［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

3

生命科学文献からの知識抽出と辞書構築
―その現状と課題―

小池 麻子
akoike@hgc.jp / 東京大学新領域創成科学研究科情報生命科学・日立製作所中央研究所

高木 利久
tt@k.u-tokyo.ac.jp / 東京大学新領域創成科学研究科情報生命科学

医学生物学分野の急激な発展と大規模実験法の開発に伴い，大量のデータを既存の文献知識を用いて解釈する必要性が指摘されている．対象となる文献量が膨大であることから，人手による文献知識の整理のみならず，自然言語処理技術を用いた効率的な情報検索／情報抽出とともに，実験解釈の自動化への期待が高い．東大高木研究室においては，医学生物系文献からの情報抽出や潜在知識の発見を目的として，自然言語処理技術を用いた情報抽出技術の開発とともに，辞書開発などの基盤整備を行っており，その成果は各種データベースとして公開している．本稿では，高木研究室において開発しているシステムの紹介とともに医学生物分野における自然言語処理の現状とその課題を概観したい．

背景

ヒトゲノムの完全解読をはじめとする医学生物学分野の急激な進展とtwo-hybrid法（タンパク質間相互作用解析）やDNA/Proteinマイクロアレイ（遺伝子／タンパク質発現解析）などの大規模実験法の開発に伴い，大量のデータが論文中もしくは各データベースに蓄積されてきた．大規模実験法により産出される大量データを解釈し，新たな知見を得るためには，文献中に埋もれた関連情報を網羅的に調べる必要がある．しかしながら，人手による検索作業は労働集約的であり，場合によっては実現不可能でさえある．たとえば，two-hybrid法によって一度に何百，何千ものタンパク質（遺伝子産物）の相互作用（protein-Aがprotein-Bと結合するという情報）が検出されるが，そのうちどれが既知かを判別することは人手では容易にできない．なぜならば，有名な遺伝子になると多くの同義語（シノニム）が存在するうえ，発表されている関連文献数も数千のレベルだからである．このような状況下において，自然言語処理による情報抽出技術や情報検索技術が近年脚光を浴びつつある．自然言語処理による情報抽出の研究としては，タンパク質間の相互作用の抽出をはじめ[1)～4)]，遺伝子と疾患の潜在的な関係性の抽出（特定の遺伝子が疾患とどのような関係にあるのか）[5), 6)]，DNAマイクロアレイの解釈（発現情報が類似である遺伝子群が文献中でどのような関係として記述されているのか）[7)]，さらには，一般化した情報抽出システムであるMedMiner[8)]などが報告されている．また，KDD-cup，TRECのゲノムトラック，

BioCreAtIvEなどのコンテストも次々と開設され，解くべき課題を研究者間で共有することにより各手法の比較を行い，生物ドメインにおける自然言語処理の問題点を明確化し，この分野の発展に貢献している．

高木研究室においては，医学生物系文献からの情報抽出や潜在知識の発見を目的として，自然言語処理技術を用いた情報抽出／情報検索技術の開発とともに，医学生物学用の専門辞書の開発などの基盤整備を行っており，その成果は各種システムとして公開している．公開システムの1つであるPRIME（http://prime.ontology.ims.u-tokyo.ac.jp）は，自然言語処理により自動抽出した遺伝子／タンパク質／ファミリ（類縁タンパク質群）／化合物の相互作用情報や機能情報と配列情報を蓄積しており，それらのコンビネーションによりパスウェイ（遺伝子／タンパク質／化合物の相互作用のネットワーク）比較・推論ができる描画ビューアも備えている．本稿では，高木研究室において開発している自然言語処理系のシステムの紹介とともに医学生物分野における自然言語処理の現状とその課題を概説する．

医学生物分野における辞書・シソーラス・オントロジ

より高度なテキストマイニングが可能か否かは，語彙に関するシソーラス（語彙集）や辞書がどの程度整備されているかに大きく依存する．残念ながら現状は不十分と言わざるを得ない．高木研究室において開発している辞書と公開されている辞書・シソーラスを紹介する．ここでは紹介しないが，専門用語の自動抽出，意味クラスの自動分類やシノニム収集の自動化などもこの分野の重要な課題である．

■遺伝子名辞書，ファミリ名辞書，機能用語辞書の構築

文献から遺伝子／タンパク質名前を自動抽出する手法は，大まかにはルールベースのheuristicsを使うもの，確率・機械学習を使うもの，辞書／シソーラスを使う方法とその混合方式がある．遺伝子の命名法は生物種ごとに大きく異なるため，精度を上げるためには，種ごとにチューニングが必要である．また，遺伝子はたくさんの同義語（シノニム）を有しているため，どの遺伝子かを特定したい，もしくは文献情報と配列情報にリンクさせたい場合，辞書を利用する必要がある．遺伝子名については，各生物種の主要研究機関のデータベースが収集・整理しているが，多くの場合不十分である．高木研究室では，主要な生物に対し遺伝子名辞書GENAを開発しており，生物種にもよるが90～95%の遺伝子名はカバーしている．GENAは各種データベースからの収集，略語抽出プログラムによる収集，主語，目的語の意味クラスを遺伝子／タンパク質／化合物に限定する動詞（phosphorylate，methylate等）を利用した収集，生物学者の知識による収集，などさまざまな手法を使って収集している．残念ながら，収集先のデータベースには相当数の誤った名前を含むため半自動的にそれらを削除しているが，さまざまな原因から自動的に除去することが困難な名前も存在する[9]．GENAにはそのほか，さまざまなサイトから収集した化合物情報も蓄積している．こちらは，MEDLINE（米国国立医学図書館が管理している文献情報データベース）のアブストラクトに出現する化合物の約85%程度が収集されている．

また，とかく遺伝子／タンパク質が着目されがちだが，文献中には遺伝子まで特定できない上位の概念（たとえば，RasはH-ras, K-rasの上位概念）が記されていることが多い．高木研究室はこれに対応するために，半自動的に階層構造を持つファミリ名辞書を構築している[9]．そのほか，遺伝子の機能を自動抽出するために機能用語を収集しているが，これについては後述する．

GENA は http://gena.ontology.ims.u-tokyo.ac.jp:8081/search/servlet/gena から，ファミリ名辞書はhttp://marine.ims.u-tokyo.ac.jp:8080/Dict/family から公開している．

■公開されているシソーラス・オントロジ

ここ数年は，急激な分子生物学分野の発展とともに研究分野の細分化が進み，分野ごとに新たな専門用語が出現するが，それらの専門用語が分野間で非共有になる傾向にある．そこで，概念の定義と概念間の関係の明確化を目的としてオントロジを作成する動きが活発である．その中でも，Gene Onotologyコンソーシアムでは主な真核生物（細胞核を持ち有糸分裂を行う生物）の研究機関が協調して，biological process（生物学的機能），molecular function（分子機能），cellular component（細胞の構成要素）についてオントロジを構築しており，遺伝子／タンパク質配列のアノテーション（生体内でどのような機能に関与しているか注釈をつけること）の標準規格になりつつある．しかしながら，このオントロジはあくまで概念の定義と分類が目的であるため，収集されている約2万用語は，これらの分野をカバーするシソーラスという観点からは不十分である．最も有名なシソーラスとしては，MeSH（Medical Subject Heading）termがある．米国NCBIがサービスしているMEDLINEでは，時間遅延があるものの全アブストラクトに論文の概要を表すMeSH termを付与している．chemistryやdiagnosisなどの83の意味ク

ラス（qualifier）とともに，Supplementary concepts（下位概念を持たない化合物情報）が計約2万（同義語11万）件，Main Headings（それ以外）が計約14万（同義語23万）件，階層構造で定義されている．そのほか，SNOMED（Systemized Nomenclature of Medicine）では部位・臓器，病理学的所見，生理機能，病名，処置・手術，薬品・化学物質，生物など11のセクションに分けて約34万概念について階層構造を構築している．疾患名に関しては，ICD（International Classification of Disease）およびICDをベースにしたdisease ontologyが構築されているが，共に名前のシソーラスというよりむしろ，疾患名の分類である．網羅性の高さでは，半自動的に収集しているUMLS（Unified Medical Language System）が群を抜く．独自に収集している語彙だけでなく，MeSH，SNOMEDなど外部の語彙情報も含まれている．UMLSでは概念を135の意味クラスに分けており，現在200万程度の用語を収集している．しかしながら，それでも網羅性としては不十分であるし，何よりも半自動で構築されたものなので誤りも多く，使用する場合は配慮が必要である．

知識抽出

どのような知識を抽出することが現在必要とされているのだろうか？　最も需要が高い情報抽出は，やはり，タンパク質／遺伝子／化合物間の相互作用およびその機能である．そのほかにも酵素の活性部位（化学反応を起こす部位）や遺伝子と疾患との関係の抽出も重要度が高い．最近では，文献中に埋もれている潜在的な知識の発見が注目されつつある．パーザ（構文解析プログラム）の精度や照応関係（代名詞や冠詞が文中のある表現を指すこと）の解消等も他の分野と同様に重要な課題である一方，研究目的にあった抽出すべき意味クラスの分類など細かいチューニングも必要である．高木研究室において行っている情報抽出技術について以下説明する．

■タンパク質／遺伝子／化合物間相互作用

タンパク質は生体内で他の生体分子と相互作用することによってその機能を発現している．したがって，生命現象を解明するためには，タンパク質／遺伝子／化合物のネットワークを解明することが必要である．ここでいうネットワークとは，“ERKがELK-1を活性化し，ELK-1がc-fosのpromoter領域（遺伝子の転写開始に関与するDNA上の特定領域）に結合して発現させることにより増殖・分化が誘導される”などの一連の作用のことである．**図-1**にEGF receptorを介したシグナル伝達系パスウェイの概略図を示すが，実際はもっと複雑である．図中のノードがタンパク質／遺伝子をエッジがタンパク質／遺伝子間の制御関係を示す．これらの相互作用情報は，大規模実験データの場合は各研究室のデータベースで管理されていることが多いが，小規模実験データは文献中に埋もれてしまっているのが現状である．現在では，文献から人手で相互作用情報を抽出し，代謝系データベース[11]やシグナル伝達系データベース[12]が構築されている．これらのデータは詳細な情報まで精度高く整理されている反面，全体としてのコンテンツ量が不十分で最新のデータも不足しがちという短所がある．また，各研究者が発見した相互作用を自ら登録できるデータベースもあるが，なかなか広く認知されないうえ，相互作用の属性に関しては各研究者が判断するには複雑であることから，提出されるデータの質のばらつきも問題である．

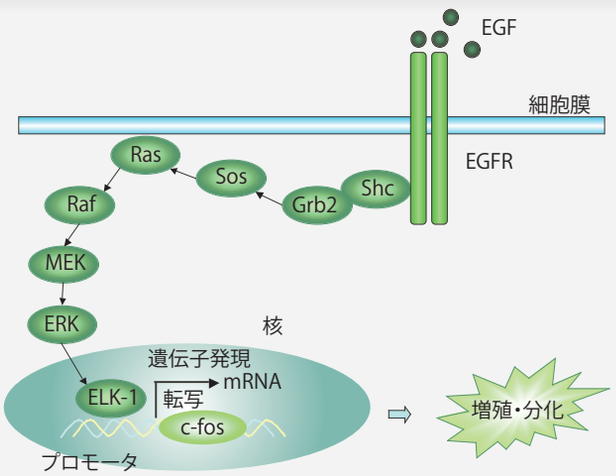

図-1　EGF receptorを介したシグナル伝達系パスウェイの概略図

■遺伝子名／タンパク質名／ファミリ名／化合物名の認識

このような状況下において，自然言語処理による情報抽出が脚光を浴びているのだが，相互作用を見つけるためには，まず遺伝子名やファミリ名を認識する必要がある．簡単そうに見えるが奥が深く，この過程の精度が，相互作用抽出に限らずテキストマイニングの精度を大きく左右するといっても過言ではない．遺伝子名認識の問題として，（1）遺伝子名の多数のシノニムの存在，（2）遺伝子名の表記揺れ，（3）一般動詞，名詞と同じスペルを持つ遺伝子名の存在，（4）多数の遺伝子で共通のシノニムを持つ曖昧性の問題（たとえばNIKはMAP3K14とMAP4K4のシノニムである），が挙げられる．高木研究室では（1）を辞書の構築により，（2）を名前の認識方法の工夫により，（3），（4）の名前を認識した後のpost-

processingによりできる限り回避している．

遺伝子の表記にはかなり揺らぎがあるので，辞書に登録されている名前を完全一致で検索するとかなりカバー率を下げてしまう．ERK-1をERK1と記述するハイフンの有無，mitogen activated ser/thr kinase 1をmitogen activated kinase 1と記述する単語の挿入／欠如，STE11/STE20をSTE11/20と記述する省略形などさまざまである．しかしながら，その揺らぎはある程度規格化（特殊文字はスペースに入れ替え等）すれば数個のルールに帰着可能である．高木研究室においては，GENAのエントリをベースにバリエーションを自動生成したのちトライ構造（辞書引きに適した木構造形式の探索アルゴリズム）にして高速に遺伝子名を認識している．このトライ構造もSTE11/20をSTE11/STE20と認識できるように工夫した構造となっている．遺伝子を認識した後，(1) フルネームと略語のペアの利用，(2) 各遺伝子と特徴的に共起する用語（keyword）のチェック，(3) shallow parser（品詞と係り受けの関係のみ付与）後の受け名詞のチェックなど多数のpost-processingを行い上述の (3)，(4) の問題を解決する[9]．この一連のプロセスにより，生物種にもよるが90%程度の精度・再現率で遺伝子名／タンパク質名まで特定できる．また，ファミリ名や化合物名においても同様のプロセスで認識している．

■相互作用情報の抽出

文献に記述されている遺伝子／タンパク質／化合物間の相互作用情報の抽出を考えた場合，相互作用には，物理的な相互作用とともに，遺伝子間相互作用がある．遺伝子間相互作用とは，“Protein-A induced the expression of gene-B mRNA.”（タンパク質Aが遺伝子BのmRNAの発現を誘導した）や“Gene-A is synthetically lethal in combination with a deletion of the gene-B.”（遺伝子Aは遺伝子Bと合成致死性を示す）などの，間接的な相互作用である．自動抽出においては，物理的な相互作用もしくはactivate, inhibit（何か他のタンパク質を介す可能性があるので直接的な相互作用とは限らない）などの明らかな制御関係に着目する傾向にあるが，高木研究室での抽出は両者を対象にしている．また，タンパク質のpromoter領域への結合，タンパク質–化合物間の相互作用なども収集対象としている．

相互作用抽出の流れは以下の通りである．(1) 遺伝子名を認識しIDに変換（遺伝子はIDで管理），(2) shallow parsing, (3)名詞句を認識するとともに，従属接続節，等位接続節，挿入句などの解析をし，ACTOR（doer of action，動作主）とOBJECT（receiver of action，被動作主）の関係を抽出する．(4) ACTORとOBJECTが特定の関係で記述されているときは，2項関係として取り出す．(5) 使用した関係（特定の動詞や名詞句）により，相互作用の物理的な種類：直接・間接，生物学的な種類：活性化，抑制，輸送，制御，その他の何らかの関係性，などに分類する．

さて，相互作用はさまざまな方法で記述される．ACTORとOBJECTは主語と目的語の関係だけにとどまらない．“activation of protein-A by protein-B”（タンパク質Bによるタンパク質Aの活性化）や“protein-B-induced protein-A（タンパク質Bに誘導されたタンパク質A）のような名詞句の中で係り受け関係で記述されるものもあれば，“protein-A is a ligand of protein-B”（タンパク質Aはタンパク質Bのリガンドである）のような，activate, inhibitなどの言葉とは無縁の形で記述されるものもある．また“the expression of protein-A causes the activation of protein-B”（タンパク質Aの発現がタンパク質Bの活性化の原因となる）のように動詞と目的語中のkeyword（expressionやactivation等）とのペアで相互作用として認識できるものもある．高木研究室においては，(1) 特定の動詞の主語，目的語で取り出すもの，(2) 動詞は特定せず，目的語の中にkeywordを認識して主語と目的語中の遺伝子名を取り出すもの（この場合は，遺伝子名とkeywordの相対位置にはかなりの制限を課す），(3) 名詞句の中で特定のフレーズを構成するものなど，さまざまなタイプで取り出している[2), 3)]．

情報表現に共通の問題だが，情報抽出にはグレーゾーンが存在する．相互作用といっても，利用する研究者によって，何を取り出したいかは大きく異なるためである．“Protein-A activates protein-B under the expression of protein-C.”（タンパク質Cの発現下でタンパク質Aがタンパク質Bを活性化させる）からprotein-Aとprotein-Cの何らかの関係を取り出したいと思うか否かは，その研究目的に依存する．我々のシステムにおいては，幅広い検索要求に応えるために，この類の関係は低い信頼度のマークを付けて一応抽出している．相互作用は，コーパスにもよるが，50%台の再現率と90%レベルの精度（正解基準は相互作用の向きと遺伝子を特定することであり，不正解例にはGENAのエラーを含む）で抽出している[2), 3)]．複数の文にまたがる，もしくは単一文中での照応関係はかなり多く，再現率を大きく下げる要因となっている．照応関係の解消は試みているが，精度を下げない効果的な方法は他の分野と同様に見つかっていない．また，shallow parserでの動詞／過去形と形容詞／過去分詞との間違い，名詞と動詞との間違いも精度・再現率を下げる要因となる．

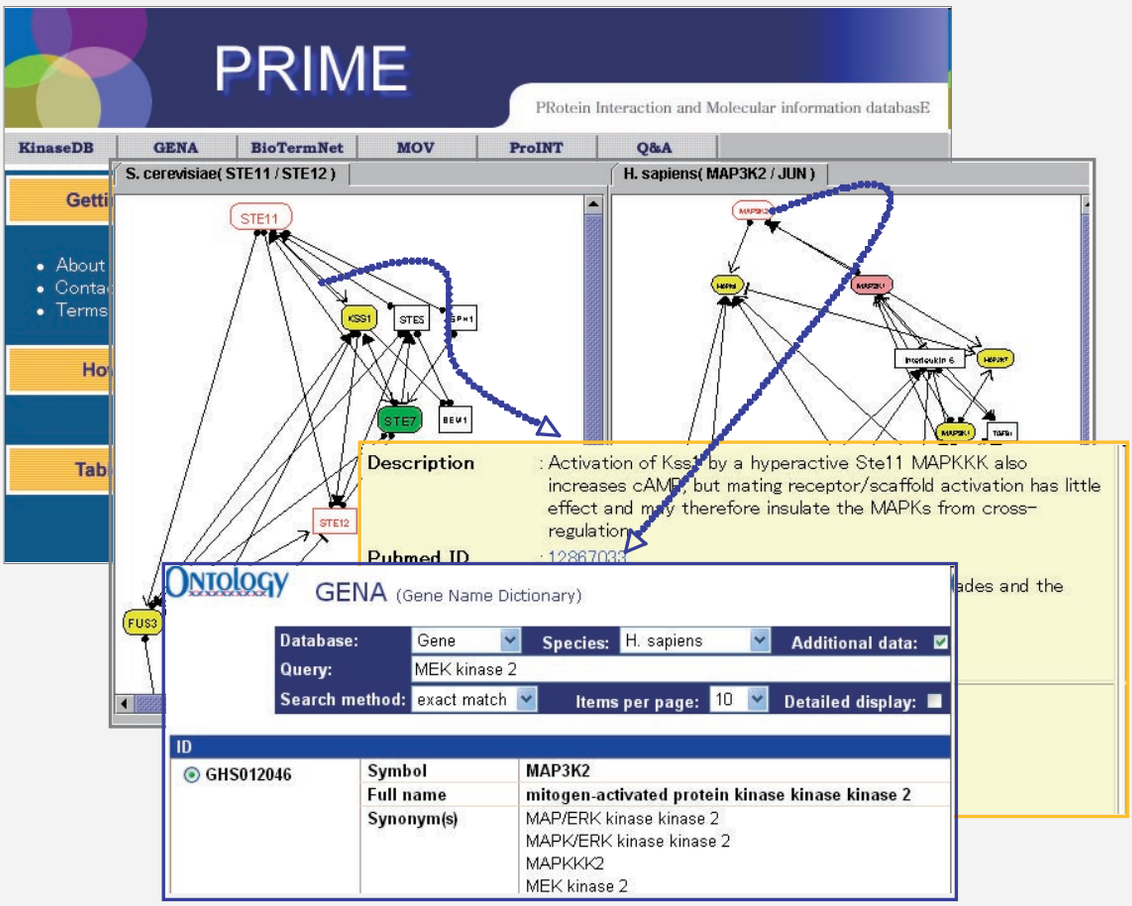

図-2　PRIME のパスウェイ表示例

現在では，主な真核生物に関して約92万（冗長で228万）の相互作用情報を抽出しPRIME databaseとして公開している．このデータベースでは，配列情報と組み合わせ，種間でのパスウェイ比較や他生物のデータを使用して，ドメイン構成（配列保存領域）や配列類似性などの条件下で対応するパスウェイを描画することが可能である．**図-2**にPRIMEで酵母（STE11からSTE12）とHuman（MAP3K2からc-Jun，STE11とMAP3K2は祖先遺伝子が同じ遺伝子（オルソロガス））のパスウェイの例を示す．このviewerではカーソルを動かすと対応するオルソロガスの遺伝子の色が変化する．エッジとノードは，相互作用の情報抽出に使用した文章と配列情報(辞書情報を含む)にそれぞれリンクしている．

■遺伝子／タンパク質の機能

従来の分子生物学では研究者が対象とする遺伝子数は数十個であり，その機能のほとんどは研究者の頭の中に整理され得る量であった．しかし，大規模実験法が導入されてから，扱うべき遺伝子数が数百以上となり機能をよく知らない遺伝子を解析する必要性が出てきた．そこで，個々の遺伝子の機能情報を文献から抽出・整理するという技術が注目されつつある．Gene ontology（GO）がよく整理されていることから，その多くはGOベースである．Gene ontologyは，主に遺伝子配列のアノテーションを目的として作成されたオントロジで，階層構造（有向非循環グラフ）を持つ．主な真核生物の研究機関において，人手で各遺伝子に該当するGO-IDを関連付けるアノテーションを行っているが，現時点では十分なアノテーションには至っておらず，自動的な機能情報の抽出が着目されている．全アブストラクト／単一アブストラクトを用いた機械学習でも予測は可能であるが，多くの生物学者は証拠となるセンテンスの抽出を望むことから，高木研究室においてはセンテンスレベルで各遺伝子にGO-IDを自動的に付与方法を開発している．

■機能用語の収集

センテンスレベルでGOベースの遺伝子機能の認識を行うためには，機能を表す用語を認識する必要がある．しかし，GOの語彙（GO-term）は，機能の分類を主な目的として構築された統制語であり，文中で使用さ

れる機能用語のカバー率は決して高くない．したがって，GOの語彙を基にあらかじめ機能用語を収集しておく方法がある．我々は，主に次の5種類の手法で機能用語を収集している．

（1）GO-termとの共起の利用，（2）GO-termとのcollocation（連語構成語）の類似度の利用，（3）ルールベースで統語的／意味的バリエーションの生成，（4）パターンマッチでの酵素名の収集，（5）動詞と専門用語のコンビネーションの作成．その他，UMLS，MeSH，WordNetなどのシソーラスを使い，下位語も追加するなど，さまざまな工夫を施している．共起では主に関連語が収集され，局所文脈の類似度では主に類似語が収集される．類似語は，常に同じ意味クラスに属するが，関連語は必ずしも同じ意味クラスには属さない．たとえば，“metamorphosis”と“metabolism”とは類似語であり，“chaperon”と“protein folding”とは関連語である．“chaperon”は他のタンパク質がfoldingすることを助けるタンパク質であり，“protein folding”は作用である．“Protein-A helps the protein folding of protein-B.”（タンパク質Aはタンパク質Bのフォールディングを助ける）でも“Protein-A is a chaperon of protein-B.”（タンパク質Aはタンパク質Bのシャペロンである）と表現してもprotein-Aには同様に，“GO:0006457:protein folding”が付与される．機能は，関連語でも類似語を使っても表現可能である．collocation 類似度の高い名詞句は，あらかじめ，各名詞句のcollocationをベクトル化し，与えられたqueryのcollocationベクトルとの類似度が高いベクトルを持つ名詞句として収集している．（1），（2）は完全自動化はかなり難しいので，候補用語が正しいか否かは最終的に博士課程レベルの生物専攻の学生に判定してもらっている．（3）においては，“apoptosis<-->apoptotic”，“transport<-->transporter”などの派生語，逆成語との半自動生成とともに，“regulation of osteoblast differentiation”<--> “osteoblast differentiation regulation”などの統語的バリエーションの自動生成を行う．また，単語ごとに収集した関連語／類似語を基に（たとえば“metabolism”-> “metabolic, metastasis, metamorphosis, reducer, reduction”）その単語を含む用語のバリエーションの自動生成を行うなど，さまざまな方法で用語を生成／収集している．

■機能の抽出

各遺伝子にGO-ID（Gene ontologyのID）を自動的に付与する過程を示す．機能の抽出の流れは，相互作用と同様以下の通りである．（1）遺伝子名，機能用語の認識，（2）shallow parsing，（3）文構造解析を行いACTORとOBJECTの関係を抽出，（4）ACTORとOBJECT（or 動詞のみ or OBJECTと動詞のペア）が特定の関係で記述されているとき，遺伝子はその機能を有すると判定する．機能の表現は実にさまざまであり，相互作用の抽出に比べて数段に難しい．基本的には，動詞にはほとんど制限をかけずにACTORが遺伝子，OBJECTが機能になるときに遺伝子機能として抽出しているが，requireなどいくつかの動詞に関しては，その逆を抽出対象としている．相互作用と同様，ACTORとOBJECTの抽出は，主語と目的語だけにとどまらずさまざまである[10]．また，“GO:0006846:acetate transport”（アセテートの輸送，GO:0006846はGO-ID）などの場合は，動詞が，“transport”，“locate”，“localize”，“translocate”，“import”，“export”のいずれかであり，目的語にacetateもしくはその下位語が存在し，かつ主語に遺伝子名が存在するとき，その遺伝子にこのGO-IDを付与する．動詞とキーワードとのコンビネーションはある程度までは自動的に行っている．そのほか，palmitoylateなどの動詞の場合は，目的語が何であれ，主語になる遺伝子の機能は自動的に決定する（この場合は"GO:0018318:protein amino acid palmitoylation"）．

ある機能に関与するか否かのグレーゾーンはかなり広い．“Protein A is highly expressed during mitogenesis”（タンパク質Aは有糸分裂中に多く発現している）からprotein A にmitogenesisを付与するか否かは前後の文脈にも依存する．また，人手によるアノテーションにおいても，キュレータ（アノテーションを行う専門家）によるアノテーション基準のばらつきはBioCreAtIvEなどでも指摘されている．

相互作用に比べると，照応関係の率も増加し，表現の多様性も広がるので遺伝子機能の自動抽出は非常に難しい課題である．特に，詳細な機能のGO-IDを付与することはかなり困難である．上位の概念のID付与を可とすると精度は90％レベルになるが，この性能評価はかなり複雑になるので，文献10）を参照していただきたい．現在では，主な真核生物に関して36万件（非冗長）程度の遺伝子–機能のアノテーションを行っており，相互作用と同様にPRIMEから公開している．

■潜在的知識の発見

今までは，かなり規定された知識抽出について紹介してきた．しかしながら，DNA/Proteinマイクロアレイなどの実験解釈の場合は，もう少し大まかな関係性が必要となることもある．一方，現在の大量な文献データを解析すると，（単独の論文に書かれてはいない）新たな関係性や知識が発見できる可能性も少なくない．ここ数年，

潜在的知識発見に向けた研究が着目されている．我々も，概念間の関係性のネットワークを解析することにより，この課題に取り組んでいる．

■概念ネットワーク

概念間の統計的な関連度をSMART-measure[13]で評価し，ユーザが入力するqueryから関連度の高い概念（関連語）を計算し，その関連語から新たな関連語を計算することの連続により概念間のネットワークが描画可能である．我々のシステム（BioTermNet: http://btn.ontology.ims.u-tokyo.ac.jp）では，この統計的なエッジの生成とともに，上述の遺伝子／タンパク質／ファミリ／化合物間の相互作用や機能情報もエッジとして組み込み可能にしている．統計情報だけでは，出現頻度が低いが重要なタンパク質間相互作用が捨てられる傾向にある一方，文献中に明記されている既知の相互作用情報だけだと概念間の関係性が発見できないこともあるため，両者は相補的な関係と言える．このシステムを利用すると“AGT-gene”->“renin activity”->“stenoses, renal arteries”->“blood pressure control”->“hypertension”のような概念間の関係が示され，以下のように，解釈可能である．“hypertension”と“stenoses, renal arteries”の患者には“blood pressure control”が必要である．“renin activity”は“stenoses, renal arteries”のときに上昇することが知られている．また，“renin”は“AGT”を“angiotensin II”に変換し（PRIME data），この相互作用は“renin-angiotensin system”と呼ばれる．このシステムでは，複数のqueryを同時に入れられるので，DNA/Proteinマイクロアレイデータで同時にクラスタリングされる遺伝子（発現が類似である遺伝子）の間にどのような繋がりがあるのか調べるのに便利である．また，連鎖解析などである疾患関連遺伝子が存在するゲノム上の領域が明らかになったときに，その中で最も表現型（外見上現れる形質）と関連性が高い遺伝子を探索するときにも有効である．しかしながら，不十分である点もまだ多い．現システムにおいては，GENA，UMLSをはじめとするさまざまなシソーラスを用いるとともに，独自に取り出した名詞句を利用している．しかしながら，上位語・下位語の関係の多くは未登録であるし，シノニムに至っては一部を除いてかなり不足している．より高度な関係を抽出しようとすると，さらに語彙情報が整備されている必要がある．まだ解くべき興味深い課題は多い．

今後の展望

医学生物学分野における情報抽出システムを紹介してきた．この種の情報抽出結果の利用は徐々にではあるが生物分野の研究者へ浸透しつつあり，短期間で実験系研究者からシステムへのフィードバックがかかるようになってきた．アカデミック，市販ともに複数の文献系システムが開発されている現在においては，情報処理としてのアイディアの斬新性だけでなく，実用に耐え得る精度を出すシステムでないと，バイオ系情報処理として評価されないフェーズに入っている．医学生物学分野における，テキストマイニングへの注目度は年々高まりつつあり，各国でプロジェクトが動き出している．また，辞書・シソーラス・オントロジの整備も進行している．今後は，これらのリソースの利用により，さらに高度なテキストマイニングのシステム開発が期待できる．

参考文献

1) Friedman, C., Kra, P., Yu, H., Krauthammer, M. and Rzhetsky, A.: GENIES: A Natural-Language Processing System for the Extraction of Molecular Pathways from Journal Articles, Bioinformatics, Vol.17, Suppl 1: S74-82 (2001).
2) Koike, A., Kobayashi,Y. and Takagi, T.: Kinase Pathway Database: An Integrated Protein-Kinase and NLP-based Protein-Interaction Resource, Genome Research, Vol.13: pp.1231-1243 (2003).
3) Koike, A. and Takagi, T.: PRIME: Automatically Extracted PRotein Interactions and Molecular NLP-based Information databasE. In Silico Biology, 5, 0003 (2004).
4) Daraselia, N., Yuryev, A., Egorov, S., Novichkova, S., Nikitin, A. and Mazo, I.: Extracting Human Protein Interactions from MEDLINE Using a Full-Sentence Parser, Bioinformatics, Vol.20, (5), pp.604-611 (2004).
5) Perez-Iratxeta, C., Bork, P. and Andrade, MA.: Association of Genes to Genetically Inherited Diseases Using Data Mining, Nat Genet. Vol.31(3), pp.316-319 (2002).
6) Daraselia, N., Yuryev, A., Egorov, S., Novichkova, S., Nikitin, A. and Mazo, I.: Extracting Human Protein Interactions from MEDLINE Using a Full-Sentence Parser, Bioinformatics, Vol.20 (5), pp.604-611 (2004).
7) Jenssen, TK., Kuo, WP., Stokke, T. and Hovig, E.: Associations between Gene Expressions in Breast Cancer and Patient Survival, Hum Genet, Vol.111 (4-5), pp.411-420 (2002).
8) Tanabe, L., Scherf, U., Smith, LH., Lee, JK., Hunter, L. and Weinstein, JN.: MedMiner: An Internet Text-Mining Tool for Biomedical Information, with Application to Gene Expression Profiling, Biotechniques, Vol.27 (6), pp.1210-1214, pp.1216-1217 (1999).
9) Koike, A. and Takagi, T.: Proceedings of HLT/NAACL BioLINK Workshop, pp.9-16 (2004).
10) Koike, A., Niwa, Y. and Takagi, T.: Automatic Extraction of Gene/Protein Biological Functions from Biomedical Text, Bioinformatics, in press (2004).
11) Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M. The KEGG Resource for Deciphering the Genome, Nucleic Acids Res. Vol.32, D277-80 (2004).
12) Krull, M., Voss, N., Choi, C., Pistor, S., Potapov, A. and Wingender, E.: TRANSPATH: An Integrated Database on Signal Transduction and a Tool for Array Analysis, Nucleic Acids Res, Vol.31 (1), pp.97-100 (2003).
13) Singhal, A., Buckley, C. and Mitra, M.: Pivoted Document Length Normalization, In Proceedings of ACM SIGIR'96, pp.21-29 (1996).

（平成16年12月29日受付）