非参数连锁分析

倪鹏生,崔 静,沈福民
(上海医科大学流行病学教研室,200032)

摘 要:非参数连锁分析是进行复杂疾病连锁分析的有效手段,本文通过拟合的数据资料,对目前广泛使用的非参数型连锁分析方法进行了探讨,为今后有针对性的选择连锁分析方法提供依据。

关键词:非参数连锁分析
中图分类号:R181.343 **文献标识码**:A **文章编号**:0253-9772(2001)04-0349-05

Non-Parametric Linkage Analysis Methods

NI Peng-sheng, CUI Jing, SHEN Fu-min
(*Dep. of Epidemiology, Shanghai Medical University*, 200032, *China*)

Abstract: We present here four non-parametric statistics for linkage analysis (APM, SIBPAL, MAPMAKER-SIB and GENEHUNTER-NPL). Using the simulated pedigrees, we introduced the usage of these methods.

Key words: non-parametric linkage analysis

非参数连锁分析方法是一种在分析前不需要对疾病或性状的遗传模式(如基因型频率、外显率等)进行确定的分析方法,与参数型连锁分析方法相比,在进行复杂疾病的连锁分析时,具有一定的优势。然而必须认识到,其中的大多数方法也是建立在假定可能的遗传模型基础上的,实际上是一种参数或半参数的分析方法。一般情况下,这种对遗传模型的估计只影响参数的估计,而不影响方法检测连锁的有效性。因此我们选取最常用的几种非参数连锁分析方法进行比较,讨论方法的适用性。

1 方法介绍

非参数分析方法的种类较多,大致可以分为以下几类:按研究的性状不同可分为分析定量性状(如SAGE-SIBPAL-HE[1])和定性性状(如:MAPMAKER-SIB[2])的分析方法;按分析的手段可分为血缘一致性(identical by descent, IBD)(如:SAGE-SIBPAL-ASP)和状态一致性(identical by state, IBS)(如:APM)的分析方法;按分析的对象可分为受累同胞对(SAGE-SIBPAL)、受累亲属对(如:GENEHUNTER-NPL[3])、寻找差异大的同胞对等;按统计手段的不同可分为 t 检验、最大似然比检验和卡方检验等;按标记位点数量不同可分为单位点、多位点分析方法。

关于血缘一致性(IBD)和状态一致性(IBS)的概念介绍:

IBS 的概念是两个同胞有相同的等位基因,就存在 IBS,因而对于图 1 中 A 和 B 可以很容易地确定 IBS。确定 IBD(同胞间相同的等位基因来源于同一个亲代的等位基因),对于图 1A 家系是比较容易的,同胞间相同的"1"等位基因其实来源于不同的亲代(第一个子女来源于母亲,第二个子女来源于父亲),然而在实际情况下,即使亲代的基因型已知(如果存在纯合子的亲代),不能确定子代间的 IBD,即使亲代都是杂合子,如图 1B 情况,也不能确定 IBD,因而需要家庭其他成员的资料,通过统计估计可能的 IBD 概率。

收稿日期:2000-07-17;修回日期:2000-11-10;
作者简介:倪鹏生(1969-),男,硕士研究生,专业:遗传流行病学。Tel:021-64174172, E-mail:ni-ps@hotmail.com

图 1 家系资料说明 IBD 和 IBS 概念

Fig.1 Examples to explain the concepts of IBD IBS

2 基本过程(以受累同胞对为例)

收集特定的受累同胞对家系资料,进行标记位点的检测,确定 IBD 或 IBS 的比例,进行统计分析,下面分别以 GENEHUNTER - NPL、APM(affected - pedigree - member)、MAPMAKER - SIB 和 SAGE - SIBPAL 分析方法为例分析比对。

3 事例分析

3.1 GENEHUNTER - NPL

GENEHUNTER - NPL 是受累亲属对分析方法,考察受累亲属间的 IBD,可进行多基因座连锁分析,有两个分析过程(pairs,all),其中 pairs 过程是将受累亲属成对分析,all 是将家系中所有受累亲属联合分析。

下面通过具体数据来进一步说明,家系结构如图 2,以这一个家系的患病情况为基础,拟合包含 40 个家系的数据资料,使得疾病符合常染色体隐性遗传模式(用 SLINK[4]软件实现)。

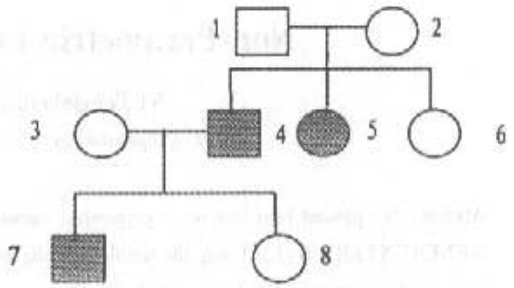

图 2 模拟的家系发病情况

Fig.2 Simulated pedigree

共使用了 4 个 2 等位基因的标记基因座(等位基因频率为 0.5),标记基因座与可能的疾病基因座在染色体上的遗传距离如图 3:

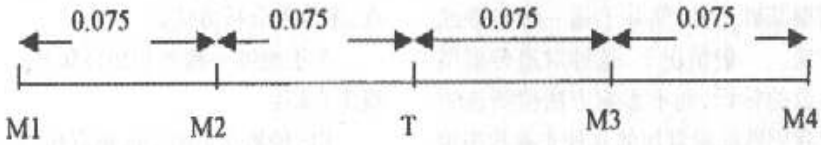

图 3 模拟的标记基因座与性状基因座(疾病基因座)的遗传距离

Fig.3 The genetic distance between the disease locus and marker loci

M: Marker T: Trait

遗传距离以重组率表示,按 Haldane 公式可进行重组率(θ)和厘摩(X)的换算,$X_{Hald}(\theta) = -0.5\ln(1-2\theta)$,从上述公式可知,T 距 M1 大约是 16.25cM。

分析方法见文献 5,分析的距离是从第一个标记基因座前 10cM 开始,到最后一标记基因座后 10cM 结束,从起始点开始每次增加 2 cM 进行分析,结果如下(表 1)。

图 4 的横坐标是图距(以厘摩表示),纵坐标是多点 LOD 值和 NPL 值,箭头表示实际的疾病基因座在遗传图谱上的位置(距离 M1 为 16.25)。

3.2 APM

APM 是受累亲属对分析方法,以 IBS 为考察对象,可以进行多基因座连锁分析。在 APM 的运算过程中根据加权与否,给出 3 种结果(不加权或等加权 f(p) = 1,频率平方根倒数加权 f(p) = 1/sqrt(p),频率倒数加权 f(p) = 1/p)。以上例模拟出的数据进行 APM 分析,结果如下(表 2):

表1　GENEHUNTER的计算结果(以M1为中心,距离为cM)

Table 1　The result of GENEHUNTER, genetic distance: cM

距离 cM	*LOD* 值	*NPL* 值 (all)	*NPL* 值 (pairs)	距离 cM	*LOD* 值	*NPL* 值 (all)	*NPL* 值 (pairs)
-10.00	7.2023	1.93382	2.00148	16.00	10.677543	3.11776	3.24954
-8.00	6.90231	2.10108	2.17257	18.00	10.378217	3.15716	3.29553
-6.00	6.23207	2.28308	2.35846	20.00	9.343969	3.21473	3.36059
-4.00	4.86028	2.48112	2.56047	22.00	6.947115	3.29080	3.44510
-2.00	1.74212	2.69666	2.78002	24.00	-1.98649	3.38582	3.54956
0.00	-INFINI	2.93129	3.01865	26.00	1.245847	3.15443	3.32191
2.00	-2.60666	2.95691	3.04983	28.00	2.607503	2.87144	3.04181
4.00	-0.27927	2.99869	3.09768	30.00	0.743689	2.61594	2.79014
6.00	-0.94066	3.05687	3.16248	32.00	-8.022923	2.38553	2.56474
8.00	-12.6208	3.13179	3.24463	34.00	-3.794640	2.19884	2.36813
10.00	6.49449	3.10720	3.22391	36.00	0.575411	2.03346	2.18866
12.00	9.24687	3.09282	3.21380	38.00	2.495728	1.88038	2.02269
14.00	10.3571	3.09635	3.22234	40.00	3.535934	1.73870	1.86923
				42.00	4.125654	1.60759	1.72733

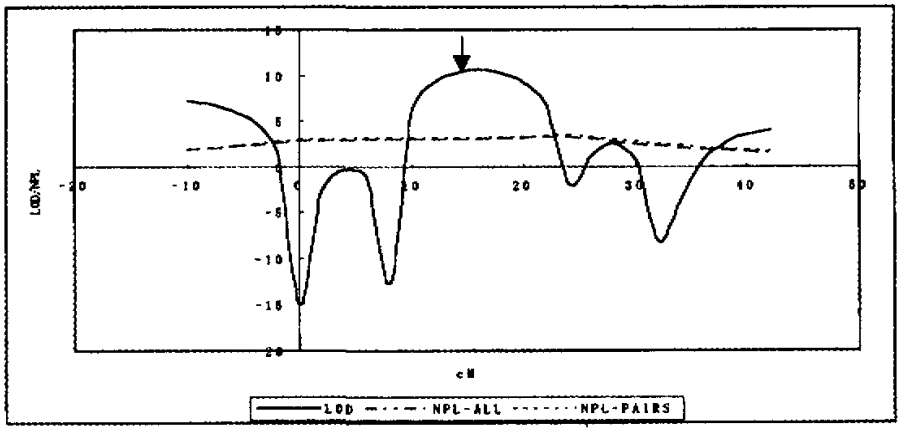

图4　GENEHUNTER多点LOD值、NPL值结果图

Fig.4　The multi-point linkage LOD and NPL scores result of GENEHUNTER

表2　APM的计算结果 Table 2　The result of APM

加权函数	家系数	患病成员	*t* 值	*P* 值
f(p) = 1	40	120	2.48928	0.00641
f(p) = 1/sqrt(p)	40	120	2.48928	0.00641
f(p) = 1/p	40	120	2.48928	0.00641

提示在此区域内存在与疾病关联并相互连锁的基因座。

3.3　MAPMAKER-SIB

MAPMAKER-SIB中有4种处理同胞对的方法,(1)每个家庭只挑选一对受累同胞,(2)一个受累同胞与其他患者同胞形成受累同胞对,(3)收集所有可能的受累同胞对,(4)对第三种情况进行加权,由于本次拟合的家系资料的特殊性,每种处理的结果是一致的。(此过程也可以在GENEHUNTER中实现)

表3　MAPMAKER-SIB的计算结果

Table 3　The result of MAPMAKER-SIB

距离 cM	*LOD*	距离 cM	*LOD*	距离 cM	*LOD*
-10.000	3.434360	8.000	4.064627	26.000	5.263476
-8.000	3.434361	10.000	4.723064	28.000	5.046867
-6.000	3.434361	12.000	5.383382	30.000	4.519600
-4.000	3.434361	14.000	5.836024	32.000	3.835469
-2.000	3.434361	16.000	6.071337	34.000	3.567365
0.000	3.434362	18.000	6.079885	36.000	3.445085
2.000	3.822678	20.000	5.900933	38.000	0.359129
4.000	4.082121	22.000	5.601391	40.000	3.254892
6.000	4.178815	24.000	5.262410	42.000	0.251346

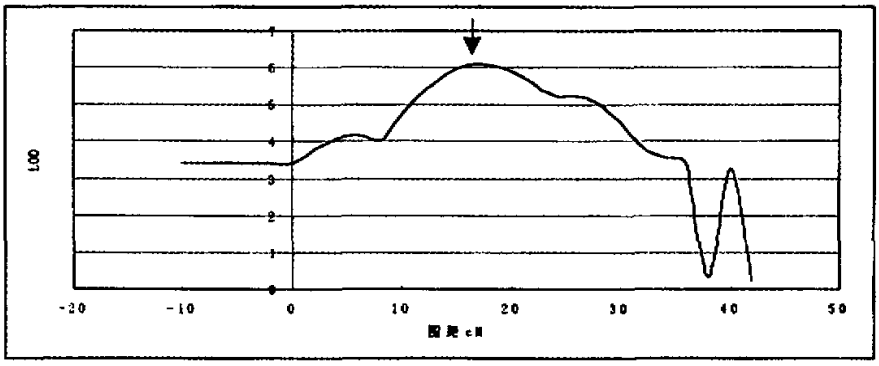

图 5 MAPAMKER－SIB LOD 值结果图

Fig.5 The linkage lod scores result of MAPMAKER－SIB

图 5 的横坐标是图距(以厘摩表示),纵坐标是 LOD 值 ,箭头表示实际的疾病基因座在遗传图谱上的位置(距离 M1 为 16.25)。可见,分析结果与实际情况相符。

3.4 SAGE－SIBPAL

SIBPAL 是 SAGE 软件中用于受累同胞对连锁分析的过程,考察 IBD 在同胞间的分布,可用于定性性状(是否患病)、定量性状(血压等)和定性性状中考虑发病年龄因素,是较为全面的进行受累同胞对连锁分析的软件,而且其使用的均数检验方法在单基因座非参数连锁分析中是效能较高的一类,其缺陷是对于家系中无同胞患病的资料无效,而且不能进行多基因座连锁分析(其新版本可能包含了这一功能)。

根据 SAGE－SIBPAL 所需的样本情况,通过 SLINK 拟合具体数据,家系结构如图 6,共拟合了 40 个家系,疾病符合常染色体隐性遗传模式。

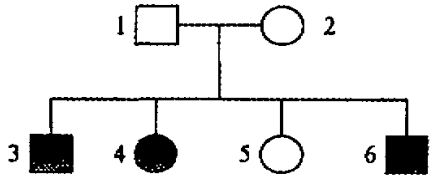

图 6 模拟的家系发病情况

Fig.6 Simulated pedigree

拟合两个双等位基因标记基因座(等位基因频率为 0.5),其中一个标记基因座(A)与疾病基因座之间的重组率为 0.075,另一个标记基因座(B)与疾病基因座之间的重组率为 0.5(不连锁)。

结果如下:

表 4 SGAE－SIBPAL 的计算结果

Table 4 The result of SGAE－SIBPAL

Locus (位点)	d.f. (自由度)	Full sib Pi Mean (全同胞 pi 均数)	t 值	P 值	Intercept (截距)	Slope (斜率)
A	118	0.53750 *	－10.1285 #	0.0000 * *	1.2085	－1.2085
B	118	0.49583	0.8694	0.806797	0.4367	0.1277

与致病位点与 A 位点连锁,而与 B 位点不连锁,与实际情况相符。

4 讨 论

对于单基因座的非参数连锁分析,建议使用 SAGE－SIBPAL,对于多基因座的非参数连锁分析,如果只包括受累同胞建议使用 MAPMAKER－SIB,如果家系中还包括其他受累亲属,建议使用 GENHUNTER－NPL(ALL)过程。

IBS 为基础的连锁分析方法效能低于 IBD 为基础的方法。

家系中受累亲属对间 IBD 能明确确定的可能

性越大,效能越高。未患病同胞在分析中的作用[6],(1)可以用作受累同胞或亲属对间IBD的推算,(2)可以当作正常同胞看待,假定与受累同胞间分享更少的IBD比例,不同的软件采取的策略不同,SAGE-SIBPAL上述两种策略均使用,研究发现,策略1可提高分析的效能,策略2应具体分析,对于外显高的疾病有用,外显低时,可能"正常"同胞与患者在基因座上的相似性更大,因而就不适用了。

当家系中患者数大于2个,研究的方法包括[6]:(1)每个家庭只挑选一对受累亲属(或同胞),(2)一个受累亲属与其他患者形成受累亲属对,(3)收集所有的受累亲属对,研究发现策略3的效能高(统计上有问题),目前较多采用策略3。

除SAGE[7]外,均为免费软件可从INTERNET上免费下载[8,9,10]。

需明确的是,非参数的连锁分析方法不是唯一用于复杂疾病基因的定位方法,目前提出的关联分析、连锁不平衡分析(linkage disequilibrium, LD)也是十分有效的手段,而且随着单核苷酸多态(SNP)基因座数量增多,分布密度加大,利用LD进行疾病定位将是十分有效的手段。

(本结果只考虑了定性性状,定量性状的受累同胞对分析方法是目前研究的热点(QTL定位),与此相匹配的软件也较多如SAGE、GENEHUNTER和MAKEMAPPER等,因此有必要进行此类研究。)

参 考 文 献(References):

[1] SAGE, Statistical analysis for genetic epidemiology, Release 3.0 [M]. Case Western Reserve University, Cleveland, 1997.

[2] Leonid Kruglyak, *et al*. Complete multipoint sib-pair analysis of qualitative and quantitative traits[J]. Am J Hum Genet, 1995, 57:439~454.

[3] Leonid Kruglyak, *et al*. Parametric and nonparametric linkage analysis: A unified multipoint approach[J]. Am J Hum Genet, 1996, 58:1347~1363.

[4] Terwilliger JD, Jurg Ott. Handbook of Human Genetic Linkage [M]. The Johns Hopkins University Press, 1994, pp.255~345.

[5] 倪鹏生,崔　静,沈福民.参数连锁分析方法[J].遗传,2001,23(1)24~28.

[6] Sean Davis, *et al*. Comparsion of nonparametric statistics for detection of linkage in nuclear families: single-marker evaluation [J]. Am J Hum Genet, 1997, 61:1431~1444.

[7] SAGE: http://darwin.cwru.edu/pub/sage.html[CP].

[8] Genehunter: http://waldo.wi.mit.edu/ftp/distribution/software/genehunter/gh2[CP].

[9] Mapmaker: ftp://ftp-genome.wi.mit.edu/distribution/software/sibs[CP].

[10] APM: http://watson.hgen.pitt.edu/register/soft-doc.html [CP].