共同利用（産業利用トライアルユース）
先端研究施設共用促進事業『みんなのスパコン』TSUBAME によるペタスケールへの飛躍
成果報告書　平成 19 年度戦略分野利用推進「計算化学手法による創薬技術開発」 i07qb

タンパク質一次構造の網羅的解析による創薬技術の開発
Development of a new drug discovery approach by comprehensive protein sequence analysis

金澤 光洋
Mitsuhiro KANAZAWA

ライフィクス株式会社
Reifycs Inc.
http://www.reifycs.com/

データベースとして存在するタンパク質のアミノ酸配列（ペプチド配列）の理論質量と、その配列が質量分析された場合のフラグメント質量の実測値と比較することで、タンパク質の一次構造の解析が幅広く行われている。しかしながらタンパク質のような巨大分子の一次構造解析においては、分析装置から得られるデータ量が膨大な上、そこから考えられる一次構造の計算が複雑かつ多様であるがゆえに、実験時の消化エラーや検討すべきタンパク質の翻訳後修飾の考慮が演算量ゆえに十分できず、解析精度を犠牲にすることも少なくない。そこで本プロジェクトでは、現実時間でより綿密な解析を行う為の高速なタンパク質一次構造計算手法の開発を目的とする。

Protein sequence analysis method using mass spectrometry is widely used for biological and medicinal researches, and that calculation method is based on the comparison between theoretical molecular weight calculated by protein sequence database and observed molecular weight of peptide fragment. It is, however, difficult in sequence analysis of large molecule such as protein, and digestion error and post translational modification are not well-cared because of huge computational effort. In this project, we develop faster calculation method for protein sequence analysis to regard further in-depth discovery.

Keywords: Protein, Peptide, Sequencing, MIS (MS/MS Ion Search), Mass spectrometry

背景と目的

本プロジェクトでは、ヒトゲノムより遺伝子領域が予測されているタンパク質のアミノ酸配列（ペプチド配列）の理論質量と、その配列が質量分析された場合のフラグメント質量の実測値と比較することで、既知のタンパク質データベースのアミノ酸配列に限定されないタンパク質の一次構造を高速に解析する方法を確立する。

DNA が翻訳されることで作られるタンパク質は生体内で多種多様な機能を果たし、疾患の解明や医薬品開発においてその構造解析は必要不可欠となっている。しかしながら、タンパク質のような巨大分子の一次構造解析においては、分析装置から得られるデータ量が膨大な上、そこから考えられる一次構造の計算が複雑かつ多様であるがゆえに、従来では多くの解析時間が必要であっただけでなく、実験時における消化エラーや検討すべきタンパ

ク質の翻訳後修飾の考慮が十分出来ず、解析精度を犠牲にすることも多く見られた。そこで、一次構造決定に必要な数日、数週間を必要とする複雑な演算に TSUBAME を用いることで 10 倍から 100 倍に加速し、従来では多大な計算時間ゆえに考慮が難しかった消化エラーや翻訳後修飾といった計算パラメーターを同時に考慮することで、高速・高精度にタンパク質の一次構造計算を行う手法の開発を提案する。

概要

弊社のタンパク質一次構造解析ソフトウェア Reifycs ProteomicSuite™ ver. 1.2 の並列化を行い高速な解析が行える TSUBAME においてそのプログラムを実行することで、従来では現実時間での検討が困難であったパラメーターによって実行し綿密な解析を実現する。Reifycs ProteomicSuite™ は Microsoft Windows 上で動作するプログラムであるゆえ、Mono (http://mono-project.com/)を TSUBAME 上に導入し解析を実施する。

結果および考察

本プロジェクトでは 1) Reifycs Proteomic-Suite™ の並列化を行うこと、2) 並列化されたプログラムを TSUBAME に移植し綿密な解析パラメーターによる解析を実行し実用化に結びつけることの 2 つを段階的に実施した。

1) に関しては、プログラムの並列化作業を行い、Table 1 にしめした条件によってタンデム質量分析器による MS/MS スペクトラムを対象にした MIS (MS/MS イオンサーチ) を用いて Intel® Core™ 2 T5600 1.83GHz 上の Windows Vista Business 32bit にて非並列化および、並列化プログラムのベンチマークを実施した。

Table 1 ベンチマークにおける 解析の条件

The # of scans	16453
Fragment Charge	1+, 2+ and 3+
Ion type	b-ion, y-ion
Enzyme	Trypsin
Fixed Modification	Carbamidomethyl (Cysteine)
Variable Modification	Oxidation (Methionine)
MS tolerance	- 1.00 Da to + 1.20 Da
MS/MS tolerance	- 0.50 to + 0.60
Database	SwissProt rel. 56.0 Homo Sapience
Missed Clearvages	2

CPU コアが 2 つを用いた並列化によって解析速度が約 61.0%向上したことがこの比較にて得られた（Figure 1）が、これは並列化のみならず、データベースと入力スペクトラムを相互比較する為のインデックスの分散による入出力の低減も解析速度向上に起因しているものと考察する。

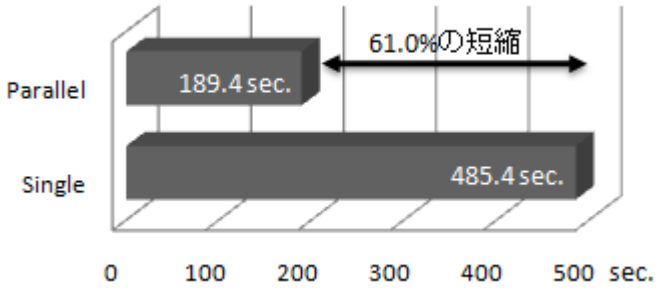

Figure 1 並列化前後の解析時間の比較

1) の並列化処理を TSUBAME に適用することで従来は検討が難しかった綿密な解析パラメーターを適用可能なプログラム実行を実現することを目的とした 2) を実施すべく、Reifycs ProteomicSuite™ の LINUX 環境下での実行テストを開始した。Mono を用いた本プログラムの実行時間は、Windows 上の処理における解析時間と比べて並列、非並列共に 60 倍以上遅く、実用的な解析時間で結果が得られなかった。この解析速度低下の原因を調査すべく Mono を用いた実行時の状況を確認したが、メモリ上の入出力処理時間の増大をはじめとした複数の問題が関連していることが想定さ

れたため、具体的な解決法を検討することを断念し解析プログラム自体のの再構築を行うこととした。平成 21 年 1 月より平成 21 年 10 月まで LINUX 環境において実行できるよう Mono を利用しないプログラムの再構築を行ったが本プロジェクト終了時においても LINUX 上で動作するプログラムは未だ完成に至っていない。

まとめ

上述の通り LINUX 環境上で処理が行えるプログラム開発が終了していない為、本プロジェクトの継続において先ずはプログラム開発の継続が不可欠である。一方、弊社製品プログラムの利用者は Windows をベースとした処理系を望んでいることから、開発継続の価値を同時に検討する必要が出ている。

実際、弊社では本課題の目的であるより綿密な解析を実施するべく、これまではデータベースを用いていた解析を予め解析を行うタンパク質を絞り込んで解析を実施させることで解析処理自身を縮小するプログラムを開発した。これにより、本課題において想定していたタンパク質のペプチド消化条件やタンパク質中のアミノ酸に結合している修飾分子候補を網羅的に検討できるような解析を実現した。とはいっても、タンパク質が絞り込めない段階での解析には意味をなさないゆえ、本課題の検討は引き続き行う必要があると考えている。