

基于单体型重构的传递不平衡检验^{*}

李彩霞^{1,2}, 黎培兴³, 方积乾¹

(1. 中山大学医学统计与流行病学系, 广东 广州 510080
2. 加州大学旧金山分校放射系, 美国 加州 CA94107-1774;
3. 中山大学数学系, 广东 广州 510275)

摘 要: 传递不平衡检验是基于家系检测疾病位点与标记位点之间连锁与连锁不平衡的经典分析方法。论文针对紧密连锁位点, 提出了单体型的传递不平衡检验方法, 并把此方法用于分析 IgA 肾病的两紧密连锁位点的基因定位数据。首先在估计核心家系的单体型频率的基础上, 重构单体型的传递/未传递的交叉分类表格, 然后通过检验此表格的对称性与边缘齐性进行传递不平衡检验, 同时, 自编 Excel 宏命令 VBA 程序, 用于家系数据单体型频率估计与重构。此方法充分利用所有家系信息, 并能处理缺失数据。C2093T-C2081T 的单体型多态性与 IgA 肾病关联。此方法推广了已有单体型传递不平衡检验。

关键词: 连锁; 连锁不平衡; 传递不平衡检验; 核心家系
中图分类号: Q-332 **文献标识码:** A **文章编号:** 0529-6579 (2007) 04-0001-05

1 研究背景

在遗传统计分析中, 连锁与连锁不平衡是两个很重要的概念。连锁考察标记位点与疾病位点的重组率 θ 是否为 0.5。如果两个基因座上的等位基因是随机关联的, 即不独立, 这种情况就叫做等位基因关联 (allelic association) 或者连锁不平衡 (linkage disequilibrium, LD)。关联通常反映了分子标记与性状功能突变之间在统计学上的非独立性 (连锁不平衡)。等位基因关联性 (或连锁不平衡性) 可以由一般的群体数据观察到, 传统的病例——对照研究是基于群体而非家系的疾病关联分析, 它通过随机选择病例和对照, 然后比较其在标记等位基因和基因型频率上的差异来说明位点与疾病的关联性。但阳性结果可能由混杂因素造成, 如不同分层人群 (stratified populations) 混杂在一起造成的虚假联系。为了克服不同分层人群混杂的影响, 相应产生了基于家庭的病例——对照研究方法。单倍型相对风险 (Haplotype relative risk, HRR) 分析 (Falk CT and Rubinstein P 1987, Tewilliger and Ott 1992)^[1-2] 与传递不平衡检验 (transmission disequilibrium test, TDT) (Spielman RS 等, 1993)^[3] 是基于家系的关联分析方法。

由 Spielman 等^[3] 提出的单个位点双等位基因的传递不平衡检验已被广泛应用。而且已经由最初

的最简单的双等位基因情形推广到复等位基因情形 (Spielman RS 1996, Sham PC 1995)^[4-5]。它通过考察多态性标记位点的等位基因的不平衡传递, 来说明标记位点与疾病位点的连锁与连锁不平衡。

传递不平衡检验需收集 Trios (父, 母, 一个受累子代) 家系数据, 每一个家系中受累子代的亲代 (父或母) 提供了一个传递的等位基因和一个没有传递的等位基因, 因此, 当要考察的位点有 m 个等位基因时, N 个 Trios 的传递不平衡检验可以整理成表 1。

表 1 N 个受累子代的 $2N$ 个亲代传递和没有传递标记等位基因数目

Tab 1 Transmitted and non transmitted alleles from $2N$ parents to N affected offspring

传递的等位基因	没有传递的等位基因				合计
	1	2	...	m	
1	n_{11}	n_{12}	...	n_{1m}	n_{1+}
2	n_{21}	n_{22}	...	n_{2m}	n_{2+}
...	...	...	...	...	...
m	n_{m1}	n_{m2}	...	n_{mm}	n_{m+}
合计	n_{+1}	n_{+2}	...	n_{+m}	$2N$

这里 n_{ij} 表示基因型为 i 的亲代 (其中传递的等位基因是 i , 未传递的等位基因是 j) 的个数。

^{*} 收稿日期: 2006-12-24
基金项目: 国家自然科学基金资助项目 (30170523)
作者简介: 李彩霞 (1970 年生), 女, 博士后; 通讯联系人: 方积乾; E-mail: caixia11@rad.pgy.ucsf.edu
(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

Shan^[3] 1995 年指出, 可通过检验表 1 数据的对称性或边缘齐性, 检验疾病位点与标记位点之间是否存在连锁或连锁不平衡。对称性检验统计量为

$$T_s = \sum_{i,j} \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}}$$

当 H_0 不连锁或连锁平衡成立时, T_s 近似服从自由度为 $m(m-1)/2$ 的 χ^2 分布。由 Spelman^[4] 提出用于传递不平衡分析的边缘齐性检验统计量为

$$T_h = \frac{m-1}{m} \sum_i \frac{(n_{+i} - n_{i+})^2}{n_{+i} + n_{i+} - 2n_{ii}}$$

Spelman^[4] 同时指出, 当 H_0 成立时, 近似服从自由度为 $m-1$ 的 χ^2 分布。

若有多个紧密连锁的位点, 考虑单体型的传递, 若每个个体的内在的单体型对可确定, 则可直接把每个单体型看成一个位点的一个等位基因, 列出上表 1 中的数据, 直接利用上面的对称性或边缘齐性检验。但实际上, 对于多个位点情形, 由于通常的观测数据是各个位点的基因型, 从而, 个体的单体型对存在不确定性。因此无法直接列出上面表 1 的传递/不传递的表格。下面在重构核心家系的单体型的基础上, 进一步估计出表格 1 中的数据。

2 方法

设一个核心家系, 其中至少有一个子代是受累(患病)个体, 则对家系 $r=1, 2, \dots, N$, 定义 H_r 为满足如下条件的 $\{(i, j, k, l)\}$ 集合: 单体型对 (i, j) 与父亲的观测基因型 g_f 匹配, 单体型对 (k, l) 与母亲的观测基因型 g_m 匹配, 单体型对 (i, k) 与受累小孩的观测基因型 g_c 匹配。即父亲的 (i, j) 中传递了 i , 未传递 j 给此受累小孩, 母亲的 (k, l) 中传递了 k , 未传递 l 给此受累小孩。定义

$$H_r = \{(i, j): \exists (k, l), (i, j, k, l) \in H_r\},$$

$$H_{nr} = \{(k, l): \exists (i, j), (i, j, k, l) \in H_r\},$$

则家系 r 可能为某个 $(i/k, j/l), (i, k, j, l) \in H_r$ 提供值为 1 的贡献, 家系 r 中的父亲可能为表格 1 中的某个格子 $(i, j) \in H_r$ 提供值为 1 的贡献, 母亲可能为表格 1 中的某个格子 $(k, l) \in H_{nr}$ 提供值为 1 的贡献。

设 $X_{ikl}(r)$ 为家系 r 的父母传递 ik 未传递 jl 提供的贡献, $Y_{ij}^f(r)$ 为家系 r 的父亲传递 i 未传递 j 提供的贡献, $Y_{ij}^m(r)$ 为家系 r 的母亲传递 i 未传递 j 提供的贡献。表 1 中 n_{ij} 可表达为

$$n_{ij} = \sum_{r=1}^N \sum_k \sum_l [X_{ikl}(r) + X_{klij}(r)] =$$

$$\sum_{r=1}^N [Y_{ij}^f(r) + Y_{ij}^m(r)]$$

从家系 r 的实际数据 $\{g_f, g_m, g_c\}$ 中, 当其连锁不能确定时, 我们不能直接得到每个 $X_{ikl}(r)$, $Y_{ij}^f(r)$ 和 $Y_{ij}^m(r)$ 的观测值, 即不能确定此家系的亲代究竟为哪一个格子提供了贡献, 但是, 我们可以知道有 $\sum_{(i,j,k,l) \in H_r} X_{ikl}(r) = 1$ 。

当 H_0 成立时, 已知家系观测基因型的条件下, 家系 r 中的父亲, 母亲, 子代潜在的单体型对恰好为 $(i, j), (k, l), (i, k)$ 的概率, 即恰好为 (i, j, k, l) 提供贡献 1 的概率为

$$P\{g_r = (i, j), g_m = (k, l),$$

$$g_c = (i, k) | g_f, g_r \times g_m, A_r =$$

$$P\{i, j, k, l | i, j \times k, l | g_f, g_r \times g_m, A_r =$$

$$\frac{P_r P_r P_r P_r}{\sum_{(i,j,k,l) \in H_r} P_r P_r P_r P_r}$$

这里 A_r 表示子代患病, i, j, k, l 表示父亲传递 i 未传递 j , 母亲传递 k 未传递 l , 因此, 不妨给 $X_{ikl}(r), Y_{ij}^f(r), Y_{ij}^m(r)$ 的估计分别为

$$\hat{X}_{ikl}(r) = \begin{cases} \frac{P_r P_r P_r P_r}{\sum_{(i,j,k,l) \in H_r} P_r P_r P_r P_r}, & (i, j, k, l) \in H_r \\ 0 & (i, j, k, l) \notin H_r \end{cases}$$

$$\hat{Y}_{ij}^f(r) = \begin{cases} \frac{P_r P_r}{\sum_{(i,j) \in H_f} P_r P_r}, & (i, j, k, l) \in H_r \\ 0 & (i, j, k, l) \notin H_r \end{cases}$$

$$\hat{Y}_{ij}^m(r) = \begin{cases} \frac{P_r P_r}{\sum_{(i,j) \in H_{nr}} P_r P_r}, & (i, j, k, l) \in H_{nr} \\ 0 & (i, j, k, l) \notin H_{nr} \end{cases}$$

此时, 仍有

$$\hat{Y}_{ij}^f(r) = \sum_k \sum_l \hat{X}_{ikl}(r), \hat{Y}_{ij}^m(r) = \sum_k \sum_l \hat{X}_{klij}(r)$$

从而, 给出估计

$$n_{ij} = \sum_{r=1}^N \sum_k \sum_l [\hat{X}_{ikl}(r) + \hat{X}_{klij}(r)] = \sum_{r=1}^N [\hat{Y}_{ij}^f(r) + \hat{Y}_{ij}^m(r)]$$

由于如此重构 n_{ij} 需要利用各单体型频率, 而各单体型频率通常未知, 需要基于样本数据进行估计。我们可以用核心家系的单体型频率估计方法^[6], 估计出各单体型频率。

Zhao^[7] 于 2000 年曾分别用了 3 种方法重构表 1 中的数据。方法 1 只利用了能确定单体型的基因型数据, 把单体型不能确定的家系不纳入分析,

从而可直接列出表 1 中的数据。方法 2 对于每个家系中的与基因型相匹配的所有可能单体型对给予相等的权重，例如，如果某亲代潜在的单体型对有 2 种可能性，则分配每一种可能性为 1/2。这很可能导致偏差，因为我们有理由认为潜在的单体型对更有可能是这些家系来源群体的单体型对频率最大的那一种单体型对。方法 3 把所有家系的父母看成来自一个满足 Hardy-Weinberg 平衡的群体的随机样本，没有利用子代信息，在估计单体型频率的基础上，重构表 1 中的数据。

3 应 用

IGA 肾病（IGA Nephropathy）是最为常见的原发性肾小球肾炎。近年来，大量研究表明 IGA 肾病的发病有遗传因素的参与。中山大学遗传教研室王一鸣教授把 Mgsin 基因，作为 IGA 肾病的候选基因，Li You-jing^[8]测定了 Mgsin 基因的 C2093T 位点与紧邻的 C2081T 位点，并对此 2 个位点的多态性与 IGA 肾病的易感性之间的进行了相关研究。其中用于分析 Mgsin 基因的家系共 232 个，其中有些患病个体的亲代基因型数据缺失，具体信息见表 2。

表 2 位点 C2093T 与 C2180T 共传递家系数据¹⁾
Tab 2 Families for co transmission of marker C2093T and marker C2180T

C2093T trios+ C2180T trios	C2093T SPF+ C2180T SPF	C2093T trios+ C2180T SPF	C2093T SPF+ C2180T trios	Total
125	56	25	26	232

1) SPF 单亲家系（single parent families）

下面我们在估计家系个体单体型频率的基础上，重构所有 232 个家系传递—未传递的 C2093T—C2081T 单体型数据，然后进行传递不平衡检验，分别采用对称性与边缘齐性检验。

利用自编 Excel 宏命令 VBA 程序 haplo computestep.s，逐步实现如下步骤：

(1) 首先导入数据至 Excel，部分数据以及输入格式见表 3。

数据的前 6 列分别为家系编号，家系中个体编号，父亲，母亲，性别，是否患病。最后两列为 C2093T 与 C2180T 两个位点的基因型。其中，基因型中的“0”代表缺失。

表 3 数据输入格式
Tab 3 Format of input data

ped	no	f	m	sex	affect	l1	l2
2	1	.	.	1	0	1/1	1/1
2	2	.	.	2	0	2/2	2/2
2	3	1	2	1	2	1/2	1/2
3	1	.	.	1	0	0/2	0/1
3	2	.	.	2	0	1/2	1/2
3	3	1	2	1	2	1/2	1/1

(2) 运行宏命令 haplo 得到每个家系的亲代单体型传递不传递的所有可能情形。见表 4。

(3) 给出 4 条单体型 (11, 12, 21, 22) 频率初值，与迭代收敛精度 10^{-7} ，运行宏命令 compute step 迭代计算单体型频率，即表 4 中的 F 。同时输出每个家系的亲代对应的 $\hat{Y}_{ij}^{(f)}$ (η), $\hat{Y}_{ij}^{(m)}$ (η), 即表 4 中的 $w(i, j)$ 。

表 4 亲代传递 / 不传递的单体型与输出
Tab 4 Transmitted/non transmitted haplotypes from parents and output

ped	no	l1	l2	haplo pairs	z1	z2	z3	tr/ntr	homo	w(i, j)	F
2	1	1/1	1/1	11/11	0	0	0	1/1	1	1	0.0633601
2	2	2/2	2/2	22/22	0	0	0	4/4	4	1	0.5827084
3	1	0/2	0/1	11/21	1	0	-1	1/3	0	0.05738	0.2151468
3	2	1/2	1/2	21/12	0	-1	1	3/2	0	0.43849	0.1387847
3	1	0/2	0/1	11/22	1	0	0	1/4	0	0.03701	
3	2	1/2	1/2	21/12	0	-1	1	3/2	0	0.43849	
3	1	0/2	0/1	21/11	-1	0	1	3/1	0	0.05738	
3	2	1/2	1/2	11/22	1	0	0	1/4	0	0.03076	
3	1	0/2	0/1	21/12	0	-1	1	3/2	0	0.52770	
3	2	1/2	1/2	11/22	1	0	0	1/4	0	0.03076	
3	1	0/2	0/1	21/21	0	0	0	3/3	3	0.19484	
3	2	1/2	1/2	11/22	1	0	0	1/4	0	0.03076	
3	1	0/2	0/1	21/22	0	0	1	3/4	0	0.12568	
3	2	1/2	1/2	11/22	1	0	0	1/4	0	0.03076	

表 4 中列 “no” 中，no=1 为家系中的父亲，no=2 为家系中的母亲，“haplopairs” 即为此亲代可能的单体型对，列 “1”，“2”，“3” 为传递单体型 1 2 3 的标识。列 “trans” 为 “传递的单体型 / 未传递的单体型”。列 “hom” 表示 “是否纯合子”。单体型 1 2 3 4 表示单体型 (CC CT TC TT)，在数据中以 (11 12 21 22) 表示。从表 4 中可以看出，表 3 中的家系 2 的双亲的单体型对 (传递的单体型 / 未传递的单体型) 可唯一确定，而家系 3 的双亲的单体型对竟有 6 种可能性。

基于第 (3) 步的数据，利用 $\hat{Y}_{ij}^{(f)}(r)$ $\hat{Y}_{ij}^{(m)}(r)$ ，即表 4 中的 $w(i)$ ，用 SAS 程序计算输出各 n_i 。SAS 结果见表 5。

表 5 SAS 结果
Tab 5 The SAS system

n12	n21	n13	n31	n14	n41
6 80964	17 2134	5 94114	3 94221	7 26534	6 64342
n23	n32	n24	n42	n34	n43
83 5696	53 3772	53 5154	30 6329	9 54337	7 68759

SAS 输出中的 n_i 即为 n_i 。利用重构估计到的各 n_i 进行传递不平衡检验。对称性检验统计量 $T_s=17.844 5$ 自由度为 6 P 值为 0.006 632 827 边缘齐性检验统计量 $T_h=21.670 4$ 自由度为 3，P 值为 0.000 076 388 从而推断 C2093 T—C2081 T 的单体型多态性与 IgA 肾病关联。

4 结 语

用于复等位基因的传递不平衡检验，通常检验的是表 1 数据的边缘齐性或对称性，其中由 Spielman 提出的边缘齐性检验方法，由于计算方便，应用非常广泛，对称性检验虽然也简单方便，但由于自由度大，检验效能较低。

当边缘齐性或对称性检验应用到紧密连锁位点时，需要构造单体型的传递 / 未传递的交叉分类表格。而由于单体型的不可观测，需在估计单体型频

率基础上重构个体的单体型，此时，应充分利用家系信息进行单体型的重构，而不应把家系数据当成群体数据或只利用连锁相能确定的家系数据。本文提出的方法充分利用所有家系信息，并能处理缺失数据。另外，自编的家系数据单体型频率估计与重构软件简单易用。

致谢：本文样本数据由中山大学遗传教研室王一鸣教授提供，特此感谢。

参考文献：

[1] FALK C T RUBINSREN P Haploype relative risks: An easy reliable way to construct a proper control sample for risk calculations[J. AnnHum Genet 1987 51: 227—233

[2] TERWILLIGER J OTT J A haploype-based haploype relative risk approach to detecting allelic associations [J. Hum Hered 1992 42: 337—346

[3] SPEIMAN R S MCGINNIS R E EWENSW J Transm ission test for linkage disequilibrium: the insulin gene region and insulin dependent diabetes mellitus[J. Am J Hum Genet 1993 52(3): 506—516

[4] SPEIMAN R S EWENSW J The TDT and other family-based tests for linkage disequilibrium and association [J. Am JHum Genet 1996 59(4): 983—989

[5] SHAM P C CURTIS D An extended transmission/disequilibrium test (TDT) for multi-allele marker loci[J. AnnHum Genet 1995 59(2): 323—336

[6] 李彩霞，黎培兴，方积乾. 核心家系数据的紧密连锁位点的单体型频率估计[J. 中山大学学报: 自然科学版, 2005 44(3): 9—11

[7] ZHAO H ZHANG S MERKANGAS K R et al Transm ission/disequilibrium test using multiple tightly linked markers[J. Am JHum Genet2000 67: 936—946

[8] LI You ji DU Yong LI Cai xia et al Family-based association study showing that immunoglobulin A Nephropathy is associated with the polymorphisms 2093C and 2180T in the 3' untranslated Region of Mesp in Genq [J. J Am Soc Nephrol 2004 15(7): 1739—1743

(下转第 8 页)

sis of airfoil flutter with structural nonlinearity [M]. Advances in Nonlinear Dynamics in China— Theory and Practice, Chapter 3. Lisse: The Netherlands, Swets & Zeitlinger Publishers, 2002

[10] 丁千, 王冬立. 用规范形直接法研究立方非线性机翼颤振 [J]. 飞行力学, 2005 23(3): 85—88

[11] 陈衍茂, 刘济科. 非线性颤振极限环稳定性判别的复数正规形法 [J]. 航空动力学报, 2007 22(4): 614—618

An Improved Equivalent Linearization Method for Strongly Nonlinear Flutter Systems

CHEN Yanmao LU Jike

(Department of Applied Mechanics and Engineering, Sun Yat-sen University, Guangzhou 510275, China)

Abstract: The limit cycle flutter of a two-dimensional airfoil with a strongly cubic pitching stiffness is studied. The linearization method is used to linearize the free oscillating equation of the airfoil in the uncoupled pitching mode. The linearized equation is addressed by the harmonic balance method and the equivalent linear stiffness of the nonlinear stiffness is obtained. The attained equivalent stiffness is implemented in the equivalent linearization method of nonlinear flutter analysis and the bifurcation chart of the flutter system is obtained. Two numerical examples are taken as illustrations to validate the proposed method. It shows that the attained equivalent stiffness is in excellent agreement with the numerical solutions and it can improve the accuracy of the equivalent linearization method. In addition, it is quite intuitive to judge whether the bifurcation is supercritical or subcritical by the proposed method.

Key words: nonlinear flutter; harmonic balance method; equivalent linearization method; equivalent linear stiffness

(上接第 4 页)

Haplotype Transmission Disequilibrium Test Based on Haplotype Reconstruction

LI Caixia², LI Peixing³, FANG Jiqian¹

- (1. Department of Medicine Statistics and Epidemiology, Sun Yat-sen University, Guangzhou 510080, China
2. Department of Radiology, University of California, San Francisco, CA 94107—1771, USA
3. Department of Mathematics, Sun Yat-sen University, Guangzhou 510275, China)

Abstract: Transmission disequilibrium test (TDT) is a classical family-based test to detect linkage and linkage disequilibrium. Haplotype transmission disequilibrium test is proposed for tightly linked loci. The method is applied to the genemapping data for Immunoglobulin A (IgA) nephropathy. Based on haplotype frequency estimation with nuclear family data, the transmission/non-transmission cross table in TDT is reconstructed. And then symmetry test and homogeneity test for the reconstructed contingency table are given to detect haplotype transmission disequilibrium. In addition, Excel macro VBA program is developed to estimate haplotype frequencies and reconstruct transmission/non-transmission haplotypes. The proposed method available information from all families including families with missing value. The haplotype polymorphism of C2093 T-C2081 T is associate with IgA. The method generalizes the existed haplotype transmission disequilibrium test.

Key words: linkage; linkage disequilibrium; transmission disequilibrium test; nuclear family