平成 27 年度　利用報告書

課題名（タイトル）：

ヒト全 SNP を網羅する E-probe の設計

利用者氏名：　〇須永　泰弘

所属：　情報基盤センター　計算工学応用開発ユニット

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

現在、ヒトゲノム上には約 1 億 3 千万箇所の一塩基多型(SNP)があり、データベースに登録されている。これまでに理化学研究所では、SNP を高い精度での検出を可能する「PCR Eprobe Melting（PEM）法」という手法を開発した。PEM 法には、SNP ごとに短い人工核酸(プライマーとプローブ)が必要になるが、この SNP ごとのプライマーとプローブのセットを検討することが研究をする上で手間となっていた。昨年、私は PEM 法に最適化した設計ソフトウェア「Edesign（イーデザイン）」を使って各 SNP を検出するためのプライマーおよびプローブの最適な配列を全 SNPs に対して計算した。当時のデータベースでは約 6000 万箇所の SNPs のみがデータベース登録されていたため、約 6000 万 SNPs 全てに対して Edesign の計算を行い、その結果 4200 万ヶ所の SNPs を検出するプライマーおよびプローブの設計に成功した。

しかしながら、残りの 1800 万ヶ所の設計を行うためには、自由エネルギー計算を含めた計算を行う必要があることが判明した。さらにすでに設計済みのプライマーおよびプローブの配列に自由エネルギー計算を考慮して再計算を行ったところ、実際の実験での SNP 決定精度の向上が認められたため、再計算を行う必要が出た。

本年度は HOKSUAI を利用したときにどの程度の計算コストが必要かを見積もることを目的とした。

２．　具体的な利用内容、計算方法

Edesign は共同研究先の株式会社ダナフォーム社から提供されたプログラムで、Primer3 をもとに PEM 法用に改変されたプログラムである。Primer3 は PCR 用のプライマーをデザインするために分子生物学の分野に置いて幅広く使用されている。（http://primer3.sourceforge.net/）

臨床での利用が検討されている 10 個の SNP を選択し最適化オプションによる速度向上を検討し、最速と思われる計算時間をクラスタ計算機と比較した。並列化は、Embarrassingly Parallel として、行った。比較に用いたサーバとクラスタ計算機のスペックは

①Xeon サーバ

CPU:Intel Xeon E5-2640@2.50GHz 12 コア×2

Memory: 48GB

②C2D クラスタ

ノード数： 500 ノード

CPU: Intel Core2Duo L7400@1.5GHz 2 コア

Memory: 2 or 4GB

である。さらに染色体 22 番の最初の 1 万 SNP を HOKUSAI の 1 ノードで計算し、16 コアと 32 コアでの計算時間の差を測定した。

３．　結果

始めに HOKUSAI での最適なオプションを検討した。Edesgin は HOKUSAI の推奨オプションである”-fast”でビルド出来なかったため、-O2 オプションから順次 –fast に近づける形で最適な物を探索した。その結果、”-O3 -Kdalign,ns,mfunc,lib,rdconv,prefetch_conditional,ilfunc,fp_contract”の計算時間が一番短く 1 割の速度向上ができた。計算結果もこれまでの物と一致したため、これを採用した。その計算時間を比較した結果を図 1 に示す。C2D クラスタは 4200 万ヶ所の SNPs を検出するプライマーおよびプローブの設計に利用した計算機であるが、残念なことに、単体では HOKUSAI の方が計算時間をさらに必要とすることがわかった。

次に、染色体 22 番の最初の 1 万 SNPs の設計を C2D クラスタで行った時と同様に HOKUSAI で計算し必要時間を計算した。図 2 に示すように、コア数に応じて計算速度が向上することが認められ、32 コアを使用しても問題なくスレッド並列が出

来ていると考えられた。

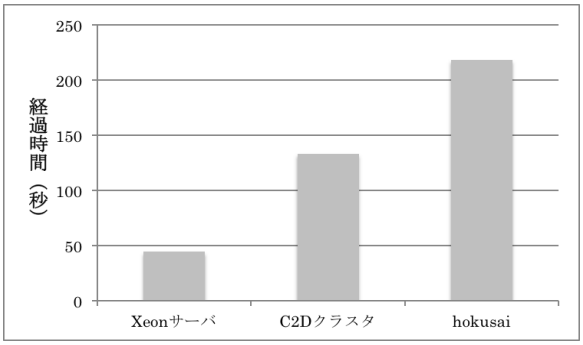

図１　選択した 10 個の SNP s に対する Eprobe を設計したときの合計の計算経過時間(秒)の比較

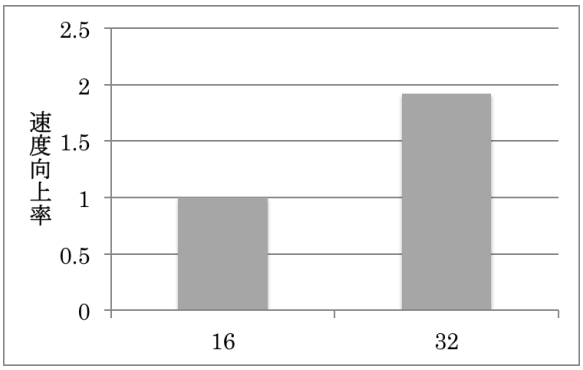

図２　HOKSAI で染色体 22 番の最初の 1 万 SNP s をデザインしたときの、速度向上率。16 コアの時を１として表現した。

４．　まとめ

HOKUSAI を用いて Edesign の計算を行うことが可能であり、32 コアをフルに利用できることが分かった。しかしながら全ゲノム計算ためには 1 回の試行では年間資源の 2.4%を必要とすることが判明した。

５．　今後の計画・展望

昨年の夏に約 1 億 3 千万箇所の一塩基多型(SNP)が報告されたため、これをすべて計算したいと考えているが、現状のままでは実行することが難しい。今後他の計算機の組み合わせも視野にいれて研究を行っていきたい。