疾患関連遺伝子同定のための高速な ARG 構築アルゴリズムの開発

新領域創成科学研究科情報生命科学専攻　修士２年　66958　重田高志
2008 年 3 月修了　　指導教員：森下真一教授

[キーワード]
関連解析、ARG

[研究の背景と目的]

近年、高度なシークエンシング技術やジェノタイピング技術の進歩により、ヒトゲノムから SNP を遺伝的マーカーとしてジェノタイプし、統計解析で疾患との関連を調べる関連解析が盛んに行われている。しかし、関連解析の手法自体は発展途上であり、SNP 数の増大による偽陽性の問題や集団の階層化による影響など多くの問題点を抱えている。

このような問題の解決策の一つとして、ARG(Ancestral Recombination Graph)を用いた手法が提案されている。これは、集団内のゲノムの過去の系統関係をグラフで表現したもので、過去に生じたゲノムの変異や組み換えに関する情報を盛り込むことができる。これを疾患関連遺伝子座のマッピングに応用する研究が近年活発に行われている。通常の手法では連鎖不平衡が用いられているが、これは過去の生じた変異や組み換えなど、非常に多くの遺伝的要素に影響を受ける点が問題となる。ARG によりその点を解決できる。

しかし、過去のゲノムの状態は分からないため、ARG を推定することは大変困難である。今まで遺伝学モデルに基づき、統計的手法で推定する研究が多く行われてきたが、計算量の問題から実用的ではなかった。そこで近年 Heuristics な手法で ARG を推定する方法(Margarita)が提案され、大幅に計算時間を短縮して推定できるようになった。しかし、依然として大規模なデータに適用するのは時間がかかる。

本研究では Margarita のアルゴリズムに基づき、より高速な手法の開発を目的とする。

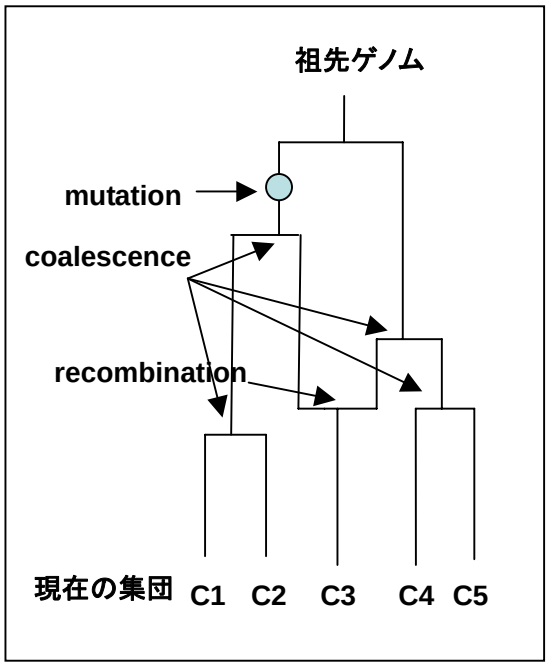

図 1 . ARG

[Margarita]

Margarita は現在観測可能な配列集合から、過去のゲノムを遡り、ARG を推定するソフトウェアである。詳細は省略するが、ここでは組み替え(recombination)をどのように定義しているかについて触れる。

配列の集合が与えられたとき、そこから最も長く共有している２つの配列を取り出す。共有箇所の端が break point となる。仮に SNP マーカーa, b の間で共有していたとすると、(a-1, a),(b, b+1)のいずれかが break point となる。

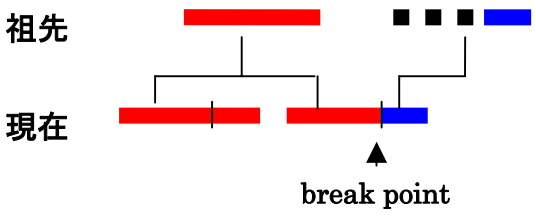

図 2　recombination

ただし、「最も長い共有配列の端」が break point であるというのは必ずしも正しいわけではないので、１割の確率で、完全に任意の共有配列を選択し、その端を break point として定義する。

[提案手法]

本研究では、上記のアルゴリズムに改良を加えた。まず、「最も長い共有配列」を求めずに、共有配列の長さが「領域の半分以上」であれば、その端を break point として定めた。さらに、「１割の確率で、完全に任意の共有配列を選択」という規則を削除し、全て同じ規則で break point を定義した。

この理由として、「最も長い共有配列の端が break point」と定義するのは、必ずしも正しいことではない上、非常に計算時間がかかるためである。この部分を修正することによって、精度を減少させずとも、ARG 構築アルゴリズムの計算時間の短縮できると考えられる。また、「１割の確率で、完全に任意の共有配列を選択」というのはあまりにも任意であり、１,2 塩基共有しているだけでも break point として定義されてしまう。これは現実的ではないと考えられる。

[実験]

データとして fregene で作成したシミュレーションデータを用いた。このソフトウェアで、塩基数１.5M/450SNP、塩基数１M/300SNP 及び塩基数 0.5M/150SNP の 3 種類の case control データ(全て case 200 control 200)をそれぞれ 50 セット用意し、マッピング精度・計算時間を調べた。マッピング精度は以下の基準で評価した。

・検出力(power)

疾患遺伝子座周辺(数 kb の窓枠を指定)に、$p < 0.05$ の p 値を持つ SNP が存在するか。

・位置(location)

最も低い p 値を持つ SNP が、疾患関連遺伝子座の近くに存在するか。

[結果]

結果の一例として表１に塩基数 1M での ARG 構築に要した計算時間の比較を記す。提案手法により、おおよそ 2 割ほど減少した。

表１. 計算時間の比較

	margarita	new
データ１	417sec	308sec
データ２	446sec	366sec
データ３	431sec	347sec

精度に関しては、先ほどのいずれの基準でもほぼ同じ精度が得られた。

塩基数 0.5M のデータに関しても同様の結果が得られた(計算時間は 6 割ほど減少)。しかし、塩基数 1.5M のデータに関しては、精度、計算時間とも元の Margarita と同精度になった。

[結論]

塩基数 1M 以内の領域に関して、ARG 構築にかかる計算時間を短縮することができ、マッピング精度も元のアルゴリズムとほぼ同精度を保つことができた。