生命科学のためのオントロジー

○高木利久、高井貴子

Ontology of Living Systems

○Toshihisa TAKAGI, Takako TAKAI

Abstract

To substantially promote the integration of heterogeneous knowledge and to further facilitate the discovery of fundamental concepts or hidden structures in biological phenomena, we study genome ontology where biological knowledge is comprehensively reorganized and both the terminology and the concepts are unified across species. We try to establish such ontology of signal transduction and to recompile a dictionary of terms in the universe of living systems.

1. 生命科学にとってなぜオントロジーが必要であるか？

情報科学におけるオントロジー研究では、診断、設計、学習支援、情報検索、自動翻訳等を目的とした知識表現が研究されてきた。生命科学がオントロジー研究の目的となったのは、分子生物学の発展とヒトゲノムやモデル生物ゲノムの配列決定の結果、知識の総量が膨大となり、計算機による解析が不可欠となったからである。

生命科学にとってなぜオントロジーが必要であるか？それは、生命科学が分子の情報を統合し、分子の協調作用として生命のメカニズムを探求する学問であることに起因する。たとえばヒトの遺伝子は少なく見積もって３万と推定されている。３万の分子を総合して考察することですら、すでに人間の認識能力を超えているのに、解析においてより重要となるのは、分子の組み合わせや他生物との比較という、一層計算量の多い仕事である。ここに、生命科学には計算機による知識処理が不可欠である原因がある。生命科学のオントロジーに課せられている問題は、機能未知の遺伝子/分子の機能のアサインを助け、分子の協調作用の解析を助けるために、既存の知識を計算機が理解できる統一された形式に変換することである。

2. 人工知能研究と生命科学

一方、計算機による知識表現から生命科学をとらえた場合、この領域はこれまでの対象領域と異なる特徴をもっている。

2.1 インスタンスが自明

オントロジーを構築する際に、対象領域におけるインスタンスと概念の境界は、決定が困難な問題である。しかし生命科学を対象領域とする場合には、インスタンスは遺伝子と生体分子であり、初めから自明である。とくにモデル生物を対象とする場合は、遺伝子の総カタログがすでに解明されているので、インスタンス集合の範囲が自明となっていることは顕著な特徴である。

2.2 知識が不完全かつ断片的でしかも急増している

機械、診断、経済活動、テキスト等の人間によって作り出された人工物がインスタンスである領域と異なり、生命科学では、インスタンス（遺伝子/分子）は人間の存在や認識とは独立に存在する。生命科学の知識とは、インスタンスが自発的にもつ内包定義を、専門家の観測によって断片的に認識したものである。専門家は未だ全容の認識には至っておらず、現時点における知識は不完全であるが、ゲノムプロジェクトの推進により、獲得される知識は急激に増加している。したがって生命科学の概念は、不完全かつ断片的でしかも急増している知識から抽出しなくてはならない。この特徴から、概念は変容の可能性を常に内含しているので、オントロジーは変容に追随できる柔軟性をもたなければならない。

2.3 オントロジーが不可欠

さらに、生命科学を従来の対象領域と比較した場合、対象領域が要求するオントロジーの不可欠性が圧倒的に高いのではないかと思われる。これまで、オントロジーは専門家に対する大きな助けとなったが、必須であるかという点についてはあまり議論されてこなかった。しかしながら、生命科学における概念が内含する膨大な計算量、不完全性と変容性に起因する複雑性を思慮すると、計算機の利用なしに概念を解析することは不可能である。さらにこの学問の発展が与える社会的影響力の大きさを考慮すると、生命科学のオントロジーの構築は必要に迫られているとさえ言える。この意味で我々は、生命科学が人工知能研究の力強い対象領域であると考えている。

3. 生命科学の概念

生命科学の概念は４種類（機能、構造、時空間、実験法）あると考えられる。

3.1 機能

生命科学における機能とは、この領域を情報科学の立場で解析した場合、生命の情報に相当する。したがって機能は生命科学のオントロジーの主題となる概念である。機能の概念とは、分子と分子集合の内包定義を専門家が観測した結果（知識）を抽象化したものである。機能概念は、観測の対象となる分子の構造の概念と、分子が存在する時空間の概念によって規定される。

3.2 構造

機能を規定する物理的実体であり、分子の内部構造と分子集合の構造である。機能と構造は常に関係しており、機能の抽象度に応じてさまざまなレベルの構造が存在する。配列、モチーフ、ドメイン、立体構造、複合体、分子間相互作用、パスウェイ、ネットワーク等の概念が含まれる。

3.3 時間と空間

機能を規定する時空間のクライテリアであり、構造と同様に、機能の抽象度に応じてさまざまなレベルが存在する。空間の概念は、細胞内の場所（核、細胞質など）、生体内の場所（系）、生物の集団、の順に階層化される。時間の概念は、細胞の時間（細胞周期など）、生体の時間（発生と老化）、生物種の時間（進化）、の順に階層化される。我々は生体内の場所のクライテリアとして臓器や組織を用いずに、系（免疫系、内分泌系など）を用いる。その理由は、臓器や組織は解剖学的観点から生命を解析した場合に機能の主体となるインスタンスであるので、したがってこの観点は、我々が目指す、分子をインスタンスとする生命科学のオントロジーとは相容れないからである。しかしながら我々が構築している生命科学のオントロジーは、将来的に臓器や組織の概念へリンクを張る予定である。

3.4 実験の手法や材料

生命科学の概念は観測を通してしか獲得することができず、観測の種類によって得られる認識が異なる場合もある。したがって観測の手法や実験に用いた材料に関する概念が必要となる。

3.5 機能と構造と時空間の関係

情報科学の立場から生命を情報（機能）の流れと捉えた場合、構造は情報（機能）の処理単位である関数に相当する。同様に時空間は、関数の変数定義域/ドメインに相当する。しかし我々の研究では、生命科学の実世界において関数（構造）の変数に相当する概念を、まだ見いだしていない。そのため現段階では、機能を規定する構造と時空間の関係を、テーブルとして定義するにとどまっている。

4. 生命科学における概念の関係

生命科学のオントロジーにおける概念の関係については、とくに推移律が成り立つ関係が重要である。なぜならオントロジーを構築する目的は、生命科学のインスタンスである遺伝子/分子の類似性を、知識を用いた推論により行うことにあるからである。

生命科学の概念で推移律が成り立つ条件とはなにであろうか？それは、機能を意図した時に概念の包含関係が成り立つことである。単に構造的に包含関係があるだけでは不足である。

例えばゲノム-染色体-DNA という関係は推移律が成立すると考えられる。これは、３個の概念が遺伝情報を担うという機能を意図した時に、ゲノムが染色体の概念を、染色体がDNA の概念を含んでいるからである。単に構造的にその一部を構築しているだけで、機能の共有性が希薄な関係、例えばDNA-グアニンは、推移律が成立しない包含関係となる。

5. オントロジー開発の例

生命科学におけるオントロジーは、領域のあるべき姿あるいはレファレンスとしてトップダウンに構築するアプローチと、領域に存在するインスタンスの解析からボトムアップに構築するアプローチがある。前者は、外延としてデータベースを作る基盤知識のためのオントロジーとも言え、後者は外延としてすでに存在しているデータを抽象化したオントロジーであるとも言える。我々は前者を主眼としてシグナル伝達系のオントロジー（Signal-Ontology）を、後者を主眼として生命科学の用語辞書（BioTerm Bank）を開発している。

シグナル伝達系は、とくにオントロジーの開発が必要な領域である。なぜなら相似の領域である代謝系ではECリスト、Biochemical Pathway Chartといったオントロジーがあり、その知識を基盤にデータベースが開発されているのに対し、シグナル伝達系は拠り所となる知識基盤が脆弱であり、それがシグナル伝達系のデータベース開発の障壁となっているからである。Signal-Ontologyは3節で述べた生命科学の概念を、シグナル伝達系という特定の領域に射影したものである。とくに構造を射影した概念である、シグナル伝達系におけるパスウェイの共通構造（Signal-Module）は、この領域に特徴的である。この概念は、分子の配列だけでなく、分子間の相互作用も、生物進化の過程で保存されてきたとする見解に基づいており、我々はSignal-Moduleをシグナル伝達系の構造の基本単位であると捉えている。

生命科学を、構成成分である分子の情報から総合的に理解するためには、この領域のインスタンスである分子のすべてをうまくマッピングできるオントロジーが必要である。BioTerm Bankで対象とする分子は、現在利用できる生物データベースから網羅的に収集したものであり、現在既知である全分子の集合と考えてよい。機械的に収集した結果50万分子となった。BioTerm Bankでは、分子をひとつひとつ手作業で分類することによって、最終的に全体をカバーするオントロジーの構築を目指す。同時に分類のログを記録しておき、ここからマッピングのルールを抽出し、計算機による自動マッピングも目指している。

表に、生命科学に関する主要なオントロジー開発についてまとめた。この中で世界をリードする活動はGene Ontologyである。Gene Ontologyは酵母、ショウジョウバエ、線虫、シロイヌナズナ、マウス、のモデル生物における遺伝子の共通な機能を定義づけている。真核生物においては遺伝子の配列と機能が生物種間で高く保存されているので、機能オントロジーの構築により、比較ゲノムに基づいて機能未知の遺伝子の機能を推定できると考えられている。最近国際ヒトゲノム解読共同体とセレラ社から各々発表されたヒトゲノム配列では、Gene Ontologyと対応づけて機能がアサインされている。

6. おわりに

生命科学のオントロジー研究はまだ端緒についたばかりであり、これから解決すべき課題は多い。我々の研究も始まったばかりである。最後に生命科学のオントロジー研究に関連した文献を紹介するのでご参考頂きたい。

Baker, P. G. et al. An ontology for bioinformatics applications. *Bioinformatics*, 15: 510-520, 1999

Giudicelli, V. & Lefranc, M. P. Ontology for immunogenetics: the IMGT-ONTOLOGY.

Bioinformatics, 15: 1047-1054, 1999

Gruber, T. R. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5: 199-220, 1993

Humphreys, H. & Lindberg, D. The UMLS project: making the conceptual connection between users and the information they need. *Bulletin of the Medical Library Association*, 81: 170-177, 1993

Karp, P. D. An ontology for biological function based on molecular interactions. *Bioinformatics*, 16: 269-85, 2000

Paton N. W. et al. Conceptual modeling of genomic information. *Bioinformatics*, 16: 548-557, 2000

Rzhetsky, A. at el. A knowledge model for analysis and simulation of regulatory networks. *Bioinformatics*, 16: 1120-1128, 2000

Schulze-Kremer, S. Adding Semantics to Genome Databases: Towards an Ontology for Molecular Biology. *Proc Int Conf Intell Syst Mol Biol,* 272-275, 1997

Schulze-Kremer, S. Ontologies for Molecular Biology. *Pac Symp Biocomput,* 693-704, 1998

Stevens, R. et al. Ontology-based knowledge representation for bioinformatics. *Briefings in Bioinformatics*, 1: 398-414, 2000

Tateishi, Y. et al. Building an Annotated Corpus from Biology Research Papers. *Proc. COLING 2000 Workshop on Semantic Annotation and Intelligent Content,* 28-34, 2000

The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nat Genet*, 25: 25-29, 2000

高井貴子, 高木利久 (2001) 生命科学のためのオントロジー. 実験医学. 19(11), (印刷中)

高木 利久 東京大学医科学研究所ヒトゲノム解析センター (〒108-8639 東京都港区白金台 4-6-1)

高井 貴子 同上

Toshihisa Takagi (takagi@ims.u-tokyo.ac.jp) Human Genome Center, Institute of Medical Science, University of Tokyo

Takako Takai (takako@ims.u-tokyo.ac.jp) Human Genome Center, Institute of Medical Science, University of Tokyo

表　インターネット上で公開されている生命科学領域のオントロジーとツール
（アルファベット順. http://ontology.ims.u-tokyo.ac.jp/Collection.html もご参照ください）

オントロジー名	中心開発機関	対象領域	URL
BioTerm Bank (BTB)	Human Genome Center	生命科学を記述する用語の総カタログ	http://ontology.ims.u-tokyo.ac.jp/BTB/
Description Logics (DL)	DL community	フレーム、意味ネットワーク、オブジェクト指向に基づくオントロジー記述言語	http://dl.kr.org/
Gene Ontology (GO)	SGD, Flybase, MGD, GXD, TAIR, WormBase, Pombase	モデル生物の遺伝子産物の機能	http://www.geneontology.org/
GeneX	NCGR (National Center for Genome Research)	遺伝子発現データの統合のためのデータモデルとオントロジー	http://www.ncgr.org/research/genex/
Human Gene Nomenclature Database	HUGO (The Human Genome Organisation)	ヒト遺伝子の名称	http://www.gene.ucl.ac.uk/nomenclature/
ImMuno GeneTics(IMGT) Ontology	CNRS (Centre National Recherche Scientifique)	免疫グロブリン、T細胞受容体、主要組織抗原複合体	http://imgt.cines.fr:8104/
INTERACTIONS Ontology	SRI International	代謝反応	http://www.ai.sri.com/pkarp/interactions.html
Knowledge Interchange Format (KIF)	Stanford University	1階述語論理に基づくオントロジー記述言語	http://logic.stanford.edu/kif/dpans.html
Molecular Biology Ontology (MBO)	Max-Planck Institute for Molecular Genetics	分子生物学の概括的な概念	http://igd.rz-berlin.mpg.de/~www/oe/mbo.html
Mouse Anatomical Dictionary	The Jackson Laboratory	マウスの発生における系統的な組織名	http://www.informatics.jax.org/searches/anatdict_form.shtml
Ontolingua	Stanford University	協調的環境におけるオントロジー構築のためのソフトウェア	http://www.ksl.stanford.edu/software/ontolingua/
PharmGKB	NIGMS (National Institute of General Medical Sciences)	医薬品の生体作用とヒト遺伝子型との関連	http://www.pharmgkb.org/
Protégé	Stanford University	専門家の知識獲得を目指したオントロジー構築のためのソフトウェア	http://www.smi.stanford.edu/projects/protege/
Protein Review on the Web (PROW)	NCBI (National Center for Biotechnology Information)	特定のタンパク質の構造と機能	http://www.ncbi.nlm.nih.gov/PROW/
RiboWeb	Stanford University	リボゾームの構造と機能	http://smi-web.stanford.edu/projects/helix/riboweb.html
Signal Ontology	Human Genome Center	シグナル伝達系パスウェイの構造と機能	http://ontology.ims.u-tokyo.ac.jp/signalontology/
STAR/mmCIF	Rutgers University	マクロ分子の結晶構造	http://ndbserver.rutgers.edu/mmcif/
TAMBIS Ontology (TaO)	University of Manchester	分子生物学の概括的な概念と概念の関係	http://img.cs.man.ac.uk/tambis/
Unified Medline Language System (UMLS)	NLM (National Librarly of Medcine)	生命科学を中心とした幅広い学問領域における文献を分類する階層化されたキーワード	http://www.nlm.nih.gov/research/umls/