3 テキストマイニングによる潜在的知識の発見支援

小池 麻子　（株）日立製作所中央研究所

医学生物学分野においては，各種研究の飛躍的進展に伴う論文の指数関数的増加により，自然言語処理技術を利用した文献からの知識抽出や，テキストマイニング技術を利用した実験結果の自動解釈などへの期待が高まってきている．一方，文献情報の大容量化，および，専門の細分化による情報の非共有化に伴い，各文献に記述されている知見を組み合わせることによる文献中の潜在的知識の発見もしくは仮説生成の可能性が議論されている．本稿では，潜在的知識発見の現状と課題を概観するとともに，当研究室で開発している潜在的知識発見システムを紹介する．

医学生物学分野のテキストマイニング

近年の文書の電子化と大容量化に伴い，多様な分野において，情報抽出，情報検索，テキストマイニング技術の重要性が増していることは周知の事実である．学術分野においてもその例外ではなく，特に，数式や化学式を使わず自然言語で知見を記述することが多い医学生物学分野で，その傾向は顕著である．医学生物学分野における抄録の多くは，NCBI（米国国立バイオテクノロジ情報センタ）から PUBMED（文献データベース MEDLINE の online 版）として無料で提供されている．抄録数は指数関数的な増加の一途をたどり，現在では 1,600 万件以上登録されている．一方，ヒトや主要モデル生物のゲノム完全解読に伴い，大規模実験手法が多数確立されてきた．従来の分子生物学の手法では実験対象の遺伝子数がせいぜい 20 ～ 30 程度であったが，これらの技術の導入により，数百，数千遺伝子の挙動が同時測定可能となった．実験結果の解釈を行うには，過去の知見や主張を考慮する必要があるため，対象遺伝子数増加により関連論文の検索と精査にも並々ならぬ時間と労力を費やすことになる．

このような問題を解決すべく，文献からの遺伝子機能情報の抽出や蛋白質相互作用の抽出，疾患情報の抽出などに関する自然言語処理やテキストマイニングの研究が盛んになり，有償／無償でさまざまな文献ベースのシステムが提供されてきている．一方，医学生物学分野における専門分野の細分化により，分野間での情報共有が難しくなっていること，また，同一分野であっても情報の大容量化により個人が記憶できる限界を超えていることなどから，異なる文献に蓄積されている知見を組み合わせることによる文献中の潜在的知識発見もしくは仮説生成の可能性が論じられるようになってきた．一般的なテキストからの知識発見とは，文書クラスタリングなどのテキストマイニング一般も含まれるが，本稿での潜在的知識発見／仮説生成とは，**図 -1** に示すように，従来の検索では難しい潜在的な概念間の関係性（知識）の自動抽出を，知識の構造化とマイニングをベースに行う手法に限定している．

以下，医学生物学文献からの潜在知識発見／仮説生成研究の歴史と現状，課題を概観するとともにその展望について論じたい．

潜在的知識発見／仮説生成に関する研究の歴史

医学生物学における潜在的知識発見／仮説生成のパイオニアとしてシカゴ大学の Swanson による研究が挙げられる．1986 年に Swanson はレイノー病に魚油の摂取が有効であることを文献情報から予測した[1]．その当時は，魚油のレイノー病への効果は知られていなかったが，その後 DiGiacomo らによって実験的にも証明された．Swanson は，レイノー病患者に，高い血液粘性，強い血小板凝集作用，および血管収縮などの血管反射に関する特徴が見られること，また，魚油，および魚油に含まれる EPA（eicosapentaenoic acid）が血液粘性，および，血小板凝集作用を下げる働きがあることを人手で文献から調べ，上述の関係性を予測した．Swanson のこのモデルは ABC モデルと呼ばれている．この場合，A-term が魚油，B-term が血液粘性，血小板凝集作用，血管反射，C-term がレイノー病となる．A-term のみ指

定して関係性のある概念を予測する方法を open discovery，A-term と C-term を指定して関係性を予測する方法が closed discovery と呼ばれている．ABC モデルは，A-B, B-C の関係を示す文書があるときに，A-C の関係を示す文書がなくともその関係性を推論する単純なモデルである．上述の Swanson の予測は closed discovery であり，ある意味では仮説生成の一部（A-term に関連する着目すべき C-term を選択すること）が Swanson の経験によって行われていることになる．Swanson は，その後もいくつかの関係性を予測し，その一部は実験的にも証明されている．これらの予測は人手で行われたが，潜在的知識発見や仮説生成の自動化がいくつかの研究グループによって試みられており，その多くは，ABC モデルを踏襲している．たとえば，Gordon & Lindsay はレイノー病の記述されている文書（R）に出現するすべての単語と連語（X）について，R 中の X の頻度，X が出現する文書中の R の頻度，および，tf-idf（term frequency-inverse document frequency）などを利用し，Swanson が人手で行った知識発見の過程を自動的に行えるか否か検討している[2]．Weeber らは NCBI の MetaMap を利用して文書中の概念を UMLS（unified medical language system）のシソーラスと意味クラスにマッピングした上で，レイノー病の文書中に出現する概念の頻度を利用した知識発見の自動化の可能性を検討している[3]．これらの方法は，いずれも open discovery では B-term を人手で選んでおり，完全な自動化を試みていない．一方，Srinivasan は，PUBMED の文献に文書分類のために人手で付与されている MeSH（Medical Subject Headings）term を潜在的知識発見／仮説生成の対象概念とし，MeSH term に付随する UMLS の意味クラスを利用した上で，意味クラスごとに各概念の tf-idf をベースに B-term や C-term を絞り込み，潜在知識発見の精度を高めようと試みている[4]．Open discovery の場合は B-term を意味クラスごとに数個に絞り，自動的に C-term まで計算しており，Swanson の予測例を比較的上位で予測している．しかし，MeSH term はそれほど大きなメタシソーラスではないので，一般的な知識発見には不向きと思われる．

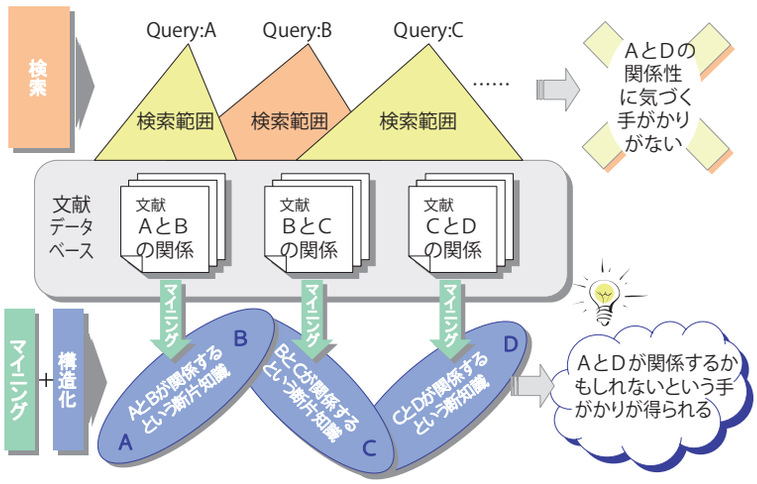

図 -1　従来の検索と潜在的知識発見支援システムの比較

潜在的な知識発見の課題

このように，文献ベースの潜在的知識発見の研究は盛んになりつつあるが，その多くは Swanson の予測の自動化の検証であり新たな関係を予測して実験的に証明されるに至った例は数少ない．仮説生成や潜在的な知識発見を自動的に行うことは困難なのである．その理由は下記のようにいくつか挙げられる．

1) **語彙の問題**：多くの概念は複数の同義語を持ち，これらを考慮しないと適切な関係性が抽出できない．また，同音異義語も多く存在し，特に略語や遺伝子名称に関しては語彙的曖昧性解消が必須である．シソーラスを用いない潜在的知識発見システムではこれらの点を考慮できないため，多くの場合意味のない関係性や概念が上位にランクされてしまう．

2) **低頻度概念の問題**：希に現れる概念でも潜在的知識発見という観点から重要なものがあるが，どのような統計的な手法を利用しても低頻度概念の関係性抽出は難しい．

3) **中間層概念選択の難しさ**：2 項間の関係性という意味においては正しいが，潜在知識発見／仮説生成という観点からは意味のない概念が多く現れる（上位すぎる概念であり情報量がほとんどない，意味クラスが不適切など）．

4) **関係性の種類指定の必要性**：関係の有無だけでなく，ある特定の関係が必要なときがある．たとえば，ある signaling pathway（蛋白質と蛋白質／化合物の相互作用）が活性化されて次の event が起こる場合は，その signaling pathway を構成している概念間（蛋白質間）の関係性（物理的相互作用）を利用しなくてはいけないが，統計的な手法だと物理的相互作用でない関係性（2 つの蛋白質の機能が似ている，発現パターンが似ているなど）も取り出してしまう．

5) **不十分な中間層の数**：始点と終点の概念の関係が遠いと，中間層が一層だとギャップが大きく解釈が難しい．

6) **ユーザの背景知識欠如の問題**：ユーザに十分な周辺知識がないと予測された関係が正しいのか否かが結果を一瞥しただけでは判断できず，そのエビデンスとなる文献をかなり読まないといけない．

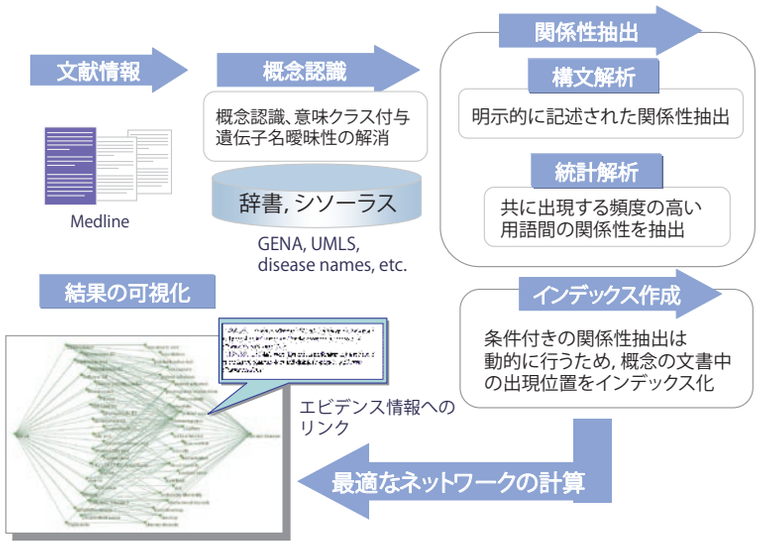

図 -2　BioTermNet の概観

7）**評価の難しさ**：closed discovery はともかく open discovery は本来正解が複数あるべきであり，どれが最も上位にくるべきか絶対的な価値基準を作ることが難しい．また，正解が数多く存在し，評価セットを準備することもかなり難しい．したがって，手法の比較が困難である．

これらの問題を解決すべく，我々は，潜在的知識発見と大規模実験結果の解釈のためのシステム BioTermNet を開発しており，以下の章でその概要を述べる．一部の機能は http://btn.ontology.ims.u-tokyo.ac.jp からアカデミックユーザには無料で公開している．

潜在的知識発見支援システム：BioTermNet

BioTermNet の概要

BioTermNet は主に MEDLINE をベースとした潜在的知識発見支援システムである．MEDLINE 以外にも OMIM や MEDLINEPlus, 薬品添付文書をベースとした概念ネットワークとのクロス検索も可能である．BioTermNet では上記の課題を以下のアプローチで可能な限り解決を図っている．

1）**シソーラスの充実化**：語彙の問題を解決するために，UMLS だけでなく我々が東大と共同で開発している GENA（gene name dictionary），disease name, pathway name などのシソーラス（MOV：multi ontology viewer として公開）を利用して，同義語を考慮した概念の認識を行っている．また，些細な綴りの多様性について（たとえば，NF kappa B, NFkappaB, NFKB），概念の認識過程で吸収するとともに，略語に関しては略語と full name との対応付けの手法を適用し，多義性に関しては，共起する用語から曖昧性の解消を可能な限り行っている（上記課題 1 に対応）．

2）**情報抽出と統計解析の併用**：低出現頻度の概念の関係性抽出，および，関係性の種類の指定が必要な場合を考慮し，構文解析による情報抽出と統計解析を併用して概念ネットワークの構築を行っている．現在のところ，関係性を厳密に区別したい蛋白質－蛋白質／化合物間相互作用情報を構文解析による情報抽出で行っている（上記課題 2，4 に対応）．残念ながら，関係性の多くは曖昧に記述される．たとえば，"cell division のときに特異的に遺伝子 A が発現している " という文章から，遺伝子 A が cell division に関与しているという関係性（示唆）を抽出しなくてはならない．ACTOR（doer of action）と OBJECT（receiver of action）との関係とは限らなく，関係性の記述の多様性から考えて，すべての概念間の関係性を構文解析で抽出することは難しい．

3）**意味クラスの利用と複数の中間層の利用**：始点と終点の概念間の関係性の遠い場合，複数の中間層の利用（たとえば，A-->B1-->B2-->C）が可能である．また，ユーザの仮説をある程度取り入れられるように，UMLS で概念に付随する，もしくは同様の基準で我々独自に付与した意味クラスを用いて中間層の概念の絞り込みを行う．たとえば，あるクエリに何らかの化合物が相互作用し，その化合物が遺伝子の発現を促進し，それにより，ある現象が起こることを仮定すると，中間層 B1 に化合物を，中間層 B2 に遺伝子のみを意味クラスとして選択し，これらを通る概念ネットワークを選択することになる．中間層の層数と意味クラスを可変とすることにより，ユーザの考えをある程度反映させることが可能となる（上記課題 3，5 に対応）．

4）**概念ネットワークの表示とエビデンスとなる文献の提示**：概念間の関係性全体を可視化することによりユーザが直感的に全体像を理解しやすいようにしている．また，探索表示された潜在的な関係が，潜在的知識発見の観点からもっともな関係か否かを容易に調べるためにエッジにリンクする形でエビデンスとなる文献を提示している（上記課題 6 に対応）．

BioTermNet では**図 -2** で概要を示すプロセスで潜在的知識発見を試みている．通常はあらかじめ 2 項関係を構文解析および統計的手法により抽出してその関連度とともに DB 化している．しかし，条件付きの関係性の場合はあらかじめ計算することができないため，概念を文書中の出現位置とともにインデックス化したものを用い，

クエリの条件に従って動的に関係性を求める．たとえば，apoptosis の条件下での蛋白質の関係性を見たいのならば，apoptosis の文書を選択した後，その文書内で蛋白質と他の概念との関連度を求めることになる．明示的に記述される 2 項関係を求める手段は複数存在する．もし，対象となる概念や概念間の関係性が明確であるならば，少々の学習セットを用意してナイーブベイズ＋ EM などの方法で自動抽出することも可能ではあるが，ここではすべての概念を対象とするため，統計値と一部構文解析を利用している．情報抽出の手法や概念の認識手法は文献 5）およびその参考文献に記述してある．

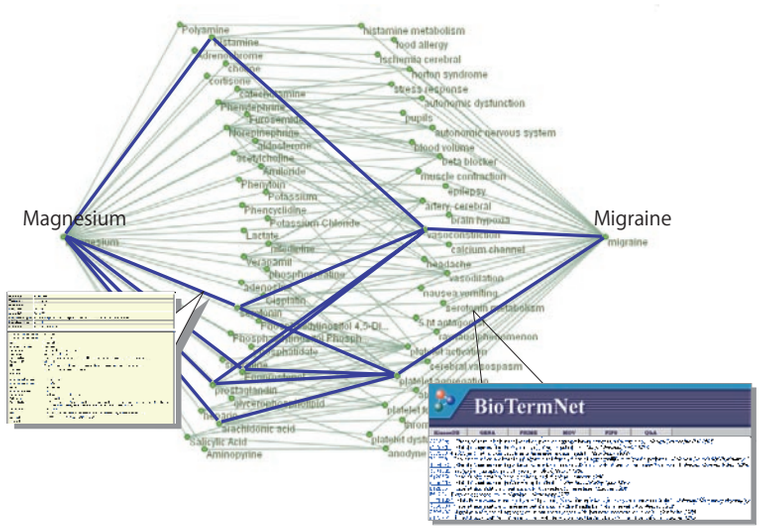

図 -3　Mg と偏頭痛の Closed discovery

最適な概念ネットワークの計算には，明示的に記述されている 2 項関係を高くスコアリングし，かつ全体のネットワーク計算にも適している統計的手法が望ましい．2 項関係の計算には，Dice 係数，Mutual information（MI）などの統計的手法，HyperGsum や Singhal の文書検索における特徴語を計算するための手法，また SVD（singular vector decomposition）により概念・文書行列の次元数を削減し，低次元空間で近接した概念が関係しているとみなす方法などがある．どの手法が最も明示的記述の抽出およびネットワーク全体の計算に適しているかは文献 6），7）で詳細に検討しているため興味がある方は参照していただきたい．以下，概略のみ述べる．概念間の関係の種類によって，明示的な関係記述と 2 つの概念の 1 文書中の出現頻度との関係は異なる．したがって，既存の手法の単純な適用だと最も効果的な統計的な手法は異なる．たとえば，蛋白質間相互作用が明示的に記述されている場合，これらの概念が複数回 1 抄録中に出現するが，蛋白質とその細胞内局在情報の関係が記述されている場合は，多くの場合は 1 抄録中に一度しか出現しない傾向にある．すなわち，蛋白質間相互作用の場合は，同一の文書で 2 つの遺伝子名（蛋白質名）が偏って出現したとしても 1 抄録中に一度しか平均的に出現しない場合は，蛋白質間相互作用である確率が低い傾向にある．Dice 係数，MI，HyperGsum などは 1 つの文献に現れる概念の出現頻度は考慮していないが，Singhal をはじめとする文書検索の特徴語計算手法の多くにおいて，1 つの文献に現れる概念の出現頻度を考慮している．蛋白質間相互作用では，Singhal の手法が最も効率がよく，蛋白質とその細胞内局在情報では HyperGsum が最もよい．下記ネットワーク全体の計算を考えた場合もこの 2 手法が潜在的知識発見という意味ではよい性能を示している．SVD はすべての概念間の関係を考慮した上での 2 概念の類似性を取り出すことになるので，2 概念に明示的な関係があるか否かという局所的な情報が重要である 2 項関係抽出においては，精度がかなり落ちる [7]．

概念ネットワークの計算では，open discovery，closed discovery ともに中間概念を介した概念間の関連度の計算することになる．その計算方法はいくつかあるが，open discovery ではたとえば 1）A→Bi, Bi→Cj（i=1,...）の各パスにおいて各ステップのスコアの相加相乗平均を計算しその最大値を Cj のスコアとする，2）Cj の次数（threshold 以上の B とのエッジ数）をスコアとする，3）Cj に繋がる各パスの相加相乗平均の和を Ci のスコアとする方法などがあり，文献 6），7）で検討している．次数情報を利用すると major な概念が高くスコアリングされすぎてしまう傾向にあり，一般的には各ステップの相加相乗平均の最大値（上記 1）を採択する方法が有効のようである．以下，Singhal の文献中の特徴語抽出の手法と上記 1）の計算方法を用いた例で事例紹介を行う．

偏頭痛とマグネシウムの関係予測

偏頭痛と Mg の関係は Swanson が 1988 年に予測をし，その後，実験的に証明されている．偏頭痛と Mg をクエリとした closed discovery でこれらの関係を探索した結果を例として**図 -3** に示す．システムの精度評価のため 1988 年以前の文献を利用している．本例では 2 層目が化合物，3 層目が化合物以外の意味クラスと設定している．また，Mg と 2 層目の関係は化合物間相互作用となるため構文解析情報のみ利用している．エッジのリンク情報を見ると以下のような解釈ができる（以下，図中青でハイライト）．Mg は serotonin 分泌，prostaglandin/epoprostenol 反応性，arachidonic acid，histamine 代謝を制御している．serotonin，prostaglandin/epoprostenol，arachidonic acid，histamine は

vasoconstriction や platelet aggregation の原因となり，vasoconstriction や platelet aggregation は migraine の原因になる．したがって，Mg の不足によりこれらの一連の働きにより偏頭痛が起こることが考えられる．医学生物学では，複数の原因によって同一の現象が起こることが多く，正解は 1 つのパスとは限らない．

偏頭痛をクエリとして open discovery を行い偏頭痛に有効な化合物を探索すると Mg は 49 番目（89571 化合物中）に，同様に有効な元素／イオン／同位体として探索すると 2 番目（3527 イオン中）にランクされる．上述のように潜在的知識の発見支援システムの評価の難しいところであるが，必ずしも Mg が 1 番目にランクされる必要はなく．実際，もっともらしい複数の概念が上位にランクされている．たとえば，dextroamphetamine が，methysergide, amphetamine, reserpine などの中間層概念を介して Mg よりも上位にランクされている．興味深いことに近年この化合物が偏頭痛に有効であることが報告されている．

医学生物学における潜在的知識発見／仮説生成はさまざまな場面で利用できる．たとえば，承認薬の効能追加の探索，連鎖解析，関連解析後の疾患関連候補遺伝子の絞り込み，DNA マイクロアレイなどの大量データの解釈，また突然変異遺伝子の表現型の解釈などに利用できる．このうち，2 例について以下紹介する．

疾患関連遺伝子探索への応用

疾患原因／関連遺伝子と物理的蛋白質相互作用を持つ遺伝子に変異が起こると，疾患原因／関連遺伝子のみに変異が存在するときよりも疾患罹患リスクが高まることがあることが知られている．このような例も BioTermNet の利用によって容易に探索できる．たとえば，BRCA1 遺伝子と物理的相互作用をする遺伝子／蛋白質を 2 層目に，BRCA1 と明示的な関係性が知られていない疾患を 3 層目に指定し，open discovery で関連性の探索を行う．BRCA1→CREBBP（物理的相互作用の関係），CREBBP→Rubinstein-Taybi syndrome（RSTS）（疾患関連遺伝子と疾患の関係），BRCA1→EP300（物理的相互作用の関係），EP300→RSTS の関係が上位にランクされる．BRCA1 は RSTS との関連性は現在のところ知られていないが，BRCA1 の変異によって RSTS の疾患罹患リスクが高まる危険性があることをこの結果は示唆している．

また，連鎖解析，関連解析などの結果のスクリーニングにも有効である．連鎖解析では，連鎖（減数分裂時に染色体の組み替えが起こるため，染色体上で隣接する遺伝子が独立に遺伝しない）を利用して，疾患家系のマーカ遺伝子の情報と罹患の有無の情報から，疾患遺伝子のゲノム上の位置を予測する．しかし，収集した家系数や浸透率（その疾患遺伝子の遺伝子型を持っている個体が病気を発症する確率）の低さにより，疾患関連候補領域がブロードな予測となり，場合によっては，予測された領域に数百の遺伝子が存在することがある．したがってこれらの領域にある遺伝子が，どのような形で対象疾患とかかわる可能性があるのか，現段階の情報から予測し，次ステップの実験計画立案を支援することが可能である．一方，関連解析においては，ヒトゲノムの解読完了と HAPMAP PJ（ヒトの病気や薬に対する反応性にかかわる遺伝子を発見するための基盤を整備する国際 PJ）の完了に伴い，マーカとしてはほぼ十分な SNP（一塩基置換）が見つけられ，ゲノムワイドな関連解析が可能となった．現在では，50 万 SNP 程度が一度に測定できるキットも市販されている．しかし，疾患患者のサンプル収集の難しさとともに，測定も十分安価でないため，関連解析に十分な検体数を実験することができないことが多い．多重検定の問題を考慮すると遺伝統計的なアプローチだけでは，スクリーニングが不十分である．したがって，疾患と関係あるか否かという遺伝統計的数値だけでなく，その遺伝子と疾患遺伝子を繋ぐ関連性情報を考慮した上で，次ステップでタイピングを行う SNP を選別していくことがコスト面からも現実的である．現在，東京大学医学系研究科徳永研究室で行っているゲノムワイド関連解析において，統合データベースプロジェクトの一環として，遺伝統計値と共に文書データも利用した疾患関連 SNP スクリーニング支援のシステムを構築中である．

マイクロアレイ解釈への応用

DNA アレイ技術の確立により，一度に数千の遺伝子の挙動の観察が可能となった．しかし，ノイズの多さもさることながら，対象遺伝子の多さゆえの実験結果の解釈の難しさも大きな問題である．この問題を解決すべく，文書データを利用して，実験結果を自動解釈する手法は活発に研究されているが，ここでは，潜在的知識発見の枠組みでの応用を紹介する．

以下は，Doxazosin という前立腺肥大に効果的な薬剤を細胞に与えたときの遺伝子発現変化から，Doxazosin がどのような疾患に効果がありそうか予測するという例である．入力が発現変化のあった約 200 遺伝子，2 層目の意味クラスを遺伝子，3 層目の意味クラスを疾患と指定し，open discovery で関連の高い疾患をランキングすると glioblastoma（グリア芽腫），neuroblastoma（神経芽細胞腫），B-cell lymphoma（B 細胞リンパ腫），astrocytoma（星状細胞腫），sarcoma of prostate gland（前立腺肉腫）など癌関連の疾患が上位にランク

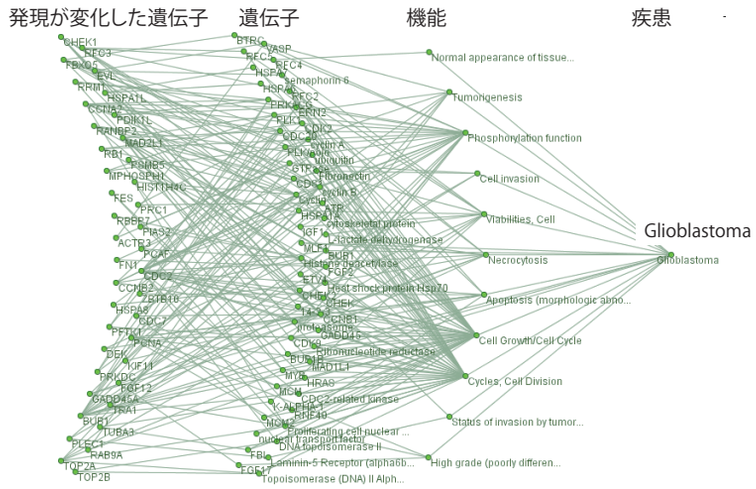

図 -4　発現が変化した遺伝子と機能および疾患との関係

される．マイクロアレイで発現に変化があった遺伝子は，cell growth, cell invasion, cell division, cell migration, apoptosis などの機能に関係しており，これらの機能に共通な疾患として癌関連の疾患があるからである[7]．上位 20 位の約半分程度は doxazosin の治療効果に関する報告がすでに行われておりもっともらしい結果が得られている．**図 -4** は 1 位にランクされた glioblastoma に関して 2 層目の意味クラスを遺伝子，3 層目を機能として描画させた例である．一般的に 2 層目に遺伝子を持ってくると類似の機能の遺伝子が増え，機能と疾患の関連，機能と遺伝子の関係が明らかになりやすい．これらのエッジにリンクした文献情報を読むと，glioblastoma とその機能の関係などが得られる．

.

潜在的知識発見研究の展望

医学生物学分野における文書ベースでの潜在的知識発見について紹介してきたが，その他の分野でも研究は行われている．例として，書誌情報を利用して，20 世紀の詩人が間接的に古代の哲学者から思想的影響を受けていることを調べる研究，および，化合物の反応経路の推定を domain knowledge や実験情報などを利用して行う研究などが挙げられる．しかしながら，一般的なデータマイニングとしての知識発見は多様な分野で盛んに研究されているのと比べ，文書ベースでの知識発見の研究は医学生物学分野に偏って多い．MEDLINE のように自由に使える膨大な文書が存在すること，また，医学生物学分野は発見すべき知識が比較的明確で，かつ，1 人の人間が計算機なしに処理するには情報が膨大になりすぎたことなどが原因と思われる．

医学生物学分野のさまざまな場面で潜在的知識発見手法は有効であり，本稿で紹介したのは一端に過ぎない．しかし，現在のシステムでは，やはりそれなりの専門知識を保有している研究者でないと使いこなせない．素人では，提示される概念の背景知識がなく，もっともらしい解が提示されているのか否かがすぐ判別できないのである．また，本稿では紹介しなかったが，この手のシステムは既知の関係性の提示にも有効である．しかし，同様に提示される概念や関係性がユーザの知識レベルにそぐわないと，ネットワーク図を一瞥しただけでは表示結果を理解できない．提示する内容をユーザの知識レベルに自動的に合わせられるような仕組みの開発が必要である．その他，提示された 2 項関係 A→B と B→C が別条件でのみ起こり，A→B→C は成立しないケース．生物現象に特徴的であるが，複数の異なるパス A→Bi→Cj, A→Bk→Cj がプラスとマイナスの効果があり A から Cj への影響が相殺されてしまうケース．同一の 2 項関係が（主に量に依存して）プラスにもマイナスにも働くケース．など解くべき個別な課題は枚挙に遑がない．潜在知識発見支援システムが，現在の検索システムと同様に万人が不自由なく利用可能な新たなパラダイムを迎えるには，このような課題を克服する必要があると思われる．

謝辞　本研究は，東京大学新領域研究科高木利久教授と共同研究で行っているものである．心から感謝の意を表します．また，本研究に関し，貴重なアドバイスをいただきました日立製作所中央研究所の丹羽芳樹博士に深謝します．

参考文献

1) Swanson, D. R. : Fish Oil, Raynaud's Syndrome, and Undiscovered Public Knowledge, Perspectives in Biology and Medicine, 30(1), pp.7-18 (1986).
2) Gordon, M. D. and Lindsay, R. K. : Toward Discovery Support Systems: a Replication, Re-examination, and Extension of Swanson's Work on Literature-based Discovery of a Connection between Raynaud's and Fish Oil. J. Am. Soc. Inform. Sci. Tech., 47, pp.116-128 (1996).
3) Weeber, M., Klein, H., de Jong-van den Berg, L. T. W. and Vos, R. : Using Concepts in Literature-based Discovery : Simulating Swanson's Raynaud-fish Oil and Migraine-magnesium Discoveries, J. Am. Soc. Inform. Sci. Tech., 52, pp.548-557 (2001).
4) Srinivasan, P. : Text Mining : Generating Hypotheses from MEDLINE, J. Am. Soc. Inform. Sci. Tech.55, pp.396-413 (2004).
5) Koike, A. and Takagi, T. : Knowledge Discovery based on an Implicit and Explicit Conceptual Network, J. Am. Soc. Inform. Sci. Tech. 58(1), pp.51-65 (2007).
6) Koike, A. : Knowledge Discovery and Hypothesis Generation from Biomedical Texts, Workshop on Information-Based Induction Sciences, pp.52-58 (2006).
7) Koike, A. : Biomedical Application of Knowledge Discovery, Literature-Based Discovery (Information Science and Knowledge Management) Springer-Verlag, in press.

（平成 19 年 5 月 30 日受付）

小池　麻子　asako.koike.ea@hitachi.com

京都大学理学系大学院修士課程終了，理学博士．日立製作所日立研究所入社，2003 ～ 06 年東京大学大学院情報理工学研究科客員助教授，現在，日立製作所中央研究所主任研究員，東京大学理学系研究科非常勤講師．