ITBL環境でのバイオインフォマティクス関連データベース整備

三菱スペース・ソフトウエア（株）

酒井　　智

1．はじめに

現在、数々の生物種ゲノム配列を読む作業が急速に進められおり、ゲノム配列のデータ量は膨大な数にのぼる。これらのデータすべてを個々の研究機関で管理することは、実質上不可能であるため、これらデータは世界の代表的な機関で一括して管理し、Webにて公開されている。研究を行うために必要なデータは、これらWebサイトからダウンロードすることで入手可能であるが、膨大なデータから必要なデータを探し出す作業は簡単ではない。できるだけ有効に必要なデータを取得し管理していくことが今求められている。

ITBLでは、スーパーコンピュータを複数使用した仮想研究所を構築することにより、数多くの計算を実行するバイオインフォマティクス分野の研究を支えている。ITBLユーザが更なる生物学研究を進めていくためには、ゲノム配列データをITBL環境にて有効に管理していく必要がある。そこで、ITBL利用推進室の業務支援として、バイオインフォマティクス研究で利用される代表的なデータベース（以下、公共データベース）のITBL利用支援環境を構築した。ここでは、ITBL利用支援環境を実現しているシステムについて紹介する。

2．公共データベース

2.1．概要

一般公開されているデータベースのうち、比較的頻繁に使用するデータベースが既にITBL環境へダウンロードされている。そして、これらのデータベースは常に最新状態に保たれるように管理されている。管理対象となっている公共データベースは以下の通りである。

・GenBank
・Swiss-Prot
・PDB(Protein Data Bank)

これらのデータベースは、全世界の研究者が実験によって決定したデータが集約されている。世界各地から集められた膨大なデータはWebにて一般公開され、自由に閲覧することができるため、研究者が独自に決定した塩基配列データをこれらのデータベースと照合させることで、データの特徴を把握することができる。現在、次々と新しいデータが決定されており、データベースのデータ量は刻々と増え続けている。

2.2．対象データベース

2.2.1．GenBank

GenBankはアメリカのNCBI (National Center for Biotechnology Information) が管理している。GenBankには、全世界の研究者が実験によって決定したDNA、及びcDNAの塩基配列データが集約されいる。2004年2月までに登録されたデータ数は約32,549,400

entriesであり、FTP経由でダウンロードすることが可能である。

なお、GenBankは「GenBank/DDBJ/EMBL/国際塩基配列データベース*」を構築している三大国際DNAデータバンクのひとつである。

【参考URL】NCBI HOME：
http://www.ncbi.nlm.nih.gov/

2.2.2. Swiss-Prot

Swiss-Protはスイスの SIB (Swiss Institute for Bioinformatics) とイギリスの EBI (European Bioinformatics Institute) が共同で管理している。Swiss-Protには、全世界の研究者が実験によって決定したタンパク質のデータが集約される。2004年9月までに登録されたデータ数は162,780 entriesである。

【参考URL】Swiss-Prot：
http://www.ebi.ac.uk/swissprot/

2.2.3. PDB (Protein Data Bank)

PDBはアメリカの RCSB (Research Collaboratory for Structural Bioinformatics) consortiumの以下の3機関で管理している。PDBには、全世界の研究者が実験によって決定した生体高分子の立体構造座標に関するデータが集約される。

- Rutgers,the State University of New Jersey
- SDSC (San Diego Supercomputer Center)
- CARB (Center for Advanced Research in Biotechnology)

【参考URL】PDB：
http://www.rcsb.org/pdb/index.html

3．検索

3.1. 概要

各データベースはデータ数が多いため、データベースを直接閲覧して特定のデータを取得することは不可能に近く、システム的支援が必要となる。そこで、データベースからのデータ取得を容易にし、ユーザの利便性を向上させるためのシステムを構築した。このシステムは以下の機能を提供する。

[1] データベース検索
検索用Web画面に入力したキーワードを使用してデータベース検索を行う。

[2] 検索結果表示
[1] の検索で取得したデータのID、及び概要をリストで表示する。

[3] データ詳細表示
[2] の各データの詳細を表示する。表示するデータはユーザが選択する。

[4] データダウンロード
[2]、[3] の画面より、データをユーザ端末へダウンロードする。

ユーザはWeb画面を操作することで、これらの機能を利用することができる。

3.2. Web画面

3.2.1. 検索パラメータ入力画面

検索パラメータ入力画面は、データベースを検索するためのパラメータ入力、及び検索を実行するための画面であり、以下の機能を提供する。

① 検索対象データベースの選択
以下のデータベースから、検索したいデータベースを選択する。

- GBSW
- Swiss-Prot
- GenBank

＊ 世界の研究者が実験によって決定したDNA、あるいはcDNAの塩基配列データをGenBank、DDBJ、EMBLの三大データバンクが、三者間で定めたデータ構築規範に沿って収集・編集し、コンピュータファイルのかたちで提供するもの

● PDB

② 検索キーワードの入力

検索したいデータのキーワードを入力する。空白を区切りとして複数のキーワードを入力することが可能である。なお、キーワードは半角英数字に対応し、それ以外の文字が入力された場合はエラーメッセージを表示する。

③ 検索の実行

選択したデータベース、及び入力した検索キーワードの情報を使用して、検索を開始する。

検索パラメータ入力画面を図１に示す。

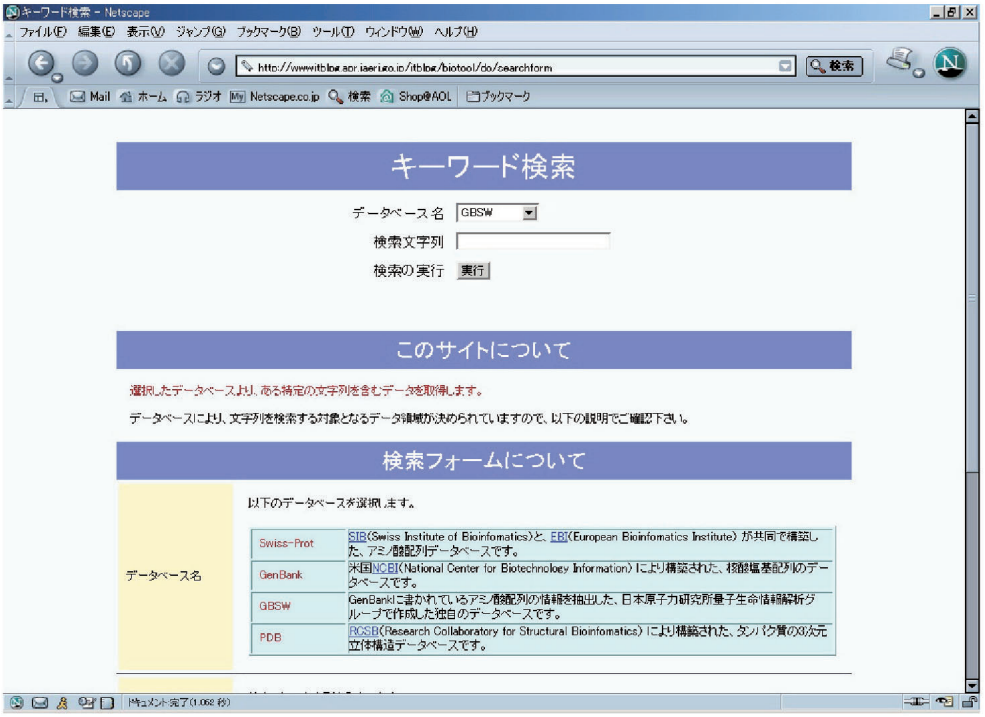

図１　検索パラメータ入力画面

3.2.2. 検索結果表示画面

検索結果表示画面は、検索パラメータ入力画面における検索の実行後に表示され、検索結果データのID、及び概要を一覧表示するための画面であり、以下の機能を提供する。

① データ詳細の表示

対象データの詳細を画面表示する。

② データのダウンロード

チェックボックスをチェックしたデータを、ユーザ端末へダウンロードする。

③ １ページで表示するデータ数の設定

１ページ毎に表示するデータ数を設定する。以下のデータ数を選択できる。（デフォルトは10データ）

- 10データ
- 20データ
- 30データ
- 40データ
- 50データ
- 100データ

④ ページの切り替え

次のページ、あるいは前のページを表示する。

検索結果表示画面を図 2に示す。

図 2　検索結果表示画面

3. 2. 3.　データ詳細表示画面

データ詳細表示画面は、検索結果表示画面である特定のデータIDをクリックした後に表示され、選択したデータの詳細情報を表示するための画面であり、以下の機能を提供する。

①　データのダウンロード

画面に表示しているデータの詳細情報を、ユーザ端末へダウンロードする。

データ詳細表示画面を次頁図 3 に示す。

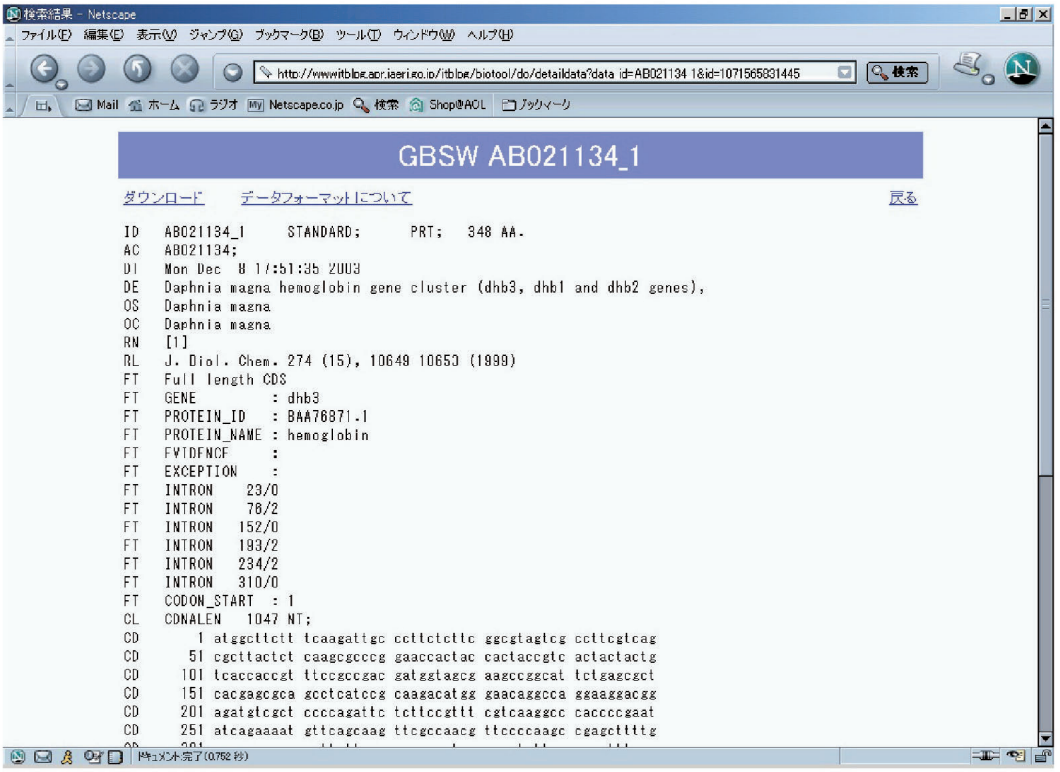

図３　データ詳細表示画面

４．システム構成

4.1. 概要

ITBLのデータベース利用支援システムは、Webサーバ、及びITBLデータベースサーバの2つのサーバで構成されている。定期的に行われるデータベースのアップデートは、ITBLデータサーバ上のデータベースに対して行われ、ユーザはWeb経由で利用可能である。その場合、ユーザはWebサーバへアクセスし、Web画面を操作することでデータを取得する。システムの構成を次頁図４に示す。

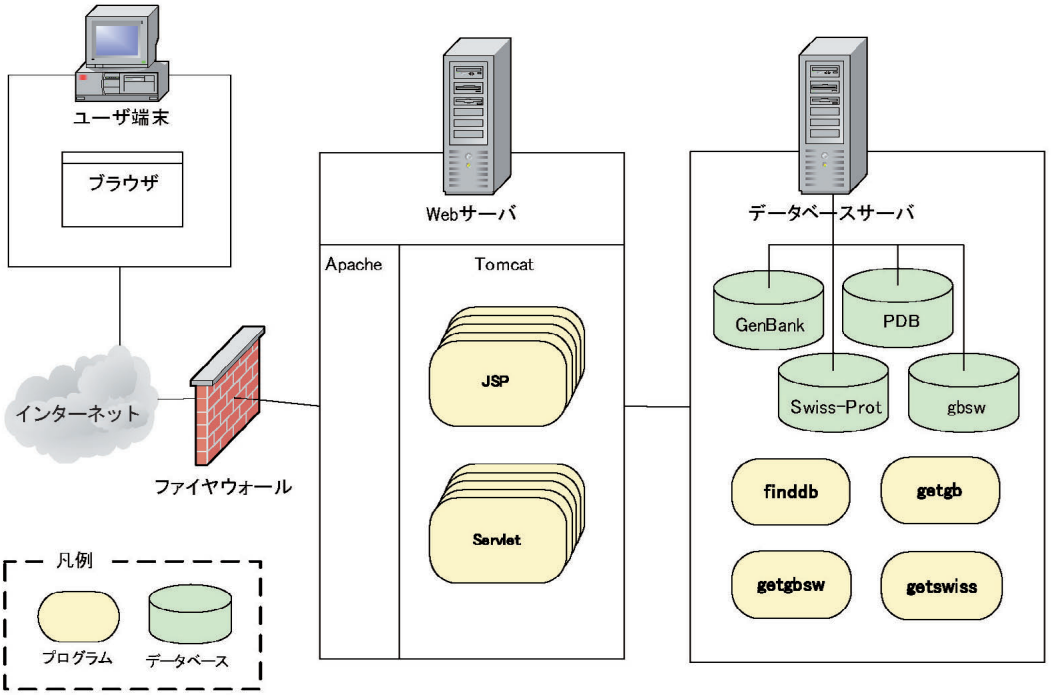

図 4　システム構成図

4.2. マシン構成

本システムを構成するマシンは以下の 2 つである。

① Webサーバ
② データベースサーバ

各マシンの仕様を表 1 に示す。

表 1　マシン仕様

Web サーバ	本体	Sunblade1000
	CPU	UltraSPARCⅢ 750MHz
	メモリ	512MB
	HDD	18.2GB
	OS	Solaris8
	IP アドレス	172.16.167.47
	ホスト名	wwwitblpg
データベースサーバ	本体	PrimePower650
	CPU	SPARC64GP 675MHz × 8
	メモリ	32GB
	HDD	2.2TB(GR730, RAID-5 構成, FC 接続) 3TB(GR730, RAID-5 構成, FC 接続)　[DB 用]
	OS	Solaris8
	IP アドレス	172.16.164.3
	ホスト名	itbltasv

4.3. フリーソフトウエア

各マシンでは、フリーソフトウエアを使用してシステムを構成している。各マシンで使用しているフリーソフトウエアを表2に示す。

表2　フリーソフトウエア一覧

マシン名	フリーソフトウエア	インストール場所
Web サーバ	Apache1.3.27	/usr/local/apache1327
	Tomcat4.1.12	/usr/local/jakarta/jakarta-tomcat-4.1.12-LE-jdk14
	Java2SDK1.4.1	/usr/local/j2sdk1.4.1_01
データベースサーバ	Java2SDK1.4.2	/usr/local/j2sdk1.4.2_01

5．おわりに

ITBLデータサーバへ定期的にダウンロードしているデータベースを研究で利用する場合、データベースへのアクセス手段が必要となる。データベース利用支援環境システムは、データ数が非常に多いバイオインフォマティクス関連のデータベースに対し、効率よく目的のデータを取得するための環境を提供している。このシステムを利用することで、研究者が研究で利用しているデータの詳細を確認することができ、また新たな研究目標を発見することが可能となる。今後、本システムに更なる付加価値を加えることで、今以上にユーザの利便性が向上し、バイオインフォマティクス分野の研究に貢献することが期待される。