

交互作用は相互作用？(6)

遺伝疾患と対立遺伝子頻度分布の関連解析



広津 千尋 | Hirotsu Chihiro

明星大学理工学部教授

■1968年東京大学大学院工学系研究科計数工学専攻博士課程修了。工学博士。東京工業大学助手を経て、1971年東京大学工学部計数工学科講師、助教授、教授。2000年定年退官、東京大学名誉教授。2000年4月より現職。

1. データの特徴と既存の方法の問題点

今回は遺伝疾患の有無を対立遺伝子（以降単にアリルと言う）の頻度分布の違いで特徴付ける問題を考える。データは疾患の有無×CAリピート数の分割表形式で得られるから、この問題はやはり本シリーズで採り上げた交互作用解析の範疇に入る。

表1はKitao et al.（参考文献[4]）によって得られたデータの一部であり、この研究分野の典型的な例である。このように、ケース（患者）とコントロール（健常者）に対して独立に取られたアリルの頻度分布の違いを統計的に解析するアプローチが採られる。高度に多型の場合には、この例、あるいはもっと典型的に表5の例のように頻度0や1のセルが頻繁に現れる。Sham and Curtis[7]ではこのように高度に多型の場合の関連解析の方法として

- ①通常の分割表に対する適合度 χ^2 検定、
- ②頻度の小さなセルを適宜併合した χ^2 検定、
- ③順次各1つのアリルを採り上げ、それ以

外をすべて併合して出来る 2×2 分割表の適合度 χ^2 を計算し、その最大を統計量とする検定、

- ④アリルのすべての2分割に対し、やはり 2×2 分割表の適合度 χ^2 を計算し、その最大を統計量とする検定、

の有意確率をシュミレーションによって評価する方法を提案し、比較検討している。なお、アリルが2通りの場合は、これら χ^2 統計量の自由度がすべて1になりこのような多様性は生じない。このSham and Curtisも含めて、市販の多くのソフトは有意確率の評価にMCMC（Markov Chain Monte Carlo）法に基づくシュミレーションを用いている。しかしながらこの領域では、極めて多くのマーカーについて検定を繰り返し、多重性の調整のため極小の有意確率を探索するのと、先に述べたようにスパースな分割表を対象とすることが多いので、正確法の適用が望ましい。また、Sham and Curtisの採り上げた4つの方法は、第3の方法を除いて総括的な検定であり、たとえ有

意な結果が得られてもどのアリルが疾患に特徴的なのか、あるいはアリル頻度分布のどのような特徴が疾患を示唆するのかといったきめ細かい情報は得られない。特に、第4の方法は統計量の特徴付けが難しく薦められない。そんな中、第3法は特定のアリルが強く関連しているときに検出力が高く、解釈も明解である。最近、この第3法に対して分割表のマルコフ性を利用したp値計算の正確アルゴリズムが得られた(Hirotsu et al.[3])ので、後で適用例を示す。以降、この第3法正確法をmax 1-to-others χ^2 法と呼ぶ。一方、このmax 1-to-others χ^2 も含め、すべての方法が対立仮説に特別の方向性を仮定しないオムニバス検定であり、ハンティンソン病のように、CAGリピートの延長が発症や重症度に関係するというような方向性があるときには効率が良くない。例えば、表1で、CAリピートの延長もしくは短縮が疾患と関連すると想定する。それは正に第4回の設定と同様であり、効率の良い検定法として最大累積(max acc.) χ^2 法が考えられ、特に2群比較の場合には正確法が得られている。最大累積 χ^2 は例えば表1で、最初に1列と2列の間、次に2列と3列の間というように次々と2分割し、分割点の両側で各列を併合してできる 2×2 分割表に対して計算した適合度 χ^2 の最大を統計量とする方法であり、過去いろいろな研究で傾向性仮説に対し高い検出力を示すことが知られている(Hirotsu et al.[1]参照)。なお、第4回ではそれらの適合度 χ^2 の最大ではなく、総計を統計量としている。一方、特定の疾患に対する関連研究の初期の

段階では、このように対立仮説を限定してしまうことは一般に難しい。その場合は max acc. χ^2 と max 1-to-others χ^2 の併用が薦められる。これは単に前者を x_1^2 、後者を x_2^2 とするとき、 $\max(x_1^2, x_2^2) \dots x_1^2$ と x_2^2 の大きい方を検定統計量とする方法であり、やはり正確な有意確率計算法が得られる。

2. max 1-to-others χ^2 及び max acc. χ^2 法の表1への適用

(1) max 1-to-others χ^2 の計算

まず第1列に注目し、それを除いたすべての列を併合して表2(a)を得る。表2(a)に対して通常の適合度 χ^2 を計算し、 $\chi^2 = 0.34$ を得る。次に、第2列に注目し、それ以外の列をすべて併合して表2(b)を得、それに対する適合度 χ^2 、 $\chi^2 = 1.53$ を得る。これを次々と行った結果は表3の第4行のとおりであり、アリル109に対し max 1-to-others $\chi^2 = 18.78$ が得られる。

表1 遺伝子座位D20S95におけるアリル頻度分布

アリル	97	99	101	103	105	107	109	111	113	115	計
統合失調症	4	30	25	10	17	90	34	37	4	1	252
健常者	5	18	28	6	41	67	5	34	5	5	214
計	9	48	53	16	58	157	39	71	9	6	466

表2 max 1-to-others χ^2 の計算

(a)			(b)				
アリル	97	99～115	計	アリル	99	99以外	計
統合失調症	4	248	252	統合失調症	30	222	252
健常者	5	209	214	健常者	18	196	214
計	9	457	466	計	48	418	466

表3 max 1-to-others χ^2 及び max acc. χ^2 の計算結果

アリル	97	99	101	103	105	107	109	111	113	115
統合失調症	4	30	25	10	17	90	34	37	4	1
健常者	5	18	28	6	41	67	5	34	5	5
max1-to-others χ^2	0.34	1.53	1.15	0.47	16.36	1.01	18.78	0.13	0.34	3.43
max acc. χ^2	0.34	0.81	0.01	0.03	6.59	3.11	1.17	2.69	3.43	—

(2) $\max \text{acc. } x^2$ の計算

一方、 $\max \text{acc. } x^2$ でも最初列1と2の間の分割を考えるが、それは $\max 1\text{-to-others } x^2$ の初項の計算と同じようになる。これを、分割点を右にシフトしながら次々と行った結果は表3の最終行のとおりであり、アリル105と107の間の分割に対し $\max \text{acc. } x^2 = 6.59$ が得られる。

(3) $\max 1\text{-to-others } x^2$ と $\max \text{acc. } x^2$ の併用

いま、 $x_1^2 = 6.59$ 、 $x_2^2 = 18.78$ であるから両者の最大は $\max 1\text{-to-others } x^2 = 18.78$ の方である。併用法の場合、検定統計量はこのように必ず $\max \text{acc. } x^2$ もしくは $\max 1\text{-to-others } x^2$ の一方に一致するが、有意確率は、この両者の最大を統計量としていることを考慮して計算するので、それぞれ単独の計算に比べてやや保守的な評価となる。しかしながら、基本の2つの統計量のどちらかが選択的に鋭く反応するような場合には、検出力の低下はごく僅かであることが示される。

これら3法の正確な有意確率は表4のようになる。この例では $\max 1\text{-to-others } x^2$ の寄与が $\max \text{acc. } x^2$ に比べてかなり大きいため、併用法の検出力の損失が小さいことが分かる。

なお、この例の場合は全体としては4染色体上で総数34の遺伝子座について同時解析を行っている。そこで、表4で得られた結果は

表4 遺伝疾患とアリル頻度分布の関連解析

方法	統計量	有意確率
$\max 1\text{-to-others } x^2$	18.78(109 vs. others)	0.000059
$\max \text{acc. } x^2$	6.59(97~105 vs. 107~115)	0.075732
<i>combined max imal x^2</i>	18.78(109 vs. others)	0.000113

0.05を34で除して、有意水準0.0015として評価しなければならない。正確法はこのように小さな有意確率を問題にするとき、また高度に多型の場合に得られるスパースな分割表を解析するときに極めて有用である。

3. 2つのマーカーにおける対立遺伝子同時頻度分布のハプロタイプ解析

親の情報を利用し、連鎖する遺伝子座でアリルの対の情報が得られる場合の解析をハプロタイプ解析という。この場合、表1のような $2 \times J$ 型分割表に替えて、 $2 \times J \times K$ 型の3元分割表が得られることになる。表5 (Oudet et al.[6]) はそのひとつの例であり、Fragile Xと健常者の遺伝子座位DXS548とFRAXAC2におけるアリルの同時頻度分布を表示している。ところが、Sham and Curtis[7]では、これを3元表として解析せず、 $2 \times JK$ 型の分割表として解析しており、それが一般的であるように思われる。しかしながら、そのやり方では重要

表5 Fragile Xと健常者の座位DXS548とFRAXAC2におけるハプロタイプ頻度分布

Fragile X ($i = 1$)

$j \setminus k$	143	145	147	149	151	153	155	157	165	169	total
192	0	0	0	0	0	0	0	0	0	0	0
194	1	3	0	0	9	16	12	0	0	0	41
196	0	5	9	1	11	1	3	0	0	0	30
200	0	0	0	0	0	0	1	0	0	0	1
202	0	0	0	0	0	0	1	0	0	0	1
204	0	1	0	1	9	3	14	0	0	0	28
206	0	0	0	0	1	0	0	0	0	0	1
<i>total</i>	1	9	9	2	30	20	31	0	0	0	102

健常者 ($i = 2$)

$j \setminus k$	143	145	147	149	151	153	155	157	165	169	total
192	0	0	0	0	0	2	0	0	0	0	2
194	1	7	5	1	17	67	7	4	1	0	110
196	0	4	6	0	3	8	1	0	0	0	22
200	0	1	0	0	0	0	0	0	0	0	1
202	0	0	0	0	0	1	1	0	0	0	2
204	0	2	0	0	3	6	2	0	0	1	14
206	0	0	0	0	0	0	2	0	0	0	2
<i>total</i>	1	14	11	1	23	84	13	4	1	1	153

表6 ハプロタイプ頻度分布（規準化定数は省略）

座位1のアリル(j)	健常者					有疾患者 (i=2)					
	座位2のアリル(k)					座位2のアリル(k)					
	1	2	3	4	5		1	2	3	4	5
モデル(a) (無3因子交互作用)											
1	1	1	1	1	1		1	1	2	1	1
2	1	1	1	1	1		2	2	4	2	2
3	1	1	1	1	1		1	1	2	1	1
4	1	1	1	1	1		1	1	2	1	1
モデル(b) (外れ値型3因子相互作用)											
1	1	1	1	1	1		1	1	1	1	1
2	1	1	1	1	1		1	1	4	1	1
3	1	1	1	1	1		1	1	1	1	1
4	1	1	1	1	1		1	1	1	1	1
モデル(c) (傾向性3因子交互作用)											
1	1	1	1	1	1		2	2	2	1	1
2	1	1	1	1	1		2	2	2	1	1
3	1	1	1	1	1		1	1	1	1	1
4	1	1	1	1	1		1	1	1	1	1

な3因子交互作用を見落すことになる。以下でこの意味を簡単な数値例（表6）で解説し、適切な解析方法を紹介した後、Fragile Xの例への適用結果を述べることとする。

まず表6(a)の例では、遺伝子座1のアリル2と遺伝子座2のアリル3が特異的、かつ乗法的に作用し、3因子交互作用はない。一方、表6(b)では遺伝子座1及び2のそれぞれ2と3の組み合わせ(2,3)が特異的であり、3因子交互作用が存在する。つまり、それぞれの遺伝子座のアリル2と3が独立に作用するのではなく、組み合わせ効果として作用する。このような意味を持つ3因子交互作用の検出は大変重要なことである。さらに、単一のマーカーの頻度分布でも考えたように、疾患がCAリピートの延長や短縮と関係するようなときの3因子交互作用のモデルとしては、表6(c)のようなモデルが考えられる。

広津他[2]及びHirotsu et al.[3]では、このよ

うな2×J×Kハプロタイプデータの生起確率 p_{ijk} に対し、対数線形模型

$$\log p_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk}$$

$$i = 1, 2; j = 1, \dots, J; k = 1, \dots, K$$

を想定し、まず最初に3因子交互作用に関する帰無仮説

$$H_0: \log p_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} \quad (1)$$

を検定することを提唱している。モデル(1)は遺伝疾患の有無(α_i)に対して、遺伝子座1(β_j)、2(γ_k)の組み合わせ効果($\alpha\beta\gamma$)_{ijk}がないモデルである。その検定法として、表6(b)のような型の3因子交互作用モデルを想定した場合にはmax 1-to-others x^2 が素直に拡張される。また、表6(c)のような型の3因子交互作用を想定した場合には max acc. x^2 が素直に拡張され、仮説が絞り切れないときにはそれらの併用法も可能である。これらはいずれも正確法であるが、その詳細及び具体的なアルゴリズムについては原著(Hirotsu et al.[3])を参照されたい。このデータにmax 1-to-others x^2 及びmax acc. x^2 の拡張法を適用した結果は表7のようであり、有意性は示唆されない。そこで表5をよく見ると、特異的に見えるセルは遺伝子座1のアリル194と遺伝子座2のアリル153に集中しており、表6(a)のケースと解釈するのが妥当である。そこで遺伝子座1, 2それぞれに対する解析に移行するが、その結果はアリル194と153が高度に有意ということになる。なお、この例では2つのマーカーは目標遺伝子の両側にある。それが3因子交互作用の見られない理由と考えている。

表7 表5のハプロタイプデータに対する3因子交互作用の検定結果

手法	統計量	特異分割点	有意確率
拡張max1-to-others x^2	4.80	(206;155)	0.49
拡張max acc. x^2	4.84	(192-204,206;143-151,153-169)	0.37
併用法	4.84	(192-204,206;143-151,153-169)	0.53

4. オッズ比の傾向性仮説検定 ($2 \times 2 \times K$ 型分割表)

前節のmax acc x^2 法は特別な場合として $2 \times 2 \times K$ 分割表におけるオッズ比の傾向性仮説 H_1 ：
 $\phi_1 \leq \phi_2 \leq \dots \leq \phi_K$ にも適用できる。ただし、

$$\phi_k = \frac{p_{11k}p_{22k}}{p_{12k}p_{21k}} \quad (2)$$

は k 番目の 2×2 表に対するオッズ比である。

例えば年齢によって層別された、喫煙習慣の有無と肺癌罹患の有無に対する 2×2 表でのオッズ比⁽²⁾が、加齢とともに上昇するというような仮説の検証には興味が持てる。このような例は幾らでも考えられるが、max acc. x^2 法は例えばある年齢を境にオッズ比が特に上昇するような傾向を検出するのに適した統計量となる。この定式化の詳細については太田他^[5]を参照されたい。

5. 終わりに

本シリーズでは毎回交互作用解析の応用例について述べた。特に通り一遍の総括検定に替えて、行あるいは列ごとの多重比較法を提案しているが、興味を持って応用していただければ幸いである。この手法は繰り返しのない2元配置で、通常の誤差平方和が得られない場合にも適用できるという利点も持っている。さらに、2回目以降は行や列の水準に自然な順序があり、単調なトレンドや凹凸といったパターン解析に興味がある場合に、累積

x^2 やその拡張型統計量を提案している。これらの統計量は変化点解析とも密接に関係しており、また機会があれば述べてみたい。

最後に、解析に用いたプログラムには提供できるものもあるので興味ある読者は連絡していただきたい。例えば2章のmax 1-to-others x^2 やmax acc. x^2 は、意味も明解であり、本論の設定以外にもいろいろと応用可能なので、オムニバスな適合度 x^2 に替えて是非試していただきたい。

*参考文献

- [1] Hirotsu, C., Kuriki, S. and Hayter, A. J. (1992) : Multiple comparison procedures based on the maximal component of the cumulative chi-squared statistic : Biometrika 79, pp.381-392.
- [2] 広津千尋・青木敏・稻田俊也・北尾淑恵(2000) : ある精神疾患と二つの近傍遺伝子座位におけるCAリピート頻度分布の関連解析について : 第68回日本統計学会年会講演報告集, pp.407-408.
- [3] Hirotsu, C., Aoki, S., Inada, T. and Kitao, Y. (2001) : An exact test for the association between the disease and alleles at highly polymorphic loci with particular interest in the haplotype analysis : Biometrics 57, pp.148-157.
- [4] Kitao, Y., Inada, T., Hirotsu, C. et al. (2001) : "A contribution to genome-wide association studies : search for susceptibility loci for schizophrenia using DNA microsatellite markers on chromosomes 19, 20, 21 and 22" : Psychiatric Genetics 10, pp.139-143.
- [5] 太田絵里・青木敏・広津千尋(2003) : $2 \times 2 \times K$ 分割表における単調仮説の検定 : 応用統計学32, pp. 107-126.
- [6] Oudet, C., Morenet, E., Serre, J. L., et al. (1993) : Linkage disequilibrium between the fragile X mutation and two closely linked CA repeats suggests that fragile X chromosomes are derived from a small number of founder chromosomes : American Journal of Human Genetics 52, pp.297-304.
- [7] Sham, P. C. and Curtis, D. (1995) : Monte Carlo tests for associations between disease and alleles at highly polymorphic loci : Annals of Human Genetics 59, pp.97-105.