公募研究：2000～2002年度

XMLを用いたゲノムデータ統合システム

●吉川 正俊[1)]　◆松原 茂樹[1)]　◆天笠 俊之[2)]　◆波多野 賢治[2)]

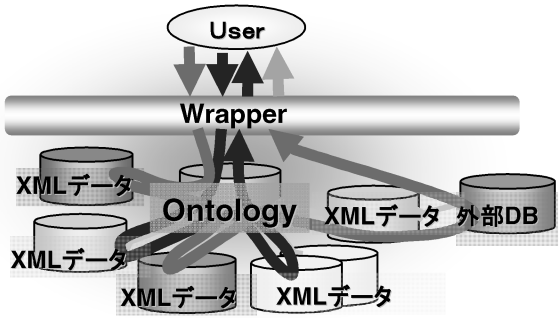
1) 名古屋大学情報連携基盤センター　2) 奈良先端科学技術大学院大学情報科学研究科

〈研究の目的と進め方〉

公開され始めている多様なゲノムXMLデータを随時システムに取り込みながら統合的に管理し，それらデータの横断的な検索やナビゲーション機能を提供する汎用的なXMLデータの構築方法，XMLデータベースの利用者インタフェースの開発，および更新が多い多様なDTDやXMLスキーマに基づくXMLデータを効率良く管理可能なデータベースシステムを開発することを目的とした．

(1) 多種類のXMLデータを効率良く管理するためにゲノムXMLに適したデータベースの基本アーキテクチャを設計・開発する．また，ゲノムデータは更新の頻度が高いために，検索，更新ともに効率よく実行可能なXMLデータ索引の開発が必要である．したがって，XMLの木構造を表現する符号化方法について研究する．さらに，更新が多い多様なDTDや XMLスキーマに基づくXMLデータを効率良く管理可能なデータベースシステムを開発する．

(2) 膨大なゲノムXMLデータを対象として情報科学の非専門家が簡易検索を行なえるようなXMLサーチエンジンの検索アルゴリズムを開発する．



〈研究開始時の研究計画〉

分子生物実験の成果としての遺伝子機能データは，必ずしも確定したもののみではなく，仮説的なデータもデータベースに登録する必要がある．そこで，XMLデータに確信度，注釈，時制などの様相を取り込んだデータモデルを構築する．また，宣言的問合せ言語を用いたビューXMLデータにも基底データと同様に様相データを追加できるようにして高階様相データの表現を可能とする．また，このデータモデルに基づくXMLデータの実装法を開発し，実際の遺伝子機能データをもとに様相XMLデータベースを構築する．

遺伝子機能データを表現するための様相XMLデータベースモデルの設計を行うとともに，変換システム，特徴量抽出など各要素技術を開発する．

・遺伝子機能データベース用様相XMLデータモデルの要件の整理

開発するXMLデータモデルの要件を，研究協力者として学内外のゲノム研究者の協力を得ながら整理する．

・様相XMLデータの論理モデルの開発

既に研究中の時制XMLデータモデルをもとに，論理データモデルを開発する．高階様相データモデルの構築に際しては，RDF (Resource Description Framework)のreificationの考え方を参考にする．

・XML変換システムの開発

様相に基づきXMLビューを構築し，ビュー上の問合せの最適化を行うアルゴリズムを開発する．データのフィルタリングなどの変換を行うシステムをXSLTやXML-QLの表現能力を補う形で開発する．

・遺伝子機能データの特徴量の開発

一次データとしての膨大なゲノムデータとは別にその特徴量を少ないデータ量で記述することにより，データの概観把握や高速検索に利用できる．自然言語データとは異なるゲノムデータに適した特徴量の開発を行う．

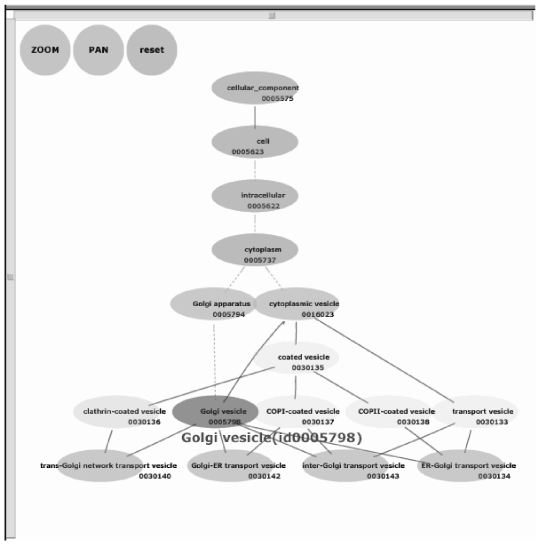
〈研究期間の成果〉

複数の異なるモデル生物種の遺伝子注釈データなど多様なゲノムXMLデータを統合的に管理し，グラフィカルブラウザを提供するシステムGeneAroundを開発した．

GeneAroundは，分子生物学者が，同じ機能を持つ遺伝子の異なる生物種間での比較，生物種横断的な検索，ナビゲーションなどを容易に行える環境を提供することを目的とした．また，システムの拡張性を確保するために，W3C標準のXML問合せ言語など標準技術を基盤にしたシステムとした．MaXML(Mouse annotation XML)など統合化語彙階層であるGene Ontologyを参照する複数のゲノムXMLデータを，Gene Ontologyを中心に統合し，XQueryエンジンを用いた検索システムを構築した．(成果公表リスト3,5,6,7,8,9)

また，約12,000個の用語から成るGene Ontologyの階層をSVG (Scalable Vector Graphics)を用いて視覚化するグラフィカルブラウザを開発し，システムに統合した．このグラフィカルブラウザは，用語間の関連をわかりやすく表示することにより，注釈データに関連付けられた用語の意味の理解を助けるとともに，用語に関連する遺伝子注釈データエントリへの検索インタフェースの役割を果たす．このグラフィカルブラウザは，以下の特徴を持つ．

■注目する用語の周辺の用語とそれらの関係を表示
■SVG(Scalable Vector Graphics)を利用し，GUIによる拡大・縮小，他用語へのトラバースが可能
■グラフィックを用いた，2次元のGO階層図示ビューワ
■単体で動作し，他データベースとの連携も容易



また，更新が多いXMLデータを管理するためのリージョン表現方法や符合化法を考案した．(成果公表リスト1,2,4,12)

XMLデータの簡易検索を可能とするシステムのプロトタイプを作成した．また，その評価を行うために，INEX (Initiative for the Evaluation of XML retrieval)国際イニシアティブに参画し，テストコレクションを作成した．

〈国内外での成果の位置づけ〉

種々のXMLデータを統合利用するための試みも各方面から研究されているが，W3C標準のXML問合せ言語を利用した柔軟な構成のオープンなシステムを指向したものは他には見当たらない．システムの拡張性を確保するために，標準技術を基盤にしたシステムを構築を行った．

GeneAroundシステムは，理化学研究所のマウス遺伝子注釈付けプロジェクトFANTOM (Functional Annotation of Mouse) において分子生物学者が実際にマウスcDNAの注釈付け作業をする際に利用されるとともに，東京大学，Clemson大学をはじめとする国内外の分子生物学者によって広く利用された．また，グラフィカルブラウザに関する論文がBioinformatics誌に掲載されたため，海外からのWebサイトへの多くのアクセスがあった．

〈達成できなかったこと、予想外の困難、その理由〉

公開されているフリーのXQueryエンジンに未実装の機能があり，その部分を補うために拡張性を犠牲にせざるを得ない部分があった．また，学生の修了や研究者の異動に伴い，Webサイトの管理が不十分となり即座にデータ更新ができなくなり，大学においてこのようなシステムを維持することの困難さを経験することになった．

〈今後の課題〉

有用性の高いシステムとするためには，より多くの種類のデータを統合する必要がある．また，利用者がより高度な問合せをできるようなインタフェースを構築するために，XML問合せ，XMLデータ格納方式の見直しによる検索の高速化やインタフェースの改良を行う必要がある．また，XMLデータベースの利用者であるゲノム研究者に，XML問合せ言語よりも簡便な問合せインタフェースを提供することが重要である．情報科学の観点からは，XMLスキーマをどの程度円滑に利用できるかが重要な点である．また，各種のデータを結びつけるメタな知識は必ずしもデータ中には明記されていないことが多いため，この部分を整理し，XQueryなどから利用可能とする必要がある．また，Gene Ontologyデータの更新に伴いGeneAroundデータの更新も必要となるが，これを自動化する一つの方法はWebサービス化であると考える．

〈研究期間の全成果公表リスト〉

1）論文／プロシーディング

1. 0304011734
T. Amagasa, M. Yoshikawa and S. Uemura: "QRS: A Robust Numbering Scheme for XML Documents", Proceedings of the 19th International Conference on Data Engineering (ICDE 2003), pp. 705-707 (2003)
2. 0304011742
江田 毅晴, 天笠 俊之, 吉川 正俊, 植村俊亮: "XML 木のための更新に強い節点ラベル付け手法", 日本データベース学会 Letters, Vol. 1, No. 1, pp. 35-38, (2002)
3. 0303311809
J. Tanoue, M. Yoshikawa, S. Uemura: The GeneAround GO viewer, Bioinformatics, 18(12), 1705-1706 (2002)
4. 0304011739
D. Dinh Kha, M. Yoshikawa and S. Uemura: Application of rUID in Processing XML Queries on Structure and Keyword, 13th International Conference on Database and Expert Systems Applications (DEXA2002), Lecture Notes in Computer Science (LNCS), Springer-Verlag, 2453, 758-767 (2002)
5. 0303311832
J. Tanoue, N. Matoba, M. Yoshikawa, S. Uemura: GeneAround: A Browsing System for Gene Annotation Using XML Technologies, The Third International Conference on Web-Age Information Management(WAIM'02), regular paper, Lecture Notes in Computer Science (LNCS), Springer-Verlag, 2419, 236-246 (2002)
6. N. Matoba, J. Tanoue, M. Yoshikawa, S. Uemura: "A System for Integration of Heterogeneous Biological XML Data", Genome Informatics 2001 (Genome Informatics Series Vol. 12), Dec. (poster) (2001)
7. J. Tanoue, N. Matoba, M. Yoshikawa, S. Uemura: "Graphical representation of Gene Ontology in Scalable Vector Graphics", Genome Informatics 2001 (Genome Informatics Series Vol. 12), Dec. (poster) (2001)
8. J. Tanoue, M. Yoshikawa, N. Matoba, S. Uemura: "An XML document management system using XLink for an integration of biological data", 9th International Conference on Intelligent Systems for Molecular Biology(ISMB2001), July (poster) (2001)
9. N. Matoba, M. Yoshikawa, J. Tanoue, S. Uemura: "Portal XML Server: Toward Accessing and Managing Bioinformatics XML Data on the Web", 9th International Conference on Intelligent Systems for Molecular Biology(ISMB2001), July (poster) (2001)
10. 0202272233
波多野 賢治, 渡邉 正裕, 吉川 正俊, 植村 俊亮: "情報検索技術を用いた部分文書構造の自動抽出", 情報処理学会

論文誌：データベース, 第42巻, 第SIG8(TOD10)号, pp. 36-46 (2001)
11. 0202272229
T. Amagasa, M. Yoshikawa and S. Uemura: "Realizing Temporal XML Repositories using Temporal Relational Databases", Proceedings of The Third International Symposium on Cooperative Database Systems for Advanced Applications (CODAS'2001), pp. 63-67, Beijing, China, April 23-24 (2001)
12. 0202272217
D. Dinh Kha, M. Yoshikawa and S. Uemura: "An XML Indexing Structure with Relative Region Coordinate", Proceedings of the 17th IEEE International Conference on Data Engineering (ICDE2001), pp. 313-320, April (2001)
13. M. Yoshikawa, T. Amagasa, D. Dinh Kha, K. Hatano, H. Kinutani, N. Matoba, J. Tanoue, M. Watanabe and S. Uemura: "On Two Query Interfaces for Genome XML Databases", IEEE Workshop on XML-Enabled Wide Area Search in Bioinformatics (XEWA), League City, Texas, December 13-14, (2000)