［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

7 企業におけるバイオNLPへの取り組み

b）連想統合による医学・生物学知識の活用ソリューション

久光　徹
hisamitu@crl.hitachi.co.jp /（株）日立製作所中央研究所ライフサイエンス研究センタ

ポストゲノム時代を迎え，創薬や臨床研究の分野では，分散したデータベースに蓄積されつづける莫大な情報を有効に活用する必要がこれまでになく高まっている．これらのデータの多くは自然言語で書かれた情報を含むため，バイオNLPは創薬から医療にかけての知識マネジメントプラットフォームのキー要素と位置付けられる．本稿では，数千万件規模のデータベースを対象とするスケーラブルな連想検索技術を核とし，大量の検索結果の内容を，リスト表示と特徴要素のネットワーク表示を併用して提示するインタラクティブな情報概観ビューア，異なるデータベースをシームレスに連結するクロス連想技術，タンパク質辞書等の辞書自動作成技術等を総合した，バイオデータベースの連想統合ソリューションについて述べる．配列検索ツールと連携すれば，配列データ等も含む，異種データが混在したデータベースに対象を拡大することが可能である．

ポストゲノム時代のデータ爆発

生物学・医学関連のデータベースにおける情報爆発はとどまるところがない．個別データベースのサイズの加速度的拡大だけではなく，遺伝子に関する網羅的研究であるゲノミクスに始まり，タンパク質に関する網羅的研究であるプロテオミクス，これらに基づき化合物の薬理作用や副作用に関して網羅的に研究する薬理ゲノミクスへと続くテクノロジーの潮流に伴い，遺伝子の配列と機能，タンパク質の立体構造や機能，タンパク質と化合物の相互作用，毒性情報等々，対象となるデータも拡大の一途を辿っている．これらデータベースは，MEDLINE（米国国立医学図書館が提供する世界最大の医学文献データベース）やOMIM（ヒト遺伝子変異と遺伝病に関する代表的なデータベース）に代表されるテキスト主体のデータベースから，ゲノムやタンパク質，化合物の，配列・構造・相互作用情報を主体とするデータベースまで，データ形式の側面からも広いスペクトルを持っている．

創薬や遺伝子診断技術の開発，さらには医学的な臨床研究に携わる研究者にとって，研究の構想段階から具体的な計画策定・遂行段階まで，最新データに基づく科学的見地からの検討とコンペティタの監視・動向予測を継続して行うことは必須である．このためには，上記の公開／商用データベース群に，公開特許とインハウスの報告書や実験結果のデータを加えた，多種多様なデータベース群を相互参照し，分散した知識を収集・統合し，分析する必要が生じる．これはきわめて知識集約的な作業であり，もはや情報システムによる効果的な支援なしでは遂行不可能な状況である．

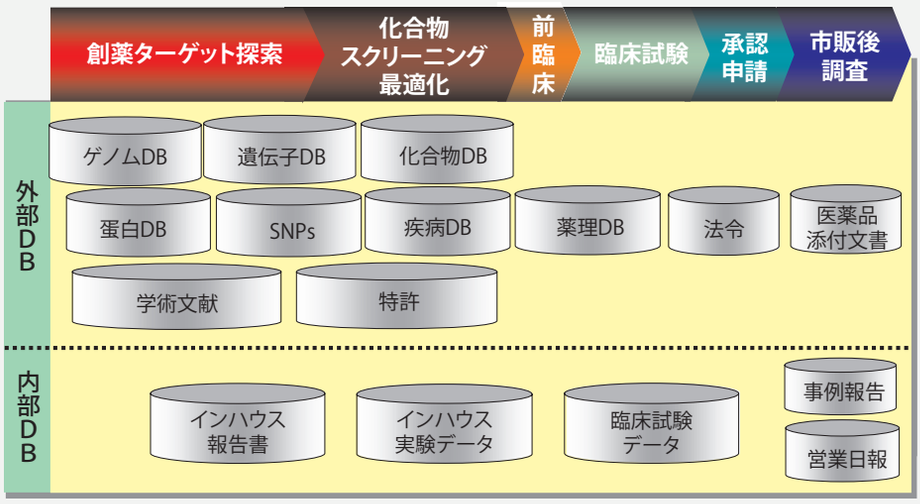

図-1　創薬プロセスチェインとデータベース

創薬プロセスチェインとバイオNLP

ゲノム情報を活用する「ゲノム創薬」への流れが加速する中，データの爆発は製薬会社にとって特に深刻な問題をもたらしている．ゲノミクス，プロテオミクス研究における網羅的アプローチにより，プロセスチェイン（**図-1**）上流部の創薬ターゲット探索において非常に多数の創薬標的候補が提供されてきたが，この「上流側でのデータ洪水」は中流域以降に及びつつある．薬品の認可の遅れは，開発コスト増やコンペティタの参入につながるため，製薬会社の利益を圧迫するばかりか死活問題ともなりかねない．このため，スクリーニング，すなわち候補化合物の絞り込みの効率化がきわめて重要となっている．

スクリーニングの効率化においては，コンビナトリアルケミストリ☆1等を用いたハイスループットスクリーニング技術や，シミュレーション技術など，既存技術のさらなる高度化が喫緊の課題であるが，創薬のさまざまなプロセスで蓄積されつづける情報を活用（検索，相互参照，フィードバック等）する必要性がますます高まっており，知識マネジメントの果たすべき役割は重大である．ここで，種々のデータベースがあるとはいえ，多くの場合，記録されるデータはテキストそのもの，またはテキストによって何らかの情報を付与された非テキストデータであり，辻井[1)]の指摘したように，ライフサイエンス分野においては，数式ではなく「言葉」で知識を語る傾向が強い．ライフサイエンス分野での知識活用においてバイオNLPがキーとなっているゆえんである．

日立製作所では，ライフサイエンス分野において，タンパク質データベース構築，医学・生物学向け情報検索システム，グリッドコンピューティング応用シミュレーション，臨床試験データ解析，遺伝子診断システムの開発等，創薬から医療にまたがる技術開発を推進しており，バイオNLPは将来的にこのプロセスチェイン全体を覆う知識マネジメントプラットフォームのキー要素と位置付けている．ライフサイエンス分野はデータの量と複雑さにおいて，NLPのフィールドとして最もチャレンジングな対象であり，そこで用いられる検索技術には，とりわけ，スケーラビリティ，速度，頑健さが要求されるため，長期に渡る開発成果を駆使して取り組んでいる[2) 3)]．情報抽出においては，ライフサイエンス分野特有の多数かつ特殊な用語辞書の構築や記述パターンの収集が必要であり，大量の情報をコンパクトかつインタラクティブに提示するユーザインタフェースも必須となる．

連想統合による知識の活用

■連想検索機能

ここでは，「連想(association)」という言葉を，情報検索の一手法である「連想検索(associative search)」での「連想」の意味で用いている．連想検索はユーザの提示した単語や文章，文書と「類似性が強いもの」「関係の強いもの」を検索する手法であり，特定の語の有無でなく，単語の分布などを総合的に勘案して行うもので

☆1　元素や官能基の組合せを利用して，化合物群を系統的かつ効率的に合成・評価する手法で，化合物の探索・最適化を網羅的かつ高速に行うことを可能とする．

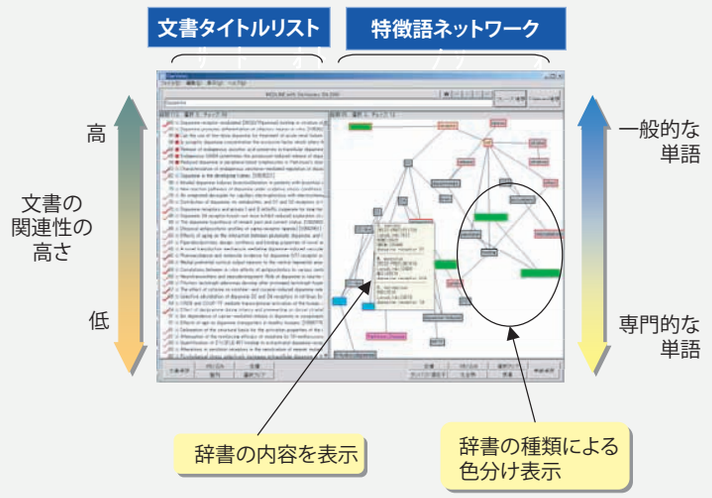

図 -2　デュアルウィンドウを持つ検索インタフェース

あり，「類似検索」や「概念検索」などとも呼ばれる．古典的なキーワード検索との顕著な差異は，ユーザが明示的にキーワード集合を特定しない／特定できない状況においても，文章（の断片），あるいは文書を丸ごと指定することにより，その内容と総合的に最も関連した文書を見つけ出す機能である．したがって，いったん何らかの興味ある文書（群）が見つかると，その文書（群）をキーとしてさらなる連想検索を行うことができるため，ユーザは「文書を選択する」ことだけに集中でき，キーワード検索に比べて格段に効率的であることが多い．キーワード検索の場合は，検索と検索の間に，「次の検索のための新たなキーワード集合を決定する」というさらなる負担が必要となるからである．

連想検索は，バイオ情報検索においてとりわけ効果的なことが多い．ライフサイエンス分野はきわめて専門用語が多い上，その数は急速に増加しつつある．しかも書き手の専門分野や慣習から統一名称で書かれない場合があり（たとえば，タンパク質は，平均して3～4個の異名を持つ[4]），キーワード検索では取り扱いが困難な特性を持っている．このような場合でも，たとえば特定のタンパク質についてユーザが関心ある論文をクエリ文書に指定すれば，タンパク質を取り巻く文脈の単語分布の類似性から，同じタンパク質が別名で表記された論文を見つけられる可能性が高い．ある意味で，文脈を用いた曖昧性解消が行われるのである．逆に，同じ表記で異なるタンパク質A，Bが表記されている場合，タンパク質Aに関する論文をクエリにすれば，やはり単語分布の総合的な評価から，タンパク質Aに関するタンパク質を含む文献が，タンパク質Bに関する文献より上位に現れる可能性が高い．

連想検索はGETA（Generic Engine for Transposable Association）☆2エンジンにより実現しており，1,300万件超の文書DB，米国特許28年分，3,500万件を超えるゲノムDB等検索をリアルタイムに行うシステムを提供している．今後，検索機能のミドルウェア化等，知識マネジメントプラットフォームの強化を行いたいと考えている．

連想検索をさらに有効活用する連携機能

大量情報のコンパクトかつインタラクティブな表示に関しては，検索結果を単に文書タイトルのリストとして表示するだけでなく，検索結果の文書群を特徴付ける“要素”をリアルタイムで抽出し，互いに関連性の強いものを連結したネットワークにより表示する機能を開発した．要素として基本となるのは単語であり，**図-2**は，文書のリストとともに「特徴単語」のネットワーク（特徴語グラフと呼んでいる）を提示するインタフェースを示したものである．特徴的な要素を抜き出す部分のスケーラビリティと速度はGETAにより保証される．「特徴度」の評価に関しては，データベースごとのチューニングが不要で，異なるサイズのデータベース間でも値の直接比較が可能な，統一的な確率尺度を開発した[5]．

文書のリストと特徴語のグラフは連動しており，グラフ上の単語をクリックすると，文書のリスト中，その単語を含む文書にチェックマークを表示し，それらを集合

☆2　国汎用連想計算エンジン GETA は，情報処理振興事業協会（IPA）が実施した平成 13 年度独創的情報技術育成事業の研究成果である．

させることができる．逆に文書をチェックすると，その文書が含む単語がハイライトされ，その文書の内容を要約する単語群を一覧できる．以上の機能を用いて，検索結果の内容の俯瞰，検索が意図どおり進んでいるかの判断，目的とする文書の発見が効率的に行われるばかりでなく，当初想定していなかったキーワードの発見も可能となる．単語は，遺伝子，タンパク質，化合物等の種類によって色分け表示が可能であり，さらにはネットワークに表示する要素として，生体機能を制御する基本となっている，タンパク質間の相互作用のような“単語同士の関係性”も含めたさまざまな対象を考えることができる．表示する対象に応じた適切な概観ビューアを選択すればよい．

情報統合を実現するためのもう1つの機能は，データベースをまたがる検索である．たとえば，興味ある論文を文献データベース中に発見した場合，その文献に記載された内容と関連する特許をできるだけ網羅的に検索したい場合など，検索対象とするデータベースを切り替え，論文を丸ごとクエリとして特許検索を続行できれば検索の流れが途切れることがない．このような検索は，第1のデータベース中でユーザが指定した文書から，特徴語グラフを生成する場合と同様のメカニズムで特徴的な要素をリアルタイムで選出し，第2のデータベースに送信することで実現される．これらの機能を組み合わせることにより，ユーザは，「連想」に基づき複数のデータベースにまたがって必要な知識を収集・統合することができる．

純粋な文書データベースでなく，1レコード内に，塩基配列やアミノ酸配列のような非自然言語のデータフィールドと，自然言語テキストによる情報が付加されたフィールドを含むデータベースの場合，検索対象となるフィールドに応じてBLAST等の類似配列検索ツールと，テキストフィールド向けの連想検索ツールの，異種の検索ツールが存在する．このような場合，クエリを配列としてまず配列フィールドの類似度でレコードを抽出し，その中のテキストフィールドに書かれた内容を概観したり，興味あるテキストフィールドを指定した後，対象データベースを文献データベースに切り替えて検索することが考えられる．逆に，テキストをクエリとして配列を求める検索も実現できる．スケーラブルなテキスト検索エンジンは，異種の検索ツールとの連携により，図-1に示したようなさまざまなデータベースを縦横に活用するソリューションを構築することが可能であり，情報統合の要となる．

☆3 生体内におけるシグナル伝達，物質輸送，転写制御，ホメオスタシスなどを実現するための，物質の相互作用ネットワーク．

展望と課題

生体物質や化合物の間の相互作用（推定される相互作用も含む）を収集し，これらを結合したネットワークを構築し，このネットワークを用いて薬理作用／副作用を発見／予測するシステムは，スクリーニング支援における強力なツールとして期待される．これは，公開されているパスウェイ☆3データベース，文献検索や文献から抽出した物質間の相互作用に関する知識に加え，実験で観測された相互作用，機械学習により予測された相互作用等を統合した，異種データが混在するネットワークから知識を抽出するシステムである．ここで，主要文献から抽出された情報は，公開パスウェイデータベースの知識に次いで「知識」の根幹を成すものであり，これに用いるバイオNLPには高い精度が要求される．このためには，高精度・高速・頑健な構文解析技術，曖昧性解消技術等，NLPにおける重要問題の解決が必要である．

将来にわたって重要な課題となりつづけるのが辞書の構築である．我々は，遺伝子名，タンパク質名については，自動構築技術を開発してきたが[4]，化合物，疾患，薬理作用／副作用と，必要な辞書の種類は増大する一方であり，それぞれメンテナンスが必要なために膨大なコストが必要となる．情報検索，情報抽出等はこれらのコンテンツに依拠するものであり，公共利用可能な辞書が現在，そして将来にわたって提供されるならば，アカデミア／産業界を問わず，バイオ研究・バイオ産業の進展にとってきわめて価値あることと考える．辞書の自動作成やメンテナンス技術等をコンペティションの課題にするなど，米国の情報検索，情報抽出コンペのような衆知を結集する仕組みも考えられよう．公的ファンディング，およびアカデミアを中心とするコミュニティに最も期待するところである．

参考文献

1）辻井潤一：ゲノム情報学と言語処理，情報処理，Vol.43, No.1, pp.36-41 (Jan. 2002).
2）藤澤浩道，絹川博之：日本語情報処理の諸相：日本語情報検索技術の系譜，情報処理，Vol.44, No.12, pp.1276-1283 (Dec. 2003).
3）高野明彦，西岡真吾，丹羽芳樹他：汎用連想検索エンジンの開発と大規模文書分析への応用，IPA 2001年度成果報告集．
4）大井洋子，大田佳宏，今一　修他：一般語との曖昧性を持つタンパク質名の自動抽出，情報処理学会自然言語処理研究会報告，2004-NL-163(4)，pp.21-28 (Sep. 2004).
5）Hisamitsu, T. and Niwa, Y.：Topic-Word Selection Based on Combinatorial Probability, Proc. of NLPRS2001, pp.289-296 (2001).

（平成17年1月4日受付）