3 生命科学分野におけるテキストマイニング

山本 泰智
情報・システム研究機構
ライフサイエンス統合データベースセンター

背景

生物の全遺伝情報（生物個体ごとに固有の並びを持つ，核酸の塩基配列情報のすべて，ゲノム）は計算機分野において喩えると，メインメモリにロードされたコードであり，必要に応じて特定の部分が CPU に読み込まれ，命令が実行されるのと同様，ゲノムの特定部分が mRNA に転写され，それに基づきタンパク質を生成するなどし，なにがしかの生物学的機能を発揮する．このゲノム中の特定の部分を遺伝子と呼び，機能を発揮することを遺伝子が発現すると呼ぶ．ゲノム情報が得られると，どの部分が遺伝子であり，それが発現すると発揮される生物学的機能は何か，といった知見を得る研究が始まる．昨今の著しい実験機器の性能向上に伴い，塩基配列情報が短時間に大量に取得されるようになった一方で，以上のような作業は多くを人手に頼っているために手間ひまがかかる．加えて，多くの研究者がさまざまな実験を行い，その結果を次々と論文で発表しているため，得られた知見は主に構造化されていない自然言語の形で集積されてゆく．研究の進展や領域の細分化に伴い，論文が発表される頻度も高くなっている．このような状況において，特定の研究課題に関連する文献を効率的に見つけ出し，上記のような情報を取得してゲノム情報と紐付けた形でのデータベースを構築するためにテキストマイニング技術を利用することが多くなっている．ただし対象となるデータの規模と種類は広く，1 つの研究室レベルですべてに対応できないため，テキストマイニング技術を用いたシステムの公開や必要なリソースの共有が行われている．本稿では生命科学分野におけるテキスト処理技術の現状，BioHackathon 2009 で議論された事項を踏まえて解説する．

テキストマイニング技術が使われる一般的な枠組み

最初に生命科学分野においてテキストマイニング技術が使われる一般的な状況を説明する．本稿でいうテキストマイニングとは，大量の文献を対象として，遺伝子名などの領域固有語を認識する固有名認識や，認識された複数の固有名間の関係を抽出する情報抽出を総称して指す．利用者はその結果を基にデータベースを構築したり仮説生成を行ったりする．その目的は 1 人の研究者では現実的に処理できない量の学術文献およびその関連情報を，計算機を利用して高速大容量に処理することで，生命科学者にとって有益な，すなわち，最終的には生命現象を解き明かすことに繋がるような結果を得ることにある．ゆえに，対象とする文献は，遺伝子の機能について記述されていたり，遺伝子の機能に基づく細胞の振る舞いが記述されていたりすることが想定される学術論文や記事である．

アノテーションとキュレーション

さまざまな生物種のゲノム情報が次々と明らかにされている昨今，たとえばヒトの体内で生じる生命現象を，遺伝子やタンパク質といった分子レベルの粒度でその仕組みを説明しようとする場合，興味を持つ遺伝子の機能について，他の生物種における遺伝的に類似した塩基配列を持つ遺伝子の機能を調べて両者を比較したり，対象遺伝子の未知の機能を推定したりすることはよく行われる．このようなとき，対象となる生物種について最新の研究成果を把握する作業は各種データベースやツールを利用して行う必要がある．

また，1 つの生物種についても，さまざまな角度から，多くの研究者が日々研究を行い，成果を文献という形で発表していることから，自身の専門としている生物種についても，これまでに得られている知見をすべて把握す

ることは困難である．たとえば疾患の原因を探ることを目的としている研究者と，代謝の仕組みを解き明かすことを目的としている研究者がそれぞれ独立して，ある遺伝子が重要な役者であることを発見したが，実際にはゲノム上の同じ領域にある同じ遺伝子であることが判明することもある．このとき，ゲノム上の位置を手がかりに遺伝子を検索し，当該遺伝子についてすでに研究されている機能などの知見を効率よく取得できれば，対象遺伝子の振る舞いについて，より深く理解することができるだろう．

このような状況で所望の知識を簡単に獲得できるよう，さまざまな生物種についてそれぞれの研究コミュニティが遺伝子の配列やゲノム上の位置，あるいはその機能について，根拠となる文献情報とともにデータベースを構築し，検索可能な形で公開していることが多い（図-1）．構築にあたっては，内容の質を保つために領域の専門家が実際に関連文献を読んで行うため，読むべき文献の増加速度に更新作業が追いつきにくいという問題がある．そこで前述の通り，あらかじめ計算機を用いて，ある遺伝子と，その遺伝子について記述している文献および具体的な記述を結びつけておき，遺伝子の機能に関するデータベースを構築・更新する際の手間を軽減しようという試みが行われている．しかしながら，ある特定の遺伝子に関する文献あるいは記述を，利用者が望む精度で検索もしくは取得することは非常に困難である．また，著者があらかじめ自身の論文に詳細な機械可読形式でのメタデータ，すなわち，そこに書かれている遺伝子やタンパク質，それらの関係に関する曖昧性のない記述子を用いた説明を加えることは現実問題として無理である．仮に今後実現できたとしても，過去の膨大な知見は自然言語のみで書かれた文献として残り続ける．

さて，あるゲノム情報が与えられたとき，その精度は対象となる生物種に依存するが，どの領域が遺伝子であるかについて計算によりある程度推定が可能であるため，最初に計算機を用いて当たりをつけた後に，実際に専門家が各種配列解析ツールなどを利用してその内容を正していく作業が行われる．この作業をアノテーションと呼び，作業者をアノテータと呼ぶ．遺伝子の情報が収められているデータベースには，計算機による推定結果と人による作業結果の双方が含まれていることが多い．また，さらに領域の専門家がさまざまな関連文献を参照するなどしてより生物学的に深い知識を付け加えたデータベースを構築する作業をキュレーションと呼び，その作業者をキュレータと呼ぶ．なお，広義の意味でのアノテーションはキュレーションを含む．したがって，生命科学分野におけるテキストマイニングは，アノテーションを支援する技術として位置づけられることが多い．

テキストマイニング技術の実際

現在，アノテーションを支援するために用いられているテキストマイニング技術は，対象分野の広さと利用者の多さ，取得・利用のしやすさ，そして処理可能なテキストデータの多さから PubMed/MEDLINE® データベースに適用される場合が圧倒的に多い．当該データベースは米国政府機関の国立医学図書館（National Library of Medicine, NLM）が維持・管理している，医学・保健・生物系の学術論文を主対象とした書誌情報のデータベースで，現在 1,800 万件以上のエントリを持つ．エントリには題目や要旨，著者名，掲載雑誌名，発表年月日などの情報に加え MeSH タームと呼ばれる統制語彙が NLM のスタッフにより内容を端的に示すキーワードとして付

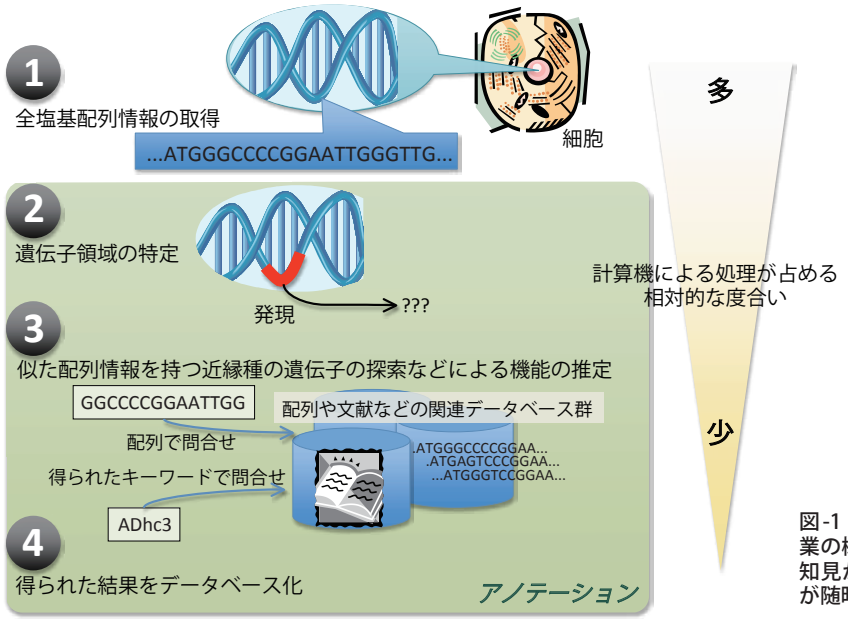

図-1　本稿で対象とするアノテーション作業の概略．遺伝子の機能などの生物学的知見が実験で明らかになれば，その結果が随時データベースに反映される．

加され，文献を検索する際の索引として利用されている．収集対象としている雑誌の発行国は米国に限らず日本を含む世界各国に渡り，そこに書かれている言語の種類は 37，雑誌の数は 5,000 を超える．日本語などの英語以外で書かれている論文の場合にはタイトルが英訳されるとともにオリジナルの論文に書かれている言語が示される．また，非常に古い論文の書誌情報も含まれており，原則 1949 年以降発表されたものが対象になっているが，現在収められているもので最も古い論文は 1865 年に発表されたものである．新規発表論文はもちろんのこと，このように古い論文についても，現在順次追加されている．このため，近年の増加率は非常に大きくなっており，去年追加された論文情報は 884,811 件と膨大な数になる．平均すると毎日 2,424 報の論文情報が追加された計算になるわけで，その大きさが実感できるのではないだろうか．

本データベースは，多くの生命科学者により PubMed 検索システム（http://pubmed.gov）を通して文献を検索する際に利用されているほか，データベースを無償で取得し，それを利用して構築したシステムを公開することが可能であるため，テキストマイニングを行う研究者や技術者にとり非常に都合のよい資源ともなっている．最近ではオンラインジャーナルが増加しており，多くの文献が PDF もしくは HTML 形式で全文閲覧可能になっているが，出版社との間の契約の問題から自由に多くのテキストデータを計算機により解析したり，その結果を公開したりすることは困難な状況である．とはいえ，Biomed Central や PLoS のようなオープンアクセスジャーナルを発行する組織が現れてきたことや，発行主体間で異なる全文データのファイル形式を 1 つの XML 形式に変換した上で多くの全文データを自由に取得可能としている PubMed Central が出現したことから，徐々にではあるが，全文データを対象とした処理システムを構築する組織も出始めている．

なお，生命科学分野における一般公開されているテキストマイニングシステムは静的および動的な使われ方に大別できる．前者は，あらかじめ，たとえば PubMed/MEDLINE 全データに対して遺伝子名や疾患名を抽出し，さらにそれらの間の関連性を抽出した結果をデータベース化しておき，利用者はそのデータベースへアクセスするような型のものである．解析処理に時間がかかる場合や，PubMed/MEDLINE データのように，それを基にしたサービスを一般公開することが可能で，多くの利用者が見込まれるデータに対して有効である．一方，後者は利用者から与えられたプレインテキストデータに対して遺伝子名や疾患名を認識し，結果をその場で返すような型のものである．一般公開ができないテキストデータを処理したいが，手元にテキストマイニングシステムがなかったり，入手や自前での開発が困難であったりする場合に有効である．いずれの型でも利用者側で開発するプログラムからアクセスして必要な情報が取得可能な Web サービスのインタフェース，すなわち Application programming interface（API）を提供するサービスが増えている．

領域固有の課題

PubMed/MEDLINE データベースを利用する場合，当該データベースは XML 形式で収められており，そこから必要な情報を取得するが，題目および要旨は 1 行ごとに分けて収められていないので，最初に行を認識する処理が行われ，続いて遺伝子名やタンパク質名などの生命現象を成り立たせるために重要な役割を担っている実体を示す領域固有語を認識する．また，同じ行の中で認識された複数の実体があれば，それらの間の関係を抽出するなど，対象実体の振る舞いに関する記述を認識する．基本は以上であるが，領域固有語の認識や，その振る舞いに関する記述の認識は，生命科学分野における対象概念の記述方法の特徴を反映し，計算機による処理において困難な課題が多くある．

遺伝子やタンパク質，疾患を示す名称は多くの同義語，多義語が存在するほか，省略された表記や研究対象領域独自の表現方法があり，任意のテキストから高精度に固有語を抽出することは難しい．さらに，抽出結果の利用目的に応じて，ある特定のテキスト中における表現を遺伝子名として認識すべきか否かの基準が変化し得ることもある．以下，同義語および多義語の実際の例を示す．

遺伝子名として MAP3K7IP3 と呼ばれるものがあるが，この正式名称は，「mitogen-activated protein kinase kinase kinase 7 interacting protein 3」であり，そして上記のほかに，NAP1，TAB3，MGC45404 という同義語を持つと遺伝子に関するデータベース Entrez Gene に書かれている．なお，MAP3K7IP3 は前述の正式名称に対し，シンボル名と呼ばれる．このように 1 つの遺伝子に複数の名称が付けられていることは多い．一方，PC というシンボル名を持つ遺伝子があり，その正式名称は pyruvate carboxylase であるが，PC として標記される概念は遺伝子に限らず，Personal Computer，phosphatidylcholine，Protein C，prostate cancer などなどさまざまである．また，シンボル名が PC である遺伝子を持つ生物種はヒト，ウマ，ウシ，イノシシ，イヌなどさまざまであり，テキスト中における PC の標記が遺伝子名であると判明しただけでは曖昧性がなくならず，さらにどの生物種の遺伝子であるかを同定することも必

要になる．

文献中における表記方法として，複数の遺伝子名をまとめて書くことがしばしば行われるが，その際に，たとえば，「PKS isoforms alpha, delta, epsilon, and zeta」と記述したり，「AKR1C1 - AKR1C4」と記述したりする．前者は PKS isoform alpha，PKS isoform delta などと列挙されるものであり，後者は AKR1C1，AKR1C2 などと列挙されるものである．計算機を用いて遺伝子名を認識する際にはこのような記述を含む文献があることに注意する必要がある．

以上のようなさまざまな課題に対処してより良いシステムを開発するために，複数の研究機関が 1 つのテキストデータを対象にしてそれぞれ独立して抽出システムを開発し，後にシステムの性能を評価し合う試みがいくつかなされている．その中で，2006 年から 2007 年にかけて開催された BioCreative 2 という評価プロジェクトでは，テキストからの遺伝子名抽出や，生物種名を特定した形での抽出などの複数のタスクが設定され，合計 13 カ国から 44 チームの参加があった [1]．現在，参加チームにより開発された複数のテキストマイニングシステムに 1 つのサイトを経由して横断的にアクセス可能になっている [2]．

これまでに発表されているシステム例

これまでに無料で一般公開されている，アノテーションに利用され得るシステムをいくつか紹介する（表 -1）．なお，ここで使う静的／動的という語の意味は，前述の通りである．

■ Whatizit

遺伝子名などの固有名認識および対応する具体的な遺伝子の同定機能を始めとする，さまざまな生命科学系のデータベース検索，テキスト処理を行うサービスで静的および動的の双方に対応している [3]．欧州バイオインフォマティクス研究所（European Bioinformatics Institute, EBI）により開発されており，静的なテキスト処理対象は PubMed/MEDLINE である．Web ブラウザ上でテキスト中の固有語をハイライトさせることなどができる（図 -2）ほか，Simple Object Access Protocol（SOAP）および，Streamed Servlet と呼ばれる API が用意されているため，プログラムからのアクセスも可能である．独自の文献検索機能などさまざまな API が提供されているので，たとえば，キーワード検索して文献集合を取得し，それに対して遺伝子名や生物種名を同定し，それらの機能一覧を取得するといったワークフローが本サイトの API 群を利用するだけで構成できる．

■ iHOP

Information Hyperlinked over Proteins の略で，Robert Hoffmann 氏により開発されている [4]．遺伝子名から検索を開始し，他の遺伝子との関連が記述されていると計

システム	静的・動的	API	ライセンス
Whatizit☆1	双方	SOAP / Streamed Servlet☆4	EBI独自規定☆6
iHOP☆2	静的	REST / SOAP / BioMoby☆5	Creative Commons Attribution-No Derivative Works 3.0 Unported
U-compare☆3	双方	UIMA / SOAP	独自規定☆7

☆1 http://www.ebi.ac.uk/webservices/whatizit/info.jsf
☆2 http://ubio.bioinfo.cnio.es/biotools/iHOP/
☆3 http://u-compare.org/japanese.html
☆4 EBI により定義されている独自仕様と思われる
☆5 生命科学分野における各種データの型とそれに対応する解析ツールの入出力に関する情報を共有し，各種解析ワークフローを容易に実現できるようにする枠組み（http://www.biomoby.org/）
☆6 http://www.ebi.ac.uk/Information/termsofuse.html
☆7 http://u-compare.org/ucompare-license.txt

表 -1　無料で公開されているサービスの例

図 -2　Whatizit 表示例．PubMed/ MEDLINE の各題目および要旨が 1 つのセルに表示され，さまざまな領域固有語がハイライト表示されている．

算機により判断された PubMed/MEDLINE 中の題目もしくは要旨中の一文群を表示する（図 -3）．遺伝子名などの領域固有語はハイライト表示され，クリックすることで当該固有語に関する情報が得られる．また，各文中に共起する他の遺伝子名との二項関係について，利用者が選択した対象文に含まれるものをグラフとして表示できる（図 -4）．本システムが表示する結果には誤りが含まれ得るため，利用者が根拠を確認しやすいよう，生物学的な実験に基づく関係があることが知られている場合はその旨表示される．

このように，自身の興味ある遺伝子と関連のありそうな遺伝子の候補について，対話的にその関係性をグラフとして描きながら仮説を生成したり，新たな文献を検索できたりする点がインタフェースとして優れている．本サービスは Representational state transfer（REST），SOAP および BioMoby 形式の API を備えている．なお，BioMoby とは生命科学系のさまざまなデータ形式とそれに対応する解析ツールの入出力に関する情報を蓄えて，利用者が容易に解析ワークフローを構築可能な環境を提供するプロジェクトである．

■ U-Compare

利用者が生命科学分野における各種自然言語処理ツールをさまざまに組み合わせてワークフローを定義し，実行可能な統合環境である [5]．東京大学，英国国立テキストマイニングセンター（National Centre for Text Mining, NaCTeM），コロラド大学保健科学センター（Center for Computational Pharmacology）の共同で開発している．自然言語処理を用いてテキストマイニングを行う際には，1 つのデータに対して複数の処理を次々に連続して実行していくが，各段階において同じ目的を果たすツールが複数提供されている場合が多い．そこで，あるテキストデータに対する一連の言語処理において，利用可能なツールのすべての組合せによる処理を自動的に実行し，組合せ方による結果の違いを簡単に比較できる環境の実現を目指して開発されている（図 -5）．このため，現時点では必ずしもアノテーション作業の目的に直接利用可能とは言えないが，Unstructured Information Management Architecture（UIMA）を利用して構築されており，SOAP 形式での API が提供されているため，たとえば，文献に遺伝子情報を付加したり，他の遺伝子や疾患との関係を抽出したりする際に，同時に複数のツールの結果

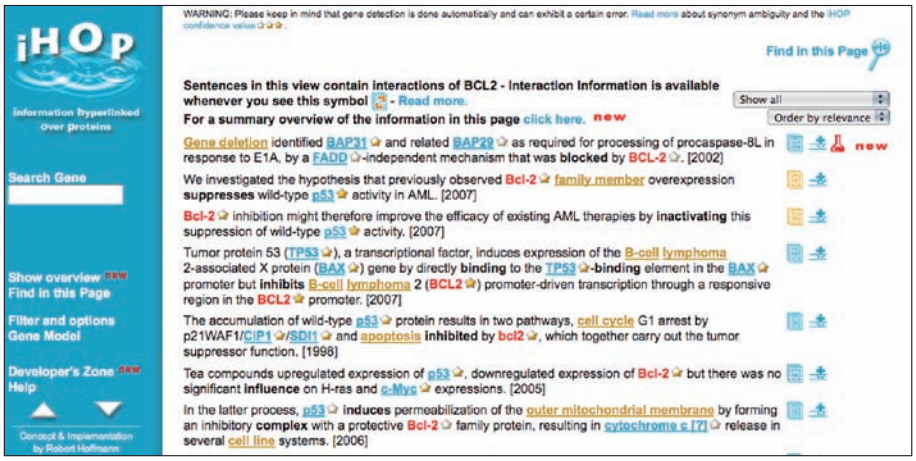

図 -3　iHOP 検索結果イメージ．特定の遺伝子に関する記述がなされている一文一文に対し，Whatizit と同様に領域固有語がハイライト表示されているほか，グラフを描画するためのアイコン（下向矢印が書かれているもの）や，一文中に共起している他の遺伝子との間に関係のあることが実験的に示されていることを表すアイコン（赤いフラスコ）がある．

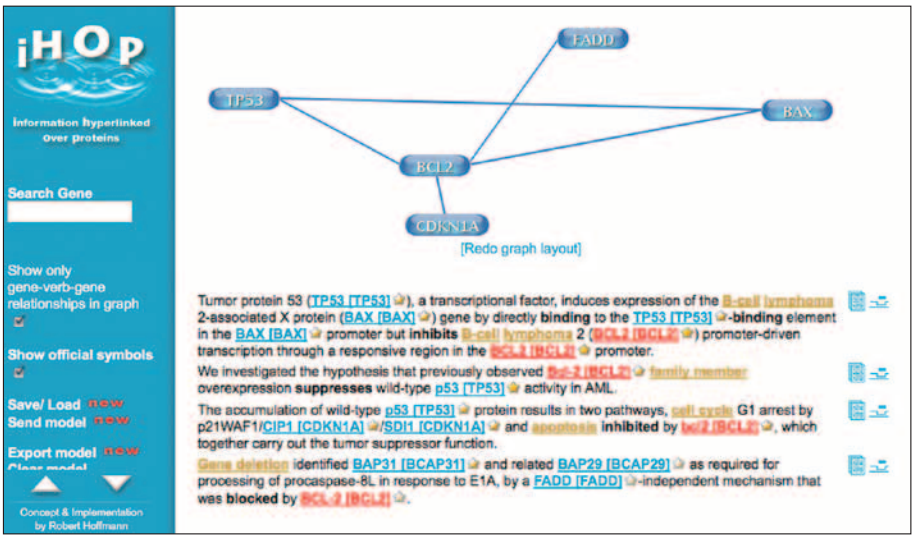

図 -4　iHOP 遺伝子ネットワーク描画イメージ．図 -3 中の下向矢印アイコンをクリックすると，対応する一文中に書かれている 2 つの遺伝子がノードとなり，両者の共起関係をエッジとするグラフが別ウィンドウ上に描かれる．アイコンをクリックするつどグラフが更新される．

を取得してその確からしさを定量的に計ることが可能なアノテーション支援システムが比較的容易に実現できるだろう．

今後の課題

これまで見てきたように，生命科学系の文献から領域固有語や，それらの間の関係を抽出するシステムが開発され一般に利用可能になっている．また，解析結果を Web ブラウザや他の GUI を利用して分かりやすく視覚化する技術も開発されてきている．そして，複数の機関により同種の機能を持つさまざまなツールが開発され，それらを比較したり，組み合わせたりすることが容易になりつつあるとともに，API も提供され始めた．

今後の方向性の 1 つは要素技術のさらなる発展である．すなわち文献に書かれている生物学的知識をさらに精度よく抽出する技術や，より多くの生物種への対応，遺伝子などの分子レベルから，細胞や組織といったさまざまな粒度，およびそれらと疾患などの生物学的現象の関係への対応，あるいは図表を含めた論文全文を対象とした抽出技術の開発であり，現在取り組まれている．必要な辞書やオントロジーの構築もある．開発された技術は今後，ゲノム情報の急激な増加に伴い，塩基配列情報の解析ツールなどと効率良く連携できることが求められるだろう．

もう 1 つの方向性は，実際にシステムを利用するアノテータや一般の生命科学研究者がさらに利用しやすい環境を実現するための技術開発である．現在アノテーション作業を行う際に多くのアノテータは PDF ファイルを取得し，印刷し，マーカーで必要な個所をハイライトしたり，データベース開発・更新用のシステムに入力するために必要な事項を手元のメモ用紙に書き込んだりしている．この作業の効率化は計算機を用いて可能であるか，さらには，可能である場合はインタフェースをどのように設計すべきであるかを詳細に検討する必要がある．Web ページの任意の個所をマークしたり注釈を付けたりするツールはいくつか開発されているが，今後は PDF ファイルに対する同様なツールを開発するとより効率を上げられるだろう．

謝辞　BioHackathon 2009 は文部科学省統合データベースプロジェクトによる支援で実施されました．本稿執筆中にコメントしていただいた皆様に感謝します．

参考文献

1）Krallinger, M., Morgan, A., Smith, L., Leitner, F., Tanabe, L., Wilbur, J., Hirschman, L., and Valencia, A. : Evaluation of Text-mining Systems for Biology : Overview of the Second BioCreative Community Challenge, Genome Biol. 2008;9 Suppl 2:S1 (2008).
2）Leitner, et al. : Introducing Meta-services for Biomedical Information Extraction, Genome Biol. 2008;9 Suppl 2:S6 (2008).
3）Rebholz-Schuhmann, D., Arregui, M., Gaudan, S., Kirsch, H. and Jimeno, A. : Text Processing through Web Services : Calling Whatizit, Bioinformatics, Vol.24, Issue2, pp.296-298 (2008).
4）Hoffmann, R. and Valencia, A. : A Gene Network for Navigating the Literature, Nat Genet, Vol.36, Issue7, p.664 (2004).
5）狩野芳伸，辻井潤一：UIMA を基盤とする相互運用性の向上と自動組み合わせ比較―国際共同プロジェクト U-Compare ―，IPSJ SIG Notes, NL (186), pp.37-42 (2008).

（平成 21 年 7 月 9 日受付）

山本 泰智
yy@dbcls.rois.ac.jp

2007 年よりライフサイエンス統合データベースセンター（DBCLS）特任研究員としてテキスト処理関連のサービス開発に取り組む．自然言語処理技術を利用したアプリケーション開発に興味を持つ．博士（情報理工学）．

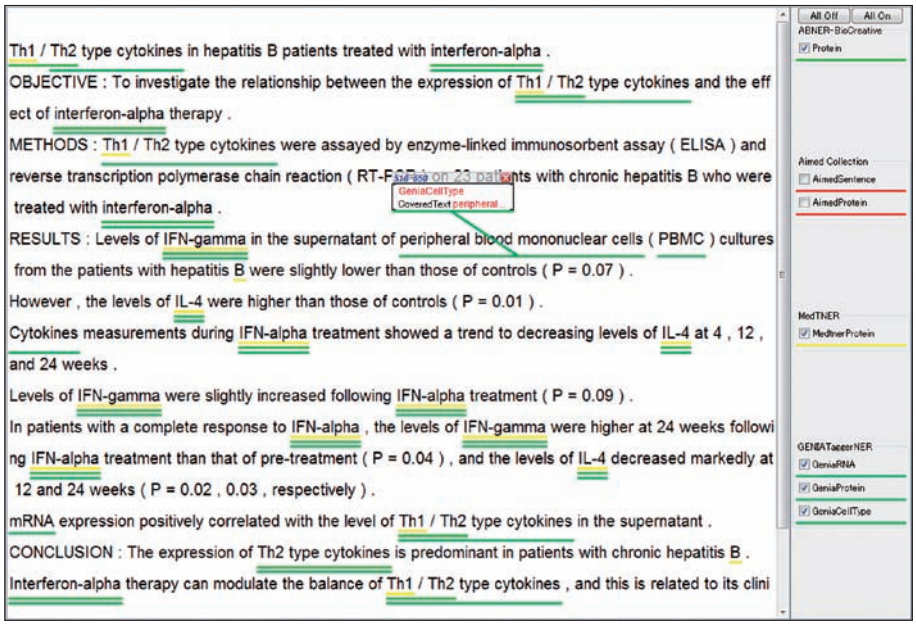

図-5　U-Compareによる複数ツールの固有名認識結果表示イメージ．同じテキストに対して複数の固有名認識ツールを適用し，その結果を視覚的に表示できる．右側のチェックボックス脇にツール名が書かれ，各ツールの結果が下線の色の違いで表示されている．