修士論文

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

66958　重田高志

指導教員　森下真一 教授

2008 年 3 月

東京大学大学院新領域創成科学研究科情報生命科学専攻

概要

近年、高度なシークエンシング技術やマイクロアレイ技術の進歩により、ヒトゲノムから SNP をジェノタイピングし、関連解析を行うことによって、疾患と関連する多型を同定する研究が盛んに行われている。しかし、手法自体には発展途上であり、様々な手法が提案されている。その中でも本研究では、ARG を用いた手法の開発に焦点をあてる。ARG は集団内のゲノムの系統関係を考慮できる長所があるものの、計算時間が大きな問題点となり、実用が困難であった。本研究では、ARG 構築アルゴリズムの一つである Margarita に改良を加えることによって、元のアルゴリズムより高速に計算可能な手法を提案した。実験結果により、塩基数 1.5M 程度の長い配列に関しては効果は得られなかったものの、塩基数 1M 以内のデータに関しては、常に計算時間を短縮することができ、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。

目次

第 1 章　序論 1
1.1 研究の背景 . . . 1
1.2 研究の目的 . . . 2
1.3 本論文の構成 . . . 2

第 2 章　関連解析及び ARG 3
2.1 関連解析の種類 . . . 3
2.2 Data quality check . . . 4
2.3 単一 SNP による解析 . . . 4
2.4 連鎖不平衡解析 . . . 4
2.5 複数の SNP による解析 . . . 5
2.6 集団の階層化（Population Stratification） . . . 6
2.7 多重検定の補正 . . . 6
2.8 Ancestral Recombination Graph . . . 7
2.9 Margarita . . . 7
2.10 Haplotype-clustering methods . . . 11

第 3 章　提案手法 13
3.1 Margarita の問題点 . . . 13
3.2 本研究でのアルゴリズム . . . 14

第 4 章　実験 16
4.1 データ . . . 16
4.2 評価方法 . . . 16
4.3 結果 . . . 18
4.4 考察 . . . 22

第 5 章　まとめ 25
5.1 結論 . . . 25
5.2 今後の課題 . . . 25

謝辞 27

参考文献 28

第 1 章

序論

1.1 研究の背景

ゲノム科学における重要なテーマの一つとして、疾患の遺伝的要因を解明することが挙げられる。これに関する研究は、高速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、大変活発に行われている。その手法として、連鎖解析 (linkage study) と関連解析 (association study) の 2 種類の手法がある。このうち、遺伝マーカーとして SNP を利用した関連解析については、複合遺伝性疾患に対して検出力が高いこと、さらに、高密度の SNP を大量にジェノタイピングする技術が大幅に進歩したことから、大変注目を集めており、大きな発見も残している。ゲノム全域から数十万もの SNP をジェノタイピングし、疾患に関連する遺伝子を同定するゲノムワイド関連解析も、近年活発行われている [5][24]。

しかし、関連解析の手法自体はまだ発展途上で多くの問題点も残されている。最も一般的な関連解析の手法としては次のような手法が挙げられる。まず、case(疾患にかかっている人々)、control(疾患にかかっていない人々) それぞれに対し、各 SNP のアレル頻度をカイ二乗検定等で比較する。頻度差に有意な違いが得られた SNP が存在した場合、その SNP と連鎖不平衡にある遺伝子座が、疾患と関連していることが推測される。しかし、連鎖不平衡の値自体は集団内で過去に生じた変異や組み換えなど、様々な遺伝的要素に左右される [19]。この点が、検出力を弱める原因となっている。

解決策として、集団内のゲノムの系統関係 (genealogy) を考慮した手法が近年提案されている。その一つとして、ARG(Ancestral Recombination Graph) を用いた手法が近年研究されている。

ARG とは、進化の過程でゲノムにおいて過去に発生した、組み換えや変異に関する情報を盛り込んだグラフである。現在観測可能なゲノム配列から、統計遺伝学的手法を用いて ARG を推定し、疾患関連遺伝子座のマッピングに適用する研究が近年活発に行われている。通常の単一 SNP による検定と比べ、ゲノムの変遷に関する情報を考慮することができるのが長所である。しかし、ARG を推定する手法については多くの研究がなされているものの、いずれもかなりの計算量を必要とすることから、ごく小さい領域にしか適用できず、あまり実践的でない。[18]。

最近、ARG の計算時間の問題を克服するため、heuristic なアルゴリズムにより ARG を推定する手法 (Margarita) が開発された [18]。これにより、計算時間は大幅に改善され、ある程度の長さをもつ領域に対しても ARG の推定が可能でになった。しかし、依然として計算時間はかかり、大規模のデータに適用することは困難である [28]。

1.2　研究の目的

本研究では、疾患関連遺伝子を同定するための ARG 推定アルゴリズムの開発に焦点をあてる。土台となる手法として、Minichiello らが提案した Margarita を基にした。この手法は大変高速であるが、必ずしも最適な ARG を求めている保証はなく、アルゴリズムにはまだ改善の余地があると考えられる。本研究では、この Margarita のアルゴリズム上の問題点を考察し、改良を加えることによって、より高速に計算できるような手法を提案することを目的とする。

1.3　本論文の構成

本論文は以下のように構成される。まず 2 章において、関連解析及び ARG に関して、近年の研究を交えながら基本事項について解説する。3 章において、本研究で採用した手法について解説する。4 章では、その手法に基づいて行った実験とその結果について示す。5 章において本研究の結論及び今後の課題についてまとめる。

第 2 章

関連解析及び ARG

この章では、SNP を用いた関連解析 (association study) 及び ARG に関して、基本的な事項についてまとめる。参考文献として、主に [1][11][6] などを参考にした。

2.1 関連解析の種類

関連解析をその規模で分類するならば、おおよそ以下のように分類できる [1]。ただし、厳密な定義はない。

- Candidate gene association study
  候補となる遺伝子や多型を特定した上で、5 から 50 程度の SNP をジェノタイプして調べる。
- Fine mapping
  1-10Mb の領域に対し、数百程度の SNP をジェノタイプして調べる。
- Genome-wide association study
  ゲノム全域から 300,000 以上もの SNP をジェノタイプして調べる。

近年、ジェノタイピング技術の発展や HapMap の成果により、Genome-wide association study が大いに注目を集めており、実績も多く挙げている [5][24]。

また、関連解析に用いるサンプルとして、主に case control study が用いられる。これは家系データを用いず、疾患の患者集団 (case) と非患者集団 (control) でサンプルを構成する方法である。この他、家系データを用いた手法もいくつか提案されている。代表的な手法として、TDT(Transmission disequilibrium test) がある [26]。これは、case の両親のデータも利用し、疾患関連座位と関連の見出された SNP に対し、真の連鎖関係が存在するかを調べる方法である。

case control のような集団のみのデータの場合、後述する集団の階層化 (population stratification) のような問題が生じる。TDT のような家系データを用いた場合、このような問題を解決できる。しかし、サンプル収集はより困難であり、ジェノタイプエラー等が存在すると大きく影響を受ける等も問題もある [16]。

case control study では、TDT のような家系データを利用したときと比べ、より少ないサンプル数で同じ検出力を得られる長所がある [17]。以下では全て case control study を仮定し、家系データについては取り扱わない。

2.2 Data quality check

データ解析を行う前に、解析に適さない SNP を除去する。具体的には、ハーディ・ワインバーグ平衡から逸脱した SNP や、call rate (各 SNP に関して、全サンプルの中で正しくジェノタイプできた割合) の低い SNP、低アレル頻度の SNP を除去する。

ハーディ・ワインバーグ平衡の検定は control で行う。計算はピアソンのカイ二乗検定で行うのが容易であるが、遺伝子型の頻度が少ない場合、カイ二乗近似が悪く、偽陽性も増える。そのため、近年では Fisher exact test が最近ではよく用いられている [30]。閾値は、$p < 0.001$ とするのが一般的である。

また、ジェノタイプできないデータ (missing genotype data) の取り扱いも大きな問題である。単一の SNP で解析を行う場合はそれほど大きな問題はないが、ハプロタイプ解析など、複数の SNP を用いて解析を行う場合は、ごくわずかなエラーが生じただけでも大きな問題となる。そこで、周辺の SNP のジェのタイピングデータをもとに、missing data のジェノタイプを推定する方法も提案されている [25]。

2.3 単一 SNP による解析

関連解析において最も一般的な手法は、case, control それぞれの各 SNP のアレル頻度もしくはジェノタイプ頻度を調べ、統計解析で頻度差を検定するという手法である。統計解析においては各アレル頻度により 2×2 分割表のカイ二乗検定で検定するのが一般的である。しかし、この手法では case, control それぞれに対し、ハーディ・ワインバーグ平衡が成立していることを前提としている問題がある。そこで、近年ではハーディ・ワインバーグ平衡を仮定しないコクラン・アーミテージ検定で検定することが多い [21]。

疾患の原因となる多型を直接観察しようとすると、候補となりうる全ての多型をジェノタイプしなくてはならない。これはゲノムを解読する技術が相当進歩しない限り、非常に困難である。しかし、次項で述べる連鎖不平衡解析により、全ての多型を直接調べなくても、疾患関連多型と強い連鎖不平衡にある SNP を調べることによって、間接的に調べることが可能である。

2.4 連鎖不平衡解析

連鎖不平衡 (Linkage Disequibrium) とは 2 つ以上の遺伝子座間にみられる現象である。ある集団において 2 つの遺伝子座に多型が存在するとする。 2 つの多型のアレルの特定の組み合わせ (ハプロタイプ) の頻度が有意に高くなるとき、2 つの遺伝子座は連鎖不平衡にあるとする。2 つの遺伝子座間に組み替えが多いほど、連鎖不平衡は失われる傾向にある。

2 つの多型が連鎖不平衡にあるとわかれば、片方の多型のアレルを調べるだけで、もう一方の多型も推定できる。そのため、領域中の全ての多型をジェノタイプしなくても、ジェノタイプされていない多型をある程度間接的に調べることが出来る。そのため、関連解析において連鎖不平衡は重要な概念である。

連鎖不平衡の大きさを調べる指標としてはいくつかあるが、よく用いられるのは D' と r^2 という 2 つの指標である。共に 0 から 1 までの値をとる。

定義を以下に示す。2 つの遺伝子座があり、片方の遺伝子座のアレルを A, a、もう遺伝子座のアレルを B, b で表す。集団中のアレル頻度及びはハプロタイプ頻度を $p(A), p(AB)$ 等で表す。このとき、まず、連鎖不平衡係数 D を、

$$D = p(AB)p(ab) - p(Ab)p(aB), \tag{2.1}$$

で表す。連鎖不平衡係数 D の取りうる値には制限が設けられているため、これを規格化し、

$$D' = \frac{|D|}{D_{max}}, \tag{2.2}$$

で表現する。ここで D_{max} は、

$$D_{max} = min(p(Ab), p(aB)), \tag{2.3}$$

で与えられる。

また、r^2 は

$$r^2 = \frac{D^2}{p(A)p(B)p(a)p(b)}, \tag{2.4}$$

となる。D' も r^2 も共に連鎖不平衡の大きさを表すが、その挙動にはいくつかの違いがある [7]。

2.5 複数の SNP による解析

複数の SNP を組み合わせて解析することで、単一の SNP で解析するより高い効果を得られることがある。ハプロタイプ解析が代表的である。

ゲノム内には組み替え頻度の非常に少ない領域がブロック状に存在している (LD block)。領域内での SNP のハプロタイプの頻度を用いて統計解析を行うことで、前述の単一 SNP による解析と比べ、大きな効果が得られることがある [4]。

LD block の定義としてはいくつか提案されており、Gabriel らの手法 [10] や、four gamete test による手法 [29] などがある。

ハプロタイプの相 (phase) を実験により直接調べることは大変困難であるため、EM アルゴリズム等を推定する方法がいくつか開発されている [14]。その中でも、PHASE [27] が最も正確であるとされている。計算時間がかかることが問題であったが、FASTPHASE [23] が開発

されてことにより、ほぼ同じ精度で高速に計算することが可能になった。しかし、ハプロタイプの解析において、相の不確かなデータの扱いは依然問題である。

統計解析を行う際、単一 SNP の検定の場合と同様、ハプロタイプ頻度により $2 \times \mathrm{k}$ の分割表により χ^2 検定を行うのが一般的であるが、この場合、ハーディワインバーグ平衡を仮定していることなどの問題点もある。そこで、回帰分析 [13] やスコア統計量 [22] を用いる手法など多くの研究例がある。

2.6　集団の階層化（Population Stratification）

サンプルとして用いた集団にアレル頻度の異なる集団が交じり合っていると、解析結果に偽陽性を引き起こす。この問題を解決する方法として、以下の 2 つの手法が一般的に良く用いられる。なお、サンプルとして家系データを用いれば、この問題は回避できる。

- Structured Association [20]
  複数の位置的に関連のない遺伝マーカーのデータから、サンプルを階層化のない亜集団に分類する。亜集団ごとに解析を行うことによって、補正を行う。
- Genomic Control [8]
  複数の位置的に関連のない遺伝マーカーについて、カイ二乗値などの統計量を得る。得られた統計量は確率分布に従うはずなので、集団の階層化が影響すると、分布が変化する。この分布の変化から階層化の影響を調べることで、統計量を補正する。

2.7　多重検定の補正

複数の SNP に対し、検定を繰り返すと、それだけ偽陽性のデータも増大する。特に、genome-wide association study のような何百万もの SNP を用いて検定する場合、この問題は大変深刻になる。

最も単純な補正方法として、Bonferroni Correction がある。これは得られた p 値に検定回数をかけた値を補正値とする方法である。例えば、50 万もの SNP を genotype し、それぞれに対し検定を行ったとすると、有意水準 α を、

$$\alpha = 0.05/500,000 = 10^{-7}, \tag{2.5}$$

とすることが多い。しかし、この方法では全ての検定が独立であると仮定しているため、かなり厳しい補正となる。そこで、FDR(false discovery rate) を制御する方法がいくつか提案されている [2][3]。FDR とは、検定で棄却される真の帰無仮説及び偽の帰無仮説を V, S とすると、

$$Q = E(\frac{V}{V+S}), \tag{2.6}$$

で定義される。すなわち、棄却された仮説のうち、真の帰無仮説がどれだけ存在するかを示す指標である。

多重検定の補正の最も優れた方法はパーミュテーションテスト (permutation test) であるといわれている。これは、case, control のラベルを入れ替えて検定を繰り返す方法である。これを繰り返すことで、新たな統計量の分布を得る。その分布のどこに位置しているかで、p 値を再計算する方法である。しかし、検定を複数回行う必要があるため、計算時間が非常にかかる問題がある。

2.8 Ancestral Recombination Graph

近年、新たな手法として ARG(Ancestral Recombination Graph) を用いる手法が開発されている。以下、ARG について解説する。

ARG とはあるゲノムの過去におけるに発生した、組み換えや変異に関する情報を盛り込んだグラフである。図 2.1 に具体例を示す。各ノードが DNA 配列に対応し、エッジがその親子関係を示している。組み換えが起きた場合、親を 2 つもつノードが存在する。また、変異に関する情報をエッジ上に盛り込むことで、過去のゲノムの変異の発生に関する情報も盛り込むことができる。完全な ARG を同定することは不可能であるが、現在観測可能な DNA 配列から、ある程度 ARG を推定することは可能である。

この推定された ARG は、疾患関連遺伝子を調べるのに有用である。推定された ARG 内で、ある枝で疾患に関連する変異が生じたとすれば、その枝以下とそれ以外で case, control の頻度に有意な違いが見られるはずである。

通常の統計解析では連鎖不平衡解析を用いる。連鎖不平衡は主に 2 つの遺伝子座間の組み換え頻度に基づくものの、変異の発生や交配のパターンなど、様々な遺伝的要素で影響を受ける [19]。一方、ARG の形状は領域内で推定される組み換えに直接依存している。この点で、通常の解析より大きな効果が得られる。

ARG 推定方法については統計遺伝学的手法を用いた方法がいくつか提案されている [31]。遺伝学モデルに基づき、ARG の尤度をベイズ推定やマルコフ連鎖モンテカルロ (MCMC) 法などを用いて求めることが試みられている。しかし、これらはいずれも大量の計算時間を必要とすることから、ごく一部の領域の適用に限られてきた。近年、heuristic な手法で ARG を推定する方法として Margarita [18] が提案された。これにより、短時間で ARG を推定することが可能になった。次の章でその詳細を解説する。

2.9 Margarita

ある配列 C の i 番目の SNP のアレルを $C[i]$ で表現する。各アレルは 0,1,$\cdot$ の 3 通りで表現し、$\cdot$ はジェノタイプした配列のいずれからも受け継がれていないために、定義できないアレルを示す。ある SNP マーカー i において、$C_1[i] = C_2[i]$ もしくは $C_1[i] = \cdot$ もしくは $C_2[i] = \cdot$ が成立しているとき、$C_1[i] \sim C_2[i]$ と表現する。また、$\neg$ を、$C[i] = 1$ のとき $\neg C[i] = 0$、$C[i] = 0$ のとき $\neg C[i] = 1$ と定義する。

長さ m の 2 つの配列 C_1, C_2 が存在するとき、共有箇所 $\{C_1, C_2\}[a, b]$ を以下のように定義

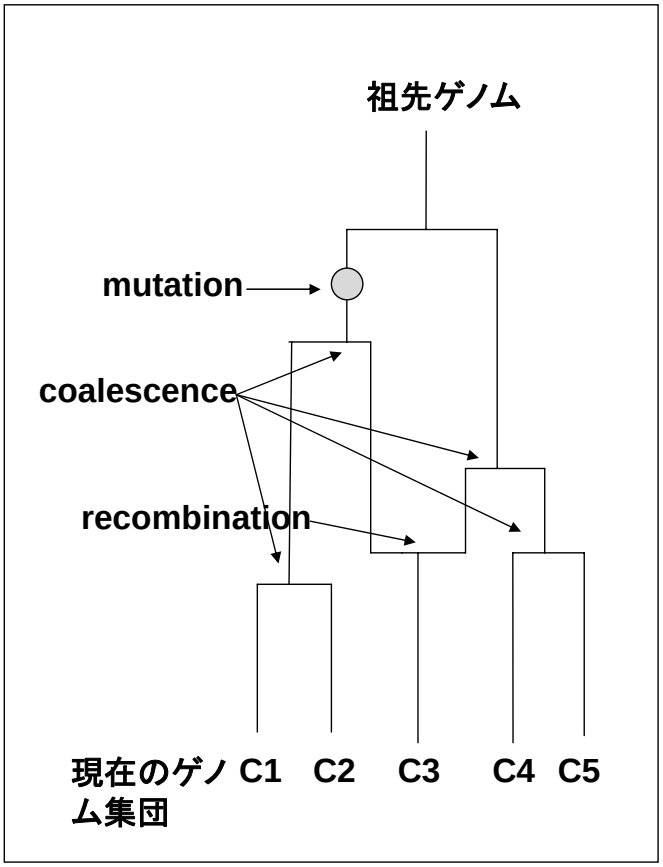

図 2.1. ARG

する。

1. $C_1[i] \sim C_2[i], a \leq i \leq b$
2. 少なくとも一つの i で $C_1[i] = C_2[i] \neq \cdot$
3. $a > 1$ ならば、$C_1[a-1] \neq C_2[a-1]$
4. $b < m$ ならば、$C_1[b+1] \neq C_2[b+1]$

ある時間 T における配列の集合を S_T で表現する。Margarita では、以下の規則を繰り返すことで S_T を推定する。

- Coalescence
  ２つの配列 C_1、C_2 に対し、全ての i において $C_1[i] \sim C_2[i]$ が成立しているとき、C_1、C_2 はその先祖において coalescence するものとする。祖先の配列集合 S_{T+1} は、

$$S_{T+1} = ((S_T \backslash \{C_1, C_2\}) \cup \{C^{'}\}), \tag{2.7}$$

  となる。ここで、$C^{'}$ は、$C_1[i] \neq \cdot$ のとき、$C^{'}[i] = C_1[i]$ が成立し、それ以外で $C^{'}[i] = C_2[i]$ が成立する。

- Mutation
  S_T 内の配列 C_1 及びマーカー i に対し、$S_T \backslash \{C_1\}$ 内の全ての C_2 に対し、$C_2[i] = \neg C_1[i]$ もしくは $C_2[i] = \cdot$ が成立しているとき、マーカー i で mutation が起きているものとする。S_{T+1} は、
  $$S_{T+1} = ((S_T \backslash \{C_1\}) \cup \{C^{'}\}), \tag{2.8}$$
  となる。ここで、$C^{'}$ は $C^{'}[i] = \neg C_1[i]$ かつ $j \neq i$ なる全ての j において $C^{'}[j] = C_1[j]$ が成り立つものとする。
- Recombination
  上の 2 つの規則どちらも満たさない場合は、recombination が起きるものとする。共有配列 $\{C_1, C_2\}[a, b]$ を選び、組み換えの起きる点 (break point)(α, β) を $(a-1, a)$ もしくは $(b, b+1)$ で定義する。S_{T+1} は、
  $$S_{T+1} = ((S_T \backslash \{C_R\}) \cup \{C_1^{'}, C_2^{'}\}), \tag{2.9}$$
  となる。ここで、C_R は C_1, C_2 のいずれかを任意に選択する。また、$C_1^{'}$ は α 以下の全ての i に対し $C_1^{'} = C_R[i]$ が成り立ち、$C_2^{'}$ は β 以上の全ての i に対し $C_2^{'} = C_R[i]$ が成立するものとする。$(a-1, a)$、$(b, b+1)$ いずれもが break point であると考えられるときは、それぞれの点で組み換えが起きているとする。

これらを繰り返すことで ARG が構築される。しかし、どの段階においても規則を満たす複数の coalescence, muation, recombination が考えられ、それゆえ、様々な ARG が考えられるそれらの中から、より最適な ARG を選択するために、以下の heuristic な規則を取り入れる。

- recombination は、mutation, coalescence いずれも起こりえないときのみ起こる。
- 複数の mutation, coalescence が起きる場合、その順序は任意。
- coalescence は、二つの配列のオーバーラップ領域が、$\cdot$ でないマーカーで少なくとも一つ共有しているしているときのみ起きる。これは、Markovian coalescent model [15] の考えに基づくものである。
- recombination の際、2 つの配列の共有配列のうち、長さが最大になるものを選択し、その端を recombination の break point と定める。これは、より長い共有配列の方が、より最近の recombination を反映している傾向にあるためである。ただし、これはあくまで傾向にすぎず、必ずしも正しいものではない。そこで、ある確率 (論文中では 1 割) で、任意に得られた共有配列の端を recombination の break point と定めることにする。
- recombination 後の最初の coalescence は、recombination の位置を定める際に選択した共有配列に基づいて行われる。

以上により ARG が求まる。しかし、真の ARG は分からないため、これを繰り返して 30 から 100 程度の ARG を作成する。ここから各マーカーにおける木を抽出すると図 2.2 のよう

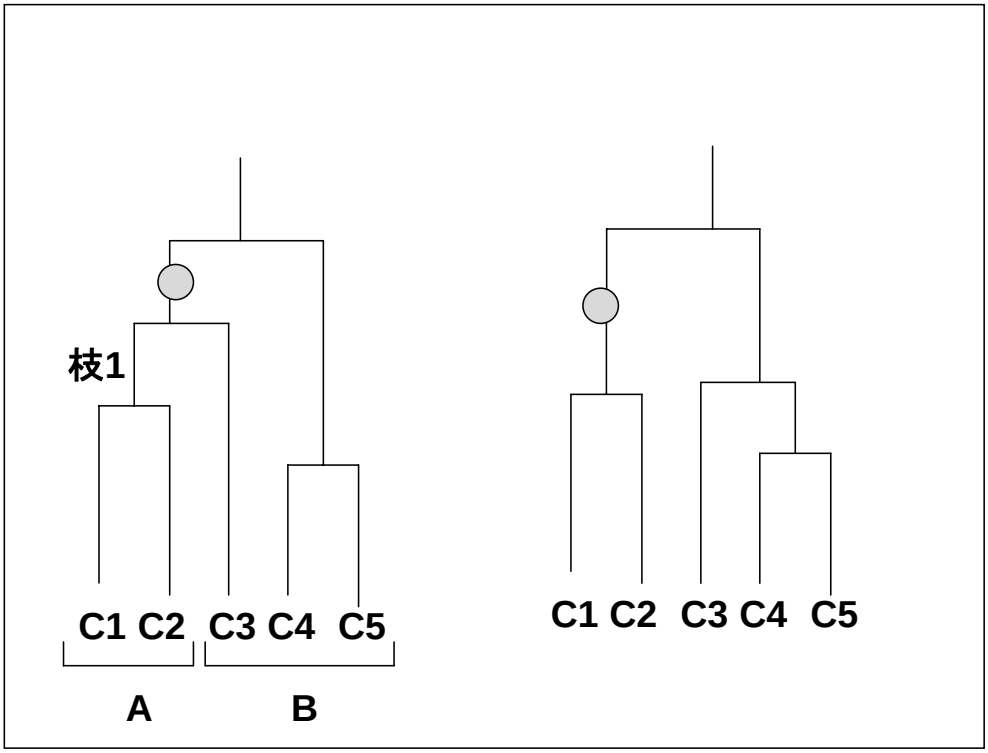

図 2.2. marginal tree

な二分木 (marginal tree) が得られる。ここで図 2.2 は元の ARG が図 2.1 の場合の marginal tree であり、領域によって 2 種類取りうる。このうちどの枝で、疾患に関連した変異が発生しているのかを統計解析によって調べる。例えば、図 2.2 の左図では枝 1 以下の集団とそれ以外の集団で、集団全体を A, B に二分できる。もし A, B で case, control の頻度が大きく異なっていれば、枝 1 において疾患と大きく関連する変異が起きたことが推測される。n 個の葉（配列の集団）が与えられたとき、このような分岐は $n-3$ 個存在する。そこで、case, control の頻度と分岐との関係を χ^2 検定で検定する。$n-3$ 個の分岐ごとに調べ、得られる統計量の最大値 (best-cut score) を各 marginal tree ごとに求める。その平均値を各マーカーにおける統計量とする。

p 値の計算にはパーミュテーションテスト (permutation test) を用いる。各サンプルの case, control のラベルをランダムに入れ替え、上記と同様に統計量を得る。これを多数繰り返すことで、経験的な帰無分布 (null distribution) をえる。この分布から、p 値が求められる。実験結果により、通常の χ^2 検定及び後述の Haplotype-clustering 手法の一つである CLADH と比較し、より高い精度が得られること、また疾患関連アレルの頻度の推定等にも有用である点が結論として得られている。

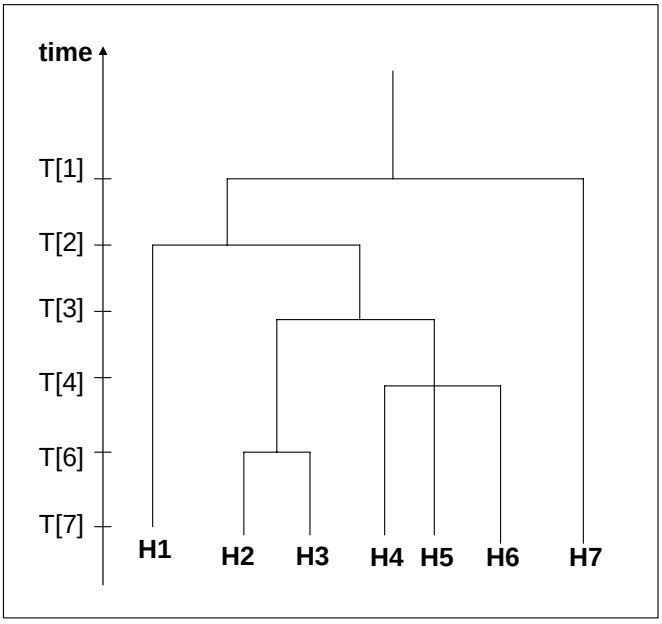

図 2.3. cladgram

2.10 Haplotype-clustering methods

ARG に関連した手法として、Haplotype-clustering methods がある。これはハプロタイプをクラスタリングし、cladgram と呼ばれる階層的なクラスターを作成する。検定はクラスターごとに行う。

クラスターの作成方法は様々な方法が提案されている。一例として、Durrant らの開発した CLADH [9] を取り上げる。

図 2.3 に cladgram の例を示す。$T[h]$ は h 個の SNP ハプロタイプのクラスターからなる。h 個のクラスターが与えられたとき、クラスター間で各ハプロタイプ間の距離の平均をとり、最小となる 2 つのクラスターを融合させる。これを一つのクラスターになるまで繰り返す。

i 番目のサンプルのハプロタイプのペアを $H_i = \{H_{i1}, H_{i2}\}$, 各ハプロタイプを $H_{ij} = \{H_{ij|1|}, H_{ij|2|}, \cdots H_{ij|M|}\}$ で表現する。ここで、$H_{ij|m|}$ は、m 番目の SNP のアレルを 1,2,0 のいずれかで表現したものである。また、0 は missing data である。さらに、m 番目の SNP のアレル 1 の頻度を q_m で表現する。　このとき、ハプロタイプ間の距離 (distance metric) を、

$$D_{i_1 j_1, i_2 j_2} = 1 - \frac{\sum_{m=1}^{M} s_{i_1 j_1, i_2 j_2 |m|} w_m}{\sum_{m=1}^{M} w_m}, \tag{2.10}$$

で定義する。ここで w_m は類似度に対する重み付けであり、$s_{i_1 j_1, i_2 j_2 | m|}$ は、

$$s_{i_1 j_1, i_2 j_2 | m|} = \begin{cases} 1 - q_m & if \quad H_{i_1 j_1 | m|} = H_{i_2 j_2 | m|} = 1 \\ q_m & if \quad H_{i_1 j_1 | m|} = H_{i_2 j_2 | m|} = 2 \\ q_m (1 - q_m) & if \quad H_{i_1 j_1 | m|} = 0 or H_{i_2 j_2 | m|} = 0 \\ 0 & otherwise \end{cases} , \tag{2.11}$$

で定義される。ハプロタイプのクラスターと疾患との関連は logistic regression によって評価する。

cladgram を用いた手法は、ARG を用いた手法と比べ、計算時間が非常に少ないという長所があり、数多くの研究がなされている。しかし、cladgram は ARG に比べ、ゲノムの変遷の大雑把な近似を表しているのに過ぎず、類似したハプロタイプや低頻度のハプロタイプの関係を扱うのが困難であるという問題点がある [18]。一方 Margarita のような ARG を用いた手法では、低頻度のハプロタイプを考慮することができ、過去のゲノムの変異や組み換えまで考慮に入れることができる点で優れていると思われる。

第 3 章

提案手法

本章では、本研究で提案した手法について解説する。それに当たって、まず前章で取り上げた Margarita の問題点について考察する。それを踏まえた上で、どのようなアルゴリズムを提案したかを解説する。

3.1 Margarita の問題点

前章で取り上げた Margarita は従来の手法に比べ、大変高速である。しかし、それでもやはり計算時間はかかり、数千、数万 SNP もの大規模な領域に対し通常のコンピュータで計算するのは困難である。

計算時間がかかる要因として、以下の 3 つが考えられる。

1. ARG 構築
2. ARG を複数個 (100 程度) 作成しなければならない
3. p 値計算の際のパーミュテーションテスト

具体的に述べる。ARG を構築するのに約 300SNP, サンプルとなる配列数が 400 のとき、CPU が 2.2GHZ のコンピュータで 1ARG あたり約 7 分程度かかる。さらに、この ARG を 100 程度作成すると、約 12 時間程度かかる。さらに、10000 回のパーミュテーションを行うと、合計で 20 時間以上の時間を要する。SNP 数が数千になると、全ての計算に数日もの時間を要する。

このうち、2 番目の問題に関しては、各 ARG 構築は完全に独立であるため、並列に計算することでかなり時間を短縮できる。もっとも、ARG 構築にかかる計算時間にばらつきが生じる場合、台数効果は得られない。全ての ARG を作成しないと、p 値の計算ができないためである。しかし、実際には表 3.1 のように、どの ARG でも計算時間にそれほど大きな変化はないので、n 台で計算すれば計算時間はほぼ $\frac{1}{n}$ になる。

p 値計算の際のパーミュテーションテストに関しては、非常に大きな計算時間を要する。大きな問題であるが、これは他の手法に関しても同様であり、いくつか解決策も考えられる。本題である ARG そのものからは話がそれるため、本研究では扱わない。

表 3.1. ARG 構築の計算時間

	全体	recombination のみ
ARG1	411.5 sec	408.9 sec
ARG2	416.9 sec	414.5 sec
ARG3	413.2 sec	410.9 sec
ARG4	412.2 sec	409.8 sec
ARG5	412.2 sec	409.9 sec

そこで 1 番目の問題について考える。ARG 構築のうち、非常に大きな計算時間がかかる部分として、recombination の break point を定める部分が挙げられる。例えば、表 3.1 のように、一つの ARG(303SNP、400 配列) を構築するのに約 7 分程度かかった場合、recombination 以外の部分では 2,3 秒しかかかっていない。つまり、計算時間の 9 割以上が recombination の break point を求めるのにかかっていることになる。これは、最長の共有配列を求めるのに大きな計算時間がかかるためである。配列の数を N とすれば、計算量はほぼ N^2 に比例する。

筆者らも述べているように、最長の共有配列を求め、その端を break point とすることは、確かに有力な手法であるかもしれない。しかし、これは必ずしも正しいわけではない。筆者らもそのことは認めており、解決策として、一定の確率で任意の共有配列を選択する、としている。しかし、この部分こそ大きな問題点があるように思える。このアルゴリズムでは、1、2 塩基のみ共有していても組み替えが起きていると見なしていることもある。この場合、大抵は、偶然共通していたという可能性の方が高いと思われる。「最長の共有配列」を求めることの根拠は生物学的にもそれほど強いものではなく、ここを修正することで、より高速に計算可能ではないかと推測される。

3.2　本研究でのアルゴリズム

前項の問題点を踏まえ、以下の本研究では、Margarita に対し、以下の修正を加えた。

1.「最長の共有配列」を求めずに、2 つの配列が十分長く (領域の半分以上) 共有していれば、その点で recombination の候補とする。すなわち、2 つの配列 C_1, C_2 共有配列を $\{C_1, C_2\}[a, b]$ とし、その長さが領域の半分以上であるときに、$(a-1, a), (b, b+1)$ のいずれかが break point となる。
2. もしそのような条件を満たす配列が存在しない場合は、Margarita と同様に最長の共有配列を選択し、break point を求める。
3.「一割の確率で任意の共有配列」の部分は削除する。

なお、coalescence 及び mutation に関しては、Margariata と同じ定義にした。

Margarita のアルゴリズムでは、最長の共有配列を求めるのに、ほぼ $O(n^2)$ 計算量かかる

(n は集団の配列の数)。本研究のアルゴリズムでは、その部分を大幅に改善できる。しかし、最悪の場合は同じ計算量がかかるため、共有配列を持ちにくい長い領域などでは、効果が減る可能性がある。

第 4 章

実験

4.1 データ

本研究で用いるデータとして、fregene [12] によって作成したシミュレーションデータを用いる。fregene とは Wright-Fisher Model によって塩基配列の集団をシミュレーションするソフトウェアであり、関連解析の手法の評価に用いられる。

データの条件を以下のように設定した。集団の大きさは一定で、塩基配列数を 20,000 とした。領域全体の mutation rate を 1.1×10^{-8} とした。さらに、全領域の内 1% を recombination hotspots として定めた。これは塩基数 2k の領域からなり、全領域に起こりうる組み換えのうち、6 割がそこで起きている。recombination hotspots とそれ以外の領域での recombination crossover rate はそれぞれ $6.56 \times 10^{-7}, 4.44 \times 10^{-9}$ とした。また、gene conversion も定め、gene conversion rate を 1.1×10^{-7} とし、長さは全て 50base pair に固定した。

また、領域の大きさから 3 種類のデータを用意した。塩基数として、1.5M、1M 及び 0.5M の大きさを選択した。ここから、minor-allele frequency(MAF) が 0.05 以上の SNP をそれぞれ 450 個、300 個、150 個ランダムに選択した。

各領域に対して、case 集団及び control 集団としてそれぞれ 200 サンプル作成した。疾患に関連する多型は、minor allele frequency が 0.05 に近いものから選択した。さらに、generative relative risk(GRR) を $GRR(Aa) = 2, GRR(AA) = 4$ と選択した。ここで、$GRR(Aa), GRR(AA)$ はそれぞれ、その genotype を持つ人の疾患の発症率から、持たない人の発症率を割った値である。

以上の条件から、case control 集団のサンプルを 50 セット用意した。そして、次項で述べる評価基準により精度及び計算時間を調べた。

4.2 評価方法

ARG を利用した Fine mapping の精度の評価は疾患関連多型を得られる検出力と、その位置の 2 つで評価する。しかし、2 章で述べたように、疾患関連多型はジェノタイプされることがほとんどない。本研究で用いるデータも、それを前提としている。そこで、以下の 2 種類の

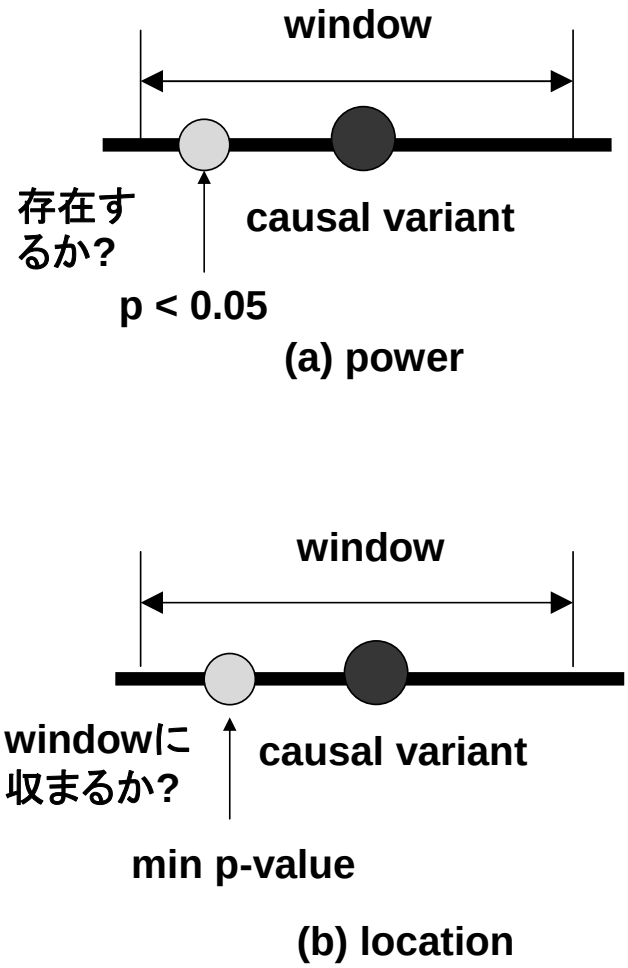

図 4.1. 評価方法

基準で評価する

1. 検出力 (power)
   疾患関連多型の周辺で、塩基数にして数十 k 程度の窓枠を定め、その範囲内で有意な SNP($p < 0.05$ とした) が存在するかを調べる。窓枠の幅は変化させてそれぞれで調べる。
2. 位置 (location)
   最も低い p 値の得られた SNP が、疾患関連多型と十分近い距離に存在するかどうかを調べる。先ほどと同様に窓枠を定めその範囲内に収まっているかどうかを調べる。

4.3　結果

塩基数 1.5M のデータの精度に関する結果を図 4.2 及び図 4.3 に掲載する。図 4.2 の縦軸は、全 50 データセットのうち、窓枠内に $p < 0.05$ の p 値が得られる SNP が存在したデータセットの割合を示している。また、横軸で窓枠の幅 (疾患関連多型からの距離) を示している。図 4.3 の縦軸は、全 50 データセットのうち、窓枠内に最小の p 値を持つ SNP が存在しているデータセットの割合である。横軸については図 4.2 と同様である。グラフのうち、”new”とあるのが本研究で提案した手法であり、元の Margarita による結果を”margarita”で示してある。また、”chi-square”とあるのは χ^2 検定で同様の実験を行ったときの結果である。同様に、塩基数 1M のデータに対する結果を図 4.4 及び図 4.5 に示し、塩基数 0.5M のデータに対する結果を図 4.6 及び図 4.7 に示す。

これらの結果から、検出力、位置いずれに関しても、提案手法は元のアルゴリズムとほぼ精度か、若干良い精度が得られていることが分かる。位置に関してはそれほど差は見られないが、検出力に関しては、元のアルゴリズムと比べても比較的良い精度が得られている。

計算時間に関しては表 4.1、表 4.2 及び表 4.3 の通りである。各領域に対して、50 個のデータセットから 10 個のデータセットを取り出し、各データセットに対して 5 回 ARG を作成して計算時間の平均値を求めた。用いた計算機の CPU の周波数は 2.2GHz である。

塩基数 1M 及び 0.5M の領域に関しては、どのデータに関しても元のアルゴリズムより計算時間は短縮されている。塩基数 0.5M の領域ではおよそ 6 割、1M の領域ではおよそ 2 割ほど短縮できている。しかし、1.5M のデータに関しては、ほとんどのデータに関して減少していない。むしろ、若干増加しているデータも多く見られる。

表 4.1. 計算時間の比較 (1.5M/450SNP)

	Margarita	New
データ 1	1330 sec	1351 sec
データ 2	1247 sec	1301 sec
データ 3	1318 sec	1320 sec
データ 4	1325 sec	1376 sec
データ 5	1284 sec	1325 sec
データ 6	1230 sec	1248 sec
データ 7	1305 sec	1299 sec
データ 8	1251 sec	1259 sec
データ 9	1200 sec	1347 sec
データ 1 0	1255 sec	1272 sec

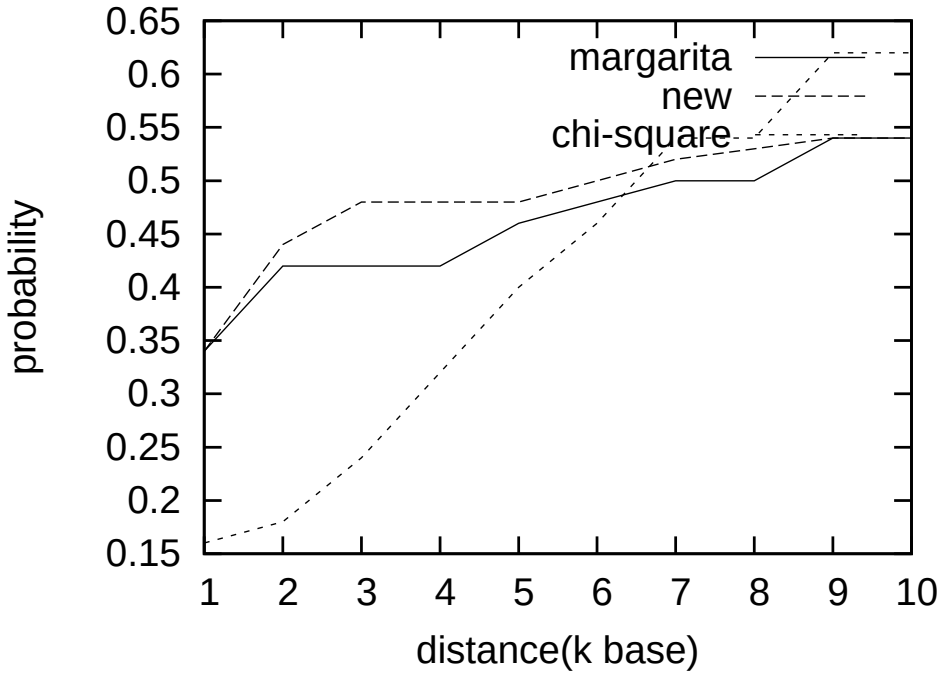

図 4.2. 検出力 (塩基数 1.5M、SNP 数 450)

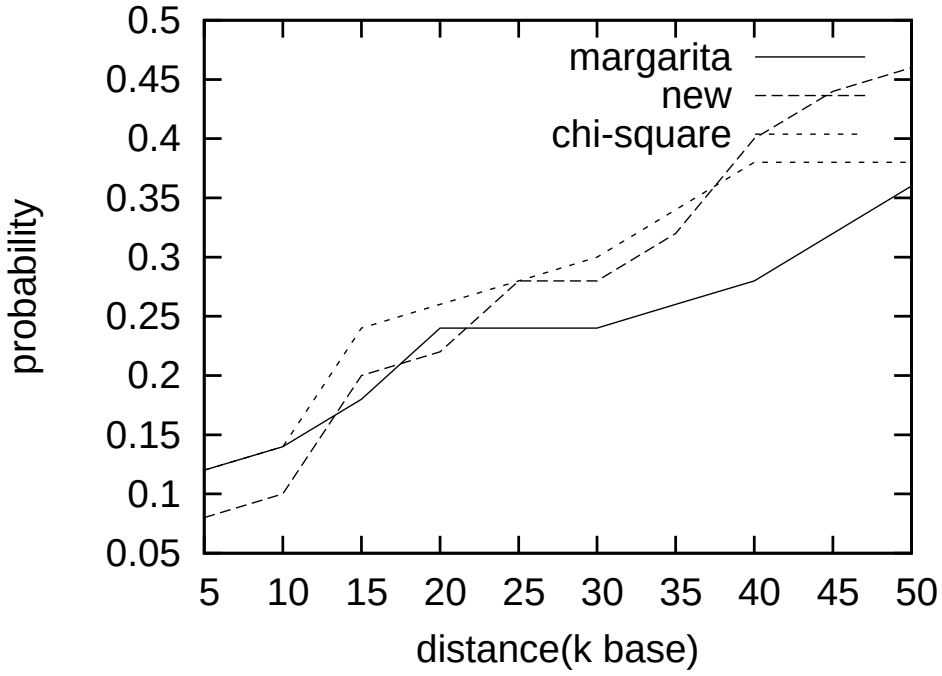

図 4.3. 位置 (塩基数 1.5M、SNP 数 450)

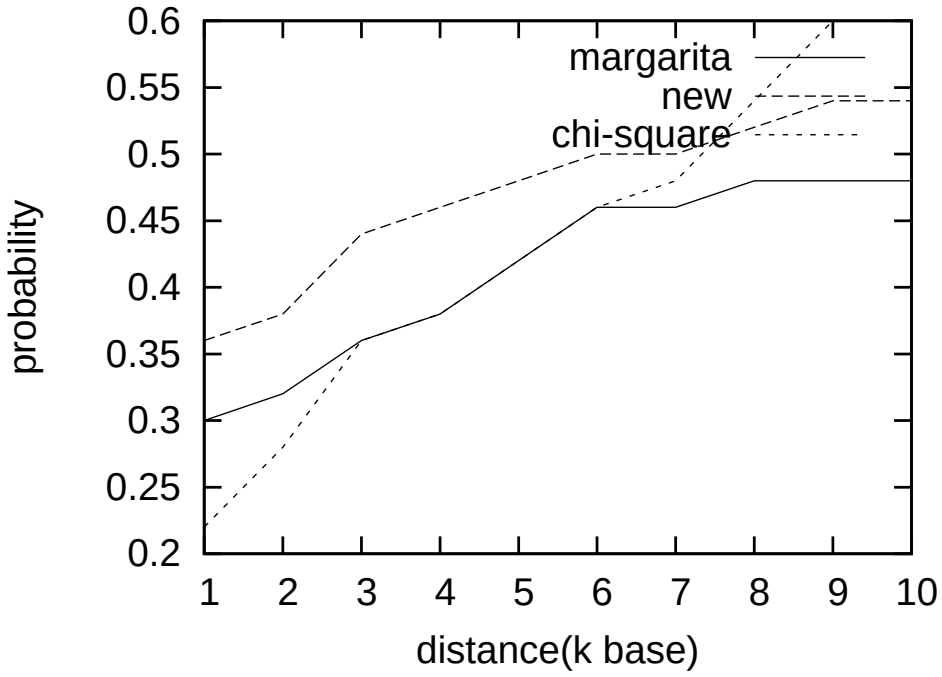

図 4.4. 検出力 (塩基数 1M、SNP 数 300)

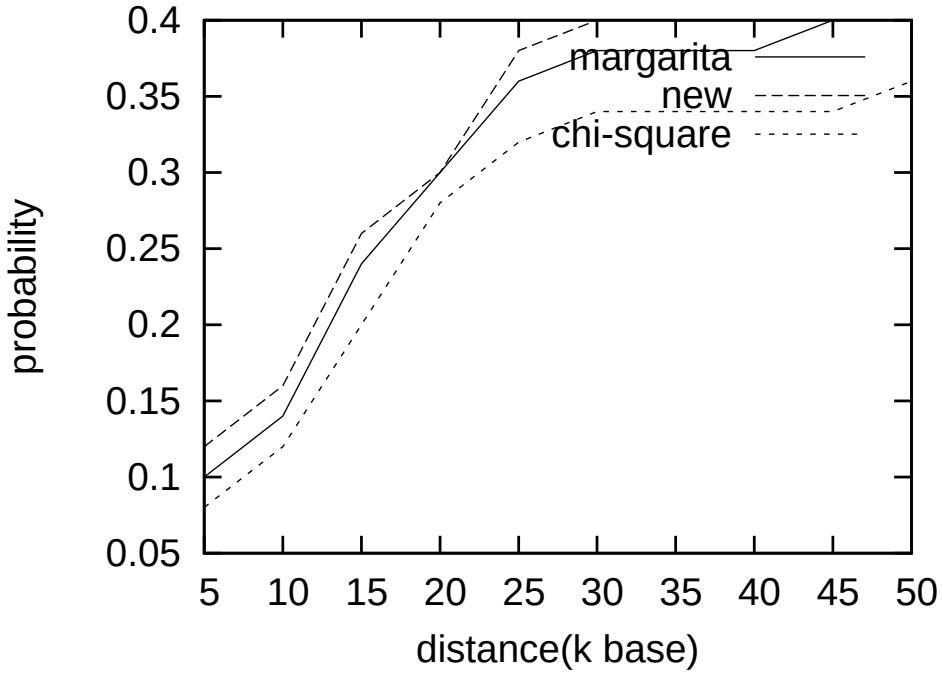

図 4.5. 位置 (塩基数 1M、SNP 数 300)

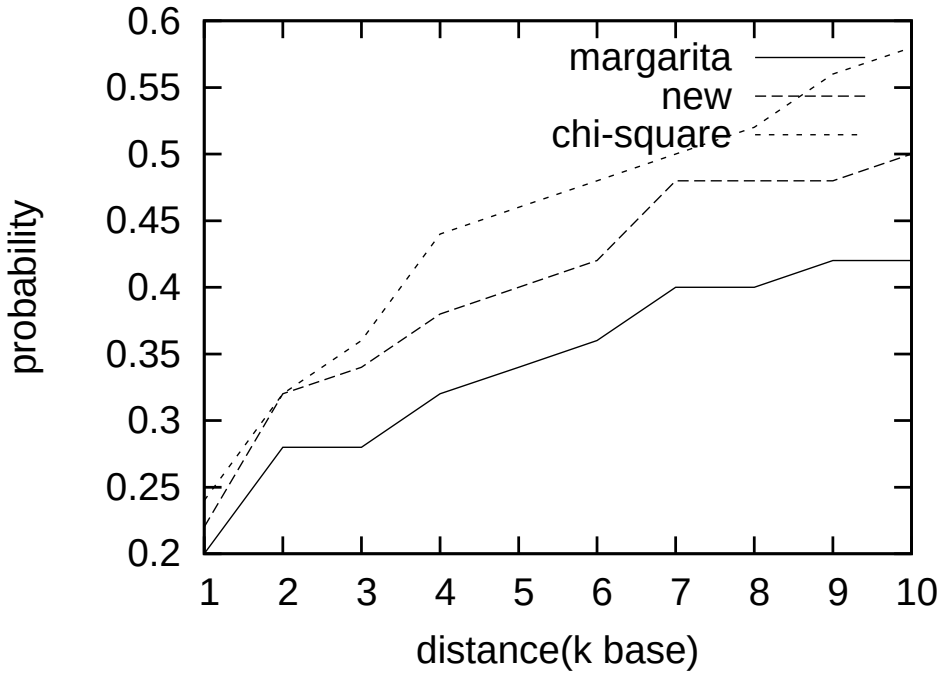

図 4.6. 検出力 (塩基数 0.5M、SNP 数 150)

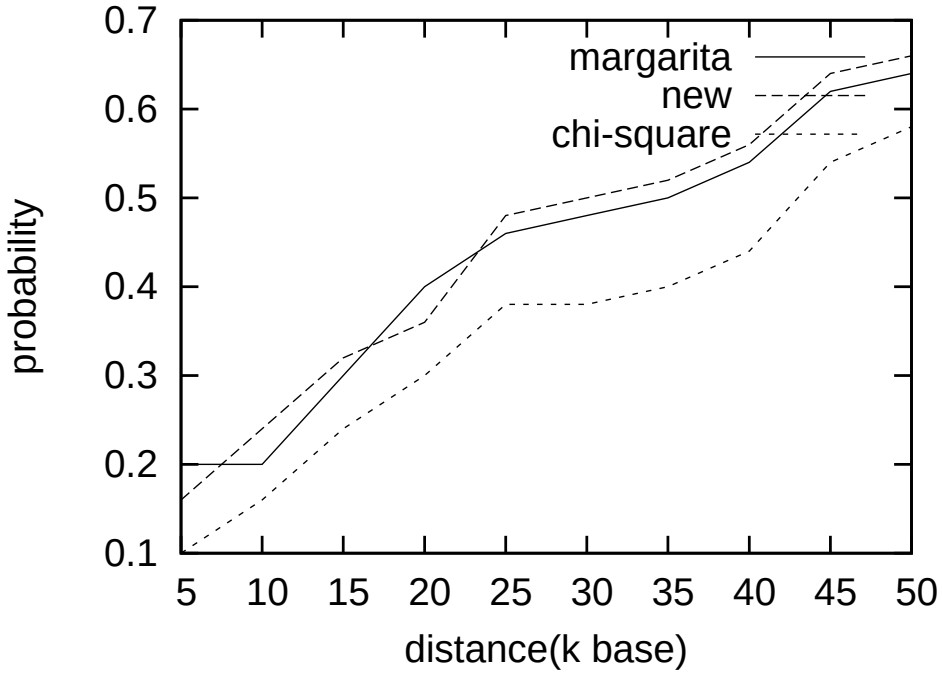

図 4.7. 位置 (塩基数 0.5M、SNP 数 150)

表 4.2. 計算時間の比較 (1M/300SNP)

	Margarita	New
データ 1	417 sec	308 sec
データ 2	446 sec	366 sec
データ 3	431 sec	347 sec
データ 4	400 sec	329 sec
データ 5	428 sec	333 sec
データ 6	421 sec	346 sec
データ 7	431 sec	347 sec
データ 8	417 sec	351 sec
データ 9	428 sec	337 sec
データ 1 0	416 sec	331 sec

表 4.3. 計算時間の比較 (0.5M/150SNP)

	Margarita	New
データ 1	59.1 sec	22.2 sec
データ 2	59.4 sec	24.2 sec
データ 3	64.4 sec	23.9 sec
データ 4	66.6 sec	24.3 sec
データ 5	59.3 sec	21.6 sec
データ 6	63.6 sec	22.5 sec
データ 7	57.5 sec	20.6 sec
データ 8	62.0 sec	20.7 sec
データ 9	57.3 sec	25.0 sec
データ 1 0	60.1 sec	21.5 sec

4.4　考察

配列が長くなると計算時間の減少が見られない原因としては以下のようなことが考えられる。まず考えられるのは、提案手法では共有配列の長さとして領域の半分以上を条件としているが、配列が長くなると、どの共有配列もこの条件を満たさないことが多くなるということである。また、アルゴリズムを変更したことにより、recombination の数が変化したことも原因として考えられる。ARG 構築の計算時間の大半が recombination にかかるため、計算量は recombination 数にほぼ比例する。アルゴリズムを変更したことにより、生じる ARG に変化があれば、計算時間も大きく変動すると考えられる。

そこで、recombination の数について考察する。表 4.4 及び表 4.5 に、ARG 内で生じている recombination の数を比較した結果を示す。計算時間の結果と同様、10 個のデータセットに関してそれぞれ 5 回 ARG を作成し、recombination 数を求めてその平均を求めた。塩基数 1.5M の場合、元のアルゴリズムの recombination の数はおよそ 10500 から 10700 程度であり、提案手法の recombination 数は 9800 から 9900 程度である。また、塩基数 1M の場合、元のアルゴリズムの recombination の数はおよそ 7500 から 7900 程度であり、提案手法の recombination 数は 7100 から 7300 程度である。ただし、元のアルゴリズムの場合、計算時間に大きく影響する recombination 数はこのうち 9 割程度である (1 割は、完全に任意の共有配列を求めているため、ほとんど計算時間はかからない)。すなわち、実際に計算時間にかかる recombination 数は、塩基数 1.5M、1M それぞれのデータに関し、およそ 9500 前後、6800 前後となり、いずれも提案手法より若干少ない値が得られる。

これらの結果から、提案手法の場合、塩基数が長くなるほど各 recombination を求めるのにかかる計算時間が上昇し、1.5M 程度になると、以前のアルゴリズムと比べそれほど変わらなくなると考えられる。すなわち、「共有配列が領域の半分以上」を満たしづらくなると推測される。また、1.5M のデータにおいて計算時間が若干増えたのは、recombination 数の違いからくるものとも推測される。

いずれにせよ、提案手法は塩基数 1M 以内のデータには有効であるが、それ以上のデータに適用する場合にはさらなる工夫が必要であると思われる。

表 4.4. recombination 数の比較 (1.5M/450SNP)

	Margarita	New
データ 1	10681	9875
データ 2	10616	9910
データ 3	10578	9878
データ 4	10691	9924
データ 5	10728	9950
データ 6	10413	9735
データ 7	10527	9753
データ 8	10662	9798
データ 9	10546	9850
データ 1 0	10505	9812

表 4.5. recombination 数の比較 (1M/300SNP)

	Margarita	New
データ１	7579	7109
データ２	7895	7279
データ３	7767	7230
データ４	7616	7128
データ５	7641	7147
データ６	7588	7183
データ７	7658	7168
データ８	7629	7153
データ９	7579	7163
データ１０	7525	7040

第 5 章

まとめ

5.1 結論

本研究では、Margarita の ARG 構築アルゴリズムの計算時間を短縮するため、改良を加えた。実験結果より、塩基数が 1.5M 程度の大きい領域に関しては効果が得られなかったものの、塩基数 1M 以内の領域に関しては、元のアルゴリズムより高速に ARG を推定でき、マッピング精度も従来のアルゴリズムと同精度を保つことができた。

5.2 今後の課題

実験結果からも分かるように、配列が長くなると、ARG 構築に要する計算時間は非常に膨大になる。今回の研究では、塩基数 1.5M 以上の配列に関しては計算時間を短縮できなかったため、長い配列に対して計算時間を短縮させることは大きな課題であると思われる。これを解決するためには、単に共有配列領域が領域の何割以下を満たしているか、ということだけでなく、固定長の長さを条件に加えるなど、さらに複雑なアルゴリズムを考えるする必要があると思われる。

しかし、塩基数 1M 程度の配列に関しても、元のアルゴリズムから大幅に短縮できたとはいい難い点もある。計算資源が限られている場合には、単純にアレル頻度を比較してから、さらに詳しく解析した方がより効率的であるかもしれない。そこで ARG 構築アルゴリズムを変える以外にも、さらなる工夫が必要であると考えられる。具体的には、作成する ARG の数を少ない数で抑えることや、パーミュテーションテストに変わる p 値の計算方法を考えることなどが対策として考えられる。特に、Margarita では ARG を 30 から 100 個程度作成して p 値を求めているが、この必要な ARG 数を見積もり、より少ない数で正確な精度を求めることができれば、かなり大きな効果が得られるのでないかと考えられる。

また、精度を上げることも大きな課題であると考えられる。本研究で提案した手法は必ずしも精度の向上を目的としたものではないものの、若干精度については向上しているものも見られる。全体として精度が向上しているとはいい難いものの、さらなる工夫により精度の向上することも可能であると期待される。recombination の break point を求める際に、heuristic な

手法だけでなく多少の遺伝学モデルを考慮することなどが解決策として考えられる。計算時間を大きく増やさない程度に、より精度の高い手法を開発することは今後の大きな課題であると考えられる。

謝辞

本修士論文作成にあたり、２年間にわたって暖かいご指導を賜わりました森下真一教授に心から感謝いたします。また、研究の途中、貴重な御意見を頂いた森下研究室の皆様に感謝いたします。さらに、実際のデータを頂き、解析に関して助言も頂いた東京大学医学系研究科人類遺伝学教室の徳永勝士教授、宮川卓様及び川嶋実苗様に感謝いたします。

参考文献

[1] D. J. Balding. A tutrial on statistical methods for population association studies. *Nature Reviews Genetics*, Vol. 7, pp. 781–791, 2006.

[2] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testting. *Journal of the Royal Statistical Society B*, Vol. 57, pp. 289–300, 1995.

[3] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, Vol. 29, pp. 1165–1188, 2001.

[4] A. G. Clark. The role of haplotypes in candidate-gene studies. *Genetic Epidemiology*, Vol. 27, pp. 321–333, 2004.

[5] The Wellcome Trust Case Control Contrium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, Vol. 447, pp. 661–683, 2007.

[6] H. J. Cordell and D. G. Clayton. Genetic association studies. *Lancet*, Vol. 366, pp. 1121–1131, 2005.

[7] B. Devlin and N. Risch. A comparison of linkage disequibrium measures for fine-scale mapping. *Genomics*, Vol. 29, pp. 311–322, 1995.

[8] B. Devlin and K. Roeder. Genomic control for association studies. *Biometrics*, Vol. 55, pp. 997–1004, 1999.

[9] C. Durrant, et al. Linkage disequibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. *The American Journal of Human Genetics*, Vol. 75, pp. 35–43, 2004.

[10] S. B. Gabriel, et al. The structure of haplotype blocks in the human genome. *Science*, Vol. 296, pp. 2225–2229, 2002.

[11] J. N. Hirshhorn and M. J. Daly. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, Vol. 6, pp. 95–108, 2005.

[12] C. J. Hoggart, et al. Sequence-level population simulations over large genomic regions. *Genetics*, Vol. 177, pp. 1725–1731, 2007.

[13] D. Y. Lin and D. Zheng. Likelihood-based inference on haplotype effects in genetic association studies. *Journal of the American Statistical Association*, Vol. 101, pp.

89–104, 2006.

[14] J. Marchini, et al. A comparison of phasing algorithms for trios and unrelated individuals. *The American Journal of Human Genetics*, Vol. 78, pp. 437–450, 2006.

[15] G. A. McVean and N. J. Cardin. Approximationg the coalescent with recombination. *Philosophical Transactions of the Royal Society of London*, Vol. 360, pp. 1387–1393, 2005.

[16] A. A. Michell, et al. Undetected genotypeing errors cause apparent overtransmission of common alleles in the transmission/disequibrium test. *The American Journal of Human Genetics*, Vol. 72, pp. 598–610, 2003.

[17] J.. Michell, et al. Comparison of statistical power between 2×2 allele frequency and allele positivity tables in case-control studies of complex disease genes. *Annals of Human Genetics*, Vol. 65, pp. 197–206, 2001.

[18] M. J. Minichiello and R. Durbin. Mapping trait loci by use of inferred ancestral recombination graphs. *The American Journal of Human Genetics*, Vol. 79, pp. 910–922, 2006.

[19] M. NordBorg and S. Tavare. Linkage disequibrium: what histry has to tell us. *Trends in Genetics*, Vol. 18, pp. 83–90, 2002.

[20] J. K. Pritchard, et al. Association mapping in structured populations. *The American Journal of Human Genetics*, Vol. 67, pp. 170–181, 2000.

[21] P. D. Sasinei. From genotypes to genes: doubling the sample size. *Biometrics*, Vol. 53, pp. 1253–1261, 1997.

[22] D. J. Schaid, et al. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *The American Journal of Human Genetics*, Vol. 70, pp. 425–434, 2002.

[23] P. Scheet and M. Stephens. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics*, Vol. 78, pp. 629–644, 2006.

[24] R. Sladek, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, Vol. 445, pp. 881–885, 2007.

[25] O. W. Souverein, et al. Multiple imputation of missing genotype data for unrelated individuals. *The Annals of Human Genetics*, Vol. 70, pp. 372–381, 2006.

[26] R. S. Spielman and W. J. Ewens. The tdt and other family-based tests for linkage disequibrium and association. *The American Journal of Human Genetics*, Vol. 59, pp. 983–989, 1996.

[27] M. Stephens, et al. A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics*, Vol. 68, pp. 978–989, 2001.

[28] I. Tachmazidou, et al. Genetic association mapping via evolution-based clustering of

haplotypes. *Plos Genetics*, Vol. 7, pp. 1163–1177, 2007.

[29] N. Wang, et al. Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population histry, recombination, and mutation. *The American Journal of Human Genetics*, Vol. 71, pp. 1227–1234, 2002.

[30] J. E. Wigginton, D.J. Culter, and G. R. Abecasis. A note on exact tests of hardy-weinberg equibrium. *The American Journal of Human Genetics*, Vol. 76, pp. 887–893, 2005.

[31] S. Zoller and J. K. Pritchard. Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics*, Vol. 169, pp. 1071–1092, 2005.