3A1-04

データ間の類似度を用いた生物学データベースの統合

Integration of Biological Database

関根　毅[*1]　平石　広典[*2]　溝口　文雄[*1]
Tsuyoshi Sekine　Hironori Hiraishi　Fumio Mizoguchi

[*1]東京理科大学　理工学部
Faculty of Sci. and Tech. , Tokyo University of Science

[*2]東京理科大学　インテリジェントシステム研究所
Intelligent System Lab. , Tokyo Univ. of Science

This paper describes Integration of Biological Database. Our system integrates two or more databases, calculates the degree of similar between data, and visualizes a PATHWAY database.

1. はじめに

近年，分子生物学のデータベース [1][2] がたくさんインターネット上に公開されるようになった．国内では京都大学化学研究所バイオインフォマティクスセンターのゲノムネットのシステムである DBGET によって，分子生物学のデータベースから一元的にデータを検索することが有名である．しかし，データは各研究機関のデータベースごとにフォーマットが規定されているためそのままでは解析に利用しにくい．そこで，データベースを統一的に扱うためのフォーマットを提案する．さらに，既存のデータベースの類似度 [3] を用いてデータ間の類似度を計算するための手法を提案する．これは同一の内容を指し示すデータはデータベースが異なっていても似たデータとして表現されると考えられるからである．そしてこれらを用いて，生物学データベースを利用しやすくするためのシステムの設計・実装を行った．

2. データベース統合モジュール

本研究で用いるデータベースを一元的に管理するためのインターフェースの定義を行う．DBGET システムは，ゲノムネットの WWW から，利用することが可能になっており，データベース内の要素を一意に示す要素をエントリと定義し，

データベース名：エントリ

とすることで，指定されたデータベースのデータを取得することができる．しかし，生物学のデータベースは，各々のデータベースでデータフォーマットが独自に定義されているため，アプリケーションの方でデータの関連などを調べなくてはならない．そこで，データベースを扱いやすくするために，データベースへアクセスするためのモジュールを図 1 のように設計した．各々のデータベースをラッピングするアダプタを定義することで，アプリケーションからは，統一したインターフェイスで複数のデータベースから情報を取得することが可能になる．アダプタは，Java のインターフェースとして定義し，データベースごとに実装する．ひとつのインスタンスはデータベース内のひとつの要素を指し示す．

次に，図 2 で示すアダプタのインターフェースのメソッドについて説明する．setEntry メソッドでは，エントリをインスタンスにセットし DBGET への URL を確定する．load メソッドで, データをダウンロードする．ダウンロードしたデータの読み込みが完了したら，Vector に格納し parse メソッドを呼ぶ．parse メソッドでは，読み込んだデータの構文解析をする．データベースごとに異なるファイル形式のため解析して，統一した形式でメモリ上に確保する．取得したデータを呼び出すには，getData メソッドか getAllData メソッドを使う．getData メソッドでは，キーを指定して一部のデータを取得する．名称などは同じタンパク質であっても生物や分野が違うと違う名前で記述されるので値を配列として返す．getAllData メソッドは，格納されているデータをすべて Vector にして返す．エントリに含まれるすべてのデータが取得できる．このインターフェースを，データベースごとに実装することで，データベースを抽象化しアプリケーションからデータを利用しやすくする．

連絡先: 関根　毅，東京理科大学理工学部経営工学科，千葉県野田市山崎　2641，Tel.04-7124-1501，j7404634@ed.noda.tus.ac.jp

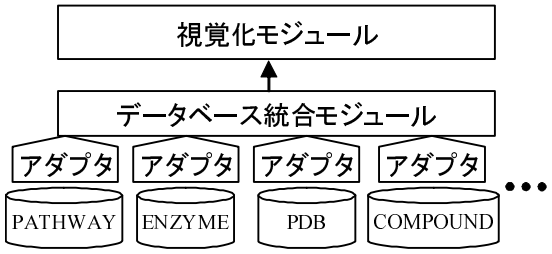

図 1: アダプタのイメージ

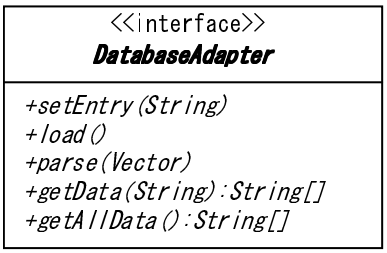

図 2: アダプタのインターフェース

3. データ間の類似性

データベースには，アミノ酸配列や遺伝子配列，代謝反応化合物など様々な種類のものがある．これらのデータベースの中でのデータは遺伝子配列なら BLAST，化合物なら SIMCOMP というように類似性を評価するシステムが開発されている．経験的に，似た配列をもつ遺伝子の機能は似ていることがわかっ

ているように，あるデータベースにおいてのデータの類似性と，他のデータベースにおけるデータの類似性にも関係があると考えられる．そこで，本研究では，複数のデータベースのデータの類似度を組み合わせることによって，データ間の類似度の計算を行う．

3.1 データベース間のリンク

本システムではゲノムネットの DBGET を用いてデータベース間のリンクを調べることにする．通常，分子生物学の分野では異なるデータベースに関連するデータがあれば，それにリンクを付加してデータベース化が行われる．文献データと文献に報告された配列データとの関連，塩基配列とそれを翻訳したアミノ酸配列の関連などが代表例である．リンク情報は

ＤＢ１：エントリー１　→　ＤＢ２：エントリー２

の形で表現され，この２項関係を用いる．ゲノムネットでは LinkDB と呼ぶ２項関係だけのデータベースをもっており，２項関係を演繹して新たな２項関係を作ることにより，多くの関連データを容易に見いだせるようになっている．本研究では，これらの関係したデータを仮想的にひとつのデータとして扱い，これらのデータの間の関係性を見出すために類似度の計算を行う．

3.2 類似度の計算

まず，個々のデータベースにおいてそれぞれのデータ間の類似度の計算を行う．このとき得られた個々の類似度を０～１の範囲にに正規化する．このとき，１のときには完全に一致し，０のとき類似性がないとみなす．そして，式１によって類似度の計算を行う．

$$\text{類似度} = \sum^{\text{すべてのデータベース}} \frac{\text{データベースごとの類似度}}{\text{データベースの総数}} \quad (1)$$

まとめると類似度の計算は次のようになる．

１．LinkDB をたどって類似度を利用する
　　データベースのデータを取り出し，
　　ひとつの仮想的なエントリとみなす。
３．あるデータベースで類似度を計算する。
３．得られた類似度を式 (1) の計算をする。
４．2，3 の操作を類似度の計算できるすべての
　　データベースについて繰り返す。

4. 視覚化モジュール

データ検索のためのインターフェースとして視覚化モジュールの設計・実装を行う．本インターフェースは，パスウェイのデータベースを元に情報検索を行うようになっている．パスウェイは，生体内での代謝経路を表していて生体内での化学反応の経路がわかる．これを調べることは病気などの研究に役立ち，ドラッグデザインなどを行う上でも重要と考えられる．

図 3 で本システムの実行画面を表示する．中央のパネル上にパスウェイの構成が表示される．パスウェイ中には，酵素や生化合物のエントリ名が表示される．右上の画面に，このパスウェイに含まれる要素がすべてツリー表示され，右下の画面には，パスウェイ上や，ツリー上の要素を選択したときに簡易的な情報が表示される．パスウェイ上のノードの一つ一つが，酵素や化合物を表している．それらを選択することで，詳細な情報を得ることが可能になっている．まず，画面上のノードを

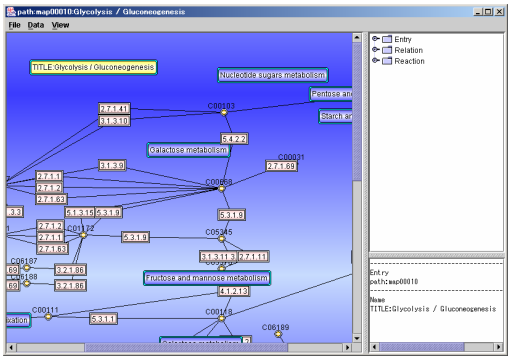

図 3: 視覚化システム

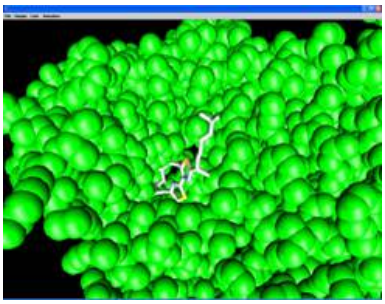

図 4: ドラッグデザインへの応用

選択すると，インターネット上のデータベースへの接続が行われ，自動的に統合されたデータの取得が行われる．

またこれらのデータを利用してドラッグデザインへの応用も行った．パスウェイ上の酵素タンパク質をドラッグターゲットとして，立体構造データを転送することによってドッキングシミュレーションを行う．(図 4)

5. おわりに

本システムにより，複数の生物学データベースを統合するためのしくみを提供することができた．また視覚化インターフェースを用いることでパスウェイと関連するデータを統合して表示することが可能になり，ドッキングアプリケーションと連携を行うことでデータを様々なアプリケーションで共有することができた．本システムを用いることで，製薬に応用するとドラッグデザインの候補物質を探すときに利用できる．

参考文献

[1] Fujibuchi, W. et al: ”DBGET/LinkDB: an integrated database retrieval system”, Pacific Symp. Biocomputing 1998, 683-694, 1997

[2] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne: ”The Protein Data Bank. ”Nucleic Acids Research, 28, 235-242, 2000

[3] Hattori, M., Okuno, Y., Goto, S. and Kanehisa, M. ”Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways”, Journal of the American Chemical Society, 125, 11853-11865 (2003)