科研費特定領域研究「情報統計力学の深化と展開」平成 18 年度研究成果発表会

可能性分割アルゴリズムによる haplotype 推定

井上 真郷 [1]
早稲田大学 理工学部 電気・情報生命工学科

1 背景・意義

様々な疾患に対して，現在の医療は臨床症状や生理学的，生化学的，病理学的検査等を組み合わせて診断，分類し，それに応じた治療を行っている．一方，多くの疾患は程度の差こそあれ遺伝的素因が関わっており，それがどの遺伝子なのかということが次々と特定されつつある．また，治療方法として最も一般的な形態である薬剤に関しても，その効果，副作用は遺伝的素因に起因することが多く，関係する遺伝子が特定されつつある．このような状況で，診断，治療のあらゆる局面で，関係すると思われる遺伝子の詳細な型 (個体差) 情報を活用することで，より適切な診断，分類をしたり，より適切な治療方法を選択しようという考え方が個別化医療 (オーダーメイド医療，テーラーメイド医療) として定着しつつある．しかし，個別化医療の前提の一つである個人の遺伝情報の特定が，測定技術の制約から完全にはできないという問題があり，これを情報処理的手法によって推定することを haplotype 推定と呼ぶ．

具体的には，ヒトは染色体に関して 2 倍体の生物で，母及び父由来の計 2 セットの遺伝情報を保持している．医学的に関心があるのは，患者がある特定の遺伝子の DNA 配列に関して例えば A 型と B 型を持っているとか，C 型のみを持っている (この場合は父母から貰った遺伝子が同じだったことを意味する) という情報である．別の医学的知見から，A 型の遺伝子があると将来臨床症状が悪化しやすいとか，B 型の遺伝子がないとある薬剤で副作用が出易い，ということが分かっていれば，この患者がどのような状態にあり (診断)，どうすればどうなるのか (治療結果の予見) ということが分かり，自ずとベストの選択をすることができるであろう．しかし，この遺伝子の型を二つ同定することが技術的に困難なのである．遺伝子の個体差は通常僅かで，もっとも一般的な種類の個体差は一塩基多型 (SNP) と呼ばれ，例えばある集団の中で，ある遺伝子の 30 番目の塩基は AGCT の内の G であることが多いが，A のこともある，といった一塩基に由来する個体差である．SNP はヒトの全染色体で平均すると約 1000 塩基対に一つの割合で存在すると言われ [1]，単体の遺伝子の多くは数百〜数万塩基対でコードされていることを考えると，一遺伝子に 10 箇所から多くて 100 箇所程度の SNP 部位が存在することになる．一箇所の SNP 部位について見ると，集団の中で最も多い塩基は G で 80%，二番目は A で 19.5%，三番目は T で 0.5%，C は観測されなかった，といった分布をしている．通常三番目，四番目の割合は極端に小さいため，情報処理上簡便のため無視することも多い．SNP 以外にも，遺伝子の長さがそもそも異なる，といったタイプの個体差もある．さて，DNA 配列の同定は polymerase chain reaction (PCR) と呼ばれる，染色体上の特定の部位の遺伝子を増幅させる方法がベースとなるが，この方法が，父母由来の遺伝子を区別できずに，両方とも同時に増幅させてしまうことが問題の本質である．増幅された遺伝子は，その後端から順に一塩基ずつ，AGCT の何れであるかを調べていける．この時に，父母由来の遺伝子が同じであれば (homozygote) 良いが，異なる (heterozygote) 塩基については，例えば G と A が両方ある，という形で検査結果が出てくる．heterozygote の箇所が 0 又は 1 ならば，この患者の持つ二本の塩基配列を同定できるが，それ以上の場合は SNP 部位の組合せになり，一意に同定できない．一人のデータからでは，どの組合せが尤もらしいかは均等で推定を行うことは不可能であるが，遺伝的に近縁の多数の個体のデータがあれば，互いに同じ塩基配列を共有している可能性が高いという仮定から，各個人について尤もらしい二本の DNA 配列を推定することがある程度可能となる．

既存の手法は概ね Bayes 推定の枠組みで説明できる．主な差異は，集団内の haplotype 頻度 (どの塩基配列がどの程度の割合で集団内に存在しているか) にどのような事前分布を割り当てるかで，1) Dirichlet 分布，2) coalescent モデル，の二つが主流である．1) は EM アルゴリズムを用いる手法が幾つかあり [2, 3]，これらは Dirichlet 分布のパラメータをある値に固定した時の MAP 解と本質的に同等である [6]．SNP の箇所数が少ない場合は良好な推定結果を返すが，計算量，メモリ必要量が 2 の箇所数乗に比例するため，箇所数が 20 程度までが限界となる．これを改善するため，問題を様々に分割して解く手法も提案されているが，使用する近似のため推定精度は低下する．2) の coalescent モデルは陽に事前確率を与えるものではなく，個々の可能性同士の遷移確率を与えるものであるため，MCMC シミュレーションと組み合わせて近似的に事後分布を得るものである [4, 5, 6]．遺伝的な変異をより忠実にモデルに取り込んでいるため，推定精度は若干上回るものと期待される．一方欠点としては，シミュレーションの計算量が膨大であること，収束判定基準の決め方が難しいなどが挙げられる．

[1]E-mail: masato.inoue@eb.waseda.ac.jp, 本文書は 2006 年 11 月 14 日公開

本研究で扱う手法は 1) の EM アルゴリズムと前提としているものが似通っているため，相性が良い．EM アルゴリズムの計算限界は様々な方法で解決が図られて来たが，ここでは全く近似を行うことなく解決する手法を提案する．直感的には，$\mathcal{O}(2^N)$ の可能性の内，互いに同等な可能性同士をグルーピングすることでより低いオーダー留めるという操作を行う．本手法は基本的に EM アルゴリズムと組み合わせることで，EM アルゴリズムと基本的に同等の答をメモリ消費量・計算量を低く抑えながら算出することが可能で，結果的に EM アルゴリズムの計算限界を遥かに伸ばすことができる．

2 モデル

本手法は必ずしも EM との組み合わせが必須ではないが，ここでは成功例として EM をベースに説明する．

haplotype 推定問題は次のように置き換えることができる．即ち，被験者 I 人がそれぞれ N 次元ベクトル (ここではカードと呼ぶ) を二枚持っているとし，i 番目の被験者の j 番目のカードを $d_{i,j}$ で表す．ここでは 1 枚目か 2 枚目かは区別し，1 枚目を母親由来，2 枚目を父親由来の遺伝子に対応させて表記する．カードに書かれている内容が同じならば，それらのカードは区別しない．N は SNP の箇所数に，n 次元目の値は n 番目の SNP 部位の塩基に対応する．各次元の値は，1, 2, 4, 8 の何れかで，1 なら集団中一番多かった塩基，2 なら 2 番目，4 なら 3 番目，8 なら 4 番目に多かった塩基を表すとする．一方，観測できるのは各人について二枚のカードの論理和 ($\oplus$ で表す) である．例えば $N = 5$ として，i 番目の被験者の二枚のカードを

$$d_{i,1} \equiv [1, 2, 1, 4, 2], \tag{1}$$
$$d_{i,2} \equiv [1, 2, 2, 1, 1] \tag{2}$$

とすると，観測データ x_i は，

$$x_i = d_{i,1} \oplus d_{i,2} = [1, 2, 3, 5, 3] \tag{3}$$

となる．これについてもカードと呼ぶが，区別する時は $d_{i,j}$ を単独カード，x_i を合成カードと呼び分ける．単独カードに書かれる個々の数字は 1,2,4,8 の 4 通りだが，合成カードに書かれる数字は 1,2,3,4,5,6,8,9,10,12 の 10 通りである．すると，haplotype 推定問題は，全員の観測データ，つまり I 枚の合成カード $\{x_i\}_{i=1,2,\ldots,I}$ から全員の単独カード計 $2I$ 枚 $\{\boldsymbol{d}_i\}_{i=1,2,\ldots,I}$(但し $\boldsymbol{d}_i \equiv [d_{i,1}, d_{i,2}]$) を推定する問題となる．

集団内で遺伝子は十分混ざり合っているとして Hardy-Weinberg 平衡を仮定すると，被験者がどの 2 枚のカードを持つかは互いに独立で，

$$P\left(\boldsymbol{d}_i \middle| \boldsymbol{y}\right) \equiv \prod_j^2 y_{d_{i,j}} \tag{4}$$

となる．h は任意の単独カードを表し，y_h は集団内で h の占める割合，$\boldsymbol{y}$ は全ての y_h を要素とする 4^N 次元のベクトルである．各人の単独カードが与えられれば，観測データは一意に決まるので，

$$P\left(x_i \middle| \boldsymbol{d}_i\right) \equiv \delta_{\boldsymbol{d}_i \in \mathcal{D}(x_i)} \tag{5}$$

となる．δ は指示関数で，条件が真の時 1，偽の時 0 となる．$\mathcal{D}(x_i)$ は，二枚の単独カードについて $d_1 \oplus d_2 = x_i$ となるような全ての並び $\boldsymbol{d} \equiv [d_1, d_2]$ の集合を表す．被験者が直接の血縁関係にないと仮定すると，各人の間でもこれらの確率は独立であり，同時分布

$$P\left(\{x_i\}, \{\boldsymbol{d}_i\} \middle| \boldsymbol{y}\right) = \prod_i^I \delta_{\boldsymbol{d}_i \in \mathcal{D}(x_i)} \prod_j^2 y_{d_{i,j}}. \tag{6}$$

を得る．

EM アルゴリズムは，未観測データを含む実現値に基づいて確率分布パラメータの最尤推定を行うもので，推定量

$$\hat{\boldsymbol{y}} \equiv \arg\max_{\boldsymbol{y}} \left\langle \ln P\left(\{x_i\}, \{\boldsymbol{d}_i\} \middle| \boldsymbol{y}\right) \right\rangle_{P\left(\{\boldsymbol{d}_i\} \middle| \{x_i\}, \boldsymbol{y}\right)}. \tag{7}$$

を得ることを目的とする．$\langle \bullet \rangle$ は期待値を表す．更に，この式は簡単には解けないことが多いので，代入を反復的に行うことで数値的に解くものである．

$$\boldsymbol{y}^{(t+1)} \equiv \arg\max_{\boldsymbol{y}} \left\langle \ln P\left(\{x_i\}, \{\boldsymbol{d}_i\} \middle| \boldsymbol{y}\right) \right\rangle_{P\left(\{\boldsymbol{d}_i\} \middle| \{x_i\}, \boldsymbol{y}^{(t)}\right)}. \tag{8}$$

上記 EM 流の対数尤度は，極大値への収束が保証されるが，最大値への収束は保証されない．アルゴリズム自体は非確率的であり，結果は初期値及び繰り返し回数 t のみに左右される．初期値に特殊な値を設定すると鞍点に収束することがあるが，これは測度ゼロと思われる．

上記 EM アルゴリズムはハイパーパラメータ $\boldsymbol{y}$ についての点推定を行うもので，$\{\boldsymbol{d}_i\}$ についての推定ではない．パラメータ $\{\boldsymbol{d}_i\}$ を点推定するには，

$$\{\hat{\boldsymbol{d}}_i\} \equiv \arg\max_{\{\boldsymbol{d}_i\}} P\left(\{\boldsymbol{d}_i\}\middle|\hat{\boldsymbol{y}}\right) \tag{9}$$

を行うものとする．最終的には，$d_{i,1}$，$d_{i,2}$ は並びではなく組として正解すればよい．

3 方法

3.1 可能性分割の概念

EM アルゴリズムの更新式を書き下すと

$$y_h^{(t+1)} = \frac{1}{I}\sum_i^I \frac{\delta_{h\in x_i} y_h^{(t)} y_{x_i\ominus h}^{(t)}}{\sum_{h\in x_i} y_h^{(t)} y_{x_i\ominus h}^{(t)}} \tag{10}$$

となる．但し，$h \in x_i$ のように表記した場合は，合成カードはそれが含む可能性のある単独カード全てを要素とする集合を表すものとする．$x_i \ominus h$ は単独カード h と合わせると合成カード x_i になるような単独カードを表すものとする (但し，後でこの定義を拡張する)．

y_h の内，どの x_i にも含まれない h については $y_h = 0$ なので最初から計算しない等の工夫をしても，観測データが heterozygote を多く含んでいる場合は忽ち計算限界に達する．即ち，合成カードが 3,5,6,9,10,12 の数字を計 n 個含んでいるとすると，x_i の要素数は 2^n となり，上記更新式はメモリ容量及び計算量の制約から $n = 20$ 程度で計算限界を迎える．これを突破するのが本論文の主目的である．

更新式の対称性から，各 y_h の値は，互いに入れ替えても区別できないような最小単位のグループに分割できそうである．この最小単位グループ内では，全ての i について $\delta_{h\in x_i}$ の値が同じであり，かつ，全ての i について $x_i \ominus h$ は互いに同じグループに属していればよい．即ち，y_h と $y_{h'}(h \neq h')$ についてこれらの条件が全く同じであれば，両者には入替対称性があり，初期値を入れ替えれば収束値も入れ替わる．

一般に，二つの haplotype の関係に，単点突然変異 (single site mutation) の関係なら「近い」，そうでなければ「遠い」とか，ある haplotype は別の二つの haplotype を 1 回交叉 (crossover) させると一致するので「近い」などといった関係を考慮しない場合，各 haplotype は上記包含可能性によってのみ区別でき，包含可能性が同一なら区別する必要が無い．この考えを推し進めると，単独カードの全ての可能性を，包含可能性に基づいて適切に分割・グルーピングして，グループ単位で情報処理をすれば十分だということになる．

3.2 定義及び構成法

このようなグループを要素として含む集合 $\mathcal{G}$ の十分条件を考える ($\mathcal{G}$ は集合の集合である)．いきなり集合内の各グループが互いに排反というのは難しいので，最終的なグループを包含するような大きなグループの存在も許した集合 $\mathcal{G}'$ を先に考える．まず，x_i に含まれるか否かを区別する必要があるので，これらを要素と認定する (これらの要素を便宜上原グループと呼ぶことにする)．

$$x_i \in \mathcal{G}' \tag{11}$$

ある要素 g が含まれていれば g の補集合 $\bar{g}$ も要素として含む必要があるので

$$\bar{g} \in \mathcal{G}' \quad (\forall g \in \mathcal{G}') \tag{12}$$

とする (これらを補グループと呼ぶことにする)．また，既存の二つのグループの交わりが存在するならば，それも区別する必要があるので，

$$g \otimes g' \in \mathcal{G}' \quad (\forall g \in \mathcal{G}',\ \forall g' \in \mathcal{G}') \tag{13}$$

とする (これらを共通グループと呼ぶことにする)．$g \otimes g'$ は集合 $g \cap g'$ を表すものとする．また，既存グループが x_i に包含されていた場合，x_i についてそれと対になるグループも区別しなければならないので，

$$x_i \ominus g \in \mathcal{G}' \quad (\forall i,\ \forall g \in \mathcal{G}') \tag{14}$$

$(a)_n$	$(b)_n$	$(a \oplus b)_n$
x	x	x
ohters		3

$(a)_n$	$(b)_n$	$(a \otimes b)_n$
x	x	x
3	x	x
x	3	x
ohters		0

$(a)_n$	$(b)_n$	$(a \ominus b)_n$
x	x	x
3	1	2
3	2	1
ohters		0

Table 1: カード同士の演算の定義．各カードに書かれている各次元の値についてそれぞれ独立に演算を行う．答に一つでも 0 が含まれる場合は，答は空集合とする．この表では簡単のため，単独カードの数字は 1,2 のみに，合成カードの数字は 1,2,3 のみに限定している．x は，各行について同一の値 (1,2,3 の何れか) を表す．

とする (これらを対グループと呼ぶことにする)．$x_i \ominus g$ は，g に含まれる何れかの単独カードと組み合わせると合成カード x_i となるような単独カードの集合を表すものとする．

これらを満たす集合 $\mathcal{G}'$ を求め，その後要素間について包含関係にあるグループは，包含している方のグループを全て取り除くことにすれば，

$$\mathcal{G} \equiv \{g|\ g \in \mathcal{G}', g' \not\subset g\ \forall g' \in \mathcal{G}'\} \tag{15}$$

互いに排反なグループのみで単独カードの全ての可能性を分割・網羅することができる．

集合 $\mathcal{G}$ は以下の手順で構成することができる．まず，$\mathcal{G}'$ は空集合とし，観測データを全て候補集合 $\mathcal{C}$ に含める．

$$\mathcal{G} = \phi \tag{16}$$
$$\mathcal{H}(x_i) \in \mathcal{C} \tag{17}$$

続いて，$\mathcal{C}$ に要素がなくなるまで，以下の操作を繰り返す．$\mathcal{C}$ より要素を一つ (c) 取り出し，補グループ

$$\bar{c} \tag{18}$$

及び $\mathcal{G}'$ の各要素 g との間で新たに生ずる共通グループ，対グループ

$$c \otimes g \tag{19}$$
$$c \ominus g \quad (\text{if } c = \mathcal{H}(x_i)) \tag{20}$$
$$g \ominus c \quad (\text{if } g = \mathcal{H}(x_i)) \tag{21}$$

が c と異なり，かつ $\mathcal{G}'$ にも $\mathcal{C}$ にも含まれていない時，これらを $\mathcal{C}$ に追加し，c を $\mathcal{G}'$ に加える．

以上の構成法により，$\mathcal{G}'$ は常に，どの要素についてもその補集合を，またどの二要素についてもそれらの共通グループ・対グループを，$\mathcal{G}'$ 又は $\mathcal{C}$ に含ませながら成長させることができ，最終的に $\mathcal{C}$ の要素は最終的には必ず $\mathcal{G}'$ に加えられる．$\mathcal{G}'$ から $\mathcal{G}$ を構成するには，定義通りに $\mathcal{G}'$ の要素を一つ一つ調べ，どの要素も包含していないなら $\mathcal{G}$ に加えるという操作を全ての要素について行えばよい．これで $\mathcal{G}$ の定義を満たす集合を構成できた．

3.3 内部表現及び構成法の改善

上記の単純な構成法は，現実的なデータサイズでは忽ちメモリ容量オーバー及び計算量爆発を起こし，EM アルゴリズムよりも却って非現実的なほどである．そこで，以下のように改善を行い，$\mathcal{G}'$ の代わりに $\mathcal{G}^*$ を構成することにする．

まず，$\mathcal{G}^*$ の要素数を減らすため，補グループを要素とすることを止め，必要な場合は随時 $\mathcal{G}^*$ の要素を組み合わせて計算により求めるものとする．

すると，$\mathcal{G}^*$ の要素であるグループは，一々単独カードを含んでいる枚数分記憶せずとも，全て合成カード一枚の情報で代用することができる．一枚の単独カードは $2N$[bit] を必要とし，合成カードは $4N$[bit] を必要とするので，要素数 3 以上のグループについてはメモリの節約になる．また，共通グループ，対グループなどの算出も楽になる．具体的には，$\mathcal{G}'$ と同様に以下を定義する．

$$x_i \in \mathcal{G}^* \tag{22}$$
$$g \otimes g' \in \mathcal{G}^* \quad (\forall g \in \mathcal{G}^*,\ \forall g' \in \mathcal{G}^*) \tag{23}$$
$$x_i \ominus g \in \mathcal{G}^* \quad (\forall i,\ \forall g \in \mathcal{G}^*) \tag{24}$$

$\otimes$ や $\ominus$ の演算結果は幸いにも必ず一枚の合成カードで表すことができる (表 1 参照).

次に，構成法を改善することを考える．先に述べた単純な構成法では，共通グループを作成する組合せ数が膨大となるが，実際に共通グループを持つことは少ないという無駄がある．また，新しく作られたグループが $\mathcal{G}'$ 又は $\mathcal{C}$ に既に含まれているかどうかの検査も時間がかかる，などの欠点が見られるので，これらを改善する．

共通グループが存在するか否かは，$\mathcal{G}^*$ の要素を包含関係が簡単に分かるような構造で保持することで，比較的容易に可能性を限定することができる．これには，feedforward 型ネットワークのような構造を採用し，出口から順に作っていくような手順を採用する．具体的には最初にルートノードを設けておき，最初に $\mathcal{G}^*$ に加えられるグループはルートノードに直接ぶら下がるノードとする．以後，既に存在するグループに包含されるグループはその下に，どのグループにも包含されない場合はルートノードの下に追加していく．この包含関係をここでは親，子と呼び，同じ親に属する子同士を同胞と呼ぶことにする (遺伝関係の親子とは関係ないので注意)．場合によっては複数の親が存在することがあるが，その場合は，親同士が更に包含関係にあれば最も小さい親以外はリンクを張らないものとする．それでも尚複数の親が残ることはよくあり，その場合はそれら全てとリンクを張る (このために，このネットワークは木構造とはならない)．また，グループをネットワークに加えていく順番によっては，既存の親子関係のリンクに新しいノードが割り込むこともあるが，これは後述する別のルールを設けてこのようなことが起こらないように工夫する．

このネットワークは，全てのノードを一巡したい場合に不都合なので，ツリー構造も同時に取り入れることにする．即ち，リンクを二種類，必須リンクと冗長リンクに分類し，必須リンクのみの場合はツリー構造になるように拡張する．リンクの種類を決定するには，新しいノードを追加する際に，親ノードの内の一つを何らかの基準で選択し，その親へのリンクを必須リンク，それ以外の親へのリンクを冗長リンクとする．必須リンクはルートノードから見て分岐することはあっても合流することはないので，木構造を達成できる．

次にグループ g の rank を

$$\mathrm{rank}(g) \equiv \log_2 |g| \tag{25}$$

と定義する．$|\bullet|$ は集合の要素数を表す．ここでは各グループは一枚の合成カードで表されているので，rank はカードに書かれている 1,2,4,8 以外の数字の数である．詳細は割愛するが，rank は色々と細かい高速化で使い勝手がよい．

rank を利用して，$\mathcal{C}$ から要素 c を抜き出す際に，rank の高いものから順に抜き出すことにする．新しく作られる共通グループや対グループの rank は必ず $\mathrm{rank}(c)$ 以下になるため，$\mathcal{G}$ に既に含まれるグループの rank は必ず $\mathrm{rank}(c)$ 以上となる．すると，先のノードの割り込みは，既にあるノードよりも高い rank のノードが加えられることによって発生する可能性が初めて生ずるものであるので，ノードの割り込みを完全に防ぐことが可能になる．また，$c = x_i$ の場合も，$c \ominus g = \phi$ $(\forall g \in \mathcal{G}^*)$ なので，計算する必要が無くなる．

以上のネットワーク構造及び追加順番を前提にすると，共通グループを作成する際，$c \in \mathcal{C}$ の相手として調べなければならないのは全ての $g \in \mathcal{G}^*$ ではなく，この後 c がネットワークに追加される時の同胞のみでよいことになる．何故なら，c の親や，親の親... は c を包含しているので，$c \otimes g = c$ となるので不要である．これ以外のグループは，c の親との共通グループを持たない場合は $c \otimes g = \phi$ であり，共通グループを持つ場合は，その共通グループ (即ち c の同胞) との交わりを調べれば十分だからである．これにより，無駄な共通グループ (空集合や既に作成してあるグループ) を作ってしまう事を大幅に削減できる．

また，新たに作られる共通グループは必ず rank が低いものになるため，$\mathcal{G}^*$ に既に存在している可能性はなく，$\mathcal{C}$ 内に同じものがあるかどうかだけ調べればよくなる．この他にも，ハッシュテーブル，並べ替えされたリストを用いて集合操作を高速化する等の改善が可能である．

次に，補グループを計算で求める方法を示す．最終的に必要なグループは，互いに排反なもののみで十分なことから，他のグループを包含するグループについて，それらのグループを除いた集合を求めればよい．具体的には，ネットワーク構造を利用して，子グループの補集合との交わりを計算すればよい．

$$g \cap \bigcup_{g' \in \mathrm{children}(g)} \overline{g'} \tag{26}$$

ルートノードは全ての単独カードの可能性を含むと見做せば，どの x_i にも含まれない単独カードのグループも求めることが可能である．これらの補集合に基づくグループは，一枚の合成カードで表現することが一般には不可能で，データによっては非現実的なメモリ量・計算量を必要とする可能性がある．一方，後述するように EM アルゴリズムとの併用に際しては，これらのグループはそもそも計算する必要がない，ということもある．

3.4 EM アルゴリズムの併用

本手法を EM アルゴリズムと組み合わせた実装例を示す．EM アルゴリズムの収束値は初期値 $\boldsymbol{y}^{(0)}$ に依存するため，ランダムな初期値で収束させることを何度か繰り返し，収束時の EM 流対数尤度

$$2I \sum_h \hat{y}_h \ln \hat{y}_h \tag{27}$$

が最大となる答を選択する方法が妥当である．この目的関数はエントロピーの負定数倍であり，EM アルゴリズムは制約条件付で分布 $\boldsymbol{y}$ のエントロピーを最小化することを目的としている．

ところで，排反化した各グループ $g \in \mathcal{G}$ について，同一グループに属する複数の単独カード h, h' について考えると，EM アルゴリズムの結果が $\hat{y}_h > 0, \hat{y}_{h'} > 0$ となることは通常有り得ない．これの直感的な説明は，エントロピーを最小化させるためにはどちらかに確率を集中させた方が良いからである (初期値を $y_h^{(0)} = y_{h'}^{(0)}$ とした場合や，このグループが原グループに一致する $g = x_i$ 場合などを除く)．このことを応用すると，EM アルゴリズムを使用するに当たり全ての単独カードを用いる必要はなく，各グループについて一枚を代表させれば十分だということになる (但し，EM アルゴリズムのダイナミクスは更新に従い複雑に相互作用するため，収束値の考察のみからは最適性は保証できない)．

更に，単独カード h の確率 y_h はより多くの x_i に含まれている程，値が大きくなる (正確には小さくならない) ことが更新式から分かる．従って，$g \subset g'$ とした時，補グループ $\bar{g} \cap g'$ に属する単独カードは g に属する単独カードに比べて同等又は確率が低く，点推定の立場からは最初から計算対象から外すのが得策である．

以上のことから，最終的には $\mathcal{G}^*$ の要素について，ネットワーク構造から，子を持っていない要素のみを抜き出して新たな集合

$$\mathcal{G}^{**} \equiv \{g |\ g \in \mathcal{G}^*, g' \not\subset g\ \forall g' \in \mathcal{G}^*\} \tag{28}$$

を構成し，$\boldsymbol{y} \equiv \{y_g\},\ g \in \mathcal{G}^{**}$ について

$$y_g^{(t+1)} = \frac{1}{I} \sum_i^I \frac{\delta_{g \subset x_i} y_g^{(t)} y_{x_i \ominus g}^{(t)}}{\sum_{g \subset x_i} y_g^{(t)} y_{x_i \ominus g}^{(t)}} \tag{29}$$

を計算すればよい．これにより，$\boldsymbol{y}$ の次元数は $|\bigcup_i^I x_i|$ から $|\mathcal{G}^{**}|$ に減らすことができ，オリジナルの EM アルゴリズムではメモリ容量・計算量の制約から解けなかったサイズの問題が解けることを期待するものである．

一般に，この可能性分割アルゴリズムを用いることで，EM アルゴリズムの正答率が期待値として低下する理由は無い (寧ろ，試行回数が制限されている状況では若干上昇するかもしれない)．一方，大規模データでの可能性は低いと思われるが，データによっては，$\mathcal{G}^*$ が殆ど全ての部分集合を含むことになるなどの理由から，本来の EM アルゴリズムよりも多くのメモリを必要とすることが有り得る．極端な例としては，$\mathcal{G}^*$ は最大で全ての合成カード 10^N 枚を含む可能性があるが，単独カードの可能性は最大 4^N 通りである．

4 結果・考察

本手法について人工データを用いて性能を検証し，EM アルゴリズムでも解けるサイズのデータについては EM アルゴリズムと同等の正答率を得た．また，EM アルゴリズムでは解けないサイズのデータについても現実的なメモリ消費・計算時間で合理的な正答率を得た．

人工データについては，以下の手順で作成した [3]．まず，予め決めた SNP 部位数を用いて，第 1 世代として 30 人分，計 60 枚の単独カードを互いに独立にランダムに作成した．単独カードの数字は 1 又は 2 を等確率でランダムに決めた．次に 100 世代を経過させた．具体的には，前の世代からランダムに二人を選び，更にそれぞれランダムに二枚の内の一枚のカードを複製して次世代の一人とした．また，確率 10^{-5} で，カードを複製する際にランダムに一つの数字を写し間違えて 1 を 2 に，2 を 1 にした (単点突然変異)．また，確率 10^{-3} で，カードの複製をする際に，数列のランダムな位置で，数列を写す元のカードを同じ親のもう一枚のカードに切り替えた (交叉)．第 2，第 3 世代では人口を 2 倍に増やし，それ以降は 1.05 倍に増やしていった．第 101 世代の中からランダムに予め決めた人数を抜き出し，haplotype 推定の対象とした．

上記人工データを用いて検証を行った (表 2 参照)．まず，EM アルゴリズムで解くことの出来るサイズ，SNP 部位数 10，被験者数 100 人のデータを 1000 セット作成し，それぞれ解かせてみた．結果，EM アルゴリズムの誤答率は 2.48%，本手法も同じく 2.48%であった．各データセットについてみると，誤答率は両手法間で多少の違いがあったが，概ね同じ結果であった．一般に誤答正答は初期値依存性があるため，原理的に必ず一致するものではない．両者は本質的に同等の性能を有している筈で，この結果は期待通りである．

	EM アルゴリズム	提案手法
平均誤答率	2.48%	2.48%
平均計算時間	0.14 秒	2.16 秒
平均 $\|\boldsymbol{y}\|$	843	819
平均 $\|\mathcal{G}^*\|$	-	13403

	提案手法
平均誤答率	0.31%
平均計算時間	77 秒
平均 $\|\boldsymbol{y}\|$	500
平均 $\|\mathcal{G}^*\|$	287700

Table 2: 人工データでの計算結果．左は SNP 部位数 10，被験者数 100 人での 1000 回の平均結果，右は SNP 部位数 100，被験者数 100 人での 100 回の平均結果を表す．$|\boldsymbol{y}|$ は EM アルゴリズムの際の次元数を表す．

次に，EM アルゴリズムでは解くことの出来ないサイズ，SNP 部位数 100，被験者数 100 人のデータセット 100 セットについて，提案手法で解かせてみた．結果，誤答率は 0.31%であった．また，CPU は Intel® Pentium4® 3.2GHz，メモリは 1GByte のシステムを用いて，計算時間は 1 セット当り平均 77 秒であり，メモリ容量も十分であった．比較的現実的なデータに近い人工データを，このようなパソコンレベルのシステムで解くことが出来たため，本手法はかなり実用性の高い手法であると思われる．

本手法では，最終的に EM アルゴリズムに渡すグループ数は小さく抑えられるものの，途中の $\mathcal{G}^*$ の要素数が多くなることが欠点である．このため，被験者数等データサイズが増加した場合は，メモリ容量の限界がボトルネックとなる可能性が高いと思われる．

単一の遺伝子について haplotype 推定する場合は，SNP 部位数は 100 程度で十分だと思われる．しかし，遺伝子によって SNP 頻度はかなりのばらつきがあることが示唆されているため [1]，予断を許さない．被験者数については今後数万程度には増える可能性が高いと思われる．また，突然変異や交叉の率についても，染色体上の部位によって大きな変動があると言われている．以上より，検証作業は今回用いた設定のみでは不十分で，広範な設定での人工データ，及び実データによる計算限界の検証が今後更に必要である．

5 結論

本研究は，haplotype の様々な可能性を推定するに当たって，観測データからは区別できない haplotype 同士をグループとしてまとめて扱うことで，メモリ不足及び計算量爆発を緩和するものである．推定性能の比較的良い EM アルゴリズムは，SNP 部位数 20 程度で計算限界を迎えるが，本手法と組み合わせることで，本質的に EM アルゴリズムと同等の性能を担保しながら，扱える SNP 部位数を 100 程度と，大幅に増加させることができた．データによっては，本アルゴリズムの方が先に計算限界を迎える可能性があるが，現実のデータでそれが起こる可能性は低いと思われる．今後，幅広いデータで，計算限界の検証を行うことが必要である．

謝辞

本研究は東京女子医科大学の鎌谷直之氏，株式会社ジェネシス・テクノロジーズの田中順治氏，及び，早稲田大学確率的情報処理研究室の進藤裕之君，千明裕君との共同研究である．

本研究は文部科学省科学研究費補助金 特定領域研究 課題番号 18079012，基盤研究 (A) 課題番号 17200016，早稲田大学定課題研究助成費等の助成を受けた．

本手法については，成果を医療分野に広く還元したいため，特許を申請しない．

References

[1] J. C. Venter, M. D. Adams, et. al: Science **291** 1304, 2001.

[2] L. Excoffier, M. Slatkin: Mol. Biol. Evol. **12** 921, 1995.

[3] T. Niu, Z. S. Qin, X. Xu, J. S. Liu: Am. J. Hum. Genet. **70** 157, 2002.

[4] M. Stephens, N. J. Smith, P. Donnelly: Am. J. Hum. Genet. **68** 978, 2001.

[5] Z. Qin, T. Niu, J. S. Liu: Am. J. Hum. Genet. **71** 1242, 2002.

[6] M. Stephens, P. Donnelly: Am. J. Hum. Genet. **73** 1162, 2003.