特集

バイオ文献活用基盤としてのXML検索技術

XML Document Retrieval Technology for Biological Literature

あらまし

ライフサイエンス分野では，ヒトゲノム解読を受けて，新たな研究フェーズ，ポストゲノム時代を迎えたと言われている。ポストゲノム時代の重要な研究目標の一つとして，遺伝子産物間の相互作用の解明が挙げられる。そのために収集・分析すべき実験データの量は，ゲノムの配列データに比べても桁違いに大きく，コンピュータによる支援が不可欠である。とくに，数値データの集合に過ぎない実験データを，生物学的知見に結びつけるには，言語情報の形で蓄積された知識資産を縦横無尽に活用する必要がある。本稿では，遺伝子産物間相互作用の解明における文献情報の役割について論じ，バイオ文献情報の活用基盤として開発したXML検索技術を紹介する。

Abstract

Now that a large amount of genome sequences, including the human genome sequence, are widely available, biomedical research is shifting to the phase called the “post-genome era.” Protein-protein interaction is an important research issue in this era for understanding the mechanisms of biological processes. An enormous amount of data, including gene expression data and genome sequences, are required to elucidate these interactions. Computers will then be used to help researchers in the analysis and interpretation of experimental data. Especially, to convert numerical data into biologically significant information, the knowledge resources that have been accumulated in various languages should be useful. This paper discusses the role of biological literature databases, especially how they help researchers interpret biological data. This paper also introduces our XML document retrieval technology, which can be used as a foundation for building biological research assistance systems.

仲尾由雄（なかお　よしお）

ITコア研究所グリッド&バイオ研究部 所属
現在，文書処理技術に基づくバイオ産業向けソリューションの研究開発に従事。

井形伸之（いがた　のぶゆき）

ITメディア研究所ドキュメント研究部 所属
現在，XML検索技術の研究開発に従事。

小櫻文彦（こざくら　ふみひこ）

ITメディア研究所ドキュメント研究部 所属
現在，コンテンツ統合技術の研究開発に従事。

ま え が き

ライフサイエンス分野においては，ヒトゲノム解読を受けて，遺伝子産物間の相互作用の解明に研究の焦点がシフトしつつある。DNAやアミノ酸配列データに加え，膨大な遺伝子・蛋白質の発現データの蓄積なども計画されつつあり，それらにより，遺伝子産物間の相互作用の解明が飛躍的に進むことが期待されている。しかし，一方で，膨大なデータの解釈を支援する情報技術には未成熟な面があり，データ量の増大が，新たな知見の創成の加速につながりにくいという現状がある。

本稿では，膨大なデータと知識とを有機的に組み合わせて活用するための基盤として開発したXML検索技術について，想定する利用モデルと現在の実現状況を紹介する。

バイオ文献情報の役割と情報構造

● 文献情報の役割

ライフサイエンス分野においては，早くから研究成果を蓄積した公開データベースの整備が進められてきた。最も著名な医学・生物学関係の文献データベースであるMEDLINEは，1971年より米国NLM（National Library of Medicine）によりオンラインサービスが行われている。MEDLINEには，4,600を超える医学・生物学関連の論文誌について，1,100万件を超える論文の抄録と書誌情報が蓄積されており，医学・生物学の研究者の貴重な情報源となっている。また，文献で報告された実験データ（ファクトデータ）のデータベース化も進められている。例えば，1988年にNLMの一部門として設立されたNCBI（National Center for Biotechnology Information）により，DNA塩基配列データベースGenBankなどの整備が進められてきた。それらのファクトデータベースとMEDLINE文献データベースとの間にはリンクが張り巡らされており，文献とファクトとを相互に参照することが可能になっている。

文献データベースは，第一義的には，研究者による関連文献の入手を支援するためのものであるが，ファクトデータを解釈する上でも重要な手がかりとなり得るものである。ファクトデータベースの登録エントリ数（DNAの塩基配列や蛋白質のアミノ酸配列の数）は，10年に5倍という爆発的なペースで増加を続けている。[1] そのため，人手による整理・体系化には限界があり，計算機による体系化の自動化・支援が強く求められている。例えば，酵母2ハイブリッドシステム（yeast two-hybrid system）という手法で蛋白質間の相互作用を解析し，相互作用に基づくネットワークを作成すると，千種類以上の蛋白質が一つのネットワークの中に取り込まれてしまうと報告されている。[2] その中から生物学的に重要な機能的連関を見いだすには，生物学の研究者が，自己の知識と直観に基づき，重要な意味を持つ相互作用を丹念に選別する必要がある。

そのような知的作業を計算機で支援する上で，文献情報は大きな手がかりとなり得る。例えば，同様の機能が文献で報告されている蛋白質の間の関係であるか，といった観点から相互作用を評価し，研究者の知識と整合性の高い相互作用を選別・提示すれば，研究者の分析効率が飛躍的に向上する可能性がある。また，実験により類似性が観察された蛋白質グループに対して，文献情報を手がかりに，共通機能を示唆する言語表現を自動付与すれば，研究者は，より多くの実験結果を総合的に理解した上で，仮説の生成や検証ができるようになると考えられる。[3] そのほかにも，文献情報をファクトデータと有機的に組み合わせて活用する様々な形態が考えられる。例えば，蛋白質の機能推定のための基本手法として，アミノ酸配列データのホモロジーサーチ（類似蛋白質を検索する手法）があるが，その検索精度が，文献情報の併用により向上したという報告もある。[4]

● 文献情報・ファクトデータの情報構造

上述の文献情報・ファクトデータには，いずれも，かなり複雑な構造を持った情報であるという点に特徴がある。例えば，MEDLINE文献データベースでは，各エントリは，エントリの識別子，出典情報，抄録に加え，検索用に人手で付与された索引語群（MeSH：Medical Subject Headings）など，他種類の情報を含む。また，ファクトデータでは，配列などのデータに加え，出典となる文献へのリンク，データの部分に対する注釈などを含むため，より複雑な構造となっている。

表-1は，GenBankの霊長類ゲノム配列データPart I（2001年6月：73,369 エントリ）の統計データである。

表-1　GenBankエントリに関する統計データ（Primate Sequence Entries，Part 1の場合）

	最小	最大	平均
文書サイズ（バイト）	1,020	644,410	6,562
タグ数	25	9,787	164

文書サイズは，タグ間の空白を除去して集計

文書サイズ・タグ数とも大きなばらつきがあり，1万近いタグを含む複雑な構造のエントリもあることなどが見て取れる。

なお，従来，これらの情報構造は，一定の物理フォーマットにより表現され，各エントリは単純ファイルの形式で流通していたが，近年，XML形式による流通が一般的になりつつある。XML形式は，ライフサイエンス分野だけでなく，広く普及しつつあり，汎用ツールの入手も容易になっているので，この流れは，今後も続くものと予想される。

XML検索技術の位置付け

前記の文献情報の役割を考慮し，今回のXML検索技術開発においては，つぎの2種類の利用モデルを想定した。一つは，検索インタフェースとXML検索エンジンのみから成るモデル（**図-1**）である。このモデルでは，XML検索エンジンは，エンドユーザの質問に対して，直接的に検索を行い，検索結果を返す。もう一つは，検索I/FとXML検索エンジンに加え，文書情報を分析するエンジンを加えたモデルである（**図-2**）。このモデルでは，XML検索エンジンは，エンドユーザの質問に対して，様々な観点から文書の分析を試みる分析エンジンに対して，分析に必要な情報を提供する役割を果たす。

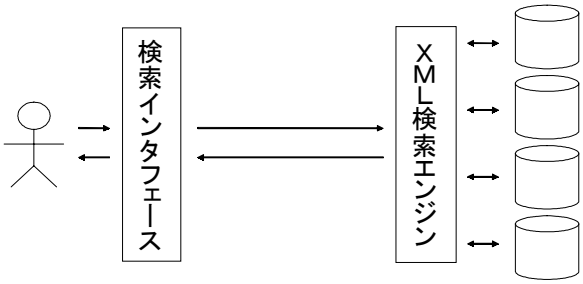

図-1　情報検索型の利用モデル
Fig.1-Typical interaction pattern between modules in an information retrieval system.

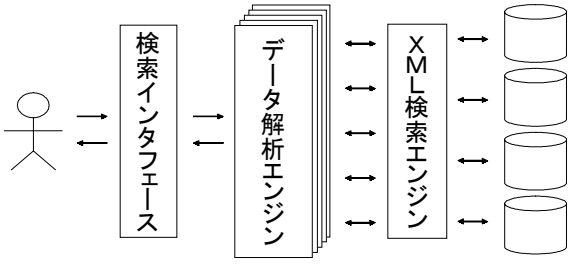

図-2　マイニング型の利用モデル
Fig.2-Typical interaction pattern between modules in a literature data mining system.

前者のモデル（図-1）は，従来の情報検索のモデルと基本的に同一であるが，構造化された文書を対象にしている点と，データの追加・更新が頻繁に行われる点に特徴がある。

前述のように，バイオデータベースのエントリは，文献情報だけでも，著者名や索引語の繰り返しがあるため，ビジネス分野で扱われるXML帳票データなどに比べ，かなり複雑な構造になっている。また，文献情報とファクトデータとを有機的に組み合わせて活用することを考えると，より複雑な構造を持つファクトデータもXML文書として格納できることが望ましい。そのため，バイオデータベース向けXML検索技術においては，複雑な構造への対応が大きな要件となる。

また，従来の情報検索では，収集・蓄積された文書集合に対して索引を一括付与するという運用形態が多かったため，検索処理に比べると，索引作成処理の速度性能は問題とならないことが多かった。一方，バイオデータベースにおいては，定期的文書追加に対応するために，運用に耐えられる時間内で，索引を更新できることも重要な要件の一つとなる。

後者のモデル（図-2）においては，要求される応答性能と検索エンジンが出力すべき情報の種類が，前者のモデルとは異なる。前者のモデルにおける検索エンジンは，人間に対してテキスト情報を返すことが基本になる。この場合，応答速度は，人間の認知速度を超えてまで速くする必要はない。一方，後者のモデルでは，分析エンジンに対して，検索結果となるテキストに関する何らかの特徴量を返すことが基本となる。例えば，文書の類似度の判定などでは，

(1)　各テキストに，それぞれの語がどの位の頻度で出現しているか，
(2)　検索対象集合において，それぞれの語がどの位の数の文書に出現しているか，

といった情報がよく用いられる。検索エンジンは，そのような情報を分析エンジンに返すことになる。

分析エンジンは，人間に対して応答を返す間に様々な分析処理を行うので，その下請けとなる検索エンジンの応答時間は，できるだけ短いことが望ましい。よって，(1)のような情報を返す場合には，テキスト情報そのものをではなく，集計を済ませた頻度情報のみを返すこ

とが望まれる。また，(2）のような文書集合に関する特徴量については，検索対象全体だけでなく，任意の文書集合について高速に取得できることが望まれる。特徴量集計の対象となる要素についても，語のように固定的な要素だけでなく，任意の語数からなる複合的な表現なども含めて，特徴量が集計できることが望まれる。

開発方針と実現状況

今回開発したXML検索エンジンXAR21（X Archiver 21st century edition ） は ， IntelligentSearch（BizSearch）などで適用実績のある高速全文検索エンジンTerass[5]をベースに，構造化文書対応モジュールを追加・拡張することで実現した。この検索エンジンは，文字n-gram索引技術により，検索条件として任意の文字列が指定できるという柔軟性と，高速性を両立していることに特徴がある。今回の開発においては，この柔軟性と高速性を生かし，柔軟で可用性の高いXML検索エンジンを実現するために，以下のような方針を策定した。

(1)　検索条件は，構造に関する条件と内容に関する条件の組合せで指定する。構造条件は，要素の親子関係に関する条件のみを有効とし，要素の出現順序に関する条件の判定は省略する（出現順序が条件通りでない場合を含め，多めの検索結果を返す)。これは，構造検索対応による索引情報の増大を抑制するための方針である。

(2)　定期的な文書追加を想定し，索引の追加作成機能を用意する。具体的には，1日あたり数KBのデータを100万件規模で追加可能な程度（数時間で処理が完了する程度）の速度で，追加文書を含めた索引の作成（追加・更新）できる性能を確保する。

現在は，これらの方針に従う開発が一通り完了し，以下のような機能を実現できた状況にある。

(1)　テキスト部に対しては，基本機能として，部分文字列一致検索機能を実現した。また，必要な箇所に対しては，完全一致検索を可能にするための機構も用意した。

(2)　属性部に対しては，基本機能として，トークン列の部分一致・完全一致検索機能を実現した。トークンとは，空白で区切られた一続きの英数字記号を指す。

(3)　テキスト・属性部とも，情報の種類（データタイプ）によって異なる検索ニーズを満たせるよう，事前に索引作成方法を指定して，照合方式を切り替える機構を用意した。照合方式としては，文字列単位の比較，トークン列単位の比較，数値比較の3種類を用意した。

(4)　文書追加に関しては，索引マージ機能を追加することで，登録済み文書数による追加速度の劣化を抑制した。例えば，200バイト程度のデータを500万件格納する際，20回に分けて分割登録すると一括登録に比べ約3倍の（累積）処理時間（17.4時間）を要していたものが，約1.2倍の（累積）処理時間（6.2時間）に短縮された。

表-2は，現行のXAR21を，UltraSPARC-II 300 MHzの計算機（1.5 Gバイトメモリ）上で動作させた場合の速度性能の概要である。この表は，前述のGenBankのDNA配列データ（表-1）や，Webページに関するメタ情報など，性質の異なるXML文書集合（平均サイズ0.2～6 KB；文書数1万～500万）を実験データとして測定した処理時間の概要を，最も強く影響を与える要因を取り上げて示したものである。

この中で，文書構造の複雑さに最も敏感なのは，(b）文書検索性能の「条件照合」である。これは，索引を参照して検索条件に合致した文書集合を確定する処理であり，マイニング型の利用モデル（図-2）においても，分析モジュールが分析の基礎情報取得のために繰り返し呼び出す可能性が高い機能である。今回採用した構造検索用索引法では，最悪の場合，検索対象とすべきフィールド（タグで囲まれた領域）の数に比例して条件照合処理が増加する可能性がある。表-2で1秒となっているのは，前述のGenBankデータに対して，索引付けの際に区別したフィールド（約4万）のすべてを対象に，6割以上の文書に出現する検索キーを与えて検索した場合の値である。GenBankデータを対象にする場合でも，「属性記述部」といった実際に想定される検索を行う場

表-2　XAR21の速度性能

(a）文書登録性能

文書解析	文書（タグ込み）1 Mバイトあたり2秒程度
文書格納	1文書あたり1～3ミリ秒程度
索引作成	索引対象部（タグ除外）1 Mバイトあたり20秒程度

(b）文書検索性能

条件照合	1検索条件あたり10ミリ～1秒程度
ID取得	1文書あたり0.1ミリ秒程度
内容取得	1文書あたり1～10ミリ秒程度

対象：平均サイズ0.2～6 Kバイトの文書1万～500万件

合には，検索対象とするフィールドが数千程度に絞られるので，条件照合は0.1～0.2秒程度で完了する。

むすび

遺伝子産物間の相互作用の解明には，膨大な実験データを生物学の知識に沿って，整理・体系化する必要がある。そのための支援技術として，膨大な実験データを，言語情報として蓄積された知識に基づき，自動的に分析・整理するIT技術が強く求められる。今回紹介したXML検索エンジンでは，そのような支援技術の実現するための基盤技術として，柔軟で高速な検索機能を実現した。膨大な実験データを膨大な知識に照らして分析するには莫大な計算量が必要である。よって，分析処理のために，分散・並列計算機環境の利用が進むと予想される。今後は，並列に分析を行う分析エンジンに対して，分析に必要なデータを高速に提供するための機能の拡張などを行う予定である。

参考文献

(1) 金久實：ポストゲノム情報への招待．共立書店，東京（2001）．
(2) 伊藤隆史：2ハイブリッド法によるタンパク質相互作用の網羅的解析．伊藤隆史，谷口寿章（編），プロテオミクス，p.122-135，中山書店，東京（2000）．
(3) 大久保公策：医学知識を機械に伝える－BOBプロジェクト．松原謙一（編），ゲノム機能，p.134-143，中山書店，東京（2000）．
(4) J. Chang et al.：Including Biological Literature Improves Homology Search. *Proc. of PSB 2001*，p.374-383（2001）．
(5) 松井くにお，ほか：大容量情報全文検索エンジンTerass. *FUJITSU*，Vol.48，No.3，p.240-243（1997）．