A-15

多遺伝子座対応可能なハプロタイプ推定アルゴリズム

A New Haplotype Estimation Algorithm Applicable to Many SNPs Inputs

下里 二郎　　甲藤 二郎
早稲田大学 大学院 理工学研究科

1. はじめに

ヒトゲノムの約30億塩基対配列の解読が報告されて以来、塩基配列情報を利用した研究への興味が高まっている。特に塩基配列中のどの部分がどのような形質と関与しているのかを特定する作業は、重要な研究の対象となっている。これを形質と配列上の対応をつけることから、形質マッピングという[1]

形質マッピングのアプローチとして遺伝統計学的手法によるものがある。遺伝統計学的ゲノム解析では2本の染色体のそれぞれに遺伝子がどのように並んでいるかを示すハプロタイプを推定することが重要である。これには、血縁関係のない個人遺伝子型データを用いてEM手法に基づく最尤評価手法[2][3]の適用が有効だが、このアルゴリズムを20遺伝子座を超える多遺伝子座に単純に適用した場合、SNPの組み合わせ数が膨大な量となり、現在の計算機で許容できる時間内にハプロタイプを推定することは困難である。

そのため、20以上の多遺伝子座の遺伝子型データを入力データとしてハプロタイプを推定するには、新たなアルゴリズムが必要である。本稿では、100遺伝子座に対応できると考えられる新たなアルゴリズムの提案を行い、既存のアルゴリズムとの実行結果の比較を行うことにより提案アルゴリズムの評価を行う。

2. ハプロタイプ推定

2. 1 遺伝子型とハプロタイプの定義

遺伝子型とは2本ずつ存在する染色体の同じ位置(遺伝子座)にある遺伝子の組み合わせのことである。この遺伝子型の組み合わせをアレルという。ハプロタイプとは、同一染色体上に存在する複数の遺伝子座におけるアレルの並び方のことである。ハプロタイプは実験的に得るのは困難であるのに対し、遺伝子型データは比較的容易に得られる。しかし遺伝子型データは組み合わされているアレルが2本の染色体のうちどちらに存在するかわからないため、全ての遺伝子座に2個のアレルがある場合、遺伝子型データから考えられるハプロタイプを全て生成すると、その数は遺伝子座数nに対して$O(2^n)$となり、遺伝子座数に対して急激に増加してしまう。そこで、得られた遺伝子型データから効率よくハプロタイプを推定する手法が必要である。

2. 2 連鎖不平衡解析の定義

連鎖不平衡とは、連鎖する2つ以上の遺伝子座で観察される現象で、異なる遺伝子座間におけるアレルの分配が独立でない現象である。ある疾患に関与する遺伝子座、およびそれと連鎖した遺伝子座(マーカー遺伝子座)との間に連鎖不平衡が存在することを利用して疾患遺伝子座の位置を推定するのが連鎖不平衡解析である。それは家系が違っても疾患遺伝子が共通の祖先遺伝子に由来するものであればその近傍のハプロタイプが現在も残っていて連鎖不平衡が形成されている(common disease common variant)[1]と考えられるためである。つまり、ある疾患患者と正常な人の集団でハプロタイプ頻度を比較した場合、患者集団に限って頻度の高いハプロタイプがあれば、その部分が疾患遺伝子である可能性が高いと考えられる。

3. 提案方式の概要

本手法の流れを図1にしめす。本手法は大きくわけて2つの処理からなる。第1の処理は、入力データから同遺伝子座以外の任意の2遺伝子座のアレル間の遷移確率をEMアルゴリズムによって求めるというものである。第2の処理は、患者の遺伝子型データから、各遺伝子座のアレルをノードと見たてて、図2のようなマルコフモデルを作成し、その患者が持ち得るハプロタイプを推定するプロセスである。ここで図2のエッジの重みには第1の処理で得られた遷移確率を用いるのだが、連鎖不平衡は隣接した遺伝子座だけでなく、全ての2遺伝子座で起こり得る現象なので他の遺伝子座のアレル全てにエッジを持たせた。

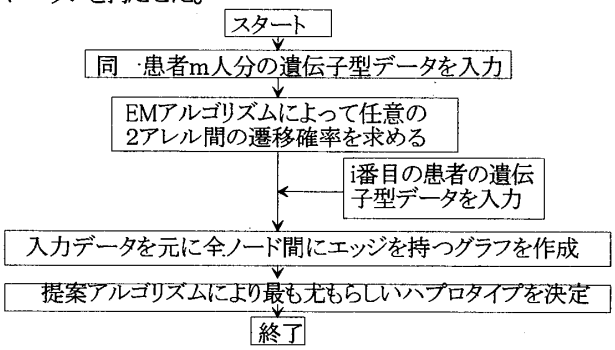

図1　提案アルゴリズムの流れ

4. EM手法による遷移確率算出についての詳細

表1のようにデータが与えられているとき2遺伝子座間の各アレルの遷移確率は以下のようにEM手法により求めることが出来る。ここで、第1遺伝子座のアレルをA/a、第2遺伝子座のアレルをB/bとし、それぞれのアレル頻度をP_A, P_a、遷移確率を$P_{AB}, P_{Ab}, P_{aB}, P_{ab}$とする。

表1　2遺伝子座間における観測データ

観測データ		観測数	Diplotype	完全データ
A/A	B/B	n0	A-B/A-B	n0
A/A	B/b	n1	A-B/A-b	n1
A/A	b/b	n2	A-b/A-b	n2
A/a	B/B	n3	A-B/a-B	n3
A/a	B/b	n4	A-B/a-b	n40
			A-b/a-B	n41
A/a	b/b	n5	A-b/a-b	n5
a/a	B/B	n6	a-B/a-B	n6
a/a	B/b	n7	a-B/a-b	n7
a/a	b/b	n8	a-b/a-b	n8

A New Haplotype Estimation Algorithm Applicable to Many SNPs Inputs
Jiro Shimosato, Jiro Katto
Graduate School of Science and Engineering, Waseda University

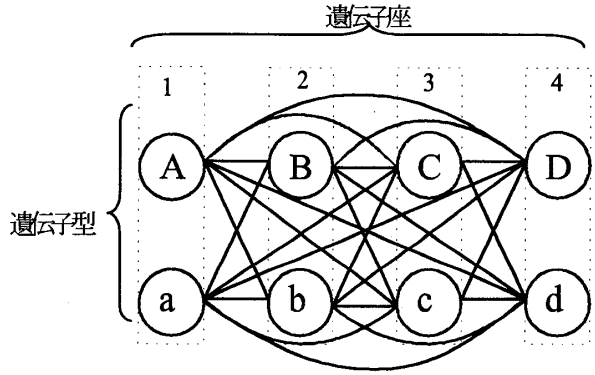

図2　遺伝子型のマルコフモデル

E-step：不完全データn0～n8とパラメータ$P_{AB}, P_{Ab}, P_{aB}, P_{ab}$から次式によってn40、n41を推定する

$$n_{40} = n_4 \cdot \frac{P_A P_{AB} \cdot P_a P_{ab}}{P_A P_{AB} \cdot P_a P_{ab} + P_A P_{Ab} \cdot P_a P_{aB}} \quad \cdots (4.\ 1)$$

$$n_{40} = n_4 \cdot \frac{P_A P_{Ab} \cdot P_a P_{aB}}{P_A P_{AB} \cdot P_a P_{ab} + P_A P_{Ab} \cdot P_a P_{aB}} \quad \cdots (4.\ 2)$$

M-step：E-stepで求まった完全データを最大にするパラメータを次式によって求める

$$p_{AB} = \frac{2n_0 + n_1 + n_3 + n_{40}}{2(n_0 + n_1 + n_2) + (n_3 + n_4 + n_5)} \quad \cdots (4.\ 3)$$

$$p_{Ab} = \frac{n_1 + 2n_2 + n_{41} + n_5}{2(n_0 + n_1 + n_2) + (n_3 + n_4 + n_5)} \quad \cdots (4.\ 4)$$

$$p_{aB} = \frac{2n_6 + n_7 + n_3 + n_{41}}{2(n_6 + n_7 + n_8) + (n_3 + n_4 + n_5)} \quad \cdots (4.\ 5)$$

$$p_{ab} = \frac{2n_8 + n_5 + n_7 + n_{40}}{2(n_6 + n_7 + n_8) + (n_3 + n_4 + n_5)} \quad \cdots (4.\ 6)$$

以上を、各遷移確率が一定の値に収束するまで行う。

5. ハプロタイプ推定アルゴリズムの詳細

本提案におけるハプロタイプ推定のアルゴリズムの概要を以下に示す。

① ハプロタイプを知りたい患者の遺伝子型データを入力
② 入力データを元に、図2のようなマルコフモデルを作成
③ スタートするノードを1つ選ぶ
④ 各ノードへの遷移確率の中で最も大きい値を持つノードを新たに選ぶ
⑤ 選ばれたノードからリンクしている全てのノードまでのエッジの積を計算する。ただし、一度選ばれたノードがある遺伝子座の別のノードは除く。
⑥ 全ての遺伝子座のノードが選ばれるまで④⑤を繰り返す
⑦ スタートするノードを変えて④～⑥を繰り返す。全てのノードがスタートするノードに選ばれれば終了
⑧ n個（n:ノード数）のハプロタイプのうち、残りのハプロタイプについても同じスコアリングを行う。両ハプロタイプの積をディプロタイプのスコアとし、ディプロタイプスコアが最も高いものをその患者が持つディプロタイプとする。

6. 実験結果

3つのサンプル集団（各32人中）において、その遺伝子型データを参照して10、14、16、18遺伝子座のデータをそれぞれのサンプルに作成した。それらを入力データとして本プログラムを用いてハプロタイプ推定を行った。また、同じ入力データに対してEM手法に基づく最尤評価手法による方式を用いたハプロタイプ推定プログラムLDSUPPORT[4]によるハプロタイプ推定を行いその結果と本手法の結果の一致した割合を表2に示す。

表2　提案手法の正解率

入力データ	サンプルA(%)	サンプルB(%)	サンプルC(%)
10遺伝子座	87.5	90.6	84.4
14遺伝子座	90.6	90.6	84.4
16遺伝子座	93.7	90.6	90.6
18遺伝子座	90.6	90.6	90.6

7. 実行時間

表3に本手法とLDSUPPORTの入力遺伝子座数と実行時間の関係を示す。LDSUPPORTでは20遺伝子座以上の入力データではメモリ使用量の問題から計算出来なかったため省略した。実験を行う計算機として、Pentium4Processor1.8GHz、メモリ256MBマシン、OSはWindows2000、プログラムの実装にはVisual C++6.0を使用した。

表3　実行時間結果

入力遺伝子座数	LDSUPPORT[s]	本手法[s]
6	0.080	0.030
10	0.551	0.090
18	880.958	0.220
20		0.971
30		3.455
50		6.679

8. 考察

表2に示したように本手法によるハプロタイプ推定は、[4]の手法による推定結果と完全に一致はしていない。これは図2のマルコフモデルに問題があると考えられる。たしかに連鎖不平衡は任意の二遺伝子座間で起こり得るものだが、遺伝子座位が離れるほどその度合いは小さくなる。そこで図2のマルコフモデルを少し変え、例えば隣接する数遺伝子座にのみエッジを持たせるようなモデルや、もしくはマルコフモデルはそのままで遺伝子座が近いノード間の遷移確率から順に高いバイアスをかける方法が考えられる。だが両手法とも、得られた入力データの遺伝子座位が実際のゲノム上でどれくらい離れているかによって結果が変わってくる恐れがあるので注意が必要である。

また、本手法では独自のスコアリングをしているためハプロタイプの尤度を算出できない。得られたハプロタイプのスコアから尤度を算出できるようにする必要がある。

9. おわりに

本稿では100遺伝子座のSNPデータに対応することを目的としたハプロタイプ推定手法の提案を行った。EMアルゴリズムによる遺伝統計学的手法とマルコフモデルにおける辺重み最大化の問題をあわせることで実行時間を大幅に削減することに成功した。

参考文献:

[1]鎌谷直之(編):"ポストゲノム時代の遺伝統計学"、羊土社、2001
[2]Excoffier,L. & Slatkin,M.: Mol.Biol.Evol., 12: 921-927, 1995
[3]Fallin,D. & Schork,N.: Am.J.Hum.Genet., 67: 947-956, 2000
[4]Kitamura,Y. et al: manuscript submitted for publication, 2001