

ポストゲノムにおけるノンパラメトリック連鎖解析

富田 誠¹

1 はじめに

ゲノム統計学におけるパラメトリック連鎖解析 (parametric linkage analysis) とは疾患の発症が特定のメンデル型の遺伝形式をとる (常染色体優性遺伝、劣性遺伝など) と仮定して解析を行う連鎖解析手法である。メンデル型の遺伝形式をとるには表現型 (疾患の発症) がただ 1 つ疾患原因遺伝子座の遺伝子型のみにより決定されており、かつその疾患原因遺伝子座の遺伝子型と表現型 (疾患の発症) の関係が浸透率 (penetrance) ² によってのみ決まることが必要である。パラメトリック連鎖解析ではこの浸透率および疾患対立遺伝子頻度を入力データの中に記載して解析を行うことになる。

これに対し、ノンパラメトリック連鎖解析ではそのような前提を全く用いることなく遺伝解析を行う。このことによりノンパラメトリック連鎖解析では浸透率や遺伝形質がはっきりしない疾患の原因遺伝子探索に利用することが可能となっている。特に最近注目されている「ありふれた病気」(common disease) ではその多くのものは遺伝原因が多遺伝子性であることが予想されている。疾患原因遺伝子座が複数個あり各遺伝子座が独立に発症に寄与しているとき、表現型 (罹患の有無) としてメンデル型遺伝形質は取り得ない。このようなポリジーン疾患の解析にノンパラメトリック連鎖解析は非常に有効である。

現在よく行われている罹患同胞対解析 (affected sib-pair analysis) では同胞 (兄弟姉妹) で罹患しているペアを集めて解析ができるので、成人以降になって初めて発症するような遅発性疾患 (家族サンプルを得たくとも、先祖側のメンバーが既に亡くなっていて遺伝子型タイピングが不可能なことが多い) の原因遺伝子探索にも非常に有効である。罹患同胞対解析や Genehunter などの NPL 解析では家系内にいる 2 人の罹患メンバー間に伝達されてくる同じ祖先由来 (= 同祖) の対立遺伝子数に着目する。家系内の罹患メンバー間に伝達されてくる同祖の対立遺伝子数と、一般の集団での家系内の同じ関係にある (発症に関係ない) 2 人での同祖の対立遺伝子数の確率分布とを比較して解析を行う。そのため、これらノンパラメトリック連鎖解析では同一家系内に最低 2 名の罹患メンバーが必要であり、そのような小家系を多数集めて解析

¹ 社団法人バイオ産業情報化コンソーシアム, tomita@jbirc.aist.go.jp

² ある遺伝子型を持っている個体が、ある表現型をとる確率 (あるいはある遺伝子型を持っている多くの個体の中で、ある表現型をとる個体の割合) を浸透率という

を行うことになる。

2 IBD について

IBD(Identity By Descent, 共有する同祖の遺伝子数) とは、2 人の間で対立遺伝子が祖先の同じものを共有する (これを同祖であるという) 状態をいう。ノンパラメトリック連鎖解析では血縁関係にある 2 人の間について IBD を考える。IBD はまた、IBD である対立遺伝子の数を示す言葉でもある。例えば血縁の繋がりが全くない 2 人では同祖の対立遺伝子を共有することは決して有り得ないので常に $IBD = 0$ である。親子間では、子供は片親から必ず 1 つの対立遺伝子を受け取っており (IBD である)、かつ残りの対立遺伝子はもう片親から伝達されてくる (IBD ではない)。よって親子間では必ず $IBD = 1$ となる。一卵性双生児ではお互いの持つ対立遺伝子は全て同じものである。よって一卵性双生児の間では必ず $IBD = 2$ である。同胞 (兄弟姉妹) 間の IBD について考える。同胞間においては $IBD = 0, 1, 2$ のいずれの場合も取りうる。同胞間の IBD は確率分布に従う。

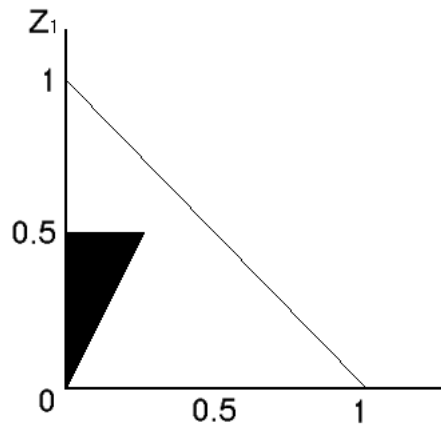
3 罹患同胞対解析

複雑な形質の遺伝的分析に対して、罹患同胞対解析はノンパラメトリック解析の手法を用いた重要なツールである。

罹患同胞対解析では、疾患と相関を示す座位については、罹患同胞対の共有する同祖遺伝子の数 (IBD: ノンパラメトリック解析の理論の節を参照) は一般的な同胞対より多いという仮説を証明する。一般的な同胞対の場合、同胞の IBD が 0, 1, 2 である確率はそれぞれ 0.25, 0.5, 0.25 である。もちろん、罹患同胞対であっても、疾患と相関の無い座位については、同胞の IBD の確率分布はこれに等しい。罹患同胞対の IBD が、この一般的な同胞対の IBD に比べてどれくらいずれているのかを推定し、一般的な同胞対の IBD と同じであるという帰無仮説を否定する。

罹患同胞対の IBD の数が 0, 1, 2 である確率を z_0, z_1, z_2 とし、その仮説の下に観測データの得られる確率 (仮説の尤度) を最大とする最尤法 (Maximum Likelihood Estimation) を用いる。それにより z_0, z_1, z_2 の最尤推定値 $\hat{z}_0, \hat{z}_1, \hat{z}_2$ を求める。さらに、最尤推定値の場合の尤度 (最大尤度) と、 $(z_0, z_1, z_2) = \alpha$, $\alpha = (\alpha_0, \alpha_1, \alpha_2) = (0.25, 0.5, 0.25)$ の仮定の下 (即ち帰無仮説) での観測データの得られる確率 (帰無仮説の尤度) との比 (尤度比) を求める。その尤度比の対数が LOD 値である。

ところで、 $z_0 + z_1 + z_2 = 1$ なので自由なパラメータは 2 つ。これらの変数と、



相対危険についての関係がわかっている（ただし、疾患がこの座位のみによる場合）。罹患者の同胞の IBD の確率分布は、IBD=0 が確率 0.25、1 が 0.5、2 が 0.25 である。ここで、 r を一般人の発症率、 λ_O を親子間（つまり IBD=1）の相対危険、 λ_M を一卵性双生児間（つまり IBD=2）の相対危険、 λ_S を同胞間の相対危険とすると、それぞれの発症する確率は IBD=0 の同胞が一般人と同じ r （なぜならこの病気はこの座位のみでできるから）、IBD=1 の同胞が $r\lambda_O$ 、2 が $r\lambda_M$ なので、同胞の罹患者の割合、 $r\lambda_S$ は、 $r\lambda_S = 0.25r + 0.5r\lambda_O + 0.25r\lambda_M$ となり、従って $\lambda_S = 0.25 + 0.5\lambda_O + 0.25\lambda_M$ となる。

(z_0, z_1, z_2) は罹患者の同胞が罹患者であった場合の IBD の事後確率なので、ベイズの定理より、

$$\begin{aligned} z_0 &= 0.25/\lambda_S \\ z_1 &= 0.5\lambda_O/\lambda_S \\ z_2 &= 0.25\lambda_M/\lambda_S \end{aligned}$$

ここで、同胞相対危険について説明すると、疾患を持つ個体の同胞が、それと全く無関係の個体よりどれくらい発症率が高いかの比を同胞相対危険 (λ_S) という。一般に、 λ_S は子供の相対危険 λ_O より大きく、ドミナンス分散（用語解説参照）が無いときのみ等しくなる。 z_0, z_1, z_2 が取り得る範囲を考える。当然、 $1 \leq \lambda_O \leq \lambda_S$ なので、上の関係より $z_1 \geq 2z_0$ 、 $z_1 \leq 0.5$ 。当然、 $z_0 \geq 0$ なので、 (z_0, z_1) の動く範囲は図の黒い範囲である。これを Holman's triangle (ホールマンの三角形) とよぶ。

ドミナンス分散が無いという事は $\lambda_S = \lambda_O$ の場合で、これは上式から、 $z_1 = 0.5$ となり、ホールマンの三角形の上辺のみを動くことになる。

ホールマンの三角形の中で動く z_0, z_1, z_2 に対して、観察データ（つまり遺伝子型）の得られる確率（つまり仮説の尤度）を最大とするような最尤推定値を求め、その尤度を $L(H_1)$ とする。帰無仮説の下（つまり病気と座位と無関係）での観察データの得られる確率（つまり帰無仮説の尤度）を $L(H_0)$ とすると、LOD 値（lod score）とは、 $\text{LOD 値} = \log_{10}(L(H_1)/L(H_0))$ と表される。

$L(H_0)$ は観察データが定まれば一定なので、 $L(H_1)$ を最大化することは LOD 値を最大化することに等しい。これを個々の家系について行うのではなく、すべての家系のデータの下での尤度を計算し、それを最大化する。それぞれの家系のデータの下での尤度をすべての家系データの下での尤度に結びつける手順は以下の通りである。

今、 i 番目の同胞対について考える。この同胞対について IBD が確定すれば良いが、通常は確率分布に従う。 i 番目家系について、観察データの下での $\text{IBD}=j$ の確率を w_{ij} とする。異なった IBD の事象は排他的で和は全事象なので、この家系について観察データの下での全仮説の尤度は $L_i = \sum_{j=0}^2 z_j w_{ij}$ 。異なった家系の事象は独立なので、すべての家系データの下での仮説の尤度は、 $\prod_i L_i$ 。帰無仮説の尤度は同様に $\prod_i (\sum_{j=0}^2 \alpha_j w_{ij})$ 。従って、位置 s における尤度比（仮説、 (z_0, z_1, z_2) の帰無仮説に対する）は、

$$Z(s) = \prod_i \frac{z_0 w_{i0} + z_1 w_{i1} + z_2 w_{i2}}{\alpha_0 w_{i0} + \alpha_1 w_{i1} + \alpha_2 w_{i2}}$$

最尤法を用いて、 z_0, z_1, z_2 は上の式が最大になるように推定される。上の式を最大にすることは、対数をとって下の式を最大にすることと同値である。

$$\log Z(s) = \sum_i \log_{10} L_i(s)$$

即ち、全体の LOD 値 = $\sum$ 各家系の LOD 値である。

4 Mapmaker/sibs の紹介

罹患同胞対解析を行うための標準的ソフト Mapmaker/sibs を紹介する。なお、Genehunter 2 に搭載されている、estimate などの罹患同胞対解析の機能は Mapmaker/sibs のものであり、NPL を用いた時のみ Genehunter 固有のアルゴリズムを用いる。Mapmaker/sibs では、罹患同胞対の IBD の数が 0, 1, 2 である確率を z_0, z_1, z_2 とし、その仮説の下に観測データの得られる確率（仮説の尤度）を最大とする最尤法 (Maximum Likelihood Estimation) を用いる。

4.1 動作環境

Mapmaker/sibs は C でかかれたプログラムで、Unix で動く。SunOS (Solaris でも動く)、DEC Alpha などで行実行可能なファイルとして提供されているが、ソースコードから Linux 環境で実行できるようにインストールすることも可能である。また Windows でも、C 言語がコンパイルできる環境であれば実行可能であるが、ここでは触れない。

4.2 入手

一般に、連鎖解析の種々のソフトウェアのリストは以下の Home Page にあり、ここを参照すればほとんどのソフトウェアにアクセス可能である。

<http://linkage.rockefeller.edu/soft/>

WWW ブラウザ (Netscape, IE など) などを用いて簡単にダウンロードできる。

sibs-2.1.tar.Z というファイルをダウンロードする。

4.3 ファイルの解凍

sibs-2.1.tar.Z は tar.Z という圧縮形式であるため、解凍しなければならない。解凍するとたくさんのファイルが生成されるため、ディレクトリを新規に作って sibs-2.1.tar.Z を移動させてから解凍する方が良いだろう。

以下、UNIX でのコマンドにて説明する。

sibs というディレクトリを作成する。

```
% mkdir sibs
```

sibs-2.1.tar.Z をこのディレクトリに移動する。

```
% mv sibs-2.1.tar.Z sibs
```

ディレクトリ sibs へ

```
% cd sibs
```

Z ファイルを解凍する。

```
% uncompress sibs-2.1.tar.Z
```

すると、sibs-2.1.tar というファイルが出現して、sibs-2.1.tar.Z は消去される。

tar ファイルを解凍する。

```
% tar xvpf sibs-2.1.tar
```

以上で解凍完了である。

(% zcat sibs-2.1.tar.Z — tar xvf - とすれば一気に解凍することもできる)

4.4 コンパイルと Mapmaker/sibs の起動

SunOS や Solaris ならば
% make
とすれば、自動的にコンパイル終了する。その後、
% sibs.sun
とすれば、プログラムが開始される。
他の UNIX 環境でも実行可能なので、添付ファイルである INSTALL を良く読み、Makefile の内容を調整してインストールを行なう。

5 数値例

ここでとあるサンプルファイルを用いて Mapmaker/sibs で解析してみた。
dominance variance は有りで計算している。
Mapmaker/sibs で得られる結果は出力ファイルとして mls.out, mls.ps, share.ps
の 3 つのファイルであり、これらはそれぞれ次のような意味を持つ。
mls.out : 座位 (一番左、即ち parameter ファイルで一番上に書かれた座位が
0) 毎の LOD 値とそれを与える z_0, z_1, z_2 の最尤推定値の表。
share.ps : z_0, z_1, z_2 それぞれの最尤推定値を座位を横軸にとって記載。(次頁
の下グラフ)
mls.ps : 座位を横軸に取った LOD 値のグラフ。(次頁の上のグラフ)

結果として、最重要視されるものは mls.ps のグラフである。このグラフ
は LOD 値を用いているが、出力される LOD 値と χ^2 統計量には次の関係が
ある。

$$\begin{aligned}\chi^2 &= -2 \ln[L(H_0)/L(H_1)] \\ &= 2 \times \ln(10) \times \text{lod 値} \\ &= 4.6 \times \text{lod 値}\end{aligned}$$

帰無仮説 H_0 は連鎖無し、対立仮説 H_1 は連鎖有りである。ゲノム統計学の
分野では通常、LOD 値が 3 以上であれば有意であると考えられている。

グラフを見ると座位 9 で非常に高い LOD 値が認められ、それ以外も全体的
に有意な LOD 値を得られている。しかし、通常のデータではこのように
顕著に結果が得られるとは限らず LOD 値を 2 以上のものを示唆的な結果と
して挙げることが多い。

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)

