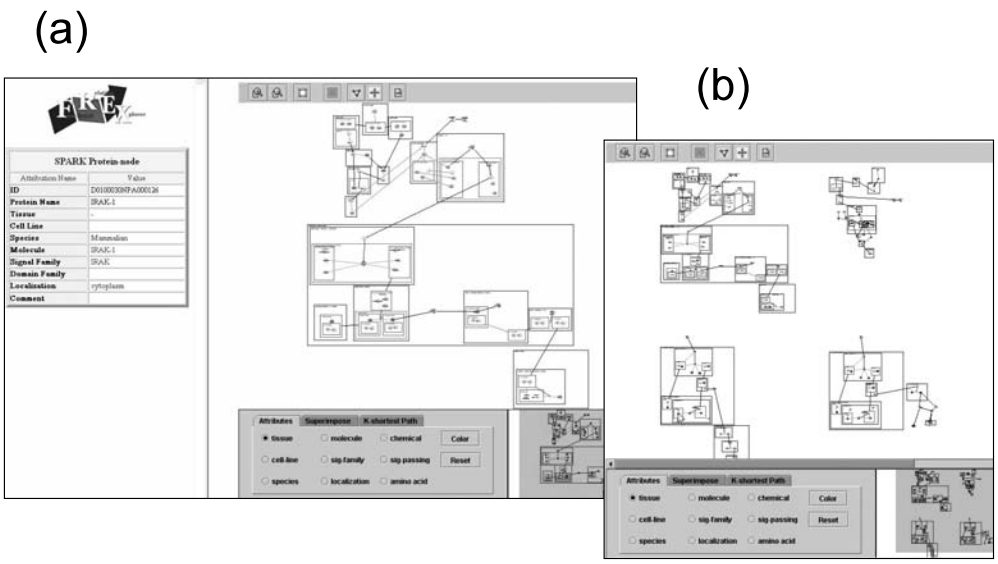

図 5 パスウェイ検索結果の表示．(a) 単一パスウェイ，(b) 4 つのパスウェイの同時表示
Fig. 5 Query results of a diagram search.

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 2 通りに表示できる．1 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（たとえば，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）．

4.4 オントロジを使用した緩和検索

本報告でユーザインタフェースの詳細を割愛している重要な機能の 1 つはオントロジに基づいた検索条件緩和である．これは，検索で指定された各項目について該当するオントロジを走査し，入力概念クラスの下位概念，兄弟概念などに検索対象を広げる機能である．検索対象は，あらかじめ指定したステップ数まで緩和することができる．緩和に関するパラメータはシステム指定のデフォルト値とユーザ指定の値のどちらかを選択する．

5. 関 連 研 究

パスウェイデータベースとしては KEGG[4]，EcoCyc[5] などの研究がある．KEGG は人間の描いた図を使用したクリッカブルマップなので本システムとは趣が異なる．EcoCyc（もしくは BioCyc☆）はフレームシステム上に構築された詳細なオントロジ定義を持った代謝パスウェイデータベースである．しかし，いずれも代謝パスウェイなどのすでに枯れた知識のデータベース化に重きが置かれている．

一方で，本報告で述べてきたシステムは論文中に記載されたパスウェイ知識一般を計算機上に表現することを目指しており，前述したように部分構造の明示化，不均一粒度の問題や不完全な知識の記述の問題などを扱う必要がある．

Gene Ontology（GO）[6] は生物学分野で非常に重要な知識基盤である．GO は主に is-a 関係で定義された統制語句の体系である．分子機能，生体内プロセス，細胞内局在部位の 3 つについての概念体系を提供しており，主立ったモデル生物種のゲノムデータベースは GO の語彙を使用して機能情報をアノテーションしている．

しかしながら，これらの概念体系は個々の機能，プロセスのメカニズムに関する知識は提供しない．我々のデータベースはメカニズムの構造情報をオントロジでアノテーションした知識を提供することで構造情報とオントロジの間のギャップを埋める役割を果たしている．

6. 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．

提案システム FREX 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すべてのオブジェクトをオントロジで意味づけている．こ

☆ http://www.biocyc.org

の結果，(1) 文献内に散見される不完全な知識の記述に構造を与え計算可能にした．(2) 不均一な粒度の記述から構成される知識を統一的に扱うことを可能にした．(3) 階層的な構造により，部分パスウェイに関する知識を明示化し検索可能にした．(4) 検索時に複数の属性とそれらが各々満たすべきオントロジ概念を指定することで，柔軟で強力な検索機能を実現した．

ユーザが入力した文字列の意図をシステムが理解する術を持っている点でオントロジベースの検索機能は従来のキーワードベースの検索と決定的に異なる．生体内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をデータベース化する際には設計段階からオントロジを強く意識したシステムにすることが重要である．

実験生物学者のワークベンチ的なユースケースを考えた場合，ユーザインタフェースおよび計算量の双方の観点から対象知識の範囲を指定できることが望まれるが，我々の提唱する知識表現手法は任意の大きさに部分構造を指定できる良い性質を備えている．また本報告では触れなかったが，パスウェイを構成するエッジには，他のエッジとの依存関係などの制約情報も含まれており，宣言的な記述形式を採用していることと合わせて，ユーザの思考過程を支援する様々な論理学ベースの推論手法が摘要可能であると考えている．

パスウェイデータベースのデータ更新は難しい問題である．理由は 2 つある．第 1 に，分子生物学論文は研究者の主観をともなった実験情報の解釈結果の記述であり，異なる論文間で内容の無矛盾性が保証されない．第 2 に，細胞のメカニズムには排他的な制御が含まれるために，知識の単調な増加が望めない．実際のデータ更新に際しては 2 つの状況が想定される．第 1 の状況は，既存知識の詳細化である．このようなデータ更新は既存パスウェイにおいて分解木の葉にあたるノードの下に新しい木を挿入する操作で対応できる．第 2 の状況は，既存の知識との整合性がとれない修正をともなうデータ更新である．この場合，データ更新には上述のような本質的な難しさがともなう．このため，データベース自体には矛盾する複数の知識の格納を認め，別途矛盾検出用の推論エンジンを構築することが望まれる．

一般に，データの収集は非常にコストのかかる作業であるため，自然言語処理技術を応用した支援システムの構築が望まれる．しかしながら，まだ難しい課題が残されている．高次知識の収集に際しては，専門用語同定，用語のクラス分類，専門用語と既存の生物学データベース間の対応関係の同定，単純な相互作用の同定，相互作用の組み合わさったプロセスの同定などの要素技術を確立する必要がある．我々は，収集に際して高度な専門性を必要とする高次知識を宣言的な形式で蓄積していくことで，生物学者だけでなく，バイオロジを対象に自然言語処理技術やその他の知識処理技術を開発しているグループにとっても有用な知識基盤を提供できると期待している．

7. 今後の課題

FREX はオントロジを用いた強力な検索機能を有している．しかしながら，現状のシステムは，あらかじめ登録された複雑な知識を検索・閲覧する機能を提供しているにすぎない．残念ながら記述された知識の更新，矛盾検出，パスウェイに関する帰納推論などのより知的な推論機能[7] は実現されていない．これらは今後の課題として残されている．

前述したように，本プロジェクトでは従来あまり形式化されてこなかった不均一な知識のデータベース化に重きを置いており，現在の FREX システムでは独自に収集したデータのみが検索対象になっている．一方で KEGG の代謝経路に関する知識のように，よく整理された公開データについては今後積極的にインポートしていくことが望まれる．

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．

謝辞 本研究は科学技術振興機構バイオインフォマティクス推進センター（BIRD，JST）および文部科学省科学研究費特定領域研究（C）“ゲノム情報科学”の支援を受けています．

参 考 文 献

1) Fukuda, K. and Takagi, T.: Knowledge representation of signal transduction pathways, *Bioinformatics*, Vol.17, pp.829–837 (2001).
2) Fukuda, K., Tsunoda, T., Tamura, A. and Takagi, T.: Toward information extraction: Identifying protein names from biological papers, *Pacific Symposium on Biocomputing 1998*, pp.707–718 (1998).
3) Eppstein, D.: Finding the k shortest paths, *SIAM J. Computing*, Vol.28, pp.652–673 (1998).

4) Kanehisa, M. and Goto, S.: Kegg: Kyoto encyclopedia of genes and genomes, *Nucl. Acids. Res.*, Vol.28, pp.27–30 (2000).
5) Karp, P., Riley, M., Paley, S., Pellegrini-Toole, A. and Krummenacker, M.: EcoCyc: Electronic encyclopedia of *e.coli* genes and metabolism, *Nucl. Acids. Res.*, Vol.27, pp.55–58 (1999).
6) The Gene Ontology Consortium: Gene Ontology: tool for the unification of biology, *Nature Genetics*, Vol.25, pp.25–29 (2000).
7) Fukuda, K. and Takagi, T.: Signal transduction pathways and logical inferences, *Proc. 2001 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences, METMBS '2001*, pp.297–303 (2001).

(平成 15 年 9 月 25 日受付)
(平成 16 年 1 月 19 日採録)

（担当編集委員　石川 博，市川 哲彦，原 隆浩，
佐藤 聡，土田 正士）

福田賢一郎（正会員）

2001 年東京大学大学院理学系研究科卒業．理学博士．同年より産業技術総合研究所生命情報科学研究センター研究員．生命科学知識の体系化を目指し，知識表現，オントロジー構築，文献からの知識抽出などの研究に従事．人工知能学会，日本バイオインフォマティクス学会，ACM，IEEE 各会員．

山縣　友紀

大阪大学大学院医学研究科修士課程終了．2001 年より科学技術振興機構バイオインフォマティクス推進センター（JST，BIRD）技術員．

高木　利久（正会員）

1976 年東京大学工学部計数工学科卒業．九州大学を経て，1992 年東京大学医科学研究所ヒトゲノム解析センター助教授．1994 年同教授．2003 年より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所属．工学博士．バイオインフォマティクスの研究に従事．