転写制御領域をゲノムワイドにスクリーニングするための転写因子結合候補領域データベース

宮崎　和典† 菅野　美津子† 伊藤　聡† 芦野　俊宏‡
(株)東芝　研究開発センター† 東洋大学　国際地域学科‡

1．はじめに

ポストゲノムにおける重要な課題の一つは、ゲノムの塩基配列から転写制御に関わる領域を同定することである。転写とは、遺伝子の塩基配列からタンパク質の設計図であるメッセンジャーRNA（mRNA）を合成する過程のことである。転写は、転写因子と呼ばれるタンパク質がゲノム上の特定の領域（転写制御領域）と結合・解離することにより制御を受ける。

これまでに、計算機によりゲノムの塩基配列から転写制御領域を予測するための基本的なツールが開発されている。例えば、転写因子が認識する特定の塩基配列を文献から収集し、データベース化したものが幾つか存在する（TRANSFAC が有名である）。また、TRANSFAC のデータに基づいて塩基配列から転写因子が結合する領域を予測するアルゴリズムも複数開発されている。現段階では、これらのデータベース、アルゴリズムを活用して、予測された転写因子結合候補領域からいかにして生体内で実際に働いている転写因子/転写制御領域を同定するかが重要な課題となっている。また、ゲノムプロジェクトによりゲノムの塩基配列が容易に入手できるようになった現状では、計算機により転写因子結合候補領域を検索可能な領域がゲノムワイドになっている。今後は、これまでより 1、2 桁以上広範囲な領域について転写因子結合候補領域を検索し、大量に検出される候補から効率よく転写制御領域をスクリーニングする手法の開発が期待されている。本報告では、このようなスクリーニング手法を開発するための基盤となる転写因子結合候補領域データベースを構築したので報告する。

2．システムの構成

本システムの構成を図１に示す。本システムは、(1)転写因子結合候補領域検索モジュール、(2)転写因子結合候補領域データベースマネジメントシステム（DB-MS）、(3)転写因子結合候補領域可視化モジュールから構成される。各モジュールの実装には、Java 2 Platform, Standard Edition ver.1.4.1 を用いた。

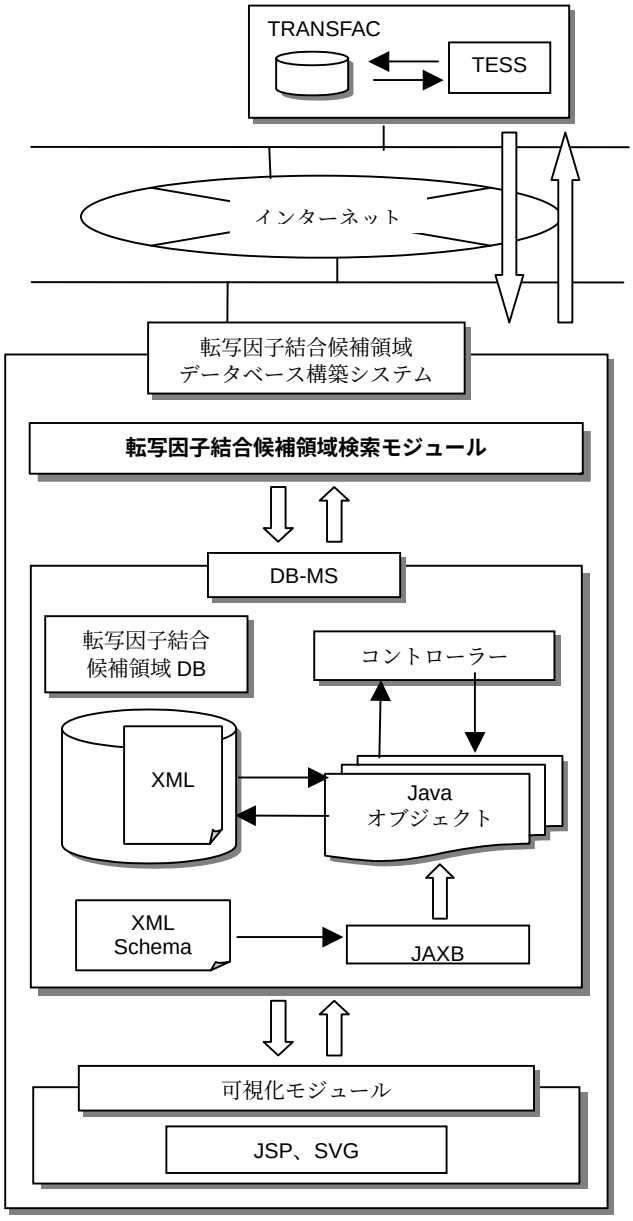

図１．システム構成

2-1　転写因子結合候補領域検索モジュール

Database system of transcriptional regulatory sites
Kazunori Miyazaki†, Mitsuko Sugano†, Satoshi Itoh†, Toshihiro Ashino‡
† TOSHIBA CORPORATION, Corporate Research & Development Center
‡ Department of Regional Development Studies, Toyo Univ.

本モジュールは、Web で公開されているツールを活用し、転写因子結合候補領域をゲノムワイドに検索して、転写因子結合候補領域データを半自動的に収集する。転写因子結合領域予測ツールとして、ペンシルベニア大で開発された TESS[1] を利用した。

2-2　転写因子結合候補領域 DB - MS

TESS による転写因子結合候補領域の検索結果は HTML ファイルとして提供される。しかし、HTML ファイルでは計算機による大量解析には向かない。このため本報告のシステムでは、データを XML で記述し直すこととした。これは、以下の 3 つの理由、(１)計算機による処理を行うためのインターフェイスが容易に入手できること、(２)データの構造化とデータ構造の拡張が容易であること、(3)テキストベースのフォーマットであること、による。本報告で開発するデータベースは、データの解析が目的で構築するものであり、解析で得られる知見も新たなデータとしてデータベースに追加することを念頭に置いている。つまり、データベース構築後にデータ構造を拡張する可能性が高く、これには上述(2)の XML の特徴が重要である。また、バイオ関連のデータは基本的にテキストベースであり、テキストの内容（塩基配列等）を目で確認することが必要となる場面が多い。このため、データを閲覧して簡単に内容を把握できるテキストベースの構造である意義は、少なくともバイオ関連のデータを扱う場合には高いと考えられる。特に、XML のように構造化して表示させることができると、単なるテキスト以上にデータの内容の理解が容易になる。

転写因子結合候補領域データの構造は、XML schema により定義した。TESS による検索結果の HTML ファイルから必要なデータを抽出し、先述の XML schema に従い XML 文書を生成し、転写因子結合候補領域 DB に格納した。

転写因子結合候補領域 DB に格納されたデータの処理、管理・更新等を行うための Java アプリケーションは、XML データバインディングを利用した。Sun Microsystems, Inc.が提供している WSDP （ Java Web Services Developer Pack 1.2）に含まれる JAXB (Java Architecture for XML Binding) v1.0.1 を利用し [2]、先述の XML schema に対応する Java プログラムを作成した。JAXB により生成された Java プログラムに加えて、これらのプログラムを統括し、XML 文書の処理を管理するための Java アプリケーション（図 1、コントローラー）を作成した。

2-3　転写因子結合候補領域可視化モジュール

本システムでは、データ可視化と可視化に伴うデータ処理ツールは JSP を利用した Web ベースで構築した。また、グラフィックスの描画には、SVG を利用した。

3．ダイオキシン応答性遺伝子の転写因子結合候補領域 DB の構築

我々は、細胞にダイオキシンを作用させた際に転写量が変動する遺伝子を 26 個同定している。これらの遺伝子は、共通の転写制御を受けている可能性があり、これらに共通する転写因子の同定をする目的で、本システムにより 26 遺伝子近傍のゲノム配列における転写因子結合候補領域の検索、データベース化を行った。この結果、約８万箇所を超える転写因子結合候補領域がリストアップされた。図２は、26 遺伝子の内の 1 遺伝子 Cyp1a1 で予測された転写因子結合候補領域のゲノム上での分布を可視化モジュールにより可視化した例である。

図２．転写因子結合候補領域のゲノム上での分布の可視化例

4．おわりに

転写制御領域をスクリーニングためのツール開発において基盤となる転写因子結合候補領域データベースを構築した。今後、スクリーニングツールの開発に応用していく予定である。

参考文献

1. http://www.cbil.upenn.edu/tess/
2. http://java.sun.com/developer/technicalArticles/WebServices/jaxb/