課題名（タイトル）：

ParaHaplo: A program package for haplotype-based whole-genome association study using parallel computing

利用者氏名： ○三澤　計治　　佐久間　俊広　　滝中　徹
坂上　和也　　斎藤　実
所属　　： 社会知創成事業　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

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

近年の DNA 配列決定技術の進歩により，大量のゲノム配列が解析されるようになりました．ゲノム配列は，個人の間に違いがあります．これを利用し，数万人に対して，ゲノム全体に渡り，患者群とコントロール群との遺伝子頻度を統計的に解析することによって，疾患関連遺伝子を発見する手法がゲノムワイド関連解析です．大人数のゲノムデータを解析するためには，高速計算が不可欠です．そこで我々は，京速コンピュータ「京」の上で，ゲノムワイド関連解析のためのプログラム ParaHaplo を開発しています. ParaHaplo は大量のゲノムデータを分割し，ユニット間・コア間のハイブリッド並列により，高速計算を実現しました。この高並列化をテストするために RICC を利用させてもらいました。

１．　具体的な利用内容、計算方法

RICC 上でコンパイルし、利用するＣＰＵ数を変えながら速度を測定し、並列度を測定しました。

２．　結果

高並列でも十分な並列性能が出ることが確認されました

３．　まとめ

ParaHaplo は

４．　今後の計画・展望

京速コンピュータ「京」での利用を目指し、今後は RICC 上でさらなるチューニングを行い高速化を実現する予定です。

５．　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

京速コンピュータ「京」での ParaHaplo の利用ではデータ入出力がボトルネックになると予想され、それに対しステージング機能を用いたデータ入出力の分割・並列化を予定していますが、今まではデータ入出力が再現できなかったので、それを踏まえた利用をさせてもらいたいと思います

６．　一般利用で演算時間を使い切れなかった理由

７．　利用研究成果が無かった場合の理由

平成 22 年度　RICC 利用研究成果リスト

【論文、学会報告・雑誌などの論文発表】

- Misawa K, Kamatani N (2010) ParaHaplo 2.0: a program package for haplotype-estimation and haplotype-based whole-genome association study using parallel computing. Source Code Biol Med 5:5
- Misawa K, Kikuno RF (2010) GeneWaltz--A new method for reducing the false positives of gene finding. BioData Min 3:6
- Misawa K, Kikuno RF (2011) Relationship between amino acid composition and gene expression in the mouse genome. BMC Research Notes 4:20

【国際会議などの予稿集、proceeding】

【国際会議、学会などでの口頭発表】

アメリカ人類遺伝学会　Washington DC. USA

A program package for haplotype-estimation and haplotype-based whole genome association study using parallel computing. K. Misawa, N. Kamatani.

【その他】