データ解析融合研究

宮野 信（東京大学医学研究所ヒトゲノム解析センター）

次世代生命体統合シミュレーションソフトウェアの研究開発におけるデータ解析融合研究開発チームの開発している技術は、飛躍的に増大しているゲノムや遺伝子発現データをベタスケールで解析する応用技術で、データ同化（data assimilation）と呼ばれる技術によりシミュレーションモデルと観測データの融合を図るモデル構築技術です。データ同化は、シミュレーションモデルと現実データとの差額を埋める技術として、地球物理的分野で発展してきています。地球規模の複雑な現象の高精度予測のために、時空間観測・計算データと最先端の大規模なシミュレーションモデルを統合し、適切な初期値・境界値やパラメータ等を実際の現象を反映するように決定する技術です。これからの技術は、ベタスケールの計算能力がlimitに達して実現できるもので、今後の生命科学・医学研究において極めて重要な技術になると考えています。飛躍的に増大するシーケンシャルデータや、DNAチップをはじめとするバイオシステム計画で得られる様々なデータを使い、データ同化技術によるデータ融合や現実のデータ解析及びシミュレーション技術を開発し、ベタスケールデータ解析とシミュレーションの融合に基づいて、個人差を考慮した医療データ最適化、ビジネスなどの開発、大規模タグ付ネットワークや遺伝子ネットワーク推定による創薬ターゲット・毒性関与リスクの選別解析等の研究を進めています。

このチームは、「大規模遺伝子ネットワーク推定とその応用」（東京大学医学研究所・宮野信）、「大規模ゲノム多型及び表型型データを関連付ける新しいアレルゲノムの開発と、妥当性、有用性的検証」（理化学研究所ゲノム医療センター・鎌倉他）、「生命体系シミュレーションのためのデータ同化技術の開発」（統計数理研究所・徳川他）に関して、「タンパク質間相互作用ネットワークの推定とその応用に関する研究」（東京工業大学理工学研究科・秋山他）の4つの研究から構成され、「肺製剤化学」という共通の視点をもって研究を進めています。以下のようなソフトウェアを開発しています。

1. ヒトの全遺伝子・転写産物を対象としたネットワーク解析を可能にする大規模遺伝子ネットワーク推定ソフトウェア
2. PPIチャレンジ：1000×1000の超大規模計算を可能にする網羅的なタンパク質間相互作用推定ソフトウェア（タンパク質ドッキング解析プログラム）
3. 1人あたり50万SNPを用い、疾患と遺伝子の関連を見発見（GWAS、薬物反応性を予測するプログラム）
4. 「個」のデータを「一般」のモデルに合理的にフィットさせるシステムのためのデータ同化を手がけたプログラム
5. 以上を統合し、統合的に利用活用をするためのソフトウェア

なぜベタFLOPSの計算が必要とされるのか、単純な計算パワーや一例から見ると、非常に困難に直面しているかを簡単に紹介します。遺伝子多型研究では、単純な50万SNPの遺伝子多型の数千年規模の計算は100ギガ―TテラFLOPSで計算できます。あるグループを50万個のSNP遺伝子型データをもとに計算するプログラムがあります。この構造が決定されないと表現型とゲノム多様性の検出や、パラメータの推定が正確でないと考えていますが、これには5000人レベルの対象を必要とすると、ベタFLOPSが必要となります。次に、データ同化の現状は、地球物理関係のモデルを工夫を凝らしたモデルで、パラメータ数を2000倍に落とし、100万次元、粒子数1500、計算時間、順位数100、1時間の計算を1.3テラFLOPSのコンピュータで1日かかってやっています。ベタスケールだとこれが1秒で可能になり、この技術をネットワーク推定や神経シミュレーションなどで、個人化に応用できるようになります。また、生物の時系列データのように時系列が少ない場合、10〜100の粒子を用いたパラメータの推定に達成することが可能になります。さらに、ネットワークという新たなインタラクションの地図を作り上げつつ、新たな困難が生じます。ネットワークの構築において、構築で実現の難しいネットワークを（10万種の構築、1%に当たる）1000種に縮小するとしても、1000×1000のネットワーク推定計算の計算に至る計算量は1000ギガ・計算であります。BlueGene（1/250）で50年かかります。これによりベタFLOPSだと2〜3月で計算できます。また、数値の遺伝子ノックダウンと薬剤応答のマイクロアレイ解析データから、遺伝子を1000遺伝子（ヒト全遺伝子の5%程度）に収め込んでベタFLOPSネットワークを推定する計算は、128CPUのPCクラスタで数週間かかっています。今後、大規模に出てくる遺伝子発現データを対象に、数万から10万の遺伝子転写産物をネットワークとして解析することは、計算パワーや手法の点で不可能になっています。そのため、新たなネットワーク解析の手法の開発とベタFLOPSの計算能力が必要になっています。

今後、数値データから数ベタFLOPSのスパコンが全世界に濃いとき、登場してきます。東京大学医学研究所ヒトゲノム解析センターでも、2009年1月からヒト型ヒトゲノムの解析を始めるため、スパークスの計算で、スパークスの計算で、10ベタFLOPSの計算を手がけています。