[文章编号] 1000-2200(2008)06-0706-04

·校庆约稿·

基于全基因组关联分析的基因(环境)交互作用统计学方法进展

吴学森

[关键词] 统计学;基因型;流行病学方法;综述
[中国图书资料分类法分类号] R 195.4;Q 343.1 [文献标识码] A

一般说来,复杂疾病(complex diseases)的发生与发展并不能完全由遗传变异来解释,而应该理解为遗传变异和环境因素共同作用的结果;每个基因与疾病之间可能只存在弱关联,并不存在主基因效应,这种弱效应更容易受到外部环境的影响;如果忽略了基因与环境之间的交互作用(包括基因-基因、基因-环境的交互作用),就无法真实、准确地描述遗传变异的效应,也就出现了对同一种疾病易感位点的研究,在不同的研究者之间产生相互矛盾结果的现象[1,2]。虽然人们已逐渐认识到研究基因-基因、基因-环境间交互作用对准确把握复杂疾病的发生与运行机制十分有益[3-5],正如 Hunter[6] 指出的那样,研究基因-基因、基因-环境交互作用对研究复杂疾病的遗传效应至关重要:(1)能增强统计遗传学检测方法的检验功效;(2)能更准确地估计影响复杂疾病的群体遗传效应和环境效应;(3)能更好地揭示复杂疾病的机制,并解释环境暴露是怎样影响到生物通路的功能;(4)通过揭示环境因素的改变如何影响到生物通路的反应,可为疾病的预防和治疗提供全新的策略。

近年来,在研究基因之间、基因与环境因子之间的统计交互作用的方法学上进展很快。从研究设计类型上看,能够用于统计交互作用的设计类型有:(1)队列研究;(2)无关个体病例对照研究(case-control using unrelated control),包括成组设计的病例对照研究和配比病例对照研究、两阶段病例对照研究(two-stage case-control)等;(3)有关个体病例对照研究(case-control using related control),包括病例父母亲对照研究、病例同胞对照研究、扩展同胞对照研究(extended sib-pair)等;(4)无关和有关个体的联合病例对照研究(case-control using related and unrelated control);(5)单纯病例研究(case-only study);(6)双生子研究(twins study);(7)连锁分析;(8)分离分析(segregation analysis);(9)连锁与分离的联合分析(combined linkage and segregation analysis);(10)不完全病例-对照研究(partial case-control study)。这些方法在分析交互作用时各有优缺点,又相互优势互补。

从分析方法上看,有叉生分析(crossover analysis)、多因素 Logistic 回归模型、多因子降维法(multifactor dimensionality reduction)和基于复合 LD(composite linkage disequilibrium)的交互作用分析方法。

1 交互作用的叉生分析法

叉生分析[7,8]是遗传流行病学研究中分析基因-环境交互作用的最基本的方法之一,群体病例对照研究、病例父母亲对照研究、病例同胞对照研究、队列研究设计类型的资料均可用叉生分析方法分析基因与环境之间的交互作用。

1.1 叉生分析 表1所示的2×4叉生分析是基因与环境因素相互作用研究中的基本研究单元,它表示基因(G)、环境因素(E)均为二分类变量而组成的4种可能的组合表。同时暴露于两因素相对于同时不暴露于两因素的危险性(比值比,*OR*)记为 OR_{ge}(简记为 A);单独暴露于基因或环境因素的危险性分别记为 OR_g、OR_e(分别简记为 B、C);两因素均未暴露的病例和对照组作为共同参比组,$OR=1$。

表1 基因(G)与环境因素(E)因素作用的2×4叉生分析

基因(G)	环境因素(E)	病例组	对照组	*OR*值	意义
+	+	a	b	$OR_{ge}=A=ah/bg$	G、E联合作用效应
+	-	c	d	$OR_g=B=ch/dg$	G单独作用效应
-	+	e	f	$OR_e=C=eh/fg$	G单独作用效应
-	-	g	h	1	共同对照

表中基因与环境联合作用的效应不仅包括两者分别作用的效应,还可能包括基因与环境作用的叠加,也可能呈现基因与环境作用的相乘效应。通过不同的模型,可以判别基于不同模型的两因素间交互作用是否存在及其大小。

那么在叉生分析中,交互作用又是怎样被度量呢?由于交互作用的存在与否,与所选择的模型密切相关,根据 Rothman[9] 提出的基于相加模型计算交互作用的指标,有以下几种:

(1)交互作用指数(the synergy index,S)

$$S=\frac{A-1}{(B-1)+(C-1)} \quad (1.1)$$

意义:当 $S=1$ 时,无交互作用;$S\neq1$ 时,基因(G)与环境(E)存在相加模型交互作用;$S>1$ 时,两因子间有正交互作用;$S<1$ 时,两因子间有负交互作用;S 的绝对值越大,基因(G)与环境(E)之间的交互作用越强。

(2)交互作用归因比(attributable proportion of interaction,AP)

$$AP=\frac{A-(B+C-1)}{A} \quad (1.2)$$

意义:AP 表示总效应中有多大比例归因于基因(G)与环境(E)之间的交互作用。AP 的绝对值越大,基因(G)与环境(E)之间的交互作用越强。

(3)纯交互作用归因比(AP^*)

[收稿日期] 2008-10-05
[作者单位] 蚌埠医学院 流行病与卫生统计学教研室,安徽 蚌埠 233030
[作者简介] 吴学森(1964-),男,博士,教授.

$$AP^* = \frac{A-(B+C-1)}{A-1} \quad (1.3)$$

意义：AP^* 表示由基因(G)与环境(E)两因素引起的效应中归因于二者交互作用的比例。

(4)交互作用超额相对危险度(relative excess risk of interaction, RERI)

$$RERI = A-(B+C-1) \quad (1.4)$$

意义：表示基因(G)与环境(E)两因素联合作用与其单独作用之和的差值，同时也表示交互作用与基因(G)与环境(E)两因素以外的因素作用之间的关系(这里公式是假定后者 $OR=1$)，如果未知因子作用很大，则所研究的交互作用就变得十分次要而没有意义。这里，RERI 即为基于相加模型的两因素交互作用值。RERI 的绝对值越大，基因(G)与环境(E)之间的交互作用越强。

从上述四项指标可以看出，上述公式均是以基因(G)与环境(E)的相加模型的交互作用为前提的。

1.2 *交互作用的假设检验* 在对基因(G)与环境(E)之间交互作用做出定量测量后，需要对其进行假设检验，以推断其是否有统计学意义。

设基因(G)与环境(E)之间交互作用效应量为 T，根据相加模型，$T=(OR_{GE}-OR_G)-(OR_E-OR)$。如果 $T=0$，说明基因(G)与环境(E)之间无相加模型交互作用；如果 $T\neq 0$，说明基因(G)与环境(E)之间有相加模型交互作用；因此，在零假设下($T=0$)，统计量

$$T_{GE-OR}=\frac{T^2}{Var(T)} \quad (1.5)$$

近似服从 $\chi^2_{(1)}$ 分布。其中，$Var(T)$ 为 T 的方差，由下式求得

$$Var(T)=\sum_{i=0}^{1}\sum_{j=0}^{1}\frac{R_{ij}(1-R_{ij})}{N_{ij}}$$

其中，R_{ij} 为基因(G)与环境(E)各种暴露组合下的率，N_{ij} 为基因(G)与环境(E)各种暴露组合下的观察人数。

Andersson 等[10]在上述叉生分析基础上，引进哑变量(dummy)，见表 2。然后再用 Logistic 回归检测基因－环境变量间的交互作用。他们认为对数据这样处理后检验功效会更高，但该文并没有进一步给予论证，因此，这种变换的可靠性还有待进一步验证。

表 2　不同暴露水平的哑变量定义

基因	环境	变量 1	变量 2	变量 3
i=0	j=0	0	0	0
i=0	j=1	1	0	0
i=1	j=0	0	1	0
i=1	j=1	0	0	1

1.3 *叉生分析的优势与局限* (1)叉生分析表作为病例－对照研究中基本的流行病学分析方法，具有信息量丰富、计算直观等优点。通过叉生分析表，不仅分别获得基因和环境因素各自的主效应，还获得基于不同模型(相加或相乘模型)的交互作用效应。(2)由于 2×4 叉生分析表只能分析单个基因(G)与单个环境因子(E)，且二者均为二分类变量时的交互作用，它无法分析每个因子是多分类或多个因素间的交互作用。(3)在研究基因(G)与环境因子(E)之间的交互作用时，若存在混杂因素的影响，则交互作用的测量结果将会被歪曲。此时，需要控制混杂因素后再进行叉生分析，以反映交互作用的真实强度。

2　交互作用的 Logistic 回归模型分析方法

Logistic 回归模型不仅可以在乘法遗传模型假定下用等位基因作为遗传变量，也可用在一定的遗传模型(如显性模型、隐性模型等)假定下用基因型作为遗传变量进行基因－基因、基因－环境交互作用的分析[11~13]。下面以基因型作为遗传变量为例说明建模方法。

2.1 *基因－环境交互作用的 Logistic 回归模型* 设 D 代表疾病，E 代表环境因素，g 表示疾病相关位点的基因型，该位点有两个等位基因：易感基因 M 和正常基因 m。人群中环境因素暴露率用 $p_{(E)}$ 表示，易感基因频率用 p_M 表示。设该位点符合 H-W 平衡，则基因型 MM、Mm、mm 在人群中的频率分别为 p_M^2、$2p_M(1-p_M)$ 和 $(1-p_M)^2$。假设环境因素和遗传因素在人群中独立存在。则可建立 Logistic 回归模型：

$$P(D=1|G,E)=\frac{\exp(\alpha+\beta_g G+\beta_e E+\beta_{ge}GE)}{1+\exp(\alpha+\beta_g G+\beta_e E+\beta_{ge}GE)} \quad (1.6)$$

人群中的基线患病率为 $P(D=1)=\frac{\exp(\alpha)}{1+\exp(\alpha)}$，$OR_g=\exp(\beta_g)$、$OR_e=\exp(\beta_e)$ 和 $OR_{ge}=\exp(\beta_{ge})$ 分别为基因、环境和基因－环境交互作用的比数比。OR_g(OR_e)即当个体不暴露于环境因素(易感基因型)时的遗传因素(环境因素)的比值比。$OR_{ge}=1$ 说明环境因素与遗传因素无交互作用，$OR_{ge}\neq 1$ 即说明环境因素和遗传因素存在交互作用。$OR_{ge}>1$ ($OR_{ge}<1$)说明环境因素能促进(抑制)易感基因的表达，或者说遗传因素能增加(降低)人体对环境因素的敏感性。利用模型中的偏回归系数，可以解释不同组合下的比数比含义。

对于基因－基因交互作用的 Logistic 回归模型，只需将其中的一个基因型变量代替式(1.6)的环境变量即可，其原理相同。

3　交互作用的多因子降维法

多位学者[14~19]提出了一种新的基因－基因(基因－环境)交互作用分析方法——多因子降维法(multifactor dimensionality reduction, MDR)。在该方法中，“因子”指交互作用研究中的变量(如基因型或环境因子)，“维”是指研究的多因子组合中因子的位点数目。MDR 以疾病易感性类型(如高危、低危)的方式建模，将研究中的多个因子看作一个多因子组合(基因型组合)，这样就把高维的结构降低到一维两水平(即高维或低维)，即降维。MDR 是一种非参数、无需遗传模式假定的分析方法，适用于病例对照研究设计。其分析原理和步骤如下(见图 1)：

第 1 步：随机将数据平均分为 10 等份，其中 9 份为训练样本，另外一份为检验样本，以便进行交叉验证。

第 2 步：从众多研究因素中选择 n 个因子，可以是 *SNP* 或分类明确的环境因子，此 n 个因子代表 n 维。

第3步：根据这n个因子中每个的观察值水平，将个体

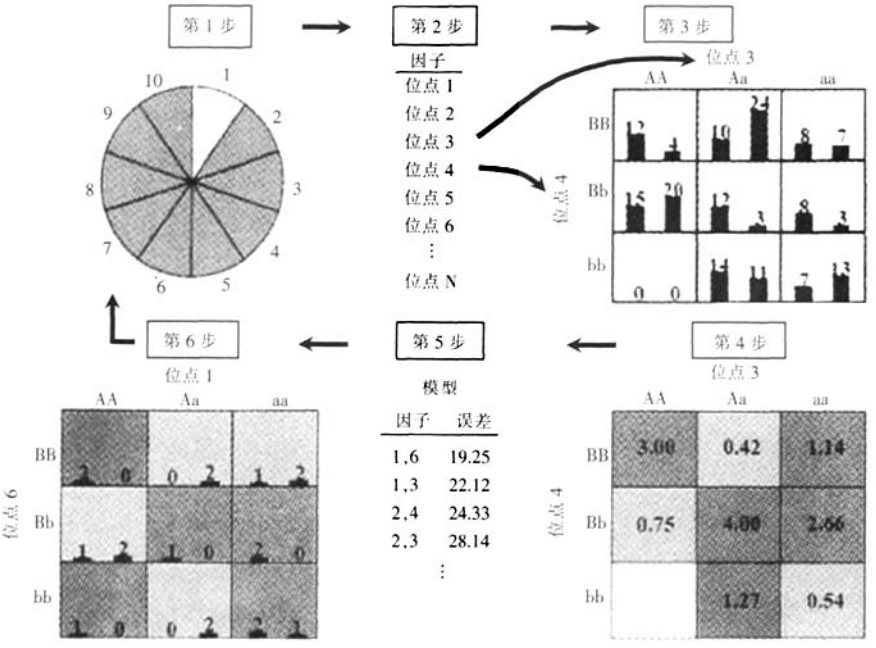

图1 MDR基本步骤示意图

划分为不同的分类，也就是图中的单元格。单元格中左侧直方图表示病例，右侧直方图表示对照。

第4步：在n维的每个多因子分类（单元格）中，计算病例数和对照数的比值，若病例数与对照数之比达到或超过某个阈值（例如≥1），则标为高危，反之则为低危。这样就把n维的结构降低到一维两水平。

第5步：多因子分类的集合中包含了MDR模型中各因子的组合。在所有的两因子组合中，选择错分最小的那个MDR模型，该两位点模型在所有模型中将具有最小的预测误差。

第6步：通过十重交叉验证评估模型的预测误差，以及单元格分配时的相对误差。也就是说，模型拟合9/10的数据（训练样本），其预测误差将通过剩下1/10的数据（检验样本）来衡量。选择预测误差最小的模型作为最终的模型，取10次检验的预测误差平均值，作为模型相对预测误差的无偏估计。由于数据分组的方式对交叉验证的结果影响较大，因此，十重交叉验证过程将重复进行10次，对n个因子可能的集合将重复进行10×10次的交叉验证。

通过十重交叉验证，在一定程度上可以避免因数据转换的偶然性，使Ⅰ类错误增大而产生假阳性结果的影响。预测误差是衡量MDR模型在独立检验的亚组中预测危险状态的指标，通过十重交叉验证的亚组中每一个的预测误差的平均值来计算。根据交叉验证的预测误差的平均值，选择最佳的n因子模型，并根据不同的因子数重复以上过程。最终筛选出最有可能存在交互作用的基因。

MDR的优势在于不需要考虑疾病的遗传模型，它利用计算机运算速度快的优势，对多个基因进行随机组合，按照上述方法找出存在交互作用的基因位点。但当主效应存在时，用MDR方法很难得到最终模型，且同样受遗传异质性的影响；它只是一种数据挖掘方法，不是严格意义上的统计方法，还无法判断它的Ⅰ类错误和检验功效。

MDR分析软件包可在 http://www.epistasis.org/mdr.html 免费下载。

4 基于复合LD的交互作用分析法

吴学森等[20]提出基于复合LD的交互作用的分析法。该方法以病例-对照试验设计为基础，基于LD计算方法，构建完全有别于以上方法的一种新型基因间交互作用的统计分析方法：(1)用两个位点（基因）单倍型的外显率（P_{AB}）与等位基因的边际外显率的乘积（$P_A \cdot P_B$）的偏差（$\delta_{AB} = P_{AB} - P_A \cdot P_B$），分别定义病例组和对照组两个位点交互作用的度量，进而综合两组交互作用度量构造检验交互作用的统计量；(2) 对于基因-环境交互作用模型的构建，则将环境（分类型变量）变量视为“虚拟位点”（例如E=1表示环境暴露，E=0表示即非暴露），则同样依据上述方法构建其模型。

4.1 基因型数据的联合概率分布及其表达 对于基因之间、基因与环境之间的交互作用统计量的构建，无论是二阶或高阶情形，均至少涉及两个变量。在本研究中，均以病例-对照试验设计为基础，个体的基因数据一律用其基因型表示。无论是病例组还是对照组，均设两个位点的等位基因分别为A,a;B,b，则它们的联合基因型分布可表述为表3的形式：

则，配子的LD系数为：$\delta_{AB} = P_{AB} - P_A P_B$；非配子的LD系数为：$\delta_{A/B} = P_{A/B} - P_A \cdot P_B$，其中，$p_{AB} = P_{AB}^{AB} + P\,AB_{Ab} + P_{aB}^{AB} + P_{ab}^{AB}$；$p_{A/B} = P_{AB}^{AB} + P_{Ab}^{AB} + P_{aB}^{AB} + P_{bB}^{Ab}$。但是，当计算病例组或对照组的$\delta_{AB}$时，需要知道双杂合子的概率$P_{ab}^{AB}$、$P_{aB}^{Ab}$。然而，当它们的相未知时，则无法确定其值，只能进行单倍型推断。由于单倍型推断总是存在误差，这给后面构造的检验交互作用的统计量带来很多不确

表 3　两位点基因型概率

	BB	Bb	bb	合计
AA	P_{AB}^{AB}	$2P_{Ab}^{AB}$	P_{Ab}^{Ab}	P_A^A
Aa	$2P_{aB}^{AB}$	$2P_{ab}^{AB}+2P_{aB}^{Ab}$	$2P_{ab}^{Ab}$	$2P_a^A$
aa	P_{aB}^{aB}	$2P_{ab}^{aB}$	P_{ab}^{ab}	P_a^a
合计	P_B^B	$2P_b^B$	P_b^b	1

定的因素，为此，该研究者尝试采用复合的连锁不平衡概念，以避免推断单倍型。复合的连锁不平衡定义为[21-23]：$\Delta_{AB}=\delta_{AB}+\delta_{A/B}$，结合上述的等式和表 3 可知，双杂合子被合并为 2（$P_{ab}^{AB}+P_{aB}^{Ab}$），其值可由实验数据通过直接计数的方法获得，从而避免了单倍型的推断。因此，不需要推断单倍型就可研究两位点的交互作用。

4.2　统计量的构造　从理论上讲，为了检验两位点间的交互作用，我们可以比较病例－对照组中的 LD 的差别，来检验交互作用的存在与否。但是，如前所述，由于在计算两位点间的 LD 时，我们遇到双杂合子的问题，一般来说，我们是无法知道双杂合子的相，亦即相未知。但我们可以通过复合 LD 来避免这个问题。

复合连锁不平衡定义为：$\Delta_{AB}=\delta_{AB}+\delta_{A/B}$

通过计算病例组和对照组的复合的连锁不平衡系数 Δ_A、Δ_N，构造如下的检验统计量：

$$T_1=\frac{(\hat{\Delta}_A-\hat{\Delta}_N)}{Var(\hat{\Delta}_A)+Var(\hat{\Delta}_N)} \qquad (1.7)$$

上式近似服从 $\chi^2(1)$ 分布。

该文通过计算机模拟进一步验证了这种方法的检验功效，并与常用的 Logistic 回归方法进行了比较，在各种遗传模型下前者的检验功效都比后者高。但也存在一些不足，如果在群体中存在的亚结构影响到致病位点间的 LD，则会产生假阳性率。

随着基因分型技术的发展，分型成本的降低，特别是随着 Illumina、Affymetrix 等基因分型技术的推广应用，基于群体的大规模 SNP 基因分型数据的关联分析方法（特别是全基因组关联分析方法）（包括基因间的交互作用）将成为多基因复杂疾病基因定位研究的主要方法。因而，近年来对关联分析方法的研究和应用一直是国际上生物计算领域的研究热点，这对统计遗传学方法的研究既是一次发展机遇，也是一个巨大的挑战。

[参 考 文 献]

[1] Andrieu N, Goldstein AM. Epidemiologic and genetic approaches in the study of gene-environment interaction: an overview of available methods[J]. *Epidemiol Rev*, 1998, 20(2): 137-147.

[2] Manolio TA, Bailey-Wilson JE, Collins FS. Genes, environment and the value of prospective cohort studies[J]. *Nat Rev Genet*, 2006, 7(10): 812-820.

[3] Ottman R. Gene-environment interaction: definitions and study designs[J]. *Prev Med*, 1996, 25(6): 764-770.

[4] Rothman KJ, Greenland S, Walker AM. Concepts of interaction[J]. *Am J Epidemiol*, 1980, 112(4): 467-470.

[5] Cordell HJ. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans[J]. *Hum Mol Genet*, 2002, 11(20): 2 463-2 468.

[6] Hunter DJ. Gene-environment interactions in human diseases[J]. *Nat Rev Genet*, 2005, 6(4): 287-298.

[7] Hallqvist J, Ahlbom A, Diderichsen F, *et al*. How to evaluate interaction between causes: a review of practices in cardiovascular epidemiology[J]. *J Intern Med*, 1996, 239(5): 377-382.

[8] Hosmer DW, Lemeshow S. Confidence interval estimation of interaction[J]. *Epidemiology*, 1992, 3(5): 452-456.

[9] Rothman KJ. *Epidemiology: an introduction*[M]. New York: Oxford University Press, 2002.

[10] Andersson T, Alfredsson L, Kallberg H, *et al*. Calculating measures of biological interaction[J]. *Eur J Epidemiol*, 2005, 20(7): 575-579.

[11] Kooperberg C, Ruczinski I. Identifying interacting SNPs using Monte Carlo logic regression[J]. *Genet Epidemiol*, 2005, 28(2): 157-170.

[12] Kooperberg C, Ruczinski I, LeBlanç ML, *et al*. Sequence analysis using logic regression[J]. *Genet Epidemiol*, 2001, 21(1): S626-S631.

[13] Ruczinski I, Kooperberg C, LeBlanc M. Logic regression[J]. *J Comput Graph Stat*, 2003, (12): 475-511.

[14] Moore JH, Williams SM. New strategies for identifying gene-gene interactions in hypertension[J]. *Ann Med*, 2002, 34(2): 88-95.

[15] Hahn LW, Ritchie MD, Moore JH. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions[J]. *Bioinformatics*, 2003, 19(3): 376-382.

[16] Cho YM, Ritchie MD, Moore JH, *et al*. Multifactor-dimensionality reduction shows a two-locus interaction associated with type 2 diabetes mellitus[J]. *Diabetologia*, 2004, 47(3): 549-554.

[17] Ritchie MD, Hahn LW, Roodi N, *et al*. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer[J]. *Am J Hum Genet*, 2001, 69(1): 138-147.

[18] Bush WS, Dudek SM, Ritchie MD. Parallel multifactor dimensionality reduction: a tool for the large-scale analysis of gene-gene interactions[J]. *Bioinformatics*, 2006, 22(17): 2 173-2 174.

[19] Chung Y, Lee SY, Elston RC, *et al*. Odds ratio based multifactor-dimensionality reduction method for detecting gene-gene interactions[J]. *Bioinformatics*, 2007, 23(1): 71-76.

[20] Wu X, Jin L, Xiong M. Composite measure of linkage disequilibrium for testing interaction between unlinked loci[J]. *Eur J Hum Genet*, 2008, 16(5): 644-651.

[21] Weir BS. Inferences about linkage disequilibrium[J]. *Biometrics*, 1979, 35(1): 235-254.

[22] Weir B. Genetic data analysis II. Sinauer Associates, Sunderland, MA. 1996.

[23] Weir BS, Cockerham CC. *Complete characterization of disequilibrium at two loci*[M]//Feldman MW. Mathematical evolutionary theory. Princeton: Princeton University Press, 1989: 86-110.