Vol.2015-MPS-103 No.56
Vol.2015-BIO-42 No.56
2015/6/25

Fisher の正確確率検定による p 値の下限値を利用した遺伝子相互作用の効率的な数え上げ

山口拓郎[†1]　　西郷浩人[†2]

1. はじめに

近年の DNA 解析技術の向上により，かつてない量の塩基配列や一塩基多型の情報がデータベース上に蓄積されてきている．しかしながら，このような遺伝子型と，形質や体質といった表現型との間の関係の多くはまだ未解明である．そこで両者の関係を結ぶ鍵の一つとして注目されているのが，単一遺伝子ではなく，複数遺伝子間の相互作用を考慮したモデルの導入である．しかしながら，全ての可能な相互作用の数え上げは NP 困難であり，効率的な発見的手法が求められる．

本研究ではフィッシャーの正確確率検定の p 値の下限値の見積もりを枝刈りに利用することで統計的に有意な相互作用の数え上げを効率的に行う方法を提案する．計算機実験において提案手法をシミュレーションデータに適用した結果を報告する．また，高次交互作用を検出する手法である LAMP (Limitless-Arity Multiple testing Procedure)との比較実験の結果も合わせて報告する[1].

2. 実験方法

本研究は分枝限定法を用いて遺伝子相互作用の探索空間を枝刈りすることで，効率的な相互作用の数え上げを目指す．ここでは，d 箇所の SNP がある特定のタイプを持つときに表現型に影響が出た場合，これを d 次の相互作用と呼ぶ．提案手法 SNP Interaction Search （以下 SIS）は以下の流れで遺伝子相互作用の数え上げを行う．

① SNP データを 0, 1 のデータに変換する．
例えば SNP における minor allele を a, major allele を A とすると、 SNP の genotype は aa, aA, AA の 3 種類に表せる．この 3 タイプ{aa, aA, AA}を{001,010,100} と変換することで，1 つの SNP を 3bit の 0 と 1 で表すことができる．この genotype を変換した行列を $X \in \{0,1\}^{n \times 3snps}$，表現型の行列を$Y \in \{0,1\}^{n}$ とおき、X, Y ２つの行列を SIS への入力とする。ここで n はサンプルサイズ，$snps$は SNP 数を表す。

② 次に全ての遺伝子の組み合わせの中から，遺伝子相互作用の候補となる組み合わせを探索する．遺伝子の組み合わせごとに考えられる相互作用の有無と 2 値で表される表現型の関係から contingency table(図 1) を考える．図 1 の各 A,B,C,D は次のように算出している.例えば $\{X_1, X_2\}$ の遺伝子の組み合わせを考えるとき，$X_1 \cap X_2 = 1$ かつ $Y = 1$ であるサンプル数を A とする.今回 contingency table を使って，以下の a), b) 2 つの枝刈りを行った．

a) contingency table からフィッシャーの正確確率検定の p 値の下限値 $f(x)$ [1]を以下の式で得ることができる．

$$if\ x \le n_u$$
$$f(x) = \binom{n_u}{x} / \binom{N}{x}$$
$$if\ x > n_u$$
$$f(x) = 1 / \binom{N}{n_u}$$

	相互作用		合計
	有り	無し	
陽性	A	B	n_u
陰性	C	D	
合計	x		N

図 1 : Contingency Table

そして今までに探索した組み合わせ数を mとおき，$f(x)m > \alpha$（α は設定した有意水準）を満たすとき，枝刈りを行う．

b) 図 1 において、A < B となるような組み合わせを枝刈りする. A < B となる時に統計的に有意と判定されても「健康の原因となる相互作用が存在する」という仮説が採択される．そのような仮説が成り立つ組み合わせには興味がないので枝刈りを行う．

③　ステップ②の探索で得た相互作用の候補である組み合わせ数をm' とおく．相互作用の候補となった組み合わせに対し Bonferroni 補正を用いて有意水準をα/m'としたカイ二乗検定を行い、「遺伝子相互作用が存在する」と統計的に有意と判定された組み合わせを列挙していく．

以上 3 ステップで SIS の探索を行った．

†1 九州工業大学
Kyushu Institute of Technology
†2 九州工業大学
Kyushu Institute of Technology

Vol.2015-MPS-103 No.56
Vol.2015-BIO-42 No.56
2015/6/25

3. 使用したデータ

今回用意したシミュレーションデータは gs2.0 [2]をもとに作成した. gs2.0 は任意の遺伝子相互作用数を組み込んだ genotype データを生成することができる.

今回 Hapmap プロジェクトから配布されているデータを gs2.0 への入力データとし,

・遺伝子相互作用数 = 3　　・case-control 比= 1:1

・サンプルサイズ = 100

・risk = 10000 （ノイズを極力発生させないため）

と設定し, extension method にてシミュレーションデータを作成した.

4. 結果

gs2.0にて作成したデータに対し、有意水準0.01としてBrute Force, SIS, LAMP 3手法の性能比較を行った. Brute Forceは図1のAが少なくとも1つ以上ある組み合わせを探索した.

今回陽性の原因と設定した, つまり正解と設定した3次の相互作用を SIS, LAMP ともに統計的に有意と判定することができた.

次に相互作用の列挙にかかった時間を図2に示す. 図2はサンプルサイズを100と固定した場合の, SNP数と時間の関係の図である. 全手法ともにSNP数が増加すると演算時間は指数関数的に増加していくが, SIS はLAMPよりも高速に相互作用の列挙を終えることができた. その1つの原因として, LAMP は最適なBonferroni補正値を探すために繰り返し探索木を作るのに対し, SISは一回の探索木を作るだけで良いことが考えられる.

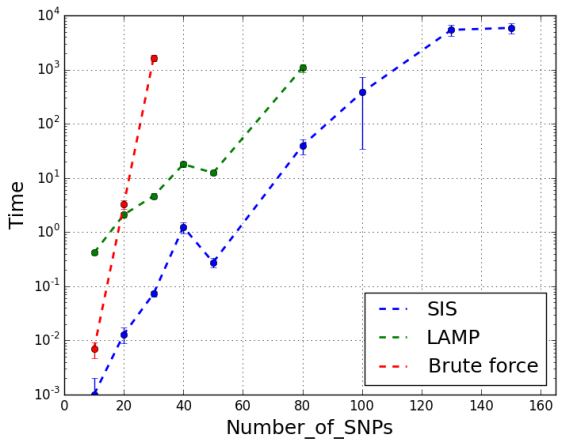

図 2：演算時間の比較
サンプルサイズを 100 と固定し, 各 SNP で 10 回のシミュレーションデータに適用した平均と標準偏差を示した。

次に表1に SNP数 = {10, 20, 30}においての, 各手法において相互作用の候補と選択された組み合わせ数と候補の中で統計的に有意と判定された組み合わせ数, SIS,LAMPでの枝刈り率を示している. 枝刈り率は 候補数/ Brute Forceでの探索数で算出した. 枝刈り率はSIS , LAMP ともに90%以上となり Brute Forceに比べ9割以上の空間の枝刈りに成功している. そして全てのSNP数においてSISはLAMPよりも候補数を小さく抑えることが出来た.

表 1：各手法の組み合わせ探索数と枝刈り率
候補数は遺伝子相互作用の候補と選択された組み合わせの数, 発見数は候補と選択された組み合わせの中で統計的に有意と判定された組み合わせ数を示す.

	SNP 数	10	20	30
Brute Force	探索数	7.22×10^3	2.16×10^6	1.18×10^9
SIS	候補数	3.43×10^2	1.08×10^3	6.91×10^3
	発見数	2.15×10^2	5.98×10^2	3.02×10^3
	枝刈り率	95.25 %	99.95 %	99.99 %
LAMP	候補数	6.58×10^2	6.81×10^3	6.10×10^4
	発見数	1.94×10^2	6.64×10^2	2.47×10^3
	枝刈り率	90.88 %	99.68 %	99.99 %

5. 考察及び今後の課題

今回,ノイズがほとんど存在しない相互作用数が 3 次の case-control 比= 1:1 のシミュレーションデータにおいて,SIS は LAMP よりも高速に遺伝子相互作用の数え上げを終えることができた. 今後はノイズを発生させたデータにおいても同様の結果を示すことができるか検証を行っていく. また現在はシミュレーションデータへの適用しか行っていないため, 実際のマウスデータへの適用も行っていく.

次に統計的に有意と判定された組み合わせの精度について述べる. SIS が候補数を LAMP よりも小さく抑えることができた原因として, SIS は図 1 の A, B において A > B となるような組み合わせは枝刈りしているが, LAMP は A < B となるような組み合わせも探索しているためと考えられる. しかし現在, 両手法共に候補数が非常に多い. 候補数を減らすことは演算時間の削減に直結するため, 更に効率よく組み合わせ空間を枝刈り出来る手法が望まれる.

参考文献

[1] Terada A, Okada-Hatakeyama M, Tsuda K, Sese J Statistical significance of combinatorial regulations.

[2] J.Li, Case Western Reserve Unicersity. http://engr.case.edu/li_jing/gs.html

[3] International HapMap Project. http://hapmap.ncbi.nlm.nih.gov