

課題名 (タイトル) :

大規模ゲノム多型データと表現型データを関連付ける新規アルゴリズムの開発と、妥当性、有用性の検討

利用者氏名 : 三澤 計治

所属 : 和光研究所 次世代計算科学研究開発プログラム
次世代生命体統合シミュレーション研究推進グループ データ解析融合研究開発チーム

1. 本課題の研究の背景、目的、関係するプロジェクトとの関係

人間は一人一人の人がそれぞれ自分だけの個性的な遺伝子を持っていて、人によって病気のかかりやすさやかかりにくさが違います。このことを逆に利用して、ある病気、たとえば糖尿病や乳がんなどの病気にかかった人を何千人・何万人と集め、病気にならなかった人をまた何千人・何万人とあつめて、その遺伝子を比較すると、病気にかかった人たちが共通に持つ遺伝子が見えてきます。このような遺伝子を疾患関連遺伝子と言います。このような遺伝子は、病気を引き起こす原因である可能性が高く、それを調べることで、病気の原因の解明が進み、さらには病気の治療法や治療薬の開発へとつなげることができます。このような研究を関連解析と言います。関連解析により、糖尿病、乳がん、リウマチ、肥満、高血圧など次々と疾患関連遺伝子が見つかってきています。

遺伝子の一人分のセットをゲノムというのですが、ヒトのゲノムは DNA 分子で 30 億塩基を両親から受け継いだ合計 60 億塩基からなっています。ゲノム全部に関連解析をおこなうことをゲノムワイド関連解析と言いますが、これは 60 億塩基×数万人の大量データに関して解析を行うので、非常に時間がかかります。しかし、患者さん達は一刻も早い治療を必要としています。

そこで、並列コンピューティング技術を使い、将来のスーパーコンピュータへの移植も視野に入れた上で、PC クラスタ上で関連解析を高速に実行できるプログラムを開発しています。今まで見つからなかった疾患関連遺伝子を見つけたためには、今までよりも大量のデータを使うことが必要であり、また、高速に実行できるプログラムは必須となります。

2. 具体的な利用内容、計算方法

今年度発表した論文の中にあるプログラム、paraHaplo に関して、RICC および FX への移植を行うためのコンパイル作業を行った。論文とプログラムは <http://www.scfbm.org/content/4/1/7>
<http://sourceforge.jp/projects/parallelgwas/releases/39695> に公開されている。

3. 結果

RICC 上での paraHaplo のコンパイルに成功した。また FX 上で、非並列化版のコンパイルに成功した。並列化版のコンパイルはまだ成功していない。

4. まとめ

RICC ならびに FX での paraHaplo のコンパイルに成功し、スパコン移植への足掛かりをつかんだ。

5. 今後の計画・展望

FX 上での paraHaplo の非並列化版のコンパイルも成功させたい。

6. RICC の継続利用を希望の場合は、これまで利用した状況 (どの程度研究が進んだか、研究においてどこまで計算出来て、何が出来ていないか) や、継続して利用する際に行う具体的な内容

FX ならびに RICC との互換性を確認しながら、手元のマシンでチューニングを行い性能を上げていきたい。

7. 利用研究成果が無かった場合の理由

RICC ならびに FX を使い始めたのが平成 22 年 2 月からだったため

