［特集］ポストゲノム時代に高まるバイオ自然言語処理への期待：バイオ自然言語処理最新事情

# 5 ゲノムデータの機械解釈

大久保 公策
kousaku@genomatrix.com / 国立遺伝学研究所生命情報・DDBJ研究センター

日紫喜 光良
t-hishiki@jbirc.aist.go.jp / 産業技術総合研究所生物情報解析研究センター

ポストゲノムデータは2～3万のヒトの遺伝子にさまざまな属性値が付与された特徴ベクトル行列である．最近数年間計算機科学の成果は特徴行列に基づく遺伝子の構造化に応用されてきた．そして今度はできあがった構造の解釈が要求されている．解釈とはこの分野の既存の知識に基づいてデータを説明することであり旧来見られた特定の法則をひたすら適応するような課題よりも人間的な課題である．そして分野の知識をどのように利用可能にしどのように適用するかという2つの点で個性の出る課題でもある．本稿では分子生物学的な考え方とそれに基づくデータの解釈について解説し，解釈を計算機に行わせようとするいくつかの試みを紹介する．

## 背景

生命にかかわる研究は常に知識に長けた研究者の注意深い観察に基づく洞察で展開してきた．しかし分子生物学の成功により枚挙的に行われた観察の結果や解釈の記述量の増大と観察の機械化がまねいたデータの量およびスコープの増大はこのクラシックな科学の手法を無効にしつつある．洞察に満ちたデータ解釈を可能にするために今この分野ではデータ解釈の機械的なアシストが切望されている．データ解釈という知識の利用は機械で部分的にでも代行できるのであろうか？　今回の特集のテーマともいえる「生物知識に関する研究」の最難関かもしれないこのテーマについて解説を加えたい．

## 分子生物学（遺伝子機能学）の現状

分子生物学や生化学といったタンパク質中心の科学はこれまで人間の体の構造と状態をすべてタンパク質の反応で説明しようとしてきた．一方これまでに人間の構造や状態に関して認識された物や事は専門用語の数に等しいとすると教科書レベルでも3万を超える（**図-1**）．分子生物学は3万種類のタンパク質の組合せで3万種類の構造や状態を説明しようというわけである．裏返すとタンパク質の機能はこれらの3万の用語で表される構造や状態への関与とその方法として説明されており，今後もそれが要求される．実際にはタンパク質の行為は「分布し認識し時に状態遷移する」ことに尽きるが，その生命活動への効果がさまざまなレベルで多様に表現されているのである．ここではタンパク質の機能を**分子機能，分布特性，細胞レベル生体レベルでの役割**（**図-2**）と分類しよう．光を感じる現象は「網膜の視細胞外節中のディスクの膜に存在するオプシンに配位した11-シス・レチナールが光を受けるとトランス型に異性化しオプシンから離れる．すると今度はオプシンが異性化しGTP分解酵素と結合しそれを活性化する．GTP分解酵素の活性

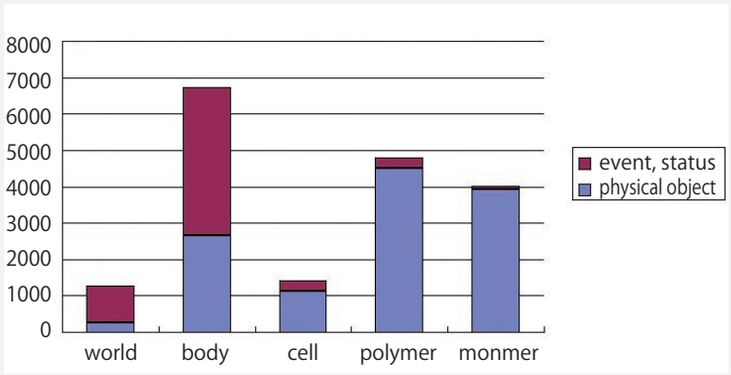


図 -1　医科学を構成する概念の分類

解剖学から内科学まで 25 冊の標準的な教科書の索引の大見出し語を集めて，形態的揺らぎを吸収した結果を人手で意味分類したもの．意味分類は物と事に分けたあと体の外のことから低分子量のものまで対象の所属する階層によって分類した．

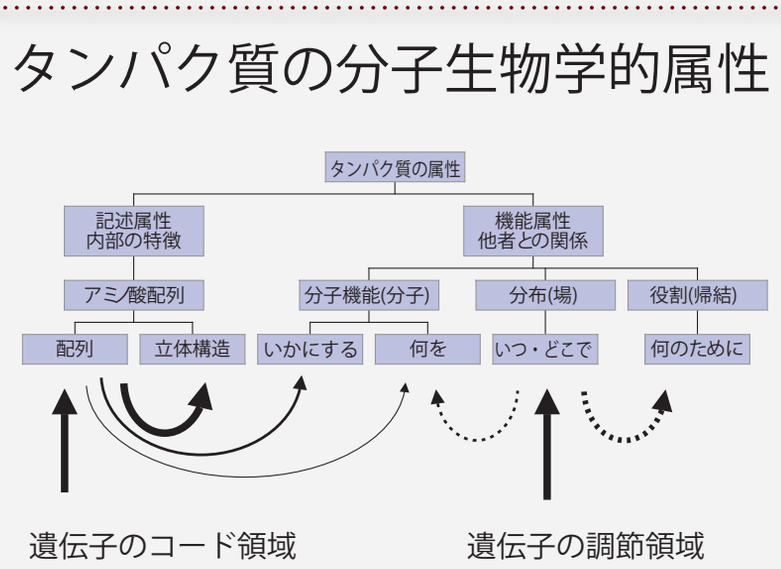


図 -2　タンパク質の分子生物学的な属性の分類

機能と呼ばれる属性は「何のために，いつ，どこで，誰に，何をしたか」というきわめて日常的な説明のための属性であることに驚く．弧型矢印とその太さは経験的な従属関係とその強さ（類似性の定理）を示している．実線は構造データに発する古典的な類似性定理，破線の部分はポストゲノムで注目される新しい測定値間の類似性の定理を示す．実際にはゲノム上の遺伝子は調節領域と構造領域に分かれておりそれぞれが独立に蛋白機能を決定しており，それ以外の属性は生体の中で両者によって自動的に決定される 2 次的な属性である．ただし，構造と調節の独立性は進化的には証明されていない課題である．

化によっていくつかの過程を経て視細胞が伝達物質を放出し，光刺激は脳へと伝わり光を感じる」と説明されているのでオプシン蛋白の機能は要約すれば「分子機能：レチナールからGTP分解酵素へのシグナル伝達｜分布：視細胞・外節ディスク｜役割：光知覚」と表現できる．

ヒトのタンパク質のうちでオプシン程度に詳細かつバランスよく機能が知られているものは１割程度に過ぎず大半の遺伝子は３つの機能のうち１つが多少知られている程度にすぎない．

## 遺伝子機能の定理

分子生物学は要素的な生命現象を多数のタンパク質やその他の分子間反応として説明し，その要素現象間の関係も説明でつなげてゆく「**説明のネットワーク**」を書き上げようとしてきた．ここで利用された分子生物学の基本法則はタンパク質の構造がDNAに記載してあるという事実（セントラルドグマ）でもメンデルの法則でもなく，「**類似性定理**」とでも呼ぶべきものである．類似性の定理とは似通った構造のタンパク質は機能的な特徴（機能属性）のうち分子機能も類似しているという定理であり（図-2実線矢印），類似構造を持つタンパク質は進化的に共通の祖先を持つという分子進化学定理の別の表現といえる．類を作って典型の特徴を当てはめる行為は人間が世界を認識する基本手法であるので特に法則としては意識されないが生物学では思考の基礎を成している．遺伝子のDNA配列やタンパク質のアミノ酸配列もそこから分子の立体的な形を算出して機能を予測するよりはむしろ，機能既知タンパク質との**類似性**を探すための属性の１つとしてもっぱら使われてきた．バイオインフォマティクスの主役が配列の類似性計算であり続けたのはタンパク質の構造の本質が配列であり，類似性定理の利用には配列の類似性計算が必要であったからである．さらに類の内包的定義すなわち対象を同類と見なす属性の明示ができ共通な機能属性と対にできた場合にはその属性は機能**モチーフ**と呼ばれ局所的な定理として収集されてきた（http://kr.expasy.org/prosite/）．このように枚挙といわれる生命科学も枚挙の結果を分類して常に抽象化し記述量を減らそうと努力しているのである．

## ポストゲノム研究

この文脈で説明すれば，ゲノム配列をあらかじめ決定してしまう行為は分かりやすい．配列によるタンパク質

の分類を全部やってしまい，これまでに知られている分子機能を類似性の定理を使って配列上類似している遺伝子にコピーするかたちで埋めようというわけである．しかし配列に従属な機能は**分子機能**であり，役割についてはこの定理は適用できない．そこでゲノムプロジェクトで同定された全遺伝子・蛋白（以下遺伝子）のセットを用いて，測定可能な“**機能属性**”についても網羅的に測定し，説明的な理解で最も重要な**細胞・生体レベルでの役割**に迫ろうというのが**ポストゲノム研究**である．機能属性のうち測定可能なものは遺伝子発現情報（どの細胞，どんなときに），細胞内の分布情報（核，細胞質，細胞膜等）そしてタンパク質と他の分子の結合情報である．これらの“測れる機能属性”の間の類似は「役割のないタンパク質は役割の行われる場に存在しない」という経験則に照らせば役割の類似性を示唆する．したがって類似した属性は類似した役割を持つという役割に関する類似性の法則が見出せるので構造から分子機能を埋めたのと同様に測れる機能属性から役割を埋めていけるはずである．以下に代表的な解釈を要するポストゲノムデータの具体例について簡単に説明する．

**遺伝子発現プロフィール**：数万の遺伝子の特定条件下（正常・疾患・薬剤刺激）での特定材料（細胞・臓器）における発現量（転写される量）に関する情報である．マイクロアレイと呼ばれるそれぞれの遺伝子に配列の一部のコピーを用意して，それらを高密度に整然と搭載したガラス板が工業的に生産されるようになって配列データを凄く勢いで産生されている．データ形式は，遺伝子行×材料列のシグナル強度（濃度）行列が基本である．解析はまず遺伝子方向のクラスタ化が行われる．ヒトの遺伝子発現データはNCBIのGeneExpressionOmnibus（http://www.ncbi.nlm.nih.gov/geo/）等に整理されている．

**蛋白結合情報**：相互作用情報とも呼ばれポストゲノムの代表的なデータである．測定方法は試験管内で行う方法と酵母細胞内で行う方法があるがいずれの場合にも1つのタンパク質に対してその他すべてのタンパク質のプールを作用させ結合が見られたタンパク質を吊り上げる（同定する）ことの繰り返し実験である．データは**結合有り無しの2値をとるタンパク質数×タンパク質数の行列**，もしくは繰り返した実験で何度再現されたかの再現回数を値に持つ行列が得られる．ヒトでは蛋白数が多いので総括的なデータはないが酵母ではかなりの数の相互作用データが存在する．それでも大抵の実験では相互作用（結合）の数は対象タンパク質の数とほぼ同数程度のきわめてスパースなデータしか得られないので，データは蛋白間をつなぐグラフとして解析されることが多い．詳細はMIPS（http://mips.gsf.de/genre/proj/yeast/index.jsp）などのサイトを参照されたい．

いずれにしても，多変量解析等ではおなじみの1～3万行の特徴ベクトル行列もしくは相関行列の形のデータがポストゲノムデータでありこのデータが作るタンパク質の構造が解釈の対象である．

## ポストゲノムデータ解釈

解釈とはいうまでもなくデータの抽象化（言語化）である．生物学的解釈ではポストゲノムデータに含まれる多数の遺伝子名称と個別のサンプル名称および数値を排除してデータを表現することである．たとえば**遺伝子発現プロフィール**の解釈は「転移性のある癌細胞では蛋白分解酵素の発現が上昇している」「増殖速度の速い細胞はDNA合成酵素の発現量が高い」などとなる．解釈には遺伝子群の抽象化，サンプル群の抽象化，属性間の関係の発見があるが，特にここでは測定値で作った遺伝子分類や遺伝子クラスタを称する述語（偏って存在する役割）を見つける作業について考える．測定は機能をある程度知っている1万程度の遺伝子とまったく機能の分からない1.5万の遺伝子を区別なく行うので，遺伝子群に与えられる機能概念は群内の機能未知のメンバに類似性の定理でコピーできるかもしれない．すでに万のオーダーに達しようとしているポストゲノムデータのセットのそれぞれが数十から数百の遺伝子群を提示しており，解釈に値しない人為的な間違いに起因するケースもかなり多いと思われている．したがって解釈は，その遺伝子群がこれまでの知識に照らしてもっともらしいか否かとその群を表現する述語を与えることが課題である．現存する代表的な機械解釈の方法を挙げて簡単に説明する．

### ■キーワード法　ジーンオントロジ（GO）

あらかじめ機能に関するキーワードを列挙しておき，それを個々の遺伝子に配ることで遺伝子を分類する方法である．遺伝子の類が与えられれば類の中に偏在するキーワードを探すことで類の機能が表現できる．機能に関するキーワードを列挙するときに間違いなく列挙しようとすると多少ともキーワードを分類しながら思い出す．さらにキーワードの間に粒度の違いがあることに気づくとさまざまなキーワードが広狭関係でつながることに気付く．つまりキーワードを完全に挙げようと思えば機能キーワードの階層分類を作ることになる．もともとアノテーション（ゲノム上に全遺伝子に関する知識をマップする作業）の生物種間の統一を目的として作られたジーンオントロジ（GO）と呼ばれる遺伝子機能に関する構造化された語彙は，測定値が作った遺伝子群の中に偏在する機能名称を探す目的に広く使われるに至って

いる．以下に少し詳細にGOに基づくデータ解釈のための遺伝子間の機能的な類似度の算出法について説明する．一般にオントロジは，概念をつないだ木の構造もしくはDirected Acyclic Graph（DAG）の形をしている．したがって，ある概念と別の概念との類似度もしくは距離を，たとえば共通の親ノードからのエッジの数などを用いて表現が可能である．この際重要なのは，階層構造の上部と下部とでは，同じエッジの数だけ離れていても意味的な距離は，前者のほうが遠いと考えるのが自然であることである（たとえば文献1）参照）．ノード間のエッジの数を数える際にはこの性質を考慮してエッジに重み付けをする．一方，オントロジの構造，特に階層の深さにはあまり意味がないという考えもある．特にGOでは，階層の深さは特定分野に関心が集中していることの反映であるという指摘もある[2]．この立場をとると，階層の深さの代わりに，概念の実データ（すでに大量の配列に対して概念がアノテーションされているのだから，そのデータセットを用いる）の中での出現確率，あるいは情報量をそれぞれのノードに割り当てる．下位のノードの出現確率は上位のノードに加算するものとする．ノード間の距離は，共通の親の中で最も出現確率が小さいノードの持つ値から計算する．この場合は階層構造の用途は構造そのものよりも，概念どうしの包含関係の判定が主となる．しかしこれらの方法で，オントロジを用いて計算した概念間の距離の妥当性を検証するのは一般的に困難である．興味ある試みとして，配列の相同性が配列に割り当てられたGO概念間の類似度に相関するかどうかが調べられた．その結果，Molecular Functionで強く，Biological Process，Cellular Componentでは弱く概念間の類似度と構造類似度に正の相関が見られた．これがたとえば類似性の法則の正当性とオントロジの妥当性を証明していることになるのか，それとも既知の機能は構造に過度に引きずられていると警告しているのかという議論は遺伝子機能の記述の根拠となっている事実に遡らなければ無意味なものであろう．

## ■遺伝子名称共起法

キーワードのマップが心理的なものであることが気になれば，遺伝子の間の関係を論文中での遺伝子名称の共起で測定する方法をとってみることができる．遺伝子×遺伝子の共起行列を遺伝子間の距離行列と見なせば，遺伝子群に対して機能的な相互関係の強さを表すような値，たとえば相互の共起の合計などを求めることができる．この名称共起法は医学系の200万を超える論文の要旨を対象に行った仕事はPubGeneいう名称で報告されている（http://www.pubgene.org/）．ただし名称共起法にはいくつかの問題がある．第1に共起行列が非常にスパースであること．彼らは1,000万以上の要旨を調べてそのうち19.2%部分に13,712の遺伝子名称を見つけたが共起関係の数は139,756であったと報告している．加えて共起相手の少ない遺伝子名称ほど数が多いというお決まりの関係が見られた．つまり関係の認められる遺伝子対が実際の数より低く評価されている可能性が高い．これは間接共起まで入れて緩和できる可能性もあるが，その正当性を評価するのは難しい．また名称の共起は必ずしも機能的な関係を表さず，構造の類似，染色体上の近傍などの構造情報を表している場合があることが予想されるということである．これもデータ内から構造的なトピックスのデータを排除することで対処可能かもしれない．この方法の限界は関係が深い遺伝子群を指摘できても関係の内容の表現ができない点である．共起関係をグラフとして表現したりすることや共起の起こっている論文のタイトルを返すことで表現を試みてもやはりさらなる解釈が要求される不完全な解釈機械である．

## ■第3の方法

第1の方法では遺伝子機能は宣言的にキーワードで表現された．第2では名称共起の頻度が遺伝子機能の関係の強さであるという仮定を基に多数の論文を観測してコードされた．第3の方法は実験研究者がデータ解釈を行うときの工程を忠実に再現することで行われる．専門家であってもよく機能を知っている遺伝子は100に満たないのが普通である．多数の遺伝子名称が作る構造を解釈するときには遺伝子名称を使ってまず教科書を調べようとする．ところが教科書には遺伝子蛋白名称のうちごく有名なものが多くても1,000程度しか書かれていない．しかもほとんどが遺伝子を大きな粒度の名称（たとえば“myc”）で呼んでいるのでデータ中で使われている固有の名称（N-myc，C-myc，D-myc）やそれぞれの配列データベース中でのIDとの対応に文脈の理解が必要なものが多い．ここで役に立つのがファクツブックといわれる便覧である．大抵は特定の機能グループの遺伝子それぞれに機能について分かっていること，構造の特徴，配列IDなどが一定の様式で記載されている．最近では配列データベース中のそれぞれに機能サマリー，関連文献などが記載されるかたちでこの機能はデータベース化された（http://bioinformatics.weizmann.ac.il/cards/）．したがって現在は教科書や便覧を経ずに配列データベースを調べさえすれば大抵の遺伝子についての機能は知ることができる．さてそれでは分野外の人間がこの配列データベースを使えば遺伝子機能のことが分かるかといえば決してそうではない．機能サマリーも関連論文も専門用語で書かれているから豊富な用語知識がなければ意味が分からない．また分野全体に対するセンス

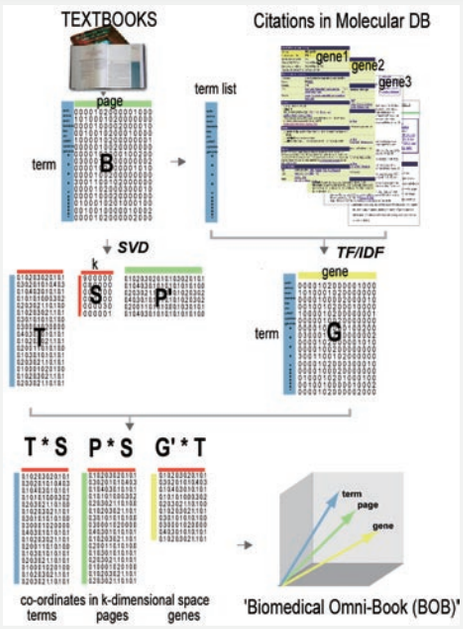


図 -3　第 3 の方法 BOB 構築のスキーマ

教科書を変えることで観点を変えられる .

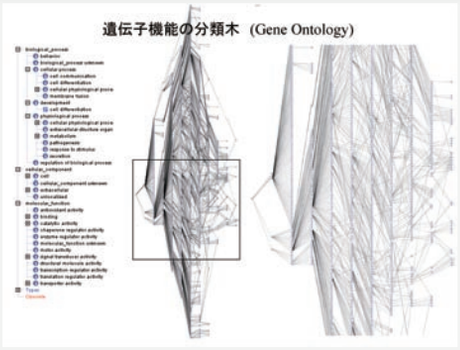


図 -4　ジーンオントロジ（GO）の木構造表現

この木は左に示した 3 つの太い枝 gene function, molecular component, molecular process のうちの最も木になりやすいはずの component の枝である．3 つの太い枝は筆者の遺伝子機能の分類との対応で表現すると，何を，どこで，何のために，となる .

がないので重要なこととありふれたことの区別が難しい．したがって正しく理解しようとすれば専門家がそうであったようにこの分野の教科書で基本的な専門用語の意味を知り，また分野のトピック構成を知り，その後に遺伝子DBを調べ，さらにリストされている文献を読むという順になる．第3の方法とはこの4つの段階をすべて踏襲する方法である．BOB（Biomedical OminiBook）と呼んでいる我々の第3の方法について少し説明する．まずBOBでは専門用語の意味データとして分野の教科書中の巻末インデックスデータを用いる．インデックスデータは代表的教科書で各頁から10～20用語程度を選び出してできた3,000用語×3,000頁程度の非常にスパースな行列である（**図-3**行列B）．これをページ区切り問題や内容の重複問題を克服するために100次元程度に次元を下げ近似して，用語の内容およびページの内容のベクトル空間を作成する（教科書空間，図-3右下）．次に機能の知れたすべての遺伝子に関するデータベースをもとに遺伝子の機能に関する記述をそれぞれについて集める．サマリーに加えて参照文献の要旨をつなげればそれぞれについて数百語から数千語の機能関連記述を付与することができる．この機能関連記述を教科書用語でインデックスすればすべての遺伝子の機能を関連用語のリストで表現した，遺伝子×用語の3,000次元の遺伝子機能行列を得る（図-3G）．この行列を教科書空間にLatent Semantic Analysis[3]に習って教科書空間に図-3にあるような手続きで写像すれば遺伝子機能が教科書空間のベクトルとして表現できる．遺伝子群が与えられると機能ベクトルの空間中のばらつきで解釈可能性を与え，解釈可能な場合にはその重心座標に近い用語や教科書ページを群の解釈として返すことができる．この方法の特徴は教科書の選択によって異なった遺伝子関係を与え異なった解釈を与える点である．解剖学の教科書を使えば遺伝子機能は解剖学用語の関係で構造化されデータ中の解剖学的に関連の深い遺伝子群が発見され解剖学的トピックとして解釈される．これは解剖学者がデータを読む態度に近いといえる．教科書の選択によって生化学的，内科学的，薬理学的等のあらゆる解釈が自在に行えるのである．

## 機械解釈の将来

現在遺伝子の機能的な近さとその内容を計算に用いる手法としては圧倒的にGOに依存する研究者が多い．オントロジで構造にされた用語を増やす作業と，遺伝子に当てはめてゆく作業は，分野全体で応援を受けて，壮大なデータができ上がりつつある（**図-4**）．この例にあやかって解剖名称や疾患症状などあらゆる概念を木の構造に宣言してゆく動きも見られる．この種の複雑な手法では仕組みが研究者に理解しやすいことが受け入れられる重要な用件であるように見受けられるが，分かりやすさに引きずられるに任せておいてもよいのであろうか．

人工知能（AI）システム間の知識共有の方法として

オントロジ（ontology）の考えが導入されたとき，オントロジは，“explicit specification of a conceptualization”[4] ──直訳すると概念化（conceptualization）の明示的な仕様書（specification）──と定義された．概念化とは，一般には，それまでばらばらに存在していた物事に対して互いの関連を見出し，1つのラベル──概念──を貼って理解できるようになることをいう．AIシステムの観点からは，「ある関心領域（area of interest）に存在すると想定されるオブジェクト，概念，その他の実体，ならびにそれらの間に成立する諸関係」[5]であり，知識を用いるすべてのシステムやエージェントが何らかの目的を持った行動を行うために持っているべき「抽象的かつ単純化された世界観（view of the world）」[4]と定義される．これらの定義から，オントロジの構成要件は単なる個々の事物の列挙だけでなく，それらの関係がむしろ重要であることが分かる．私たちは，概念間の類似度を知るために大規模なオントロジが有用であるかどうかについては懐疑的である．現実のオントロジでは，距離の妥当性の検討で提示した例で分かるように，概念間の関係は十分には記載されていない．そして，たとえ，概念間の関係を拡充する努力が続けられているとしても，関係が充足したオントロジを大規模に造るのは困難であろう．オントロジの作成にどのような困難が伴うかは，たびたび指摘されてきた（たとえば文献6））．すなわち，医学・生物学全般をカバーするならば，(1) オントロジのサイズが大きくなりすぎ，関係の管理が困難になる．(2) 概念間の関係が状況に依存するようになり，オントロジに定義された関係だけでなくもとの文献を参照する必要が出る．(3) 概念の変化に対応してオントロジの形を変えるのが困難になる．(4) 関係の一貫性を保てなくなる，などの問題点が挙げられる．

以上から，私たちは，1つのオントロジで医学生物学のすべての分野をカバーし，しかもそれを概念間の距離の計算に使おうとする企図に組しない．もちろんオントロジのすべての役割を否定するわけではない．むしろオントロジが文献6）で述べられているように「オントロジ的曖昧性の粒度の柔軟な調節を可能にする」性質を備えているならば，有用であると考える．つまり，意味の異同の判断は，多くの場合表記揺れの程度で判断できるが，それで説明できず，専門知識を用いて判断すべきものがある．たとえば同じタンパク質にまったく異なる名称がある場合がそうである．そして，同じ対象がどのような概念でどの程度細かく分類されているかは，専門分野によって異なるので，オントロジはまずそういう問題を解決すべきだということである．GOの本来の目的はアノテーションのための用語の統制controlled vocabularyであり[7), 8)]，その利用者はそのことをよく理解する必要がある．

ポストゲノムデータの自動解釈という課題を抱えることで筆者らは，機能とは何か？理解とは何か？解釈とは何か？などこれまで自身で行ってきた研究の部分を自問することでモデル化する機会を得た．計算機科学の生命科学への参入による実りは，まさにこのようにして「生命の物質的な理解」の方法について明示し共有し問い直す機会を与えることかもしれない．計算機に理解させるために進む分野の知識の整理は計算機よりもむしろ人間の教育に役に立つのかもしれない．100％の宣言とまったくの統計による手法との間を埋める多数の折衷案を経て計算機が文書を読んで理解するという理想型へと向かう機械解釈課題は，いずれにせよ生命科学の知識とこの分野の思考の双方を詳らかにすることで，さらなる展開へと導くことになりそうである．

**参考文献**


1) Lee, S.G., Hur, J.U. and Kim, Y.S.: A Graph-theoretic Modeling on GO Space for Biological Interpretation of Gene Clusters, Bioinformatics, 20(3): pp.381-388 (2004).
2) Lord, P.W. et al.: Investigating Semantic Similarity Measures Across the Gene Ontology: The Relationship between Sequence and Annotation, Bioinformatics, 19(10): pp.1275-1283 (2003).
3) Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R.: Indexing by Latent Semantic Analysis, Journal of the American Society for Information Science, 41, pp.391-407 (1990).
4) Gruber, T.R.: A Translation Approach to Portable Ontology Specification, Knowledge Acquisition, 5(2): pp.199-220 (1993).
5) Genesereth, M.R. and Nilsson, N.J.: Logical Foundation of Artificial Intelligence. Palo Alto, CA: Morgan Kaufmann (1987).
6) Tsujii, J.: Thesaurus or Logical Ontology, Which do We Need for Mining Text? in Fourth International Conference on Language Resources and Evaluation, Lisbon, Portugal: The European Language Resources Association (2004).
7) Schulze-Kremer, S.: Ontologies for Molecular Biology and Bioinformatics, In Silico Biol, 2(3): pp.179-193 (2002).
8) Ashburner, M. et al.: Gene Ontology: Tool for the Unification of Biology, The Gene Ontology Consortium, Nat Genet, 25(1): pp.25-29 (2000).


（平成17年1月12日受付）