2D2-02

# バイオポータルプロジェクトにおけるオントロジ構築と利用について

Construction and Utilization of Ontologies in the BioPortal Project

荒木 次郎[*1] Jiro Araki　川本 祥子[*2] Shoko Kawamoto　小林 悟志[*2] Satoshi Kobayashi　水田 洋子[*2] Yoko Mizuta　出宮 スウェン・ミノル[*2] Sven Minoru Demiya
ムリアディ ヘンドリー[*2] Hendry Muljadi　白井 康之[*1] Yasuyuki Shirai　市吉 伸行[*1] Nobuyuki Ichiyoshi　伊藤 武彦[*1] Takehiko Ito　北本 朝展[*2] Asanobu Kitamoto
武田 英明[*2] Hideaki Takeda　藤山 秋佐夫[*2] Asao Fujiyama

[*1](株) 三菱総合研究所 Mitsubishi Research Institute Inc.　[*2]国立情報学研究所 National Institute of Infomatics(NII)

We constructed ontologies for several disciplines such as molecular biology, cell biology, and biochemistry. Further, we integrated these ontologies by linking associated tree nodes. Brief structures of ontologies are based on tables of contents of textbooks for each discipline. Detail structures are constructed according to term co-occurrence in MEDLINE.

## 1. はじめに

バイオサイエンス分野は、近年ヒトゲノム解読プロジェクトなどに代表されるように多くの網羅的解析プロジェクトが進められ、莫大なデータが生産されている。しかし、これらのデータはその意味が明確になり、データ間の関係性が判明しているとは限らず、生命の全体像を理解するための情報の断片にすぎない。バイオ研究者はこの莫大なデータの中から必要な情報を探し出し、断片間の関係をつなぎ合わせることに苦心している。

また世間ではニュースなどで BSE や鳥インフルエンザウイルス、再生医療などの話題が取り上げられることが多く、一般社会に最新研究を分かりやすく提供する必要性が高まっている。しかし、ニュースなどは個々の話題については比較的分かりやすく書かれているものの、その話題が生命全体のどこに位置付けられるのかが理解されておらず、その話題の重要性や関連事項が分かりにくい。

以上のような背景から、バイオポータルプロジェクトでは、オントロジを知識の基盤に位置付けて、文献情報や、データベース、ニュース記事などをオントロジ上に関連付けることにより、容易にこれらを検索することができるシステムの開発を行なっている。

本稿は、バイオサイエンス分野を適切にナビゲートするためのオントロジ構築とその利用方法について論ずる。

## 2. バイオポータルプロジェクト

バイオポータルプロジェクトは、主に以下の目標に沿って研究を進めている [藤山 04][川本 04]。

- 一般の方が最新のバイオサイエンス研究成果に興味をもち、より深い理解につながるように学習支援する仕組みを構築する
- 研究者がバイオサイエンスにおける莫大なデータの中から、必要な情報を容易に収集し、知識を発見することができるようにする

連絡先: 荒木次郎, (株) 三菱総合研究所 先端科学研究センター, jiro@mri.co.jp

そのため、1) 基盤となる用語辞書、オントロジの整備、2) オントロジを利用した検索システムの開発、3)DNA・アミノ酸配列データ解析 WWW サービスの開発、4) 最新研究などを紹介するコンテンツの提供、などを行なっている (図 1)。

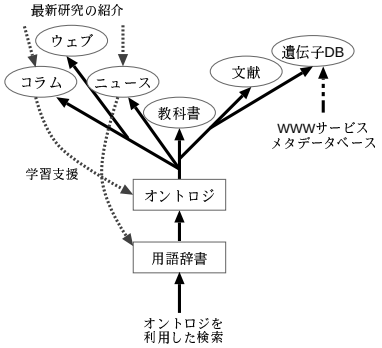


図 1: バイオポータル全体像

バイオポータル利用者は、用語辞書を利用して知らない用語の意味を調べたり、オントロジをブラウジングすることで関連用語を調べ、さらにはそれらの用語をキーワードにして文献や遺伝子データベースなどを検索することができる。またバイオサイエンス分野の最新研究を紹介するニュースやコラムに記述されている用語のリンクをたどることで、その研究の背景にある知識を学習することができる。

現在、既に完成している用語辞書や英文コンテンツの日本語キーワード検索などを試験公開している [BioPortal HP]。


図 2: バイオポータルホームページ

## 3. オントロジの構築

### 3.1 オントロジの要件

まずバイオポータルプロジェクトにおいて必要とされるオントロジの要件を検討する。

バイオポータルでは、研究者のみを対象とするのではなく、一般の方にもバイオサイエンスの最新研究を分かりやすく紹介し、それを機会にして興味をもって学習する仕組みを提供することを目的としている。そのため、オントロジで扱う知識は、研究者レベルから一般レベルまでをカバーする必要がある。

またバイオサイエンスが扱う分野は、分子生物学、細胞生物学、生化学など広範囲に渡る。これらの分野は、扱うスケールが異なるなど、同じ現象ではあるが異なる観点から説明している場合が多い。さらに、生物は個々の組織や現象が独立に機能しているのではなく、複雑に関係し合うことで機能するため、全体像を広く理解することが必要とされている。

ここまでは、オントロジを利用する側の視点からの考察であるが、次にオントロジを構築する側の視点から考えると、やはり構築する上での現実的な問題として、構築の手間という問題がある。特にバイオサイエンスは日々進歩しており、一度構築してもすぐに古くなるという問題もある。そのため、出来る限り構築の手間がかからない方法をとる必要がある。

以上をまとめると、オントロジ構築の要件は次のようになる。

1. 一般から研究者まで、また分子生物学、細胞生物学、生化学などさまざまな分野からアクセスでき、分野横断的に探索することができる。
2. できるだけ構築の手間がかからないように、既存のものを組合せたり、自動生成する。

### 3.2 オントロジ構築手順

前節のような要件を満たすオントロジとして、我々は次のようにオントロジを構築した (図 3)。

1. オントロジのおおまかなツリー構造として、教科書の目次構造を利用する [おおまかな構造化]。
2. ツリーの枝には、その枝に対応する教科書の章/節に含まれるインデックス (索引) をぶら下げる。
3. インデックス間は、医学文献データベース MEDLINE の用語共起性を利用して関係性を自動生成する。
4. また教科書インデックス以外の共起用語をオントロジの枝に追加することで、教科書レベルの用語と専門用語とを関係付ける。(同様のことを、ニュース記事などの一般向けコンテンツに適用することで、教科書レベルの用語と一般レベルの用語あるいは最新用語との関係付けが行なえる。)
5. 教科書には分子生物学、細胞生物学など複数の分野の教科書を用い、各々のツリーを構成する。ツリー間あるいはツリー内の枝には、インデックスの共通性から関連性の深さを計り、リンク関係を結ぶ。

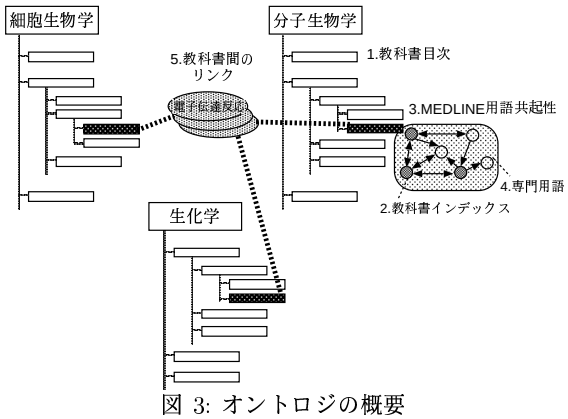

図 3: オントロジの概要

現在、次の 3 つの教科書をオントロジ化している。

- ウォーレスの現代生物学 (Biosphere)
- 細胞の分子生物学 (Molecular Biology of the Cell)
- イラストレイテッド ハーパー生化学 (Harper's Illustrated Biochemistry)

それぞれの教科書の階層別章/節数を表 1 に、インデックス数を図 4 にまとめる。なお、インデックスは教科書ごとの表記ゆれを手作業で修正し統一した。

表 1: 各教科書の階層別 章/節数

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 現代生物学 | 7 | 48 | 401 | 399 |
| 分子生物学 | 5 | 25 | 111 | 1094 |
| 生化学 | 7 | 54 | 462 | - |

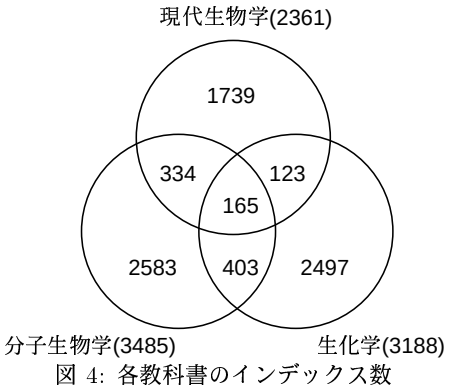

図 4: 各教科書のインデックス数

次に、MEDLINE(1991～2003 年) 約 400 万文献のアブストラクト中での用語共起性をもとに、教科書インデックス及び専門用語間の関連性の自動生成結果をまとめる。

用語関連度の評価には、Jaccard 係数を用いた。

$$Jaccard\ 係数\ (用語\ 1,2) \equiv \frac{用語\ 1,2\ 両方を含む文献数}{用語\ 1,2\ いずれかを含む文献数}$$

Jaccard 係数の閾値を下げればより多くの関連用語をつなぐことが可能であるが、意味的に関連のない用語が誤ってつなが

る可能性があり、また複雑になり過ぎることから、ある程度信頼性が保て、用語数の限定できる 0.1 を閾値とした。

その結果、教科書インデックス 7844 個のうち、文献中に含まれ、かつ他の用語と関連性があるインデックスは 3004 個あった。さらにこれらのインデックスと直接あるいは他の用語を介して間接的に関連のある専門用語を 8305 個抽出することができた。

図 5 はこれらの用語の共起ネットワークの抜粋である。色の付いたノードが教科書インデックスを表し、それ以外は新たに文献中から抽出された用語である。癌に関連する用語がうまく関連付けられ、また教科書では扱われていない用語も抽出することができていることが分かる。

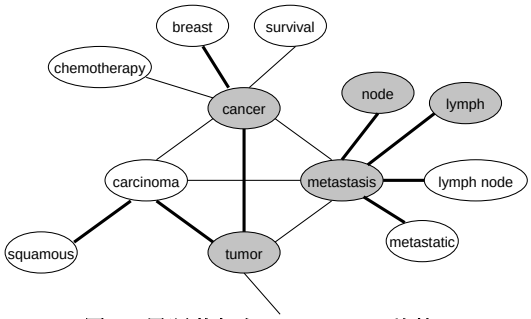


図 5: 用語共起ネットワーク (抜粋)

## 4. オントロジの利用

構築したオントロジは主に文献検索などの検索ナビゲーションに利用することを想定している。検索ナビゲーションには、以下の 2 通りがあり得る。

- 検索語の具象化
  漠然とあるテーマ (例えば「光合成」) に関連する文献を調べたい場合や、適切なキーワードが思い浮かばない場合に、オントロジの階層を下ることにより、テーマを絞り込んだり、具体的なキーワード (例えば「電子伝達反応」) を選択することができる。

- 検索語の抽象化
  分野さえ不明な用語を調べたい場合に、階層を昇ることでその用語がどのテーマに関連するかを調べ、そのテーマに含まれる別の関連用語を調べたりすることができる。

このようなユーザの検索動作を検討するために、現在図 6 に示すようなユーザインタフェースを作成し、バイオ研究者に試行してもらっている。

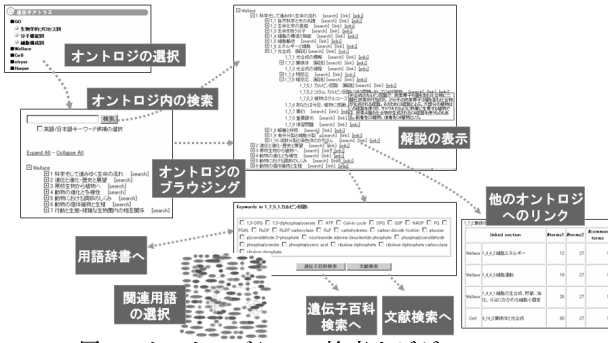


図 6: オントロジによる検索ナビゲーション

## 5. まとめ

本研究では、研究者がバイオサイエンス分野の莫大なデータの中から必要な情報を容易に収集すること、一般の方がバイオサイエンス分野の用語などを検索し容易に学習することができるように、バイオサイエンス分野を広くカバーするオントロジを構築した。このオントロジは、分子生物学、細胞生物学、生化学などの異なる分野のオントロジから構成され、分野内だけでなく分野横断的に、検索語に具象化や抽象化が可能となっている。

今後は、バイオ研究者の協力のもと、この分野特有の問題を洗い出し、オントロジを利用した検索方法のチューニングを行なっていきたいと考えている。

## 謝辞

本研究は、文部科学省科学技術振興調整費による「次世代バイオポータルの開発研究」の一環として行なわれたものである。

## 参考文献

[BioPortal HP] http://www.bioportal.jp/

[藤山 04] 藤山, 他: 日本語によるバイオ情報利用システム-バイオポータル-の開発, 2004 年度日本分子生物学会年会 2PB-300, 2004.

[川本 04] 川本, 他: バイオポータルプロジェクトにおける日本語専門用語辞書及びオントロジの構築と利用について, 2004 年度日本分子生物学会年会 2PB-301, 2004.