生物学データベースの統合および視覚化のための研究

関根　毅†　　平石　広典‡　　溝口　文雄†
東京理科大学　理工学部　経営工学科†
東京理科大学　情報メディアセンター‡

1.はじめに

近年，分子生物学のデータベース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．また,ゲノムネットの DBGET[1]システムによって,世界中のデータベースから一元的にデータを検索することが可能になった．しかし，データフォーマットは統一されていないためそのままでは解析に利用しにくい．本研究では，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を提案する．生体内でのタンパク質の反応の経路を表したものをパスウェイといい，ゲノムの機能解析の場面などでもパスウェイの情報は重要である[2]．また，大規模なデータを効果的に表示する技術である情報視覚化技術[3]も利用しやすくなっている．

そこで本研究では，パスウェイデータベースを中心としたデータの統合および視覚化を行っていく．

２．データベース統合モジュールの設計

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して，データベース名：エントリとすることで，データを取得することができる．生物学のデータベースは，データフォーマットが統一されていないので，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，アクセスするためのモジュールを　図１のように作成した。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．データベースには，それぞれ適切なアダプタを作成することによってアクセスする．

Integration and Visualization of Biological Databases
†Tsuyoshi SEKINE, Fumio MIZOGUCHI
Faculty of Sci. and Tec., Tokyo University of Science
‡Hironori HIRAISHI
Information Media Center, Tokyo University of Science.

このため，データベースを追加したときは，アダプタを追加するだけでよく，拡張性が高い．

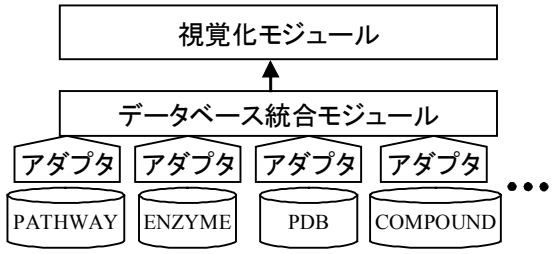

図１　統合モジュールの概念図

次に，アダプタの設計について述べる．アダプタは，Java の抽象クラスとして定義し，データベースごとに実装する．そして，ひとつのインスタンスがデータベース内のひとつの要素を指し示すものとして定義する．

```
1:import java.util.Vector;
2:public abstract class DatabaseAdapter
3:{
4: void setEntry(String entry);
5: void load();
6: void parse(Vector lineData);
7: String[] getData(Strin key);
8: Vector getAllData();
9:}
```

図２　アダプタの抽象クラス

ここで，アダプタの抽象クラスのメソッドについて説明する．４行目の 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．５行目の 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．６行目の 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，７行目の getData メソッドか８行目の 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．

このように，データベースを抽象化して表すことによって，どのデータベースも統一した形式でデータを表現することが可能になり，アプリケーションからデータを利用しやすくなる．

３．視覚化モジュールの設計

ゲノムネットの PATHWAY データベース(http://www.genome.ad.jp/kegg/pathway.html)では，ホームページ上でパスウェイの情報を取得するためのシステムが存在し，ノードをクリックすることでその情報へのリンクを得ることができる．我々は，よりインタラクティブな操作を実現するために，情報視覚化技術を考慮したパスウェイの表示システムを設計した．

図３に本システムの画面を表示する．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．より詳細な情報は，パスウェイの画面をダブルクリックすると図４のようなウインドウ群が表れ，詳細情報が表示される．

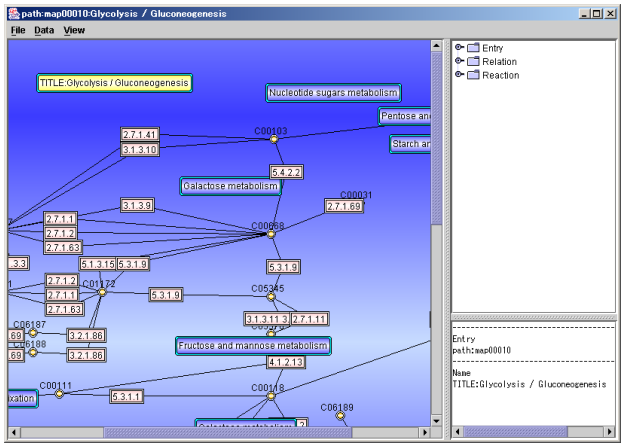

図３　パスウェイの表示

次に，詳細情報のウインドウについて説明する．まず，画面上の酵素を選択することで図４左上の画面が表示され，酵素についての情報が表示される．酵素の情報は ENZYME データベースにあり，EC 番号で一意に識別される．詳細情報は，酵素名，EC 番号，分類などのデータを表示する．

図４右上の画面は，酵素タンパク質の立体構造である．酵素タンパク質の詳細情報の中に立体構造が含まれるものは，立体構造データベース PDB のエントリが記述されていので，PDB からデータを取得して，立体表示を行う．

画面上の生化合物を選択することで，図４下の生化合物に関するデータのウインドウが表示される．ウインドウの左側のパネルには，この物質の構造式が表示され，右側に化学名，化学式，関係のある化学反応といった詳細情報が表示される．

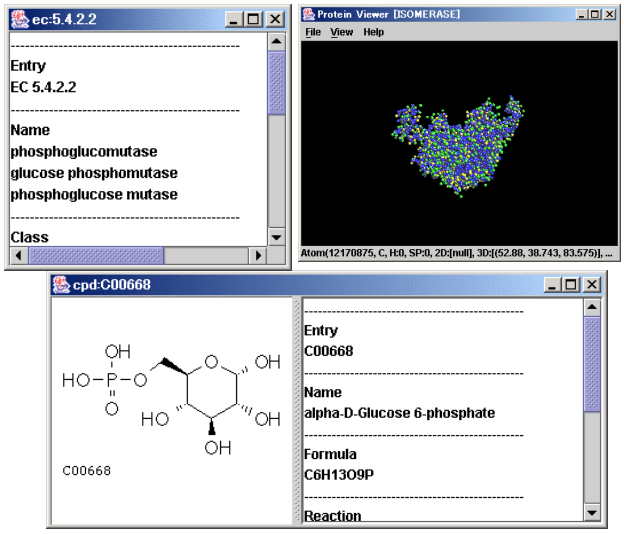

図４　詳細情報の表示

４．おわりに

本システムにより，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った．本システムを用いることにより，例えば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の手助けになりうる．今後は，データを表示するだけでなく，BLAST や FASTA といった解析プログラムとも連動していく．

５．参考文献

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

[2] 池内 俊彦, 畠中 寛: “絵とき タンパク質と遺伝子”, オーム社 (1996)

[3] Ricardo Baeza-Yates and Berthier Ribeiro-Neto, “Modern Information Retrieval”, Addison-Wesley (1999)