2003－MPS－46　(10)
2003／9／19

データグリッド 技術を用いた異種分子生物学データベースの連携手法

細川 卓哉[†]　高坂 貴弘[†]　遠里 由佳子[†]
伊達 進[†]　下條 真司[††]　松田 秀雄[†]

複数の分子生物学データベースを統合的に利用する要求が高まっている．分子生物学データベースはデータ量が膨大である上，その更新頻度が極めて高い．よって，データベース更新の面から，各データベースを物理的に統合することはコストが大きい．また，分子生物学データベースはそれぞれのデータベースごとに全く異なるフォーマットを有するため，それらを単純に結びつけることは容易ではなかった．そこで，本研究では，フォーマットの異種性を排除するためのデータ標準形式を設計し，仮想的に計算機資源を共有するグリッド技術を用いることにより，複数の分子生物学データベースを分散環境に保持したまま，それらを動的に連携させるためのシステムを開発し，その評価を行う．

A fedaration method for molecular biology databases using data grid thchnology

Takuya Hosokawa,[†] Takahiro Kosaka,[†] Yukako Tosato,[†]
Susumu Date,[†] Shinji Shimojo[††] and Hideo Matsuda[†]

The demand to use various molecular biology databases integratively is increasing. Molecular biology databases have enormous amount of data, and they are updated very frequently. Thus, considering database updating, it costs too much to construct their integrated database. Since, molecular biology databases have quite different formats, it is not easy to federate them. Therefore, in this research, we design a standard data format to resolve their heterogeneity. By using grid computing technology, we construct the system for dynamically federating various molecular biology databases, and we evaluate it.

1. はじめに

分子生物学の分野では各種生物の遺伝子やタンパク質の情報を解析する研究が盛んに行われている．それに伴い，タンパク質に関する大量の情報が複数の異なるデータベースに格納されてきている．格納されている情報は，DNA 塩基配列，タンパク質アミノ酸配列，発現パターン，立体構造，機能など多岐に渡る．現状ではこれらの情報は，記述形式の異なる複数のデータベースに断片的に分かれて格納されており，同一タンパク質に関する一連の情報を一度の検索で統一的に得ることはできない．しかし，タンパク質に関するこれら一連のデータは相互に密接に関連しており，単独でその機能を果たすわけではない．よって，タンパク質に関する多様なデータをタンパク質単位で統合して解析することが重要であると考えられる．

従来，分子生物学データベースはそれぞれのデータベースごとに全く異なるフォーマットで表現されていたため，それらを単純に結びつけることは容易ではなかった．そこで，本研究では，各分子生物学データベースに対し，XML をベースとするデータ標準形式を設計し，これを用いることにより，各データベースの円滑な連携を図る．

また，分子生物学データベースはデータ量が膨大である上，その更新頻度が極めて高い．このためデータベース更新の面から，各データベースを物理的に統合し，新しい一つの統合データベースを構築することは，非常にコストが大きい．よって，各データベースはそれぞれ個別に分散環境に保持したまま，検索時にそれらを動的に結合するための手法が必要とされる．

そこで，本研究では，実際に各データベースを分散環境に置き，それらを検索時に動的に結合させるための手法を提案する．ネットワーク上に分散した計算機資源を仮想的に共有するグリッド技術を用いることにより，各分子生物学データベースを個別に分散環境に

† 大阪大学 大学院情報科学研究科
Graduate School of Information Science and Technology, Osaka University
†† 大阪大学 サイバーメディアセンター
Cybermedia Center, Osaka University

保持したまま，それらのデータベース群を検索時に動的に連携させることが可能となり，ユーザは，個々のデータベースを意識することなく，関連するタンパク質に関する一連の情報を一度の検索で統一的に得ることが可能となる．

2. データベース連携の方式

現在，文部科学省 IT プログラムの一貫として大阪大学で行われているバイオグリッドプロジェクト [1] の中で，多数のバイオデータベースをデータグリッド技術を用いて連携するためのシステムが開発されている．システムの全体像を図 1に示す．

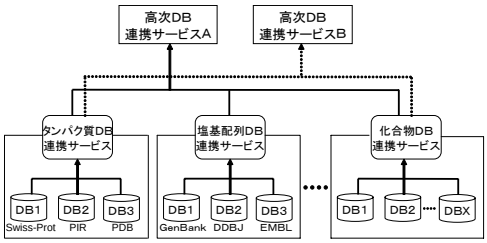

図 1 バイオデータベース連携
Fig. 1 federation of biology databases

このシステムでは，データベース連携を二つのフェーズに分けることにより，効率的なデータベース連携を目指している．膨大な数のバイオデータベース群は，タンパク質に関するデータベース群，DNA 塩基配列に関するデータベース群，化合物に関するデータベース群といったカテゴリーに分類することができる．一つ目のフェーズでは，これらカテゴリー毎の連携サービスを構築する．例えば，タンパク質に関するデータベース群はタンパク質 DB 連携サービスにより一元管理され，これにより，複数のタンパク質関連データベース群を，一つの仮想的なタンパク質統合データベースとして扱うことが可能となる．

次に，二つ目のフェーズでは，カテゴリー毎に構築された仮想統合データベース間の連携を行う．一つ目のフェーズで，カテゴリー毎に仮想的な統合がなされているため，この段階では個々のデータベースを意識する必要はなく，アプリケーション側が必要とするデータに応じて，各連携サービスへアクセスすればよい．例えば，タンパク質と化合物をターゲットとしたアプリケーションは，タンパク質 DB 連携サービスと，化合物 DB 連携サービスを連携する高次の DB 連携サービスを構築して利用する．

本研究では，図 1におけるタンパク質 DB 連携サービス部分に焦点をあて，タンパク質アミノ酸配列データベースである Swiss-Prot[2],PIR[3]，タンパク質立体構造データベースである PDB[4] を対象とし，タンパク質単位での連携検索システムの構築を目指す．

3. XML による各データベース書式の標準化

各分子生物学データベースのフォーマットの異種性を排除するため，XML を用いたデータ標準形式を設計する．

本研究で対象とするデータベース，Swiss-Prot，PIR，PDB それぞれについて，各データベースにおいて同一の内容を持つ部分を，同一のタグで表現する．さらに，各データベースの区別は，各データベースを示す名前空間 (sp,pir,pdb) を付けることにより行う．

本研究で用いた標準形式データについて，その概要を表 1に示す．例えば，この表における生物種は，Swiss-Prot 上では OS，PIR 上では ORGANISM と表されている．しかし，OS，ORGANISM がそれぞれ，生物種を表しており，かつこれらがともに同じ内容を表していることを判別するためには，専門的な知識が不可欠である．連携対象となるデータベースが二つだけであれば大きな問題にはならないが，さらに多くのデータベースを連携させようとする際，それら全てのフォーマットを熟知しておくことは困難である．

本研究で用いる標準形式では，OS，ORGANISM はどちらも同じ内容 (生物種) を表すため，データベースに関わらず，/entry/organism で表される要素で統一的に記述することにする．どのデータベースからの情報であるかは，名前空間から識別する．このように決めることで，各データベース間で共通する情報を対応付けることが出来き，さらにはそれにより，各データベースに固有のデータが何かを明示することが可能となる．

本研究では，各データベースに関して，このような標準形式への変換を行い，変換されたデータに対して検索を行うものとする．

表 1 XML 標準形式
Table 1 XML standard format

格納されるデータ	標準形式データでの位置
ID	entry/id
アクセッション番号	entry/accession
タンパク質名	entry/name
タンパク質別名	entry/alt-name
遺伝子名	entry/gene
生物種	entry/organism
文献情報	entry/reference
機能	entry/function
酵素 EC 番号	entry/EC_num
キーワード	entry/keyword
クロスリファレンス	entry/xref
アミノ酸配列	entry/sequence
配列のフィーチャー	entry/feature

4. 連携検索システムの構築

本研究では，2章で紹介したタンパク質 DB 連携サービスを想定した上で，タンパク質単位での連携検索システムを構築する．

初めに，各データベースに分散した同一のタンパク質に関するデータについて，それらのデータが同一のタンパク質に関するものであることを判断するための基準を定義し，次に，実際に構築する連携検索システムについて説明する．

4.1 同一データの統合基準

本研究では，各分子生物学データベースに分散した同一のタンパク質に関する情報を，一つのタンパク質に関するデータとして一つのエントリへ統合する．よって，各データベースに分散したデータを，同一のタンパク質として統合するための基準が必要となる．本研究では，各データベースが所持するタンパク質アミノ酸配列が全く同一であるものを，同一のタンパク質と判断し，一つのエントリへと統合するものとする．

一般に，タンパク質アミノ酸配列は，'B', 'J', 'O', 'U', 'X', 'Z' を除いた 20 種類のアルファベット数百文字以上からなる文字列で表される．数百文字以上の文字列が，完全に一致することを毎回調べるのはコストが大きい．そこで，本研究では，各データベースの XML 標準形式への変換段階で，タンパク質アミノ酸配列を CRC64 を用いて 16 文字のキーへと符号化したもの (以降 checksum と呼ぶ) を作成し，タンパク質アミノ酸配列と一緒に，標準形式データ中に含めておく．本研究では，このタンパク質アミノ酸配列の checksum が完全に一致するものを，同一のタンパク質であると判断し，一つのエントリへと統合するものとする．

4.2 連携検索システムの構成

連携検索システム全体の構成を図 2に示す．システムは，データサーバ部，連携サーバ部から構成され，データサーバ，連携サーバ間の通信により，各データベースの連携を行う．データサーバ，連携サーバ間の通信には SOAP を用いる．以下それぞれについて説明する．

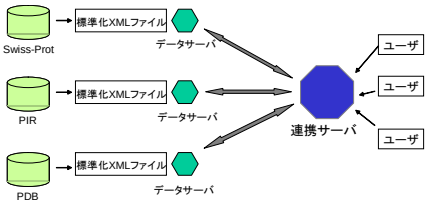

図 2 連携検索システム
Fig. 2 unified search system

4.2.1 データサーバ部

データサーバ部は，各データベースごとに存在し，それぞれのデータサーバは，各データベースの標準形式データを保持する．各データサーバは，連携サーバ部から送られてくる検索要求に対し，それぞれが所持するデータを連携サーバ部へと返す．

4.2.2 連携サーバ部

連携サーバ部は，ユーザからの検索要求に対し，各データサーバへ検索キーワードを送る．それぞれのデータサーバから返されてくるデータに対し，同一タンパク質に関するデータの統合を行う．この統合結果をユーザ側へ返す．ユーザからはこの連携サーバ部しか見えておらず，ユーザは個々のデータベースを意識することなく検索要求を出すことができる．

4.3 処理の流れ

以下に本連携検索システムの処理の流れを示す．

Step 1. : ユーザが検索キーワード (タンパク質名，遺伝子名等) を入力する．
Step 2. : 連携サーバ部は，受け取った検索キーワードを，各データサーバへ送信する．
Step 3. : データサーバ部は，受け取った検索キーワードにヒットするエントリを検索し，そのタンパク質アミノ酸配列の checksum を返す．
Step 4. : 連携サーバ部は，各データサーバから返された全ての checksum から重複を取り除き，残った checksum を再び各データサーバ部へ送信する．
Step 5. : 各データサーバ部は，受け取った checksum に対応するエントリを連携サーバ部へ返す．
Step 6. : 連携サーバ部は，送信した checksum に対して各データサーバ部から返されたエントリを，同一タンパク質に関するデータとして統合する．
Step 7. : 連携サーバ部によって統合されたデータが，ユーザへ検索結果として表示される．

以上の 7 つのステップにより，ユーザ側は，個々のデータベースを意識することなく，複数のデータベースからの情報をタンパク質単位で統合した検索結果を得ることが可能となる．

5. 提案手法の評価

提案手法の有効性を評価するため，従来手法と比べた検索コストを検討する．本研究で検索対象とする各データベースの総エントリ数は，Swiss-Prot(122564 エントリ)，PIR(283308 エントリ)，PDB(20815 エントリ) である．

具体的な例として，ある遺伝子名で検索する場合を考える．本システムを用いない場合，最も効率的に行った場合でも，最低以下のような作業が必要となる．

(1) Swiss-Prot に対して遺伝子名で検索
(2) 検索結果のエントリから，PIR,PDB へのクロ

スリファレンス情報を得る．

(3) PIR へのクロスリファレンス情報から PIR のエントリを得る．

(4) PDB へのクロスリファレンス情報から PDB のエントリを得る．

さらに，このようにして得られた 3 つのエントリを相互に参照する必要があるが，得られたエントリから必要なデータを抽出するには，3 つのデータベース全てについて，詳細に記述形式を知っておく必要がある．さらに，ある遺伝子名に対してヒットするエントリが複数個あれば，その個数分だけ上記の作業を繰り返さなければならない．

これに対し，本提案手法を用いたシステムでは，ユーザは，連携検索システムに遺伝子名を入力するのみでよい．その結果，Swiss-Prot,PIR,PDB 全てのデータベースからその遺伝子名に関する検索結果を集め，それらのうち，同一のタンパク質に関する情報は，一つのエントリにまとめた状態で返してくれる．さらに，返されるデータは，個々のデータベースに依存しない共通のフォーマットをもつため，一つ一つのデータベースのフォーマットの違いを意識しながら，それらを見比べる必要もない．

このように，ユーザは，クロスリファレンス情報をたよりに繰り返し複数のデータベースへアクセスする必要もなく，フォーマットの異なる 3 種類のデータベースからのエントリを，相互に見比べる必要もなくなり，検索に要する手間の面で，大幅な改善を図ることが出来る．

次に，検索に要する時間について検証する．始めに，本検索システムの実装環境を以下に示す．なお，本検索システムは，データサーバ，連携サーバともに，同一のマシン上に構築した．

- OS: Redhat Linux 9
- CPU: Pentium4 (2GHz)
- Memory: 1GB

5 個の遺伝子に対し，それぞれ 10 回の試行を行い，その平均実行時間を算出した (表 2)．

表 2 実行時間 (秒)
Table 2 execution time (sec)

遺伝子名	ヒットするエントリ数	出力を開始するまでの時間	出力が終るまでの時間
p56	2	1	2
HGF	3	1	4
MRP2	8	1	9
ADH2	34	1	32
ADH	65	2	60

表 2より，どの遺伝子名に対しても，検索開始から 3 秒以内に結果が出力されている．また，検索に要する時間は，ヒットするエントリ数に比例している．よって，検索結果のエントリ数が少ないほど，検索に要する時間は減少する．また，検索結果のエントリ数が多い場合，最終的に全てのエントリを出力するまでには時間がかかるが，結果は逐次的に表示されていくため，出力されたものから，順次，結果を見て行くことが出来る．

本システムを用いない場合に要する手間，時間を考えれば，表 2 に示した実行時間で結果が表示されれば，本システムは非常に有効であると考えられる．

6. おわりに

分散環境化にある複数の分子生物学データベースを，データグリッド技術を用いて動的に連携させるための手法を提案した．本システムを用いることにより，複数のデータベースに分散したタンパク質に関連するデータを，一度の検索で統一的に得ることが可能となる．

今後の課題として，さらに多くの分子生物学データベースを，システムに組み込んでいくことが挙げられる．

謝辞 本研究に協力していただいた日立ソフトウェアエンジニアリング (株) の方々に感謝の意を表す．本研究は文部科学省科学技術振興費主要 5 分野の研究開発委託事業の IT プログラム「スーパーコンピュータネットワークの構築」の一環として実施された研究成果の一部である．

参考文献

1) Biogrid Project, http://www.biogrid.jp/
2) Boeckmann B., Bairoch A., Apweiler R, Blatter M.C., Estreicher A., Gasteiger E., Martin M.J., Michoud K., O'Donovan C., Phan I., Pilbout S., Schneider M. : The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003, *Nucleic Acids Research*, Vol.31, No1, pp. 365-370, 2003.
3) Wu C.H., Yeh L.S., Huang H., Arminski L., Castro-Alvear J., Chen Y., Hu Z., Kourtesis P., Ledley R.S., Suzek B.E., Vinayaka C.R., Zhang J., Barker W.C. : The Protein Information Resource, *Nucleic Acids Research*, Vol.31, No1, pp. 345-347, 2003.
4) Westbrook J., Feng Z., Chen L., Yang H., Berman H.M. : The Protein Data Bank and structual genomics, *Nucleic Acids Research*, Vol.31, No1, pp. 489-491, 2003.