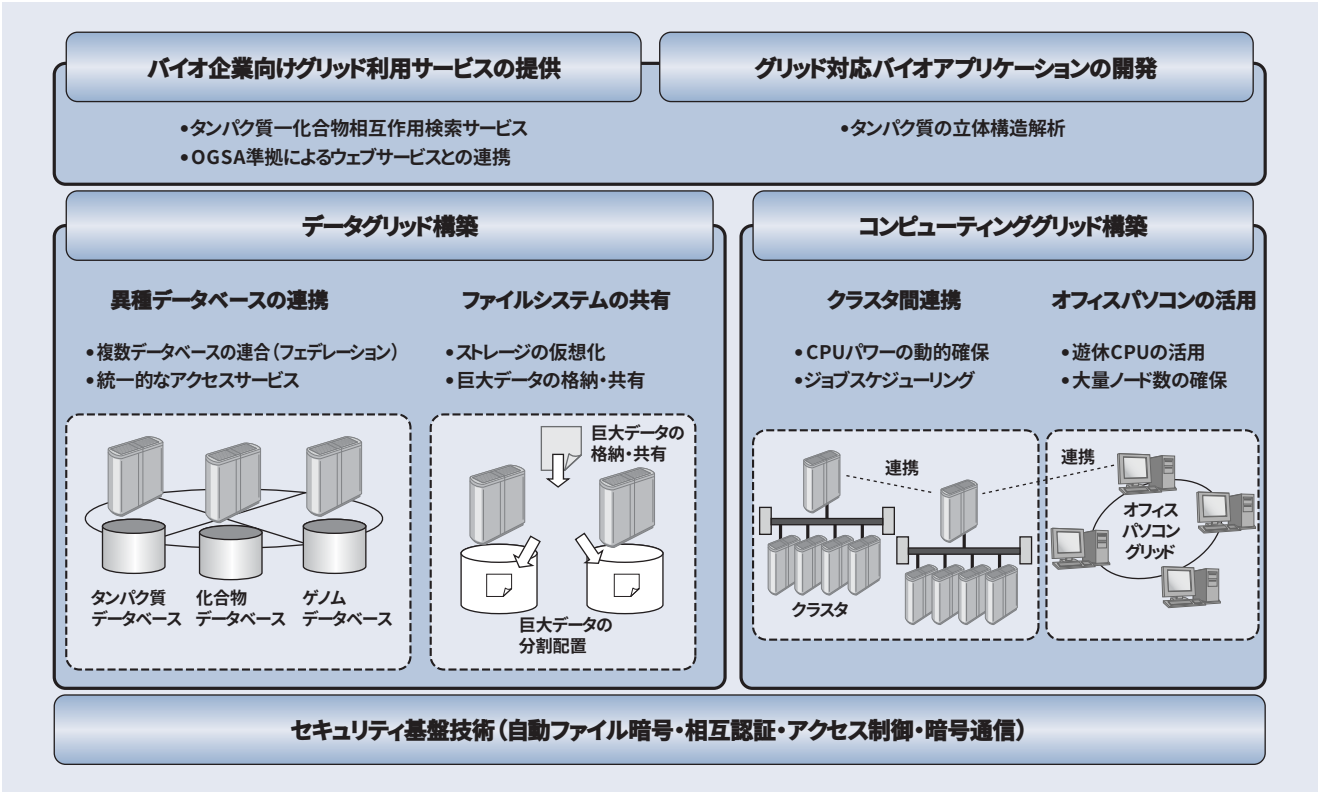
# 先端研究を加速するバイオグリッドコンピューティング

## Bio-Grid Computing for Promoting Advanced Researches

古舘丈裕 *Takehiro Furudate*　　若月謙太郎 *Kentarô Wakatsuki*




注：略語説明　OGSA（(Open Grid Services Architecture；GGF（Global Grid Forum）が策定した，グリッドシステムで提供するサービスの標準を定めた仕様）
CPU（Central Processing Unit）

**日立ソフトウェアエンジニアリング株式会社のバイオ分野向けグリッドソリューションの概要**

高いセキュリティレベルを確保する技術を基盤として，創薬に必要なバイオデータベースを仮想的に統合するデータグリッドと，オフィスパソコンを利用したコンピューティンググリッドによる低コストで高速なバイオ アプリケーション サービスを提供する。

**地理的・組織的な壁を越えて計算機資源を共有するグリッド技術は，科学技術分野を中心に発展し，ビジネス分野への応用が期待されている。日立ソフトウェアエンジニアリング株式会社は，グリッド技術をライフサイエンス分野に有効に適用する手法について大学と共同研究を進めている。大阪大学との事例では，タンパク質，化合物，疾病などの既存のデータベースを仮想的に統合するためのバイオデータグリッドを構築し，創薬のためのタンパク質―化合物間相互作用検索サービスを試作した。北里大学との事例では，Linux※1)クラスタとオフィスの机上Windows※2)パソコンで構成したコンピューティンググリッド上でタンパク質立体構造解析プログラムを稼動するように改編し，コストパフォーマンスの高いタンパク質立体構造解析システムを構築した。これらをセキュリティ製品と組み合わせることで高いセキュリティを確保し，今後バイオ分野でのソリューションとして注力していく。**

※1）Linuxは，Linus Torvaldsの米国およびその他の国における登録商標あるいは商標である。
※2）Windowsは，米国およびその他の国における米国Microsoft Corp.の登録商標である。

## 1 はじめに

ライフサイエンス分野では，タンパク質の立体構造予測，分子動力学シミュレーション，タンパク質と化合物との相互作用検索など，膨大な計算パワーや大規模データの保持，複数のデータベースを横断しての検索が必要とされる。

膨大な計算量を伴う処理を現実的な時間で実行するためには，高性能な計算設備が不可欠であるものの，予算やスペースなどの都合で必ずしも導入できるとは限らない。そのため，既存の計算資源の有効利用という観点から，計算機ごとの負荷の平坦(たん)化や，机上パソコンの活用が求められる。

一方，データ共有に関しては，同一データの再生成や重複保持の回避，テラバイト級のデータの共有といったストレージリソースの効率化と，既存データベースの統合利用が求められている。バイオ分野では，公開されているものだけで500種類以上のデータベースがあり，分子，ゲノム，タンパク質，臓器，疾病などの情報を互いに関連づけて活用する要求がある。しかし，データベース間の整合性がないため，研究者がそれぞれデータを収集して，関連するものを突き合わせているのが現状である。スムーズな協業体制を確立し，先端研究を加速するためには，このような問題の解決が不可欠である。

このような課題にこたえるため，日立ソフトウェアエンジニアリング株式会社(以下，日立ソフトと言う。)は，グリッド(格子)技術をライフサイエンス分野に有効に適用する手法について大学と共同研究を進めている。

ここでは，日立ソフトが共同研究を通して取り組んできたグリッド事例について述べる。

## 2 ライフサイエンスでのグリッドの利用

近年，科学技術分野を中心として，地理的・組織的な壁を越えて計算機資源を共有するグリッド技術が注目されている。グリッドということばはパワーグリッド(電力網)に由来しており，電力の供給源を意識することなく電気が使えるのと同じように，コンピュータをネットワークに接続するだけで遠隔地にある計算資源を利用できるようにする環境の実現をコンセプトとしている。

グリッドの利用目的による分類として，データの共有を主眼とするデータグリッド，CPU(Central Processing Unit)の計算パワーを共有するコンピューティンググリッド，実験装置の遠隔操作やテレビ会議など，リモートデバイスのリアルタイム利用に着目したアクセスグリッドなどがあげられる。ライフサイエンス分野では，大規模計算や大規模データへの需要が高く，大学などの研究機関を中心として，比較的早くからグリッド技術の研究・活用への取り組みが行われてきた。日立ソフトは，大学との共同研究を通して，データグリッドとコンピューティンググリッドの構築技術と，グリッド上で高いセキュリティ技術を付加したサービスの研究に取り組んでいる。

## 3 日立ソフトのグリッド構築事例

### 3.1 バイオデータグリッドの構築

バイオ グリッド プロジェクトは，文部科学省の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，大阪大学を中心として，2002年から始められたものである[1)]。プロジェクトには，日立ソフトが参加するデータ グリッド グループを含め，グリッド基盤技術，コンピューティンググリッド，およびデータオンライン解析の四つの研究グループがある。

データ グリッド グループは，創薬をターゲットとした高度情報検索，データベースの異種性の解消，データグリッド基盤システムの開発を研究テーマとしている。創薬のプロセスでは，タンパク質と化合物の膨大な組み合わせの中から，特定のタンパク質と相互作用を持つ化合物を検索する必要がある。

そのため，500以上の公開データベースと一部の商用デー

**表1 対象データベース**

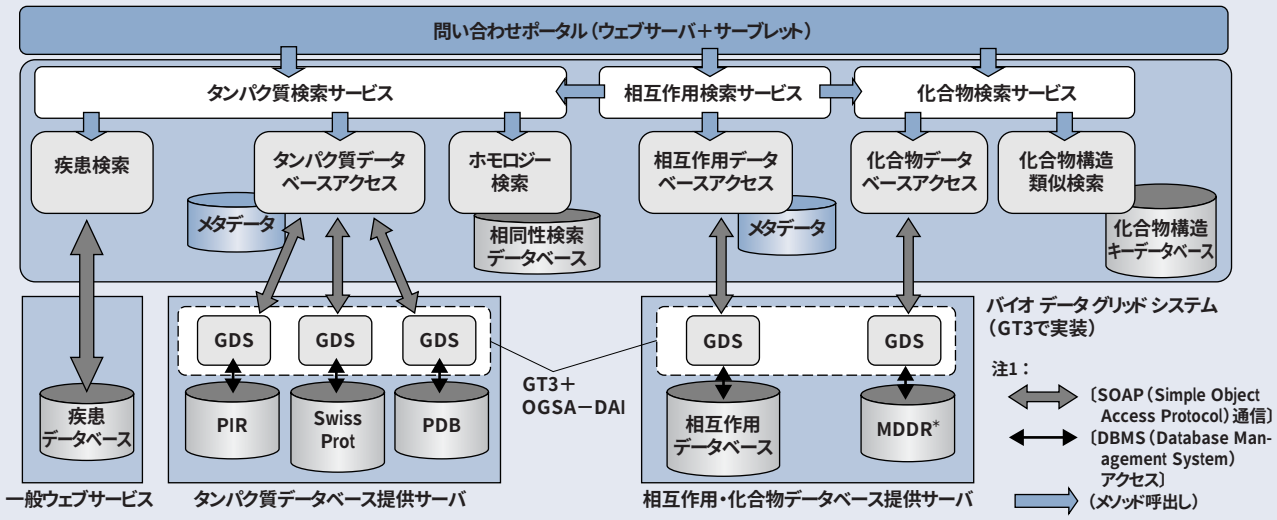
創薬に必要となる11種類のデータベースを利用する。現状では複製したデータにメタ情報を追加して利用している。

| 分野 | 対象 | データベース名 |
|---|---|---|
| 医学 | 疾患名 | Medical Encyclopedia |
| | | LITerature DB |
| 薬学・化学 | 化合物 | MDL Drug Data Report(商用) |
| 生命科学 | タンパク質 | Swiss Prot |
| | | PIR |
| | | PDB |
| | ゲノム | DDBJ |
| | 相互作用 | ENZYME |
| | | GPCR-DB |
| | | NucleaRDB |
| | | LGIC DB |

**表2 提供サービス**

単独または複数のデータベースを対象とした，よく利用される検索機能をウェブサービスとして提供する。

| 提供サービス | 内　容 |
|---|---|
| タンパク質検索サービス | ●Swiss Prot IDからの検索<br>●名称からの検索<br>●配列からのホモロジー検索<br>●疾患名からの検索 |
| 化合物検索サービス | ●名称からの検索<br>●構造類似検索(予定) |
| タンパク質－化合物相互作用検索サービス | ●指定タンパク質に相互作用する化合物の検索<br>●指定化合物に相互作用するタンパク質の検索 |




注2：略語説明ほか　GT3（Globus Toolkit 3；Globus Allianceが提供するグリッド構築ツールキットの第3版），
OGSA−DAI（Open Grid Services Architecture−Data Access and Integration；GT3とともに動作し，DBMSの差異を吸収するデータ アクセス サービスを提供する。），
GDS（Grid Data Service；OGSA−DAIの一部で，DBMSに直接アクセスするモジュール），MDDR（MDL Drug Data Report）
＊MDDRは，米国MDL社の開発薬品化合物データベースである。

**図1 バイオ データ グリッドのシステム構成**
タンパク質検索などの各サービスはGT3上で動作するウェブサービスとして実装され，SOAPを用いてデータベースサーバと通信する。各データベースに含まれるデータのカテゴリーと所在情報をメタデータとして保持する。

タベースの中から，創薬に必要となる11種類のデータベース（**表1**参照）を仮想的に統合するバイオ データ グリッドと，その上で動作するアプリケーションとして，相互作用検索サービスを試作した。このシステムで提供するサービスを**表2**に示す。

バイオ データ グリッドの構築には，GT3（Globus Toolkit 3）を使用した。GT3はGlobus Alliance[2]が提供する，グリッドミドルウェアを構築するためのツールキットであり，GGF（Global Grid Forum）で策定されたOGSA（Open Grid Services Architecture）に準拠する参照実装である。OGSAでは，ウェブサービスをベースにグリッドシステムに必要なサービスが規定されている。そのため，バイオ データ グリッドはウェブサービスとの親和性が高く，既存のウェブサービスとの連携が可能である。

また，データベース管理システムごとの差異を吸収するために，OGSA−DAI（OGSA−Data Access and Integration）を使用している。OGSA−DAIは英国e-Scienceプロジェクトで構築されたツールであり，GT3と組み合わせて利用することができる。

今回対象とした11種のデータベースを関連づけるメタデータを作成し，これらをまたがって検索を行える環境を構築した。バイオ データ グリッド システムのソフトウェア構成を**図1**に示す。ウェブサーバ上のサーブレットからバイオ データ グリッドのサービスを呼び出し，各サービスでは複数のデータベースを参照して問い合わせ結果を返す。システムは，対象データの種別や各データベースの所在情報をメタデータとして管理し，自動的に適切なデータベースへ問い合わせを発行する。あるいは，外部で運営されている一般的なウェブサービス（ここでは疾病データベースサービス）を利用する。

タンパク質−化合物相互作用検索の典型的な問い合わせでは，サーブレットと各サービス・データベース間とのデータ転送量は40 kバイト程度であり，ブラウザ上でのクリックから10〜30秒程度で問い合わせ結果を表示できている。

これらの技術を応用し，今後はバイオ グリッド プロジェクトのテストベッド環境上でサービスを公開していく予定である。

## 3.2 タンパク質立体構造解析システムのグリッド化

コンピューティンググリッドに関しての事例について以下に述べる。

日立ソフトは，北里大学薬学部と協業体制を築き，同大学が開発したタンパク質立体構造解析プログラム“FAMS（Full Automatic Modeling System）”[3]を対象として，遊休パソコンを活用した低コストのコンピューティンググリッド上で性能を確保する手法の研究に取り組んでいる。オリジナルのFAMSはLinuxクラスタ上で動作するプログラムであり，ホモロジーモデリングにより，アミノ酸配列だけがわかっているタンパク質の立体構造を予測する。プログラムは並列に実行されることから，性能を確保するためには十分なノード数を用意する必要がある。

しかし，一般的な企業でクラスタを利用する場合，スペースや保守費用，あるいはセキュリティ上の理由により，ノードの追加や他の組織のクラスタとの連携が制約されることが想定される。そのため，この事例では，企業や研究所のオフィス内で多数使用されている事務用・研究用パソコンを活用する手段を採用した。これらのパソコンのCPUの空き時間を利用

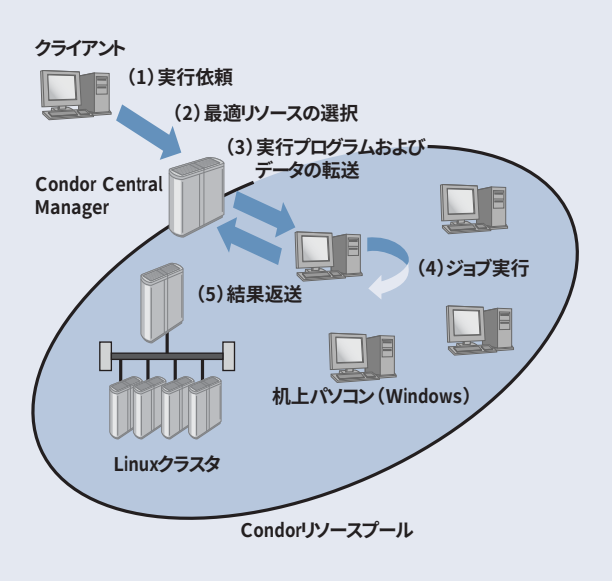



**図2 LinuxとWindowsの混合リソースから成るグリッド例**
クラスタと机上パソコン全体を計算リソースとして，FAMS（Full Automatic Modeling System）のジョブを並列実行する。

することで，低コストでCPUリソースを確保できると考える。

現状では，オフィスパソコンのOS（Operating System）としてはWindowsが採用されている場合が圧倒的に多いため，FAMSをWindowsに移植し，LinuxとWindowsの混合リソースから成るグリッドを構築した（図2参照）。

グリッドミドルウェアとしては，米国のウィスコンシン大学が開発したCondor[4]を利用した。Condorは，LinuxとWindowsの両方に対応していること，簡易的なジョブスケジューラを持っていること，さらに，ジョブを実行しているマシンで本来の利用者が作業を始めると，実行中のジョブを中止する機能を持っていることから採用した。この機能により，本来の利用者の業務に支障を来さずに，CPUの空き時間だけを利用することができる。

Condorを通してWindowsパソコン上で動作するようにFAMSを改編し，日立ソフトのライフサイエンスセンタ内の50台のオフィス用パソコンを用いてグリッドを構築した。FAMSで必要とされるデータは数十ギガバイトにもなることから，このシステムでは，割り当てたジョブが必要とする部分のデータを逐一転送している。また，ジョブの実行時間が長くなると利用者の復帰によってキャンセルされる確率が高くなるため，ジョブをできるだけ細分化している。Linuxクラスタ上でNFS（Network File System）によってデータを共有している場合と比較して，データ転送とジョブの中断が発生する分だけ，性能の低下は避けられない。しかし，オフィスに数多く存在するパソコンを利用して台数を確保することで，全体的な実行時間の縮小が期待できる。実験の結果では，100台程度のノード数が得られれば，50台程度のクラスタと同等の性能が得られることがわかった。

今後は，さらに大規模なグリッドとした場合の検証を進め，最も効果的な運用ノウハウを蓄積し，コストパフォーマンスの高いグリッドサービスの提供に結び付けていく考えである。

## 4 おわりに

ここでは，日立ソフトが大学との共同研究を通して取り組んできたバイオグリッドコンピューティングについて述べた。

これらを実際にビジネスに展開していくためには，特にセキュリティ面で幾つかの課題をクリアしていく必要がある。例えば，製薬企業では，どのデータベースを検索したかという情報でさえも重要な情報であるため，ログが取られていないことを保証する仕組みが必要である。また，遺伝子データなど，データ自体が個人情報を含んでいる場合は，その取り扱いには細心の注意を払わなければならない。

日立ソフトは，既存のセキュリティ製品で培ってきたノウハウを生かして，付加価値の高いグリッド構築ソリューションの提案に努めていく考えである。

なお，データグリッドの研究の一部は，文部科学省科学技術振興費研究開発委託事業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」によるものであり，ご助力いただいた大阪大学松田秀雄教授をはじめとする研究グループと，コンピューティンググリッドの研究にご協力いただいた北里大学梅山秀明教授をはじめとする関係各位に感謝の意を表する次第である。

### 参考文献など

1) 伊達進，外：バイオグリッドプロジェクト「スーパーコンピュータネットワークの構築」，情報処理，Vol. 44，No. 6（2003.6）
2) The Globus Alliance, http://www.globus.org/
3) Homology Modeling Service のホームページ http://www.pharm.kitasato-u.ac.jp/fams/
4) Condor Project, http://www.cs.wisc.edu/condor/

### 執筆者紹介

**古舘丈裕**
1991年日立ソフトウェアエンジニアリング株式会社入社，ライフサイエンス本部 バイオインフォマティクス開発部所属
現在，グリッド関連システムの開発に従事
情報処理学会会員
E-mail：furudate @ hitachisoft. jp

**若月謙太郎**
2002年日立ソフトウェアエンジニアリング株式会社入社，ライフサイエンス本部 バイオインフォマティクス開発部所属
現在，グリッド関連システムの開発に従事
E-mail：kwakatsu @ hitachisoft. jp