報道発表資料

2004 年 1 月 15 日
独立行政法人 理化学研究所

グリッド技術の相同性検索サービスを運用開始

- 計算資源の統合的利用による高速化 -

独立行政法人理化学研究所（野依良治理事長）は、バイオインフォマティクスを推進することを目的とした OBIGrid※1 において、複数の会員組織（表 1）が運営する計算資源を、グリッド※2 技術によって統合構築した遺伝子・タンパク質相同性検索システム「GridBlast」を開発しました。理研ゲノム科学総合研究センター（和田昭允センター所長）ゲノム情報科学研究グループ（小長谷明彦プロジェクトディレクター）の小西史一研究員らにより設計・構築されたものです。

この「GridBlast」は、市販のパーソナルコンピュータ（Celeron 1300MHz1CPU 1GB MEM）を使用して約 88 日間かかる処理を、3 つの組織（理研ゲノム科学総合研究センター、日本電気株式会社、東京工業大学）が所有する計算機（合計 230CPU）を統合することにより、約 8 時間で 253 倍（GridBlast 内の複数の Celeron の処理速度の平均と比較して）の処理速度を実現させました。このようなグリッド技術により、高速な大量相同性検索を願う研究者に提供できるとともに、サイト提供組織が増えることが期待できます。また、ライフサイエンス分野におけるポストシーケンス時代の大量情報処理に対応するシステム構築をする上での貴重な機会となります。

「GridBlast」は、ウェブサイト（http://www.obigrid.org）で 1 月 16 日より運用実験を開始し、利用ユーザの参加登録をすることで「GridBlast」が利用可能となります。また同時に「GridBlast」の検索サイトの登録も募集します。

1. GridBlast とは

GridBlast は異なるネットワーク構成、異なるジョブ管理システム、異なる計算性能を持つ計算機群を、グリッドシステム構築ミドルウェア※3「Globus Toolkit※4」を用いて統合したもので、単一システムイメージとして NCBI(National Center for Biotechnology Information)の BLAST※5 による相同性検索サービスを提供することができます。相同性検索とは、配列の類似性に基づいて、配列が似ていれば機能も同じであろうという推測により、遺伝子やタンパク質の機能を予想する方法です。グリッド技術を利用することにより、利用者は高速化に必要な並列処理、分散処理を全く意識することなく数千から数万個の配列の検索を行うことができます。

GridBlast は現在、理研ゲノム科学総合研究センター、日本電気株式会社、東京工業大学の PC クラスタ計算機※6 を用いて運用されています。また、相同性検索実行に利用される BLAST には、日本電気の製品である HomologySearcher(PC クラスタ用にチューニングされた並列型 BLAST)をベースに、株式会社 NEC 情報システムズが Grid 対応を行ったプログラムを用いています。

2. 研究成果

GridBlast では、複数利用者からの相同性検索要求を、利用可能なサイトの処理

能力に応じて、自動的に分割するなどの操作を行い、全サイトでの検索完了時間が揃うように検索クエリーのデータサイズ調整をしています。クエリーとは、処理要求(問い合わせ)を文字列として表したものです。例えば、ある遺伝子の DNA 配列情報を入力すると利用可能サイトの処理能力毎に DNA 配列情報の量を自動的に配分調整します。このように適切に配分されたエントリーセット（遺伝子の場合はその 1 サイトに投入された DNA 配列情報の量）による実行結果をまとめて、相同性検索の結果として提供します。複数のサイトからの結果の回収と結果の再構成による処理時間を必要としますが、複数の組織が所有する、これらの計算システムを 1 つのアプリケーションとして透過的に利用できることが、グリッド技術により構築されたアプリケーションのメリットとなります。

サイト数の増加に伴う効果を確認するために、1 サイトあたりの利用 CPU（Central Processing Unit：中央演算装置）数を 16 とし、サイトの数を 1 から 12 まで変化させた場合、1 サイトで実行した結果は 4.43 時間でしたが、12 サイトを利用して合計 192CPU を利用した場合の結果は 0.55 時間となりました。これは、1 サイトを利用した際の処理時間の約 8 分の 1 になります。時間当たりの処理能力は、1 サイトあたりの場合 223 エントリーでしたが、12 サイトを利用した場合は、1541 エントリーとなり約 7 倍の処理速度の向上となりました。（図 1）更に、大規模な計算に対するベンチマーク（処理速度を計測する試験）として、FANTOM※7 2.1 のアミノ酸配列(29,941 エントリー)の重複のない（non-redundant）アミノ酸に対し、繰り返し回数を 3 回として PSI-BLAST※8 を実行した結果、8 時間 19 分で完了しました。これは、1 時間当たり、3084 エントリーの処理能力があることを示します。ベンチマークを実行した際の、総 CPU 数は 230 であり、5 サイトの異なる PC クラスタ計算機システムに対して処理された結果です。

また、GridBlast には、サイトの計算資源の不調による実行の失敗時にエントリーの再投入などのフォールト・トレランス機能（障害対策機能）も実装しています。

3. 利用について

利用方法としてはウェブサイトを用意しており、オリジナルの NCBI-BLAST と操作性は変わらず、これまで使っていたオプションはほとんど利用可能となります。また、複数のユーザからの利用もサポートしており、実行中の状況などは、ユーザごとに管理され、過去に実行した結果は 1 週間保存されています。さらに実行終了時にメールによる通知機能も備えているため、長時間の検索に関しても便利です。また、このような利用時のウェブコンテンツは SSL（Secure Sockets Layer：通信内容を暗号化するための通信手順)によって暗号化されるため、安全に利用することができます。

対象ユーザは大規模の相同性検索を必要とする研究者層で、エントリー数として数千～数万クエリーを想定しています。利用可能なデータベースは、重複のない（non-redundant）核酸塩基配列データベースや、アミノ酸配列データベース等を配備しており、目的に応じて様々種類のデータベースを準備することができ、各サイトで最新でかつ同一のデータベースを利用することができる体制が取られています。

具体的には、現在 OBIGrid のウェブサイト(http://www.obigrid.org)の上で、OBIGrid 利用ユーザ参加登録すると、GridBlast を利用することができます。また、GridBlast の検索サイトも同時に登録できます。

4. 研究意義と今後の展開

近年、遺伝子やタンパク質の配列データは爆発的な増加傾向にあり、従来からの手法による検索では、その処理時間が長大化するという問題が顕在化しています。そこで、PC クラスタ計算機などを使った並列処理が問題の解決方法のひとつとして注目されてきましたが、1 組織が所有する計算資源には限界がありました。グリッド技術は、このような組織が所有する計算資源を集積して問題を解決することができ、バイオインフォマティクスやライフサイエンス研究のさらなる発展に貢献するものです。

今後は、GridBlast の運用試験を通してその品質を高め、培った技術を他のバイオインフォマティクスツールのグリッド化を視野にいれた研究開発を推進します。

（問い合わせ先）
独立行政法人理化学研究所 横浜研究所
ゲノム科学総合研究センター
ゲノム情報科学研究グループ
ゲノム解析用コンピュータ研究開発チーム
研究員　　小西　史一
Tel ： 045-503-9602 / Fax ： 045-503-9613

（報道担当）
独立行政法人理化学研究所 広報室
Tel ： 048-467-9272 / Fax ： 048-467-4715
Mail ： koho@riken.jp

<補足説明>

※1 OBI(Open BioInformatics) Grid

文部科学省科研特定領域研究ゲノム情報科学の「ソフトウェア高速化および共有化委員会と企業コンソーシアムである並列情報処理イニシアティブ（IPAB : http://www.ipab.org/）」が母体として開発中のグリッドであり、国内バイオインフォマティクス関連の大学・研究機関・企業の 27 サイトが参加し、インターネット上に VPN 装置によって構築された国内最大規模のグリッドである。

※2 グリッド

ネットワーク環境下の計算機やストレージ等の資源や情報を、所有する組織を超えて、安全に・安定して・情報サービスを享受できる基盤技術。この技術により、異機種間接続環境が可能となり、大規模な仮想計算機をひとつのシステムとして利用することができる。

※3 ミドルウェア

オペレーションシステム（OS）上で動作し、アプリケーションソフトに対して OS よりも高度で具体的な機能を提供するソフトウェア。OS とアプリケーションソフトの中間的な性格を持っている。

※4 Globus Toolkit

米国アルゴンヌ国立研究所と南カリフォルニア大学の共同開発による、グリッドシステムの構築を容易にするためのツールキットである。現在グリッドシステムを構築する際にデファクトスタンダードとして広く普及し、世界中のグリッドプロジェクトで採用されている。開発者は、Globus で用意されている関数を呼び出すことで、グリッドアプリケーションを構築する。最新版は Globus Toolkit 3.0.2 であり、ウェブページ（http://www.globus.org/）からダウンロードすることができる。

※5 BLAST

核酸やタンパク質の相同性検索を行うツールとして、米国の NCBI（National Center for Biotechnology Information）で開発され、公開されているアプリケーション。

※6 PC クラスタ計算機

クラスタとはデータ通信において末端制御装置とそれに接続されている末端の総称で、その末端装置の計算機のこと。

※7 FANTOM

理研ゲノム科学総合研究センター遺伝子構造・機能研究グループの主導により結成された、マウス遺伝子の機能解析を行う組織（FANTOM コンソーシアム;国内外のゲノム科学、生物学などの専門家）が共同で、機能アノテーション情報を付与したマウス cDNA クローン。に 2002 年 4 月 29 日から 5 月 5 日にかけて行われた FANTOM2（Functional ANnoTation Of Mouse） Cherry Blossom 国際会議では 60,770 個のマウス完全長 cDNA クローンが解析され、その成果は 2002 年 12 月の「Nature」に発表された。

※8 PSI-BLAST

Position Specific Iterative BLAST と呼ばれ、アミノ酸配列の類似性を最初に実行した検索結果もとに自動的作成する反復処理を行なうことで、重み行列を変化させることで、類似の配列の検出感度を高めた BLAST である。

http://www.ncbi.nlm.nih.gov/BLAST/

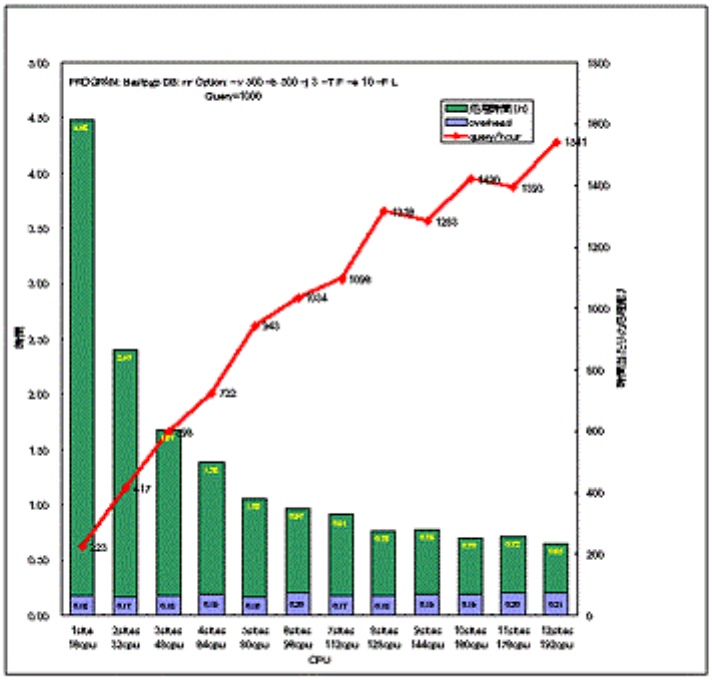
OBIGridの会員組織

接続中サイト	接続準備サイト
理研ゲノム科学総合研究センター	インテック・ウェブ・アンド・ゲノム・インフォマテックス株式会社
北陸先端科学技術大学院大学	CTCラボラトリーシステムズ株式会社
徳島大学	株式会社ベストシステムズ
東京工業大学大学院	
同志社大学知識情報センター	
三井情報開発株式会社	
日本ヒューレット・パッカード株式会社	
九州大学大学院	
大阪産業大学	
和歌山大学	
株式会社富士総合研究所	
九州大学大学院農学研究院	
国立遺伝学研究所生命情報・DDBJ研究センター	
住商エレクトロニクス株式会社	
統計数理研究所	
西日本電信電話株式会社	
琉球大学	
日本アイ・ビー・エム株式会社	
日本電気株式会社	
独立行政法人科学技術振興機構	
株式会社三菱総合研究所	
日本新薬株式会社	
東京医科歯科大学生命情報学	
東京工業大学	
東北大学学際科学国際高等研究センター	

以上の機関（順不同）による２７サイト

（表１）

OBIGrid内のPCクラスタ群を使い、最大１２個までのジョブ分割による処理能力の向上を示したものである。
実験条件formatdbによって処理された1,581,064個の重複のないアミノ酸配列データベース(nr)に対して、
1000個のPSI-Blast検索を実行した結果である。



（図１）

実験条件　Mycoplasma pneumoniae peptide sequence (278 letters)
blastpgp -d nr -v 500 -b 500 -j 3 -T F -e 10 -F L
Database has an All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF