生命情報処理における多重比較法に関する研究

M2016SS007 丸山拓也

指導教員：松田眞一

1　はじめに

堀内 [2] では, 多重比較法の最近の動向を調査し, FWER ではなく FDR の制御を対象とする研究が盛んであったと述べていた．10 年たった今，多重比較法の動向が気になった私は多重比較法の最近の動向を調査した．特に，生命情報処理で扱われている多重比較法を調査した．ゲノムワイドなデータが容易に観測できるようになり，遺伝子毎，あるいは，変異または 1 塩基多型 (SNP) 毎に検定を行う解析も頻繁に行われている．この解析の中で問題になるのが検定結果の偽陽性であり，これを抑えるために Bonferroni 補正をはじめとする多重検定補正法が利用されている．しかし，単一の遺伝子や SNP ならまだしも，複数の組み合わせを考えると近似が甘くなり，補正後に有意な結果が現れなくなる問題となっていた．そこで，質的変数のみ扱われるデータに対しては無限次数多重検定法 (LAMP) が活躍し，ゲノムワイドなデータに対しても現実的な実行時間で多重検定を実行可能にさせた．LAMP は計算機を用いて解析されているが「R」では実現されていないため「R」でプログラムを作成して，現実データで解析時間，人工データを用いたシミュレーションを行って検出力と制御力の性能評価を行う．

2　生命情報処理について

近年，マイクロアレイの技術の発達により，非常に多くの遺伝子データを得ることが可能になった．これにより，生物学，特に分子生命学に関連する様々な分野の研究が情報工学の技術を用いてなされるようになった．それらの分野は総称して生命情報処理 (バイオインフォマティクス) と呼ばれ活発に研究が行われている．本研究では特に，生命情報処理の中でも SNP の有無によって発症が起こるかどうかについて解析を行う．SNP は A(アデニン)，G(グアニン)，T(チミン)，C(シトシン) の 4 つのリボ核酸から 2 つ組み合わせてできるものである．この組み合わせにより SNP の有無が生じる．マイクロアレイデータは，遺伝子数，すなわちデータの次元数 M が非常に大きく，一方データ数 N が比較的小さい ($N << M$) ことが特徴である．N 人の被験者から，M 箇所の SNP に対して疾患を発症している被験者特有の変異を調査し，複数の変異を観測し，各変異の有無と関連しているかを検定する状況を考えるのに用いるのが多重比較法である．1 つの変異に着目したとき，変異がある被験者とない被験者の間で，発症の有無に有意な差が認められるかの調査は分割表を用いてできる．(瀬々・浜田 [6] 参照.)

表 1 被験者と変異の関係

	変異 1	変異 2	...	変異 M	発症
被験者	s_1	s_2	...	s_M	c
t_1	1	0	...	0	あり
t_2	0	1	...	0	なし
...					...
t_N	0	1	...	1	あり

表 2 各変異について 2×2 の分割表で表したもの

	変異あり	変異なし	合計
発症あり	y	$n-y$	n
発症なし	$x-y$	$N-n-x+y$	$N-n$
合計	x	$N-x$	N

2.1　多重比較法

表 1 のような変異が M 個存在し，この中から発症に対して統計的に有意な関連を示す変異を網羅的に調べるには，複合仮説の場合は $m = 2^M - 1$ 個の独立した検定が必要になる．しかし，マイクロアレイのようなデータでは検定数が莫大なため偽陽性の発生確率が 1 に等しくなる．そこで，多重比較法では「複数の検定において，いずれか 1 つ以上で偽陽性が起こる確率」である Family-wise error rate(FWER) や「有意であると判定した帰無仮説の中で本当は正しかったものの期待割合」である False discovery rate(FDR) を用いて偽陽性を制御している．どちらの制御を用いるかは研究目的によって決まる．例として，主要な研究目的が新しい遺伝子を見つける探索的なものであり，検証のための研究がその後に行われる場合を考える．この場合，関連を同定する検出力が高いのが望ましく，その後の研究で誤りがあっても訂正できる．したがって，この場合には有意な検出力に占める誤った判定割合を示す FDR を用いるのが望ましい．一方，検証のような場合は誤判定が重大なので FWER を用いるのが望ましい．(西山 [5] 参照)

2.2　FWER と FDR の定義

m 個の帰無仮説の検定する問題を考える．(m_0 は未知の真の帰無仮説の数とする．)R は棄却される仮説の数で，U，V，S，T は観測不可能な確率変数である．

表 3 複数の仮説検定における過誤の数

	非有意	有意	計
真の帰無仮説	U	V	m_0
偽の帰無仮説	T	S	$m-m_0$
計	$m-R$	R	m

ここで FWER と FDR は

$$FWER = Pr(V > 1) \tag{1}$$

$$FDR = E(V/R) \tag{2}$$

と定義される．(堀内 [2] 参照.)

2.3 FWER を制御する多重比較法

FWER を制御する手法は大きく分けて FWER の上限を理論的に計算し α 以下に抑える方法および，リサンプリングを用いて FWER を制御する方法の 2 つがある．本研究では前者の方法のみを扱う．取り上げたのは，Bonferroni 法と，ステップアップ法を利用する方法として Hochberg 法，ステップダウン法を利用する Holm 法，周辺分布を用いて FWER を制御する Tarone 法である．これらの手法は瀬々・浜田 [6]，Tarone[7] を参照されたい．

3 FDR を制御する多重比較法

FDR を制御する方法としてステップアップ法と呼ばれる BH 法や，Adaptive BH 法，2 段階線形上昇手順法 (以降 TST 法とする)，BY 法などがあるが堀内 [2]，松田 [3] に述べられているため省略する．本研究では，BH 法，TST 法，BY 法を扱っていく．

4 無限次数多重検定法

今までの多重検定補正法の問題を克服する手法として，無限次数多重検定法 (Limitless Arity Multipletesting Procedure ; LAMP) を紹介する．転写因子や SNP などの因子の組み合わせの検定を考えた場合，2 つの問題が存在する．1 つ目は膨大な計算時間である．100 個の因子が構成する全通りの組み合わせを網羅すると，10^{100} 通りを考えなければいけなくて，因子数に対して検定数が指数関数的に爆発する．2 つ目は，過剰な多重検定補正である．LAMP は以上の問題点を解決するために，前者の問題に関しては頻出パターン列挙アルゴリズムを利用する．また，後者の問題に関しては Tarone 法を利用する．変異情報における頻出パターンとは，被験者が頻繁に有する変異の組み合わせになる．頻出パターン列挙において，各列 (変異) のことをアイテムと呼ぶ．さらにアイテムの集合をアイテム集合と呼び，アイテム k 個有するアイテム集合を k-アイテム集合と呼ぶ．あるアイテム集合 I に着目したとき，I の中のアイテムをすべて有する被験者の数をサポートと呼び，サポートを $x(I)$ で表す．(瀬々・浜田 [6] 参照)

4.1 頻出パターン列挙

頻出パターン列挙問題とは，与えられたデータベース中に一定回数以上出現するパターンを列挙する問題であり，購買履歴情報の解析などに用いられる．(瀬々・浜田 [6] 参照)

本研究では深さ優先探索を行う Eclat 法を用いた．

4.2 最小サポートと p 値の下限

頻出パターン列挙と Tarone 法の接点となる，最小サポート λ と FWER の上界の関係を説明する．p 値の下限は

$$l(x) = \frac{\binom{n}{y}\binom{N-n}{x-y}}{\binom{N}{x}} \text{ただし } y = min\{x, n\}$$

と表せる．また，$x \geq 0$ において x について単調減少する関数として以下の $f(x)$ を定義する．

$$f(x) = \begin{cases} l(x) & (x \leq n \text{ のとき}) \\ l(n) & (x > n \text{ のとき}) \end{cases} \tag{3}$$

最小サポート λ と FWER の上界の関係を説明する．仮説 H_I に対応する p 値を $p(I)$ とする．全アイテム集合を L，最小サポート λ 以上のアイテム集合の族を L_λ，$m_\lambda = |L_\lambda|$，$\delta' = f(x)$ とする．

$$\begin{aligned} FWER &\leq \sum_{I \in L} \Pr(p(I) \leq \delta') \\ &= \sum_{I \in L_\lambda} \Pr(p(I) \leq \delta') + \sum_{I \in L \setminus L_\lambda} \Pr(p(I) \leq \delta') \\ &= \sum_{I \in L_\lambda} \Pr(p(I) \leq \delta') \\ &= m_\lambda \delta' \end{aligned} \tag{4}$$

$m_\lambda \delta' < \alpha$ を満たす限り λ を改善できる．アルゴリズムの中では，x を大きい方から順に減らす手続きを行っている．(瀬々・浜田 [6] 参照)

4.3 LAMP のアルゴリズム

LAMP のアルゴリズムは以下の手順で行う．

手順 1: 閾値 $\lambda = n$ とする．

手順 2: 最小サポートが λ 以上の頻出パターンを列挙する．この族を L_λ，$m_\lambda = |L_\lambda|$ とする．$L_\lambda = \phi$ の場合は $\lambda = \lambda - 1$ として手順 2 を行う．

手順 3: $\delta' = f(\lambda - 1)$ を求める．

手順 4: $\delta' m_\lambda \leq \alpha$ ならより大きな δ' で $FWER$ を α 以下に抑えれる可能性があるため $\lambda = \lambda - 1$ にして手順 2 へ戻る．それ以外なら手順 5 へ進む．

手順 5: α / m_λ を補正後の有意水準 δ としてアイテム集合の族 L_λ 内の各アイテム集合の p 値をもとめ，δ 以下の仮説を棄却する．(瀬々・浜田 [6] 参照)

4.4 作成した LAMP

作成した LAMP はアイテム列挙に R パッケージ「arules」にある eclat 関数を用いてプログラミングした．実験環境は OS は Windows7，CPU は Core i5，メモリは 4G を使用した場合，250 列の大規模データに適応することができた．

5　実データ解析

LAMP と他の多重比較法の性能と解析時間を調査するためにまずは 2 値変数のみのデータを用いる．R パッケージ「SNPassoc」に入っている case-control study 用のサンプルデータ SNP を実データとして解析する．全個数 157 である．

5.1　解析準備

データの SNP は 35 座位あるが，橋本 [1] が行っていたように有効な座位を選択して解析を行った．変異にあたる変数は，snp10001，snp10005，snp10008，snp10011，snp10015，snp10019，snp10020，性別，casco にした．また，コレステロールの値を四分位数でわけて 4 変数にした．比較対象は，BH 法，TST 法，BY 法，Bonferroni 法，Holm 法，Hocheberg 法，Tarone 法，LAMP，多重検定補正なしの 9 つの分類で，どの手法も複合仮説を検定していく．各アイテム集合から p 値を求めるのは自作プログラムを用いる．また，LAMP，Tarone 法は自作のプログラムであり，BH 法，BY 法は堀内 [2]，TST 法は松田 [3] のプログラムを用いる．他の補正法は p.adjust() 関数を用いて補正後の p 値を求める．

5.2　結果・考察

要旨の枚数の関係上，結果を載せるのは BH 法，多重検定補正なし，Tarone 法，LAMP のみにする．

表 4 SNP データでの棄却数と時間 (秒数) の比較

手法名	BH 法	なし	LAMP	Tarone 法
時間 (秒)	1058.73	1059.21	0.12	1055.75
棄却数	12	89	20	20

SNP データを解析した結果，棄却数は他の手法，LAMP，Tarone 法という順番で多くなった．解析時間は LAMP は非常に優れている．棄却した仮説はその他の手法の棄却した仮説に加え，LAMP と Tarone 法は snp10001 とコレステロール値 (第三四分位以上) を棄却することができた．

6　シミュレーション

シミュレーションを行って LAMP の検出力について調査を行う．

6.1　シミュレーションの準備

本研究のシミュレーションの流れは, 30 人から SNP7 つを採取し，発症と関連のある SNP を見つけられるかという状況を作成する．つまり，仮説数 7(複合の場合 127) の両側検定を行っていく．

(シミュレーションの目的)

· FWER，FDR を制御できる手法が 0.05 以下に保てているかを確認する．

· 偽の帰無仮説をどのくらい棄却できるのかを確認する．

· 発症と関連のある変異とそうでない変異が混合しているエラー仮説を偽の帰無仮説と対立仮説の立場から真の帰無仮説とみなして分けたときの Tarone 法と LAMP の差を確認する．

(シミュレーションのパターン)

1. 仮説間が独立な場合

＜パターン 1＞

2. 仮説間がすべて正の相関である場合

＜パターン 2＞ 直線型，＜パターン 3＞ 一点集中型，＜パターン 4＞ 二点集中型)

3. 仮説間に負の相関がある場合

＜パターン 5＞ 直線型，＜パターン 6＞ 一点集中型，＜パターン 7＞ 二点集中型)

4. パターン 4 の仮説数を 10 にした場合

＜パターン 8＞

(シミュレーション方法)

相関のある二項分布は守・久門 [4] を参照してベイズ流の二項分布として作成した．扱う手法は単体仮説に対して実データ解析で用いたの 9 つの手法，複合仮説に対しては Tarone 法と LAMP を用いる．これは，Tarone 法と LAMP は複合仮説で有効性を発揮すると考えたためである．また，変異の影響をロジスティック回帰で考えて，真の帰無仮説での発症確率を平均 0.5, 発症に関係のある変異をすべて含む偽の帰無仮説での発症確率を平均 0.622 とした．これは遺伝子の変異と疾患の発症の関連を見つけるのが困難な状況にするためである．発症と関連のある SNP の組み合わせを 7 通り作成し，パターン 1 は確率を 0.1, 0.25，0.5，0.75，0.9 の 5 通り，パターン 2 からパターン 8 は確率は 0.5，相関を ±0.1，±0.25，±0.5，±0.75，±0.9 の 5 通りずつ行い，全組み合わせ 35 通り行った．各パターンのシミュレーション回数は 1000 回である．

7　シミュレーションによる評価と考察

シミュレーションの評価は要旨の枚数の関係上，エラー仮説を真の帰無仮説とした場合の Tarone 法 (複合) と LAMP(複合) のみ表にしてまとめる．他の手法の結果については本論を参照されたい．表で用いる「棄却数の差が最大」は，発症と関連のある変異，相関すべてが同じ条件のときの Tarone 法 (複合) と LAMP(複合) の棄却した偽の帰無仮説の数の差を示す．「棄却数の差の勝利数」は，35 通り中 Tarone 法 (複合) と LAMP(複合) の偽の帰無仮説の棄却数が多かった方を勝利として数えていく．

7.1　＜パターン 1 独立＞

どのパターンでも各手法は偽陽性を 0.05 以下に抑えることができていた．Tarone 法 (複合) と LAMP(複合) もどちらも FWER を制御できていた．LAMP は保守的であるため，Tarone 法に棄却数の差で勝つことができなかった．

表 5 パターン 1 での LAMP と Tarone 法の比較

手法名	Tarone 法	LAMP
FWER の最大値	0.043	0.026
棄却数の差が最大	8	2
棄却数の差の勝利数	9	0

7.2 ＜パターン 2,3,4 正の相関＞

どのパターンでも各手法は偽陽性を 0.05 以下に抑えることができていた．パターン 2 の直線型では，LAMP と Tarone 法の複合仮説の場合を比較したとき，Tarone 法の方が優れている場面が多く見られた．パターン 3 の一点集中型では，Tarone 法の方が勝利数は多く，LAMP のほうが保守的な場面が何度も見られた．パターン 4 の二点集中型では，どの手法も保守的な結果になったが，Tarone 法と LAMP の勝利数は差あった．以上の結果から，正の相関があるときは棄却数の面で差があり，Tarone 法の方が有効な場面が多く見られた．

表 6 パターン 2 での LAMP と Tarone 法の比較

手法名	Tarone 法	LAMP
FWER の最大値	0.038	0.024
棄却数の差が最大	29	6
棄却数の差の勝利数	10	0

表 7 パターン 3 での LAMP と Tarone 法の比較

手法名	Tarone 法	LAMP
FWER の最大値	0.032	0.023
棄却数の差が最大	56	5
棄却数の差の勝利数	9	0

表 8 パターン 4 での LAMP と Tarone 法の比較

手法名	Tarone 法	LAMP
FWER の最大値	0.034	0.029
棄却数の差が最大	17	5
棄却数の差の勝利数	10	0

7.3 ＜パターン 5,6,7 負の相関＞

どのパターンでも各手法は偽陽性を 0.05 以下に抑えることができていた．パターン 5 の直線型では，正の相関よりも棄却数の差はない結果になった．パターン 6 の一点集中型では，どの手法も保守的な結果となった．パターン 7 の二点集中型もパターン 6 と同様にどの手法も保守的な結果となったが Tarone 法のほうが棄却数が多いときがあった．全体的にみると Tarone 法と LAMP はそれほど差は生じなかったが，Tarone 法が有効な場面が何度かあった．

表 9 パターン 5 での LAMP と Tarone 法の比較

手法名	Tarone 法	LAMP
FWER の最大値	0.034	0.032
棄却数の差が最大	4	3
棄却数の差の勝利数	2	0

表 10 パターン 6 での LAMP と Tarone 法の比較

手法名	Tarone 法	LAMP
FWER の最大値	0.037	0.026
棄却数の差が最大	11	6
棄却数の差の勝利数	13	0

表 11 パターン 7 での LAMP と Tarone 法の比較

手法名	Tarone 法	LAMP
FWER の最大値	0.042	0.032
棄却数の差が最大	7	2
棄却数の差の勝利数	16	0

7.4 ＜パターン 8 仮説数 10＞

表 12 パターン 8 での LAMP と Tarone 法の比較

手法名	Tarone 法	LAMP
FWER の最大値	0.060	0.020
棄却数の差が最大	2	3
棄却数の差の勝利数	2	2

仮説 10 の場合は実験回数が 100 回のため誤差は多いが，Tarone 法は FWER を制御することができなかった．仮説数が多くなると LAMP と Tarone 法の差が明確になった．

8 おわりに

LAMP の特徴が 2 点明らかになった．1 つ目は計算時間の早さである．2 つ目は保守的なところである．しかし，エラー仮説に対して有力である可能性がある．本研究では，仮説数が少ないため，Tarone 法の方がよく見えるが，増えていくと LAMP の方がよくなる可能性がある．単体仮説では Tarone 法と LAMP は保守的な面がみられ向いてないことがいえる．その他の手法は，BH 法や TST 法，Holm 法は優れた結果を出していることがわかった．BY 法や Bonferroni 法は保守的過ぎることが改めてわかった．LAMP は大規模でも検定できるため，量的変数がどの分布か把握し，正確な方法で質的変数にした列数が多くなってしまう場合でも検定することができるため量的変数への応用に期待される．

参考文献

[1] 橋本登，『関連解析における多変量 QTL への拡張』，南山大学大学院数理情報研究科修士論文，2009.

[2] 堀内賢太郎，『多重比較法の最近の動向の研究 (Web による多重比較法の実現)』，南山大学大学院数理情報研究科修士論文，2007.

[3] 松田眞一，FDR の概要とそれを制御する多重検定法の比較，『計量生物学』，**29**(2)，125-139，2008.

[4] 守真太郎・久門正人，ネットワーク上の相関のある二項分布，『情報処理学会論文誌：数理モデル化と応用』，**2**(1)，22-36，2009.

[5] 西山毅，『実践でわかる! R による統計遺伝学』，丸善出版，2016.

[6] 瀬々潤・浜田 道昭，『生命情報処理における機械学習 - 多重検定と推定量設計』，講談社，2015.

[7] Tarone，R. E. ‘A modifed Bonferroni method for discrete data’，Biometrics，**46**(2)，515-522，1990.