K13 デスクトップPCを用いたグリッドコンピューティングによるバーチャルスクリーニングシステム構築及び運用実績報告

（NTT データ[1]、アクセルリス[2]）○黒須雅隆[1]、Hongwei Huang[2]、盛崇[2]

１．はじめに

オフィスや家庭にあるような一般的な PC の余剰リソースを有効活用する PC グリッドコンピューティング（以下単に PC グリッドとする）技術は 1999 年に始まった SETI@home[1)]により一般的に知られるようになった。PC グリッドとは High Performance Computer のネットワーク連結を、一般のデスクトップ PC などに転用することによって、構築・運用面において低コストでスパコン並みの性能を実現可能にしたものである。

現在この技術を用いて世界的に大小様々な 30 以上のプロジェクトが米国を中心にボランティアベースで行われている。中でも、Folding@home[2)]、Intel-United Devices Cancer Research Project[3)]のように、近年タンパク質の構造が急速に明らかになりつつあるバイオ分野でのプロジェクト実績が蓄積されつつあり、特に米国で発生した炭素菌テロに対応する国家的なプロジェクトとして 2002 年に Cancer Research Project が一時的に標的をがんから炭素菌に変更して運営されたのは記憶に新しい。

NTT データでは、この PC グリッドによる国内初のインターネット実証実験[4)]を 2003 年に終え、大規模分散コンピューティングシステムの開発・運用ノウハウを蓄積した。

このノウハウを用いて、近年その性能が重要視されている分子ドッキング手法のシミュレーションソフトを持つ Accelrys 社と大規模バーチャルスクリーニングシステムの共同開発を行い、理化学研究所　ゲノム化学総合研究センター　タンパク質構造・機能研究グループ　計算プロテオミクス研究チーム[5)]の協力を得て、PC グリッド及び分子ドッキング手法を基盤とした *in silico* 手法の実証を目的とする「医薬品の作用・副作用の分子メカニズム解析」プロジェクトを実施した。

今回は、このプロジェクトから得られた、グリッドシステムの性能、運用課題について報告する。

２．プロジェクトの概要

疾患治療に用いられるほとんどの薬品の直接的な働きは、一定の種類のタンパク質（群）に作用してその働きを調整することである。ところが、現在広く用いられている既存の医薬品の中には、どのタンパク質にどの程度の強さで作用しているかがよく把握されていないものが多く存在する。最近は特定のタンパク質を標的として設計された医薬品も増えてきているが、その場合でも、人体中に存在するそれ以外の膨大な種類のタンパク質のそれぞれに対してどのような作用を及ぼすかが実際に確認されているわけではない。いずれの場合も、薬品の直接的な作用メカニズム、つまりどのタンパク質の機能をどのように調整しているかは十分に解明されていないのが現状である。

本プロジェクトでは、計算プロテオミクス研究チーム　松尾チームリーダーにより選別された人体への影響が大きいと想定されるタンパク質

kurosum@nttdata.co.jp

と既存薬品との作用を PC グリッドと分子ドッキング法を基盤とした *in silico* 手法を用いて網羅的に評価し、タンパク質と薬品の作用・副作用のマトリクスを作成することとした。（図１）

（本プロジェクトのタンパク質及び既存薬品の選出理由、及び解析結果については、「プロテオーム規模での *in silico* 親和性フィンガープリンティング」を参照）[6)]

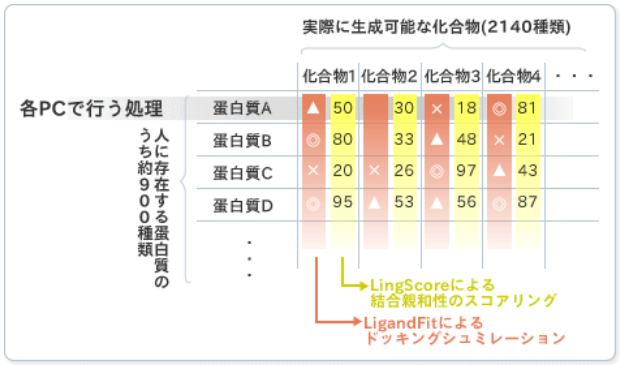

図１．マトリクスイメージ図

３．プロジェクトの検証環境

ドッキングを行うタンパク質とリガンドは、MDDR データベースより選出した既存薬剤 2140 種と NCBI の nrPDB から抽出したヒトタンパク質 561 個、ヒトプロテインキナーゼのホモロジーモデル 282 個を使用した。

分子ドッキングシミュレーションには、Accelrys 社の LigandFit[7)]を使用した。LigandFit とは、タンパクとリガンドの親和性を高速に行うソフトウェアである。3 次元構造化されたタンパク及びリガンドのデータを用いて、結合部位の自動定義→高速ドッキング→親和性評価までを行う。今回のプロジェクトのように大量のリガンドをドッキング／スコアリングするバーチャルスクリーニングのほか、個別のタンパク質―リガンドのドッキング状況の解析などにも使用される。

PC グリッドのミドルウェアには、NTT データの cell computing®[8)]を使用した。

cell computing®は、ジョブの受信から処理結果の返信までを自動で行うクライアントソフトをインストールしたデスクトップ PC とこれらを管理するセンタサーバによって構成される。また、TCP-IP を利用した通信と冗長配信やタイムアウト設定などのジョブスケジューリング機能によって、インターネットを利用した大規模な PC グリッド環境を手軽に構築できる点に特徴がある。

今回 PC グリッドのクライアントとして、セキュアなネットワークで接続された大学、企業で構成されるデスクトップ PC400 台を使用した。（図２）

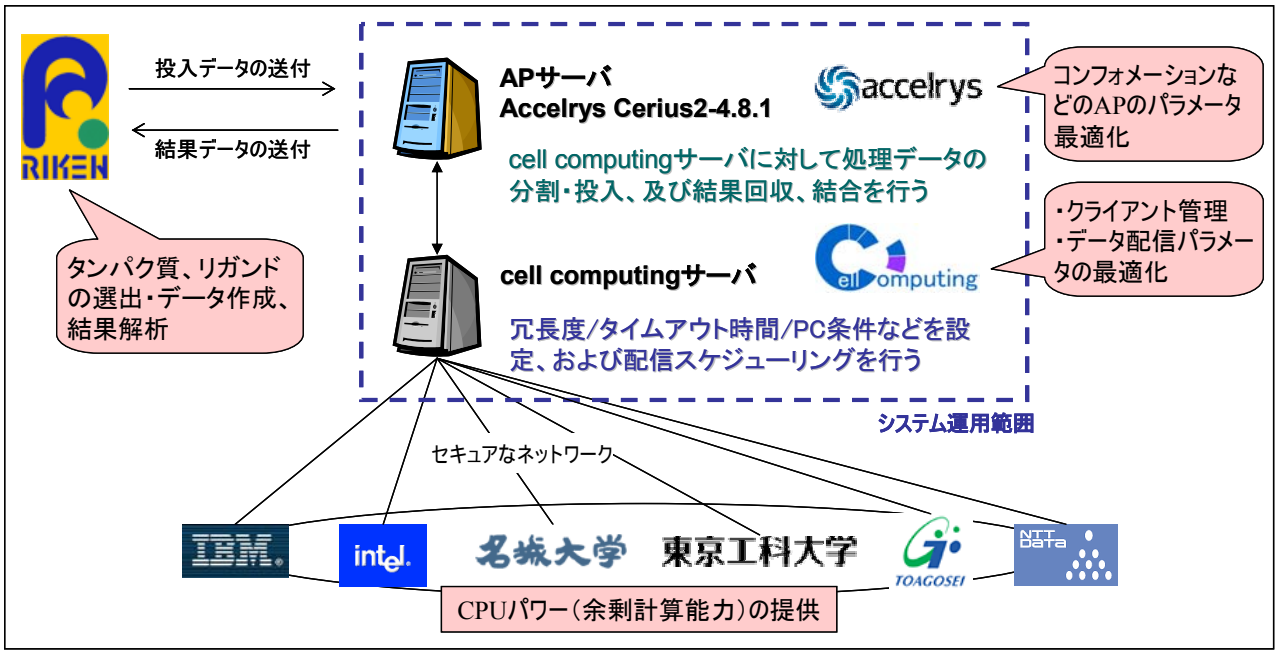

図２．実験の担当範囲及びシステム構成

４．運用結果

解析は実験開始の 2003 年 8 月 1 日から 2004 年 1 月 23 日、176 日間で行った。

解析を行うクライアント PC の性能は、CPU クロック数にして 200MHz のものから 3GHz までと広範囲に渡った。（図３）

また、大学、企業の諸事情により、提供可能なクライアント PC の稼働時間が限られていたため、平均稼働台数を算出したところ一日約 150 台となった。

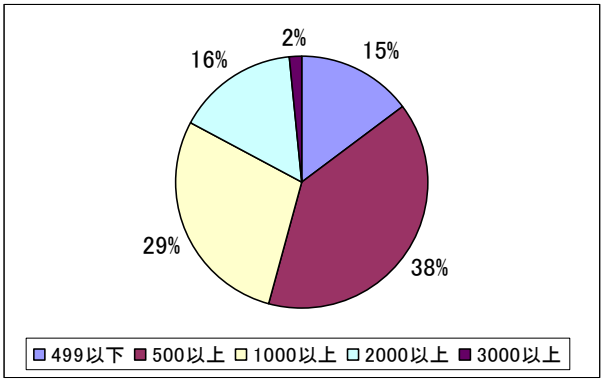

図３．クライアントの CPU クロック数割合

PC グリッドシステム全体の処理性能は、最高性能 350GFlops 相当、平均では 107GFlops 相当と開きがでたが、これは実験中に高性能なサーバ群を持つ大学の計算機センターの一時的な参加があったことが原因としてあげられる。

期間中に解析したタンパク質の総数は、ヒトタンパク質(561 個)+ヒトプロテインキナーゼのホモロジーモデル(142 個)の計 703 個(ドッキング回数：1504420 回)となった。（ただし、タンパク質の結合部位の再設定や薬剤に対して大きすぎるタンパク質については結合部位を分割したため実行数は 828 個分となったので実際のドッキング総回数は 1771920 回となった。）なお、このデータを解析するのにかかった CPU 時間（デスクトップ PC1 台で処理を実行した場合の計算時間）は 27 年 218 日 13 時間 39 分 48 秒であった。

５．運用に関する考察

<不安定な環境での計算効率化について>

各クライアント PC のジョブ処理時間をガントチャートで表示したところ、クライアント PC が接続されているネットワークの速度や CPU 能力によって処理時間に大幅な差が生じていることが分かった。（図４）

特に LigandFit は CPU 負荷が高く、クロック数が 400Mhz のマシンと 2Ghz のマシンとでは処理時間に 5 倍の差があり、結果として処理能力が低いマシンによってジョブ全体の処理時間が引き延ばされていることが分かった。

このため、全クライアント PC の半数を占める 800Mhz 以下のマシンを対象に、LigandFit を配信せず軽量な LigandScore のみを配信する設定を行ったところ、設定前と比較して 1 つのジョブに対して 30%の処理速度の効率化が見られた。

また、このように PC 性能のばらつきによる影響が大きかったことから、大学の計算機センターのように高性能かつ均質なマシンを多数所有している場合は、PC グリッドを用いることで容易に高い処理能力を実現できることがわかった。

<クライアントの維持管理について>

本プロジェクト完了後、計算資源提供者に対しアンケートを実施した。アンケート結果の全体の傾向として、システム面におけるクレームは少なく、リソース提供に対するインセンティブやプロジェクト進捗のフィードバックに対する要望が多かった。また、このような要望はトップダウンで導入されているグループでは少なかったことから、そうでないグループではリソース提供に関するコンセンサスを得てシステムが導入された後も、リソース提供者に対するモチベーション維持の仕組みが必要であることがわかった。

<大規模ドッキングについて>

短期間にタンパク質の結合部位特定作業を大量にこなす必要があり、結合部位の大幅な指定に

よるヒットミス（リガンドがタンパク質にヒットしない状態）が多数発生したことから、結合部位の見直しによる処理データの再投入が行われた。また、当然ではあるがタンパクとリガンドの形状が一定ではないため、1 ドッキングあたりの処理時間の予測が困難であり、最適なスケジューリングを行う上で問題となった。

6．まとめ

本プロジェクトが実施された背景には、1）PC グリッド技術の一般化、2）分子ドッキングシミュレーション技術の進歩、3）タンパク質の構造解析が進み、1 つの薬剤に対する複数のタンパク質の影響を解析できる背景が整ったことという 3 つの要素がある。

PC グリッドについては単純並列処理よりさらに幅広い課題に対応するためのバイオ分野でよく用いられている MPI を用いた処理方式の検討や、分子ドッキング手法については *in vitro* との結果の整合性確認といったように、それぞれの分野においてまだ検討課題は残されているが、今後さらに多くのタンパク質の構造が明らかになるにつれ、今回のように手軽に大規模計算が行える計算基盤へのニーズが増えていくことが予想される。

そのようなニーズに対して、PC グリッドを利用したバーチャルスクリーニングは、短期間かつ安価に手段を提供することができることが今回のプロジェクトで実証されたといえよう。

謝辞

実験におけるタンパク質及びリガンドの選定に協力頂いた理化学研究所　ゲノム化学総合研究センター　タンパク質構造・機能研究グループ計算プロテオミクス研究チーム　松尾洋リーダーと佐藤氏、及びクライアント提供に協力頂いた名城大学、東京工科大学に深謝する。また、同様に協力頂いた日本アイ・ビー・エム株式会社、インテル株式会社、東亜合成株式会社に感謝する。

References and Notes

[1]http://setiathome.ssl.berkeley.edu/

[2] http://folding.stanford.edu/

[3] http://www.grid.org/projects/cancer/

[4] http://www.cellcomputing.jp/example/test3.html

[5] http://www.riken.jp/r-world/research/lab/genome/protein/index.html

[6] Y.Matsuo Accelrys Japan User Forum 2004

[7] http://www.accelrys.com/cerius2/c2ligandfit.html

[8] http://www.cellcomputing.co.jp

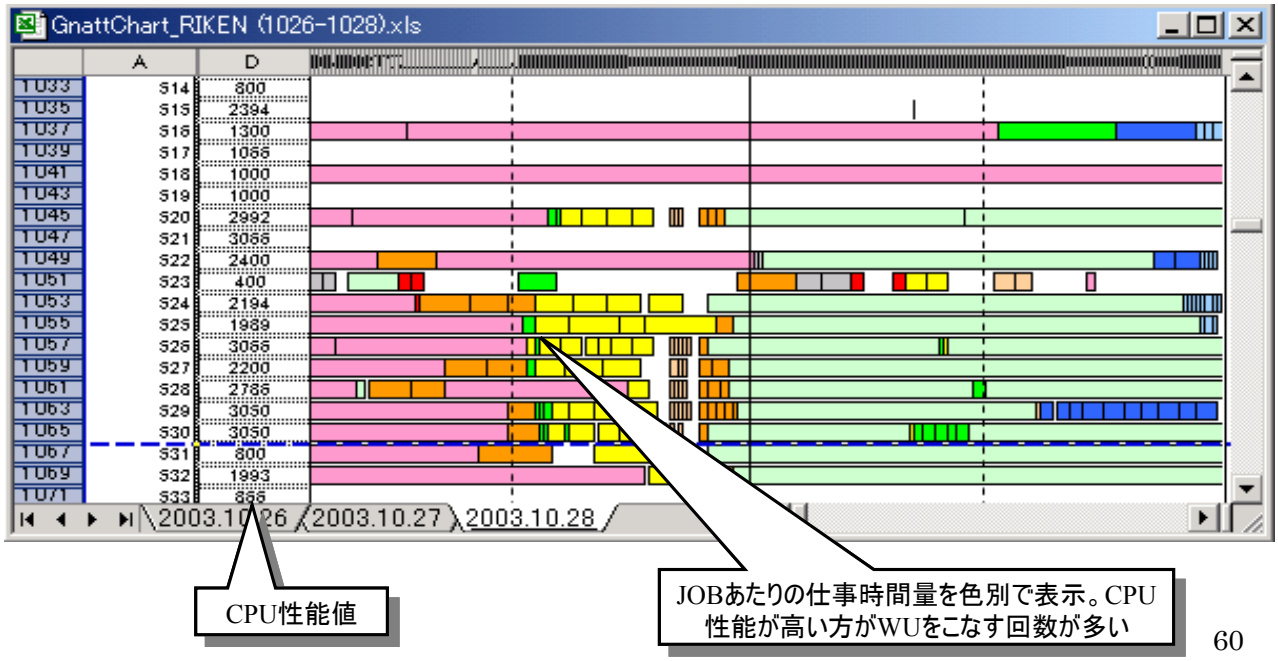

図４．クライアント PC 毎のジョブ処理時間