·基础理论与方法·

应用多因子降维法分析基因-基因交互作用

唐迅 李娜 胡永华

【摘要】 **目的** 介绍在遗传流行病学病例对照研究中，应用多因子降维法(MDR)分析基因-基因交互作用。**方法** 简述 MDR 的基本步骤、原理及其特点，并结合研究实例说明在病例对照研究中如何应用软件进行 MDR 分析。**结果** 相对于传统的统计学方法，MDR 是一种无参数、无遗传模式的分析交互作用的方法，理论和实例研究均表明其分析交互作用具有较好的效能，目前已成功应用于散发性乳腺癌、心房颤动和原发性高血压等疾病的研究。**结论** MDR 能够应用于病例对照研究进行基因-基因交互作用的分析，且具有较传统的统计学分析方法无法比拟的优势。

【关键词】 病例对照研究；多因子降维法；基因-基因交互作用

The application of multifactor dimensionality reduction for detecting gene-gene interactions *TANG Xun, LI Na, HU Yong-hua. Department of Epidemiology & Biostatistics, School of Public Health, Peking University, Beijing* 100083, China

Corresponding author: HU Yong-hua, Email: yhhu@bjmu.edu.cn

【**Abstract**】 **Objective** To introduce the application of Multifactor Dimensionality Reduction (MDR) method for detecting gene-gene interactions in genetic case-control studies. **Methods** A brief overview on basic steps involved in the implementation, theoretical details, available software as well as the use and features of the MDR method were discussed based on a practical research case. **Results** Advantages of MDR were compared to the conventional statistical approaches, showing that MDR method was a novel, nonparametric, genetic model-free approach that was developed specifically for detecting gene-gene interactions. Theoretical and empirical studies suggested that MDR was having reasonable power for detecting gene-gene interactions. Applications of MDR method had found the evidence of gene-gene interactions in several diseases such as sporadic breast cancer, atrial fibrillation and essential hypertension. **Conclusion** MDR method could be used for detecting gene-gene interactions in genetic case-control studies as having great advantages versus the conventional statistical approaches.

【**Key words**】 Case-control study; Multifactor dimensionality reduction; Gene-gene interactions

心脑血管疾病等多基因病，并不遵循普通的孟德尔遗传模式，很可能受到多个基因位点及环境危险因素的影响，而产生复杂的高阶交互作用[1]。单核苷酸多态性(single nucleotide polymorphism, SNP)作为第三代遗传标记，已广泛应用于遗传流行病学研究。通常所采用的传统的参数统计方法，如 logistic 回归模型或广义线性模型，分析病例对照研究设计中众多 SNP 之间的基因-基因交互作用，其过程繁琐，并且对模型参数的结果很难解释，况且大多数的多基因病的基因型和表型之间并非线性关系。logistic 回归模型的参数估计，可能会产生较大的误差而导致Ⅰ类错误增大。另外，采用传统的回归模型分析基因-基因交互作用，可能会导致Ⅱ类错误的增加，而使效能降低[1]。同时，在研究多位点之间基因-基因交互作用时，每增加一个 SNP 位点，所需的样本量将呈指数倍增加，考虑到基因型频率，即使样本量较大，数据分布在高维空间中仍显得相对稀疏，很可能出现某些基因型组合没有观察值，这种情况称为“维度困扰”(curse of dimensionality)[2]。

2001 年 Ritchie 等[3]首次提出了多因子降维法(multifactor dimensionality reduction, MDR)，“因子”是交互作用研究中的变量(如基因型或环境因素)，“维”是指研究的多因子组合中因子(如基因型)的数目，以疾病易感性分类(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(即高危或低危)，即为“降维”。这是一种非参数、无需遗传模式的分析方法，适用于病例对照研究或患病不

基金项目：国家“十五”科技攻关课题资助项目(2001BA703B02)

作者单位：100083 北京大学医学部公共卫生学院流行病与卫生统计学系

通讯作者：胡永华，Email: yhhu@bjmu.edu.cn

一致同胞对设计，只需具备各位点的遗传数据（例如SNP），即可进行基因-基因交互作用的分析，而无需其他特殊条件。与其他传统的统计学建模方法相比，其优点在于可以大大降低建模所需的自由度，MDR 方法的主要特点是：①并不需要指定遗传模式（显性或隐性遗传）和交互作用模型（线性或非线性模型，加法或乘法模型）；②结合 MDR Software 程序包[4]，可以识别多个 SNP 位点之间的高阶交互作用。

基本原理

MDR 方法实际上是一种组合划分方法（combinatorial partitioning method, CPM）[5]的扩展，虽然所针对的结局变量的类型不同，CPM 要求连续变量，而 MDR 针对的是诸如疾病状态等分类变量，但它们都是采用数据降维的策略，以解决在有限的样本量条件下，分析高维数据之间交互作用的问题。

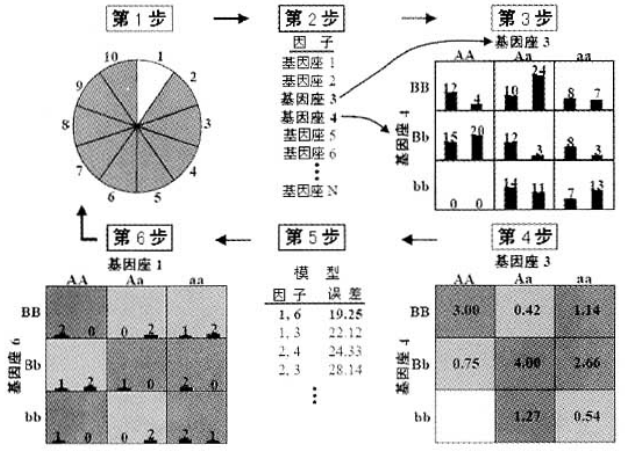

图1　MDR 基本步骤示意图[6]

1. MDR 分析的基本步骤[7]：如图 1 所示，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 1 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素，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，将个体划分为不同的分类，也就是图中的单元格，单元格中左侧条带表示病例，右侧条带表示对照；第 4 步，在 n 维的每个多因子分类（单元格）中，计算病例数与对照数的比值，若其病例、对照数之比达到或超过某个阈值（例如≥1），则标记为高危，反之则为低危，这样就把 n 维的结构降低到一维两水平（即，高危或低危）；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，选择个体错分最小的那个 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该模型的预测误差，以及单元格分配时的相关误差。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，作为模型相关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 的交叉验证。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。

2. 模型评估与检验：交叉验证（cross validation）和置换检验（permutation test）是评估 MDR 模型统计学意义的两个重要手段。交叉验证一致性通过以下方法衡量[3]：对每次的十重交叉验证，比较同一个位点/因子组的验证次数。如果因子组合只发生在一个亚组中，为最小值 1；如果所有 10 个亚组确定的都是相同的位点/因子组合，则为最大值 10。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类误差增大而产生假阳性结果的影响[1]。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，其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。最佳模型的假设检验可以通过使用不同的随机数进行置换检验，来评估交叉验证一致性和预测误差估计值的大小，确定该模型与那些无关联的模型相比是否更合适。

3. MDR 软件分析：目前最新的 MDR Software 程序包（版本0.6.1）是基于 Java 程序编写的源代码开放的免费软件（可在 http://www.epistasis.org/mdr.html 免费下载）[4]，用户可以通过图形用户界面（GUI），在多种操作系统下均可进行 MDR 分析，使得此法的运算过程大大简化，并自动给出预测误差的结果，作为模型内部真实性的估计。Windows 操作系统下可直接点击 mdr.jar 文件运行程序，但必须先安装 Java 2 Runtime Environment（JRE），这

可以从 Java 的网页上(http://www.java.com/)免费下载安装。

(1)MDR 分析的数据格式及编码:

列:attributes,代表属性(变量),即所研究的多个因子(遗传位点或环境因素),例如X1～X20 共 20 个研究因子。

行:instances,代表记录,即研究的样本(患者和非患者),例如共有 400 人的样本量。

最后一列的 Class 分类变量用于区分患者(病例)和非患者(对照),一般而言,用Class=1表示患者(病例),Class=0表示非患者(对照)。同胞对或配对资料亦然,但必须一行病例与一行对照间隔输入,分析此类资料时需选中 paired analysis 选项。Ratio 表示病例/对照(Class1/Class0)数目的比值,例如在配对资料中Ratio=1。

对于实际数据,可参考 MDR 程序自带的 MDR-SampleData.txt 文件的数据格式进行输入。除分类变量(Class)外,其他所有值均可用字符串表示,通常用变量值 0,1,2 表示诸如"GG"、"GC"、"CC"基因型比较方便,亦可采用其实际的字符串标签"GG"、"GC"、"CC"表示。字符串变量,若不采用数值编码,则不能包含空白字符串,如 Tab 键或空格键。例如,"AA"是有效变量,而"A A"则不合要求,这将被视为两个变量值而产生多于标题行所代表列的数量。

MDR 数据文件的要求:①数据文件以记事本文件(.txt)格式保存,修改数据文件时必须关闭 MDR 窗口和程序后,在记事本文件中进行修改并保存;②数据的首行必须是一组以 Tab 键分隔的标题,每一个代表一列的属性;③数据行的分类变量(Class)必须置于数据集的最后一列,其赋值必须为 0 或 1,通常将未患病(对照)设为 0,而将患病(病例)设为 1;④文件中的所有变量之间必须以 Tab 键分隔。

(2)分析参数配置:数据文件载入(Load Datafile)后,可以点击 MDR 主窗口上方的"configuration"标签配置分析参数,包括随机数种子、属性数目范围、交叉验证数目等。一般而言,采用程序默认的设置即可对病例对照研究的数据进行分析(run analysis)。但需要注意的是,所加载的数据文件若是配对资料的形式(数据文件内病例与对照交替输入表示它们是匹配的),例如,在配对病例对照或患病不一致同胞对设计中,必须选中 paired analysis 选项,并假定是1∶1匹配(配对)的资料。设定该选项的结果就是在交叉验证的过程中配对的对子将始终在一起分析。该选项的默认值是未选中,当导入的是非配对的资料时不能选择该选项,否则在 MDR 分析中将产生错误的结果。

(3)结果输出:分析完毕后,将在 summary table 中显示每个等级的最佳模型(属性组合)、训练样本或检验样本的准确度(即正确分类数与所有分类的记录总数的比值)、符号检验及 P 值(即检验准确度>0.5的数目,并通过非参数符号检验计算 P 值)、交叉验证一致性(即在某一特定交叉验证中,指定的属性组合被选为最佳模型的次数)。

选中不同的模型将在图形模型表(graphical model)中显示所选模型的详细情况,图 1 中每个单元格里的左侧条带表示病例(Class=1),右侧条带表示对照(Class=0),条带上方的数字表示例数。图中深灰色的单元格代表的组合是超过比率阈值(ratio threshold)的,而浅灰色的单元格表示没有超过阈值的组合,白色单元格表示没有数据的单元格。该图形模型表必须以 .eps 的后缀名文件保存,这种基于矢量图形的文件可以在 Microsoft Word 中使用,亦可转换为 JPEG 格式的图像文件后使用。由于图形组合随着维度的增加而增多,可以在此表中选择查看有限数目的维度情况,或查看属性组合水平的完整的维度模型,需要注意的是,图形表的右下方的"limit dimension"选项默认选择只显示 3 个维度,只有取消选择该选项,才会显示完整模型,否则每一页都将单独保存。

最佳模型表(best model)则显示了所选模型的详细结果,并列出了模型性能的不同评价方法,主要包括准确度、灵敏度、特异度、OR 值、χ^2 值、$Kappa$ 值等。

(4)模型判读:在所有等级的最佳模型中,应选择符号检验显示有统计学意义的模型,且交叉验证一致性(CV consistency)越大越好。例如,分析 MDR-SampleData.txt 文件的数据,三位点(X1,X6,X8)和四位点(X1,X2,X6,X8)的模型均有统计学意义($P=0.001$),但三位点模型的交叉验证一致性更大(10/10),且检验样本的准确度也更高(0.8713),故选择该三位点模型作为最佳模型。

当然,MDR 程序还提供了一些附带的高级功能选项,例如,对众多属性进行预处理的过滤器(filter),它可以用来对属性进行筛选,从而减少分析中所需考虑的属性或变量数目。目前可以使用的过

滤器有 Relief F 值统计量[8]、χ^2 值统计量和 OR 值统计量。当两个或以上的属性或离散分类变量之间存在条件依赖性(例如交互作用)时,Relief F 值统计量比 χ^2 值统计量等单变量过滤器更胜一筹;而当 MDR 分析的属性存在独立主效应时,则宜使用 χ^2 值统计量过滤器;OR 值统计量通常用于二分变量的比较,当属性为多分类(超过两水平)时,MDR 将计算每个可能的 OR 值,并报告最大的 OR 值。

另外,为配合 MDR Software 程序包使用,程序开发者还提供了单独的 MDR-Data Tool Software 和 MDR-Permutation Testing Software 程序,分别用于转换 MDR 数据文件的格式和对 MDR 分析进行置换检验。

实例分析

Tsai 等[9]在中国台湾人群中研究肾素血管紧张素系统(RAS)的基因在心房颤动发病中的作用。该研究选取病例、对照各 250 例,并对其年龄、性别、左心室功能紊乱和心脏瓣膜疾病的发生进行了匹配。首先对 *ACE* 基因(I/D)、*AT1R* 基因(A1166C)和 *AGT* 基因(T174M, M235T, G-6A, A-20C, G-152A 和G-217A)的 8 个多态性位点进行了单个位点的关联研究,结果显示 *AGT* 基因的三个位点与疾病存在阳性关联:M235T($P<0.001$),G-6A($P=0.005$)和G-217A($P=0.002$)。

采用 MDR 方法分析此 8 个多态性位点的交互作用发现(表 1),最佳模型包含了 *AGT* 基因的 2 个位点(T174M, M235T)和 *ACE* 基因的 1 个位点(I/D),此三位点的模型的预测误差为 37.26,经 1000 次置换检验发现交叉验证一致性和预测误差都有统计学意义($P=0.001$)。而同时发现的四位点的模型也具有统计学意义($P=0.01$),其包含了上述三位点模型中的全部位点,以及另一个G-6A位点,且此G-6A位点在单个位点关联研究时也存在阳性关联,但由于与三位点模型相比,此模型的交叉验证一致性较低(8.4),且预测误差较大(39.42%),故选择三位点模型作为最佳模型。因此,该研究提示 *AGT* 基因 T174M、M235T 位点与 *ACE* 基因I/D位点之间,可能存在基因-基因交互作用。另外,虽然单个位点的关联研究提示 *AGT* 基因的三个位点可能存在主效应,但在 MDR 交互作用模型中只包含了其中的一个多态性位点(M235T),这也说明了在心房颤动发病中基因-基因交互作用的重要性。

表1 MDR 分析心房颤动中多位点交互作用的模型[9]

多因子组合中位点的数目及其组合	交叉验证一致性	预测误差
2 位点:M235T/A-20C	8.9	42.36
3 位点*:T174M/M235T/ID	10.0*	37.26*
4 位点:T174M/M235T/ID/G-6A	8.4#	39.42#
5 位点:T174M/M235T/ID/G-6A/G-152A	4.5	40.29
6 位点:T174M/M235T/ID/G-6A/G-217A/AT1R	3.9	42.91
7 位点:T174M/M235T/ID/G-6A/G-217A/AT1R/A-20C	6.7	44.03

* 经 1000 次置换检验 $P<0.001$; # 经 1000 次置换检验 $P=0.01$

讨 论

基因-基因交互作用的分析是目前遗传流行病学研究的热点问题之一,通常应用 logistic 回归分析时,若采用前进法选择变量,由于只检验那些具有统计学意义的独立主效应的交互作用的变量,而存在局限性。多基因病的位点变异存在交互作用,但其主效应很可能没有或者很小,这将会被排除出方程;而采用后退法选择变量时,包括所有主效应和交互作用项的完整模型可能会需要太多的自由度;逐步回归法相对而言更灵活,但其同样也会受到需要太多的自由度的限制。并且,维度困扰也可能会导致 logistic 回归模型中参数估计的错误。

作为一种非参数、无需遗传模式的分析方法,MDR 方法以疾病易感性分类(高危、低危)的方式建模,Hahn, Moore[10]证明了 MDR 所创造的是区分高危、低危个体的理想的判别分类模型。MDR 方法选择合适的基因型组合,检验所有可能的多位点基因型的组合,并报告最佳分类的组合,所采用的基因型组合的分类策略与贝叶斯分类(naive Bayes classifier)相似。因此,贝叶斯分类是 MDR 分组的基础,这也为 MDR 方法奠定了重要的理论依据。交叉验证和置换检验是评估 MDR 模型统计学意义的两个重要手段。MDR 方法基于小样本考虑,采用留一校验交叉验证(leave-one-out cross-validation)的方法,可以更好地获取选择模型的无偏估计,该算法的正确分类率(correct classification rate)的结果相当好,事实上这种算法早已被广泛应用于计算机、信息学等领域[11]。

Hahn 等[4]对 MDR 方法进行模拟数据的研究,结果表明,所选择的具有最低的预测误差和最高的交叉验证一致性的最佳模型,包含了正确的功能性 SNP 位点,并且置换检验表明,交叉验证一致性和预测误差都具有统计学意义($P=0.001$)。Ritchie

等[12]估计了基因型错分、缺失数据、拟表型以及遗传异质性对 MDR 效能的影响。模拟数据的结果显示，病例、对照数各为 200 例时，检测两位点交互作用，MDR 的效能达到 80% 以上。即使在 5% 的基因型错分和 5% 缺失数据的情况下，对大多数模型该结果仍然真实。但需要注意的是，随着考虑因素的增加，分类错误在减少，这可能是由于高阶模型对数据过度拟合。并且当模型大小增加，预测误差也相应增加。因此，虽然高阶模型过度拟合数据，但其预测能力变差。

MDR 方法适合对病例对照研究或患病不一致同胞对设计进行2～6 个基因位点或环境因素的交互作用分析，目前已成功应用于散发性乳腺癌、心房颤动和原发性高血压等疾病的研究[7]，但这也只是为研究遗传流行病学交互作用提供一种可选择的方法或策略。固然，它也有一些不足之处：当主效应或已知的协同作用存在时，用 MDR 方法很难得到最终的模型，例如 MDR 提示最佳模型为四因子模型，但它并不能明确是四因子之间都有交互作用，还是两组单独的两因子交互作用，抑或是两个主效应加上另外两因子的交互作用等[6]。并且 MDR 同样也会受到遗传异质性的严重影响[12]，必须引起注意。此外，等位基因关联或连锁不平衡对 MDR 效能和Ⅰ类错误的影响还未知，这特别是在评估位点内交互(显性、隐性)时更重要。提供关于效能和样本量的详细说明也很重要，比如进行 3 个、4 个，甚至 10 个位点交互作用的研究需要多少数据？一般认为，几乎没有任何一种方法可以理想化地用于所有情况下的数据分析，而 MDR 更可能成为得到一致结果的几种方法之一[7]。

在后基因组时代，遗传流行病学研究的主要目标是了解各基因的功能，其中包括基因-基因、基因-环境之间复杂的交互作用。虽然目前尚不能奢望能够完全解释全部的基因-基因交互作用，但至少可能对多基因疾病中相对重要的一些交互作用予以探讨，这也将有助于今后对多基因疾病更全面的认识。当然，对于简单的基因-基因的统计学交互作用的研究，并不一定必然能够得到生物学交互作用的结果。

参 考 文 献

1 Moore JH. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Hum Hered, 2003, 56: 73-82.

2 Moore JH, Ritchie MD. The challenges of whole-genome approaches to common diseases. JAMA, 2004, 291: 1642-1643.

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

4 Hahn LW, Ritchie MD, Moore JH. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics, 2003, 19: 376-382.

5 Nelson MR, Kardia SL, Ferrell RE, et al. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. Genome Res, 2001, 11: 458-470.

6 Coffey CS, Hebert PR, Ritchie MD, et al. An application of conditional logistic regression and multifactor dimensionality reduction for detecting gene-gene interactions on risk of myocardial infarction: the importance of model validation. BMC Bioinformatics, 2004, 5: 49.

7 Moore JH. Computational analysis of gene-gene interactions using multifactor dimensionality reduction. Expert Rev Mol Diagn, 2004, 4: 795-803.

8 Robnik-Sikonja M, Kononenko I. Theoretical and empirical analysis of Relief *F* and RRelief *F*. Mach Lear J, 2003, 53: 23-69.

9 Tsai CT, Lai LP, Lin JL, et al. Renin-angiotensin system gene polymorphisms and atrial fibrillation. Circulation, 2004, 109: 1640-1646.

10 Hahn LW, Moore JH. Ideal discrimination of discrete clinical endpoints using multilocus genotypes. In Silico Biol, 2004, 4: 183-194.

11 Jelinek F, Mercer R. Probability distribution estimation from sparse data. IBM Technical Disclosure Bulletin, 1985, 28: 2591-2594.

12 Ritchie MD, Hahn LW, Moore JH. Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. Genet Epidemiol, 2003, 24: 150-157.

(收稿日期：2005-09-29)

(本文编辑：张林东)