社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2005－DBS－136 (5)
2005－ＦＩ－ 79 (5)
2005／5／19

バイオインフォマティックスにおける最近のオントロジーの研究開発について

高井貴子
東京大学大学院情報理工学系研究科
taka@bi.is.s.u-tokyo.ac.jp

発表要旨

生物学のゴールは，生物を分子の協調システムとして計算機上に再構築することであり，その基盤となるオントロジー開発を紹介する.

生物学の機能概念オントロジー

高井貴子

(人工知能学会誌 (2004) 19(2), 137-143. に掲載)

1. 生物学でオントロジーが必要とされている

1.1 Conceptual Biology

最近生物学では「Conceptual Biology が必要だ」といわれている[Blagosklonny 02]。生物学者はこれまで概念を生成してきた。しかしこれからの生物学者は「概念を解析する」能力が必要とされる、という提唱である。ここに象徴されるように、生物学における概念の処理について注目が集まっている。本稿で解説するGene Ontologyは、生物学の概念を対象としたオントロジーである。2000 年に初版が発表され、現在では世界中の生物学者と情報科学者に広く利用されている。わずか 3 年間の間に関係論文は 100 報を超えた。

1.2 生物学と概念

生物学は、記述的学問であると言われる。対象(生物)を観測し、その結果を解釈して概念化し、記述に残す。その記述が学術文献として蓄積される。すなわち生物学者の仕事は新しい概念を生成することであると言える。生物学の学問領域には、これまで蓄積された膨大な概念が存在する。生物学者とは、この膨大な概念を知識として身につけており、そこから独自の仮説を構築することができ、その仮説を支持する実験を立案し、実施することができる専門家である。仮説生成に利用できる概念をいかに豊富に利用性の高い形で身につけているかが、生物学研究における鍵となる。

生物学は、Conceptual Biology というまでもなく、もともと概念の解析と生成を繰り返しながら発展してきた学問である。概念を解析する能力は生物学者たる必要条件であり、採り上げるまでもない当然の行為であった。しかし状況の変化をもたらしたものは、ゲノム研究であった。

1.3 ゲノム研究

生物学の目標は、生物がなぜ生きることができるのか、生命の原理とはなにか、という問いに答えることである。生命の原理を親から子へ伝える実体がゲノムである。猫の子は猫となり、猫から鳩は生まれない。また猫の子は、生まれながらにして猫として生きるために必要な性質を身に付けている。猫が猫であるための情報は、親猫から受け継いだものだ。生物は遺伝情報により設計されるが、その遺伝情報全体をゲノムと呼ぶ。ゲノムはゲノムの複製という形で、その生物の性質を決める情報の一式を親から子へ伝える。ゲノムは DNA という化合物に姿を変えて、生物の細胞内に存在する。DNA の並び順を調べて、そこに暗号化された遺伝情報を読み解く研究をゲノム研究と呼ぶ。ゲノム研究は生物学の部分領域であるが、現在の生物学研究課題の主流をなしている。

ゲノム研究の特徴は、遺伝情報全体を対象とすることである。それ以前の生物学では、局所的な遺伝情報が主な研究対象であったので、一度に処理する概念は通常 2 桁程度のオーダであった。しかし遺伝情報全体を対象にしようとした途端、一度に処理する概念のオーダは 4 桁から 5 桁に跳ね上がり、人間の処理では追いつかなくなり、計算機の助けが必要となった。これが Conceptual Biology が提唱された背景である。

ゲノムの遺伝情報は、タンパク質という機能をもつ遺伝子産物の形に変換される。この変換プロセスは遺伝子発現と呼ばれる。個々の遺伝子発現は、時空間的に高度に制御されており、個体の形成と自己組織化を実現している。遺伝子発現の際に、遺伝子産物を作る鋳型として使われる DNA 領域を遺伝子と呼ぶ。ゲノムの中で遺伝子である DNA 領域を発見する問題は、完全ではないものの、大方解決している。遺伝子の数はヒトやマウスで 3 万個、ハエで 1.4 万個、酵母では 0.6 万個と見積もられた。ゲノム研究における次の課題は、各々の生物が持つ遺伝子一式がどのようにしてその生物の機能を決めているのか、という問題を解明することである。

ヒトのゲノムは 30 億個の DNA から成るが、そのうち遺伝子である領域は僅か 5%でしかない。その上ひとつの遺伝子は数個の断片に分かれてゲノム上に存在している。そのため、遺伝子からタンパク質を作る前に、ゲノム上の遺伝子の部分のみをコピーして断片を張り合わせ、鋳型を完成する作業が必要となる。完成された鋳型は mRNA と呼ばれる。mRNA を鋳型にタンパク質が作られる。近年ゲノム研究では、遺伝子発現量、すなわち細胞内の mRNA やタンパク質の存在量を、網羅的かつ組織的に観測し、それが生物の機能をどう決めているのか解析しようとする研究が進んでいる。遺伝子の発現量は、数値で表現されるので計算可能である。一方、生物の機能は概念であり、そのままでは計算不可能である。生物学の概念を計算可能に変える必要が生じる。これが Gene Ontology の開発目的である。

2 生物学における概念とそのオントロジー

2.1 生物学の概念

生物学ではどのような概念が使われているのであろうか？答えは人工物と何ら変わらない。構造物の概念と、その機能の概念である。生物学における構造物概念は、組織、器官、細胞、細胞内器官、分子といったレベルの階層構造をもつ。生物学における機能概念も、構造物概念と同様の階層構造をもつ。なお人工物と生物の違いは、前者では設計原理の存在が前提であるのに対し、後者では設計原理の解明が学問の目的であり、いまのところ設計原理は部分的にしか解明されていないことである。

2.2 生物学で開発されたオントロジー

表1に、生物学の概念を対象として開発されたオントロジーについてまとめた。Mouse Anatomical Ontology[Blake 03]、Digital Anatomist Foundational Model[Noy 02]、CYTOMER[Matys 03]は、構造物の概念を対象としたオントロジーである。その他は、機能概念に重点を置いたオントロジーである。Gene Ontology[Consortium 01]が機能全般を対象としているのに対し、IMGT[Lefranc 03]は免疫機能、INTERACTION Ontology[Karp 00] は代謝機能、PharmGKB[Oliver 02]は医薬品の生体作用機能にドメインを限定している。TAMBIS[Stevens 00]は、Gene Ontologyと同様に機能全般を対象としたオントロジーであるが、Gene Ontology が抽象度の低い下位概念に焦点を当てているのに対し、TAMBIS は上位概念に焦点を当てたトップオントロジーである。TAMBIS では has-Component、is-Structure-of とい

った生物学における関係概念について解析し、その制約について定義している。その他のオントロジーでは、概念間の関係を is-a 関係と part-of 関係に限定している。

UMLS[Humphreys 93]およびそのサブセットである MeSH は、医学生物学の学術文献の抄録データベースである Medline のキーワード・インデックスである。Medline は古く 1950 年代から整備され、現在 1,400 万件の抄録を格納している。Medline の Web 版は PubMed と呼ばれる。

2.3 Gene Ontology

Gene Ontology[Consortium '01]の開発目的は、遺伝子産物(タンパク質)の機能の記述に必要となる概念を定義することである。真核生物に共通な機能を中心に、できるだけ多くの機能を対象とすることを目指している。Gene Ontology は Molecular Function 、Biological Process 、Cellular Component の 3 種のオントロジーの集合体となっている。Molecular Function は遺伝子産物単体の機能のオントロジー、Biological Process は複数の Molecular Function で構成される上位の機能のオントロジー、Cellular Component は細胞を構成する構造物のオントロジーである。

Gene Ontology では、すべての概念は遺伝子産物との関係を基準として定義されている。この点が Gene Ontology の特徴であり、一般のオントロジーと異なることに注意しなくてはならない。Gene Ontology における概念間の関係の定義は、概念の本質属性という観点から考察されたものではなく、「下位の概念に関係付けられている遺伝子産物は、上位の概念においても同様に関係付けられなければならない」という規則に基づくものである。この規則は true path rule と呼ばれ、Gene Ontology のすべての概念に適用されている。 また Gene Ontology の 3 種のオントロジーは独立しており、3 種のオントロジー間における概念の関係付けはなされていない。オントロジー間を関係付けるものは遺伝子産物であるよう設計されている。

概念と遺伝子産物の関係は"Associated Gene"関係として定義されている。この関係の意味をもう少し明確にすると、Molecular Function と Biological Process に格納されている機能概念に対する遺伝子産物の関係は「主体」である。Cellular Component に格納されている構造物概念に対する遺伝子産物の関係は「位置する」である。ここで注意しなければならないことは、機能の主体となったり構造物に位置したりする実体は遺伝子産物(タンパク質)であるのだが、"Associated Gene"関係で関係付けられている情報は「遺伝子」の情報である点だ。生物学では、遺伝子産物の機能は同時に遺伝子の機能として捉えられており、かつ遺伝子が機能を決定する一義的な因子であるという考えが主流となっていることから、遺伝子と機能が、遺伝子産物を飛び越して、直接結びついて論じられることが少なくない。これは生物学における暗黙の了解となっている。既存のデータベースにおいても機能は遺伝子に関係付けられている例が多いので、この慣行に従った方が既存のデータベースとの交換性からも便利である。以降この解説においても、「遺伝子の機能」という表現を「遺伝子産物の機能」と同じ意味で用いることにする。

Gene Ontology が格納する概念数は 16,600 を超えたところである(2003 年 11 月 24 日版)。Gene Ontology は生物の遺伝子をどの程度カバーしているのであろうか？真核生物の中で最も単純な生物のひとつである酵母については、全遺伝子(0.6 万個)がカバーされている。ハエでは 1.4 万個のうち 60%の遺伝子、マウスでは 3 万個のうち 40%強の遺伝子がカバーされている。

2.4 Gene Ontology が提供する情報

2.4.1 概念の定義

Gene Ontology において概念は、概念名、ID、定義、別名、概念間の関係、を定めることにより規定されている(図 1A)。概念定義は自然言語で記述されており、論理表現を用いた内部構造の定義はなされていない。概念間の関係は is-a 関係と part-of 関係に限定されている。推移律は is-a 関係と part-of 関係のいずれについても成り立つとしている。その他の論理的制約は定義されていない。したがって、Gene Ontology の概念を推論手続きで参照するのは便利ではない。概念の参照には概念の ID を用いるか、あるいは概念名と定義に対するテキスト検索が用いられる(http://www.godatabase.org/cgi-bin/go.cgi 参照)。その意味で Gene Ontology は、概念を形式的に定義するというより、データとして格納するに近い。

概念の階層関係は DAG(有向非循環グラフ)構造であるよう制限されている。概念階層木は、Molecular Function、Biological Process、Cellular Component 各々にひとつずつ構築されている。is-a 階層木と part-of 階層木に分かれておらず、ひとつの階層木に両方の関係が格納されている。そのため is-a 関係が定義されていない概念が存在する。1 割近い概念が is-a 関係を欠いている[Yeh 03]。概念階層木では多重継承が許されているが、前述の true path rule が守られることを前提としている。

2.4.2 付属する情報

遺伝子は機能以外に配列、構造、発現量、疾病の原因といった様々な属性を持つ。これらの属性を対象としたデータベースが既に多数構築されているので、それらとのデータ交換性をサポートするために、Gene Ontology では遺伝子を仲介としたインデックスファイルを提供している。現在 15 の主要なデータベースに対してインデックスファイルを提供している。

遺伝子に関するもう一つの重要な情報は、その遺伝子を報告した文献の情報である。Gene Ontology における遺伝子産物と機能の関係付けも文献に基づいて行われており、その文献情報は Gene Ontology に格納されている(図 1B)。文献情報は、生物学者が Gene Ontology の内容を確認したい時やより詳しい情報を調べたい時に役立つ。しかし文献情報のより強力な利用法は、自然言語処理の技術を用いた、文献から遺伝子と機能の関係を機械抽出するツールの開発である。文献情報は機械学習の正解集合を与える。情報抽出については、「3.3 機能の自動アノテーション」の章で詳しく紹介する。

最後に evidence code について説明する。Gene Ontology に格納されている遺伝子と機能の関係は、文献に基づくものの他に、配列や遺伝子の発現といったその他の情報の類似性から推測されているものがある。手段によって関係付けの信頼度が異なるので、この情報を付記する必要がある。手段

は evidence code としてコード化されている。Gene Ontology には信頼度に関する記載はない。信頼度を判断するのはユーザであるとされている。

2.5 オントロジーとしての Gene Ontology

人工物についても生物についても、機能概念がドメイン知識の中で重要な位置を占めることには変わりない。人工物の機能概念については、現象をプロセスの連鎖として捉える立場(プロセスオントロジー)と、入出力を持つ装置が主体的に行っている処理の連鎖として捉える立場(デバイスオントロジー)がある。また機能の主体、客体、主体の振る舞い、役割、文脈依存性、機能分解、方式の分離、等の解析がなされている[Kitamura 03]。Gene Ontology では、これらの解析はなされていない。

Biological Process は Molecular Function の組み合わせにより構成される機能であるので、両者の概念の間には機能分解関係があると考えられる。しかし Gene Ontology ではオントロジー間の関係は解析されていない。

他にも Gene Ontology は、is-a 関係の完備、論理表現による概念の定義、論理的制約の定義、意味の異なる part-of 関係の分離など関係概念の解析、といったオントロジーとして必要とされる性質を満たしていない。その理由について、Gene Ontology の見解は次である。「できるだけ多くの概念を収集すること、収集した概念の間に何らかの関係(is-a 関係か part-of 関係のいずれか)を与えること、概念に遺伝子を関係付けること、を優先している。開発者の人的資源に余裕がないので、その他の解析には労力を割けない」。この見解は、生物学ドメインからの要求を最優先し、オントロジーとしての完備性は後回しにする、という Gene Ontology のスタンスを示している。このスタンスの是非は、Gene Ontology の実際の利用状況と生物学への貢献度から判断されるべきであろう。

一方 Gene Ontology のスタンスを了解するとしても、増え続ける概念の整合性を保つことは、オントロジーとしての完備性の整備なしにはいつかは破綻する危険があると指摘されている[Wroe 03, Yeh 03, Smith 03]。

3. Gene Ontology の利用

Gene Ontology は実際にゲノム研究で幅広く利用されている。

3.1 機能の専門用語語彙としての利用

Gene Ontology が登場して以降、新たにゲノム配列が決定された場合には、機能のアノテーション(遺伝子のデータベースに機能に関する属性を登録すること)には Gene Ontology の語彙を用いることが慣習となっている。

最近発表されたある病原虫のゲノム解析を例として挙げる[Verjovski-Almeida 03]。マンソン住血吸虫はヒトの肝門脈に住み着き死に至らしめる病原虫である。このゲノム配列が最近決定された。遺伝子は 14,000 個、そのうち発現している遺伝子は半数の 7,000 個と見積もられた。

マンソン住血吸虫による健康被害を防ぐためには、治療薬やワクチンの開発が必要である。これまでの研究から、病原虫の遺伝子の中で治療薬やワクチンの標的となるものは特定の概念に関係付けられることが分かっている。その概念とは、「受容体」、「酵素」、「チャネル」、「輸送体」といった Molecular Function の機能概念、また「細胞表面」という Cellular Component の構造物概念である。これらの概念と Gene Ontology を頼りに、マンソン住血吸虫における治療薬およびワクチン開発の候補となる遺伝子が探索された。

これまで機能について、他の生物種で解析された結果を別の生物種へ網羅的に適用することは容易ではなかった。Gene Ontology はこの種の適用に対して計算機で支援する道を拓くとともに、語彙の基準としてアノテーションの標準化を促進し、アノテーションの再利用性の向上に貢献している。このように Gene Ontology は、ある生物のゲノム配列が決定された後、それを医薬開発へ応用するために必要となる機能アノテーションを迅速化し、応用研究の促進に貢献している。

3.2 遺伝子発現量の解析

ある細胞における遺伝子発現量のパタン(DNA マイクロアレイ実験、DNA チップ実験により取得される)と機能の相関性の解析は、Gene Ontology の開発当初から目標として掲げられてきた。現在この種の解析に Gene Ontology を利用することは当然となっている。

Gene Ontology の利用は、結果の解釈として用いる場合と、機械学習の正解集合として用いる場合がある。前者の場合は、事前知識を用いずに遺伝子発現データのクラスタリングを行い、その結果を Gene Ontology で解釈する[Hvidsten 03, Zhong 03]。後者の場合は、Gene Ontology が与える遺伝子の機能の類似性と遺伝子発現量の類似性との間には相関があると仮定し、発現量の類似性から、遺伝子の機能を予測するものである[Hvidsten 03]。これは後述の「3.3 機能の自動アノテーション」を目指す研究のひとつでもある。なお生物学においてこの仮定は正しいと考えられている。

3.3 機能の自動アノテーション

Gene Ontology が機能アノテーションの基準として受け入れられているだけでなく、Gene Ontology を用いてこのアノテーション作業を自動化する研究が急ピッチで進んでいる。

機能アノテーションの自動化は、現在壁にぶつかっている。現在の方法では約 2/3 の遺伝子しかカバーできず、30%強の遺伝子が機能アノテーションを欠いたまま残ってしまうという壁である。この原因は、遺伝子の類似性を測る(計算可能な)因子が配列しかないことにある。遺伝子の配列類似性と機能の類似性の間に良い相関があることから、遺伝子の配列の類似性に基づく機能推定は優れた成果を上げてきた。しかしこの方法の欠点は、配列が類似する遺伝子のクラスタが得られても、そのクラスタの中に機能アノテーションされた遺伝子がない限り機能推定ができないことである。

そこで、機能の類似性を測る、配列以外の(計算可能な)因子が求められていた。Gene Ontology が格納する機能・遺伝子・文献という組み合わせは、まさに求められた因子を与えるものである。つまり Gene Ontology は、ある遺伝子と機能のペアに対して、その関係性を報告した文献集合を定義していることになり、文献から遺伝子と機能の関係を機械抽出する研究における機械学習の正解集合を与える[Raychaudhuri 02, Raychaudhuri 03, Chiang 03]。同様に、Gene Ontology

が外部データベースとのデータ交換性をサポートするインデックスファイルを提供していることから、外部データベースが格納する遺伝子の属性と組み合わせると、機能・遺伝子・遺伝子配列、機能・遺伝子・遺伝子発現量、機能・遺伝子・タンパク質間相互作用、機能・遺伝子・ドメインおよびモチーフ配列、という情報の組み合わせを得ることができ、これを機械学習の正解集合とすることで、さまざまな属性データを使った機能推定の可能性が拓けた[Troyanskaya 03, King 03, Jensen 03, Lagreid 03, Thomas 03, Lord 03, Letovsky 03, Hvidsten 03]。機械学習の手法としては、最大エントロピー法、ベイズ推定、決定木、マルコフモデル、が試されている。

3.4 データベース統合

ゲノム研究におけるデータベース統合は、これまで遺伝子の配列を基準として進められてきた。Gene Ontology の登場により、遺伝子の機能を基準としたデータベース統合が可能となった。配列情報はシンボルのリストであるため、機械処理には適しているが、人間が直感的に解釈できる対象ではない。一方機能は、遺伝子の属性の中で人間が直感的に最も解釈し易い属性である。機能を基準としたデータベース統合は、生物学者にユーザフレンドリーな利用環境を与え、統合データベースの利用性を格段に高めると期待される。

さらに機能を基準としたデータ統合は、ゲノム研究がゲノムに暗号化された遺伝情報をすべて読み解いた暁に、その情報を格納する枠組みとして不可欠であると考えられている。

Gene Ontology の開発母体となっているゲノムデータベースである SGD（酵母）[Weng 03]、Flybase（ハエ）[Consortium 03]、WormBase（線虫）[Harris 03]、MGD（マウス）[Blake 03]、TAIR（シロイヌナズナ）[Rhee 03]は、Gene Ontology を基準としたデータ統合を実現している。また Onto-Tool[Draghici 03]、Go Fish[Berriz 03]、DAVID[Dennis 03]、GOblet[Hennig 03]といった、機能を基準としたデータベース統合を支援するツールが次々と開発されている。

3.5 Gene Ontology の概念の解析

Gene Ontology を DAML+OIL[Wroe 03]、フレーム表現（Protege 上に実装）[Yeh 03]、Inductive Logic Programming (Prolog)[Badea 03]といったオントロジー記述言語に移植する研究が行われた。これらの研究では Gene Ontology がオントロジーとして不完全である問題が取り上げられている。

DAML+OIL と Protege への移植においては、is-a 関係が完備されていないことが問題となった。その対応策として、ルート直下に抽象概念を追加する措置が取られた。さらに Cellular Component については、手動による is-a 関係の追加や、いくつかの概念を「場所の概念」から「構造物の概念」に変換することによって、下位概念との関係を part-of 関係から is-a 関係に変換する措置が取られた。

Molecular Function、Biological Process、Cellular Component 間の関係の整備が必要であることも指摘された。その目的で、Part-of-Process（Molecular Function の属性で Biological Process との関係を示す）、Occurs-at-Component（Molecular Function と Biological Process の属性で Cellular Component との関係を示す）という関係概念が追加された。さらに part-of 関係を解析し、Part-of-Complex (Component/Integral-Object の関係)、Contained-Within（Biological Process の概念間における時間推移の包含関係）といった関係概念が追加された。

概念の内部構造の解析も着手された。Gene Ontologyでは、ある生物種に特有な機能概念は概念名に"(sensu 生物種名)"を付けて表現されている。移植後、"(sensu 生物種名)"は属性に移された。

哲学オントロジーの立場からさらに踏み込んで、Gene Ontologyが意味論的完全性を獲得するために必要な措置についても論じられている[Smith 03]。下位分類を行う基準の明確化が必要であり、その概念が一時的な存在であるのか継続する存在であるのか（Continuants and Occurrents）、全称限定であるのか存在限定であるのかを厳密に扱うとともに、概念間の依存関係を明確にすべきであるとしている。この措置を上位概念について施すだけでも、全体の見通しが大きく改善すると考えられる。

4. Gene Ontology 開発の経緯と運営

Gene Ontology は SGD（酵母）[Weng 03]、Flybase（ハエ）[Consortium 03]、MGD（マウス）[Blake 03]のゲノムデータベースの開発グループが、自分達の研究の推進を目的として始めた活動である。ゲノム研究としては珍しく、政府主導の国家プロジェクトではない。途中から TAIR（シロイヌナズナ）[Rhee 03]と WormBase（線虫）[Harris 03]が加わっている。当初は米国におけるどちらかと言うとプライベートな活動と見なされていた Gene Ontology であるが、その後欧州連合におけるゲノム研究の拠点である EMBL が参加し、EMBL が開発するデータベース群との交換性が整備された頃から、Gene Ontology は世界標準として認識されるようになった。

Gene Ontology は、上記のデータベース開発グループから各々2-3 人が集まり、2 週間に一度ミーティングを持ち、そのミーティングにおける合意で、概念が採択され概念間の関係が決定されるそうである[Cherry 私信]。また半年に一度公開ミーティグが開かれる。常に利用者の意見を吸い上げる窓口が開かれており、生物学者の利用性を念頭に置いた開発を続けている。Gene Ontology の更新版は毎月リリースされる。

生物学において酵母、ハエ、線虫、シロイヌナズナ、マウスという 5 生物をカバーするということは、地球上の真核生物の全種類をカバーするということとほぼ等しい意味を持つ。なぜならこれらの生物は真核細菌、昆虫、線虫類、植物、ほ乳類を代表する生物であり、この 5 生物のことが分かれば、それらが代表する他の生物のこともおおよそ分かったも同然と考えられているからである。

5. さらなるパラダイムシフトへ

生物学とは、生物がなぜ生きることができるのか、という問いに答えようとする学問である。問題の解決を目指して、さまざまな手段がこれまで用いられてきた。その手段に応じた新しい融合領域が生み出され、その度に生物学者はパラダイムシフトの洗礼にさらされてきた。分子生物学、構造生物学、生物化学、生物物理学、生物工学、生物情報科学はその例で

ある。そして Conceptual Biology が新たに生まれようとしている。

Gene Ontology の登場は、この新たなパラダイムシフトの一部に過ぎない。Gene Ontology は生物学の概念の一側面を対象としているに過ぎない。今後、生物の多様な概念のさまざまな側面を対象としたオントロジーが開発されなければならない。また概念間の推論、概念を表現する言語、概念の可視化についても、研究が進められなければならない。

生物学というドメインは、これまでオントロジー研究が対象としてきたドメインとは少し異なる特徴を持っているように感じる。生物学でオントロジーは必要とされている。概念を処理できなければ今後の生物学の発展はあり得ない、そこまで意識が高まっている。その高まりが、Conceptual Biology という提唱に、Gene Ontology が驚くほど急速にドメインに受け入れられ広く利用されている現実に、現れている。生物学の概念の集合は、生物学という学問そのものである。また生物が生きるという不思議を解き明かすための素材となる知識の宝庫である。その概念構造を研究するということは、生物学研究そのものに他ならない。これは生物学が記述的学問であった故に生じた幸運なのかもしれない。

Gene Ontology は、生物学にオントロジー研究が取り込まれた第一歩であると言える。Gene Ontology の開発グループが自ら表明しているように Gene Ontology はオントロジーというより"controlled vocabulary"に近いと言えるであろう。今後の課題は、生物学をドメインとして情報科学的に完備されたオントロジーを開発できるのか、そのオントロジーは生物学のどのような問題を解くことができるのか、それが生物学の発展に真に貢献するのか、という未知の問題に向かうことであろう。未開の原野ではあるが、ドメインの専門家からの熱い要求と、学問の中核であるかもしれないという幸運な特徴は、オントロジー研究の新しいドメインとして十分魅力的なのではないであろうか？

謝辞

本稿に対し有益なコメントを頂きました大阪大学産業科学研究所 溝口理一郎教授に感謝申し上げます。

文献

[Badea 03] Badea, L.*Pac Symp Biocomput*, 565-576, (2003).

[Berriz 03] Berriz, G.F., J.V. White, O.D. King, and F.P. Roth, *Bioinformatics*, 19 (6), 788-789, (2003).

[Blagosklonny 02] Blagosklonny, M.V. and A.B. Pardee, *Nature*, 416 (6879), 373, (2002).

[Blake 03] Blake, J.A., J.E. Richardson, C.J. Bult, J.A. Kadin, and J.T. Eppig, *Nucleic Acids Res*, 31 (1), 193-195, (2003).

[Chiang 03] Chiang, J.H. and H.C. Yu, *Bioinformatics*, 19 (11), 1417-1422, (2003).

[Consortium 03] Consortium, F.*Nucleic Acids Res*, 31 (1), 172-175, (2003).

[Consortium 01] Consortium, *Genome Res*, 11 (8), 1425-1433, (2001).

[Dennis 03] Dennis, G., Jr., B.T. Sherman, D.A. Hosack, J. Yang, W. Gao, H.C. Lane, and R.A. Lempicki, *Genome Biol*, 4 (9), R60, (2003).

[Draghici 03] Draghici, S., P. Khatri, P. Bhavsar, A. Shah, S.A. Krawetz, and M.A. Tainsky, *Nucleic Acids Res*, 31 (13), 3775-3781, (2003).

[Harris 03] Harris, T.W., *et al.*, *Nucleic Acids Res*, 31 (1), 133-137, (2003).

[Hennig 03] Hennig, S., D. Groth, and H. Lehrach, *Nucleic Acids Res*, 31 (13), 3712-3715, (2003).

[Humphreys 93] Humphreys, H.L., D: *Bulletin of the Medical Library Association*, 81 170-177, (1993).

[Hvidsten 03] Hvidsten, T.R., A. Laegreid, and J. Komorowski, *Bioinformatics*, 19 (9), 1116-1123, (2003).

[Jensen 03] Jensen, L.J., R. Gupta, H.H. Staerfeldt, and S. Brunak, *Bioinformatics*, 19 (5), 635-642, (2003).

[Karp 00] Karp, P.D., *Bioinformatics*, 16 (3), 269-285, (2000).

[King 03] King, O.D., R.E. Foulger, S.S. Dwight, J.V. White, and F.P. Roth, *Genome Res*, 13 (5), 896-904, (2003).

[Kitamura 03] Kitamura, Y., Mizoguchi, R., *Expert Systems with Application*, 24 (2), 153-166, (2003).

[Lagreid 03] Lagreid, A., T.R. Hvidsten, H. Midelfart, J. Komorowski, and A.K. Sandvik, *Genome Res*, 13 (5), 965-979, (2003).

[Lefranc 03] Lefranc, M.P., *Nucleic Acids Res*, 31 (1), 307-310, (2003).

[Letovsky 03] Letovsky, S. and S. Kasif, *Bioinformatics*, 19 Suppl 1 I197-I204, (2003).

[Lord 03] Lord, P.W., R.D. Stevens, A. Brass, and C.A. Goble, *Bioinformatics*, 19 (10), 1275-1283. (2003).

[Matys 03] Matys, V., E. *et al. Nucleic Acids Res*, 31 (1), 374-378, (2003).

[Noy 02] Noy, N.F., M.A. Musen, J.L.V. Mejino, and C. Rosse, *SMI-2002-0925*, (2002).

[Oliver 02] Oliver, D.E., D.L. Rubin, J.M. Stuart, M. Hewett, T.E. Klein, and R.B. Altman, *Pac Symp Biocomput*, 65-76, (2002).

[Raychaudhuri 03] Raychaudhuri, S. and R.B. Altman, *Bioinformatics*, 19 (3), 396-401, (2003).

[Raychaudhuri 02] Raychaudhuri, S., J.T. Chang, P.D. Sutphin, and R.B. Altman, *Genome Res*, 12 (1), 203-214, (2002).

[Rhee 03] Rhee, S.Y., *et al.* W. Beavis, T.Z. Berardini, G. Chen, D. Dixon, *Nucleic Acids Res*, 31 (1), 224-228, (2003).

[Smith 03] Smith, B., J. Williams, and S.-K. S.: The Ontology of the Gene Ontology, *Proceedings of AMIA Symposium 2003*, (2003).

[Stevens 00] Stevens, R., P. Baker, S. Bechhofer, G. Ng, A. Jacoby, N.W. Paton, C.A. Goble, and A. Brass, *Bioinformatics*, 16 (2), 184-185, (2000).

[Thomas 03] Thomas, P.D., *et al.*, *Genome Res*, 13 (9), 2129-2141, (2003).

[Troyanskaya 03] Troyanskaya, O.G., K. Dolinski, A.B. Owen, R.B. Altman, and D. Botstein, *Proc Natl Acad Sci U S A*, 100 (14), 8348-8353, (2003).

[Verjovski-Almeida 03] Verjovski-Almeida, S., R. *et al.*, *Nat Genet*, 35 (2), 148-157, (2003).

[Weng 03] Weng, S., Q. *et al.*, *Nucleic Acids Res*, 31 (1), 216-218, (2003).

[Wroe 03] Wroe, C.J., R. Stevens, C.A. Goble, and M. Ashburner, *Pac Symp Biocomput*, 624-635, (2003).

[Yeh 03] Yeh, I., P.D. Karp, N.F. Noy, and R.B. Altman, *Bioinformatics*, 19 (2), 241-248, (2003).

[Zhong 03] Zhong, S., C. Li, and W.H. Wong, *Nucleic Acids Res*, 31 (13), 3483-3486, (2003).

デバイスオントロジーに基づくシグナル伝達機能オントロジーの構築

(SIG-SW&ONT-A302-07.より抜粋，一部改変.)

高井貴子*1，溝口理一郎*2
1) 東京大学大学院 情報理工学系研究科，2) 大阪大学 産業科学研究所

要旨
生物のゲノムと機能の関係を解析するためには、機能データのデータベース開発が必要であるが、これまで機能を記述する首尾一貫した形式がなかったため、データは蓄積されるものの比較や推論が不可能であった。本研究ではデバイスオントロジーを応用してシグナル伝達機能のオントロジーを構築し、これを基準とすると、機能の首尾一貫した記述が可能であることを見出した。

情報科学と生物学の融合分野であるバイオインフォマティクスにおいて、人工知能研究が注目されている。生物学は記述的学問であると言われる。生物学の特徴は知識が概念として蓄積されることである。生物学の急速な発展から、蓄積されたる概念の量は膨大となり、計算機の助けなしには処理が困難となってきた。その結果、生物学の概念を対象とした人工知能研究とくにオントロジーが注目されている。

生物学は生物の機能を研究する学問であるので、概念の中でも機能概念が特に重要である。本研究では、生物学の概念の中で、シグナル伝達に関する機能概念を対象とする。シグナル伝達とは、生物と環境とのコミュニケーションおよび生物の内部のコミュニケーションに必要となる機構である。環境から病原菌が侵入すればその排除を行う機構や、生体を構成する多数の細胞の働きがバラバラにならないように調節する機構は、シグナル伝達機能の代表例である。

生物学においては、ゲノム解析研究の進展により、生物が持つ全遺伝子が明らかとなった。しかしながら、遺伝子が分かれば生物の機能が分かるわけではない。生物の機能は、人工物と同様に、複数の反応の連鎖で達成される。したがって、機能を達成するための連鎖反応と遺伝子との関連を解析することが重要である。

シグナル伝達は、生体内で起こる連鎖反応の一種であり、上記のコミュニケーションに関係するものを指す。反応の成分となるのは、遺伝子を鋳型として作られるタンパク質、金属イオン、低分子量の化合物である。生物学ではこの連鎖反応をパスウェイと呼ぶ。

生体内パスウェイを対象としたデータベース開発においては、シグナル伝達パスウェイと対峙する代謝パスウェイについて、機能とパスウェイの首尾一貫した書式が見出され、それに基づくデータベース開発、および機能やパスウェイの比較解析が行われている。一方、シグナル伝達パスウェイについては、機能とパスウェイのデータは蓄積されているものの、首尾一貫した記述に基づいていないため、比較や推論ができないだけでなく、データベース統合や検索にも障害となっている。

本研究では、シグナル伝達における機能とパスウェイをデバイスオントロジー[來村，溝口(2002) オントロジー工学に基づく機能的知識体系化の枠組み人工知能学会誌，**17**，61-72]の枠組みに当てはめ、機能オントロジーを構築した。さらにこの機能オントロジーを基準とすると、首尾一貫した機能とパスウェイの記述が可能であることを見出した。

これまでシグナル伝達パスウェイの記述が統一されていなかった理由は、「シグナル」という概念が定義されていなかったため、シグナルの流れとしてパスウェイをモデル化できず、コンテキスト非依存性概念である振る舞いレベルでモデル化していたことにある。シグナル伝達にはB0-B3の異なる種類の振る舞いが混在しており、それらを混在したまま記述していたため、統一した書式がとれなかった。一方、代謝パスウェイも実は振る舞いレベルでモデル化されているのだが、代謝パスウェイにはB1的振る舞いしか存在しないため、そのままで統一した書式が可能となっていた。シグナル伝達のモデル化においては、この事実に気づかずに、代謝パスウェイにおけるモデル化をそのまま適用したため、記述の統一ができなかったと考えられる。

一般には化学反応を対象にした現象のモデル化にはデバイスオントロジーではなくプロセスオントロジーが用いられるが、我々は本研究の目的にはデバイスオントロジーが相応しいと判断した。シグナル伝達パスウェイをプロセスオントロジーで記述することは可能である。しかしながらその結果は生物学が求めている方向と一致しない。生物学研究のゴールは分子と生物機能の関係を解析することである。したがって機能は分子の作用として記述されなければならない。プロセスオントロジーは分子の状態変化そのものの注目しているため、主体がなく、分子の機能を定義する目的に相応しくない。一方、デバイスオントロジーはまさにこの記述の枠組みを与えるものである。本研究は、これまで人工物の機能を対象として確立されてきたデバイスオントロジーの枠組みが、生物の機能にも適用可能であることを示し、その適用が生物学研究の進展に有用であることを示した。

<u>参考</u>

Takai-Igarashi, T. and Mizoguchi, R. (2004) Cell signaling networks ontology. In Silico Biol. 4, 81-87.

Takai-Igarashi, T. and Mizoguchi, R. (2004) Ontological integration of data models for cell signaling pathways. Genome Informatics, 15, 255-265.

Takai-Igarashi, T. and Mizoguchi, R. Development of an ontology for systems functions of signal transduction, submitted.

高井，溝口，高木 (2003) デバイスオントロジーに基づくシグナル伝達機能オントロジーの構築，人工知能学会，セマンティックウェブとオントロジー研究会，SIG-SW&ONT-A302-07.

http://athos.biochem.s.u-tokyo.ac.jp/CSNO/