·基础理论与方法·

复杂疾病病因研究中基因间交互作用分析：基于基因型传递不平衡的多因子降维法

李娜 唐迅 陈大方 胡永华

【导读】 介绍复杂疾病病因研究中分析基因-基因交互作用的一种新方法：基于基因型传递不平衡的多因子降维法(MDR-PDT)。文中简述 MDR-PDT 的基本原理、步骤及特点，并结合研究实例说明其应用过程。MDR-PDT 是原始 MDR 的扩展，可用于多种结构类型的核心家系资料分析基因-基因交互作用。**结论** MDR-PDT 具有非参数、无需遗传模式的特点，并能充分利用家系中多个家庭成员的信息，在复杂疾病病因研究中分析基因-基因交互作用具有良好的效能。

【关键词】 多因子降维法；核心家系；复杂疾病；基因-基因交互作用

Identification of gene-gene interactions related to the etiology of complex disease: a multifactor dimensionality reduction-genotype pedigree disequilibrium test *LI Na, TANG Xun, CHEN Da-fang, HU Yong-hua. Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing* 100083, China

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

【**Abstract**】 To introduce the application of a multifactor dimensionality reduction-genotype pedigree disequilibrium test (MDR-PDT) for detecting gene-gene interactions in the etiology of complex disease. A brief overview on the basic theory, implementing steps and features of MDR-PDT were described, and a practical research case was demonstrated to application of MDR-PDT in nuclear family studies. The MDR-PDT approach was the extension or development of conventional MDR method which could be used for detecting gene-gene interactions in families of diverse structure. **Conclusion** MDR-PDT was a new nonparametric and model-free method which might use additional family members in the nuclear families and had a good power to identify gene-gene interactions.

【**Key words**】 Multifactor dimensionality reduction; Nuclear family; Complex disease; Gene-gene interactions

冠心病、2 型糖尿病、原发性高血压和哮喘等复杂疾病的发生受多个微效基因和环境因素的影响，并普遍存在基因-基因、基因-环境交互作用[1]。如今借助先进的分子生物学技术，检测和识别大量的疾病候选基因已并非难事，但在此基础上如何正确评价复杂疾病病因模式中多个基因或环境因素间的交互作用(即多因子交互作用)，合理剖析疾病发生的遗传基础，已成为目前遗传流行病学研究所面临的重要问题[2,3]。多因子降维法(MDR)是新近发展的一种的非参数、无需遗传模式的高阶交互作用分析方法，自 2001 年 Ritchie 等[4]提出该方法以来，已广泛应用于 2 型糖尿病、高血压、哮喘、多发性硬化症、恶性肿瘤等多种复杂疾病[5-10]，并显示出良好的效能[4,11]。但是，原始 MDR 通常只能用于匹配的病例对照研究和患病不一致的同胞对研究，而不适用于具有父母或者多个同胞信息的核心家系研究。然而，后者是遗传流行病学研究中常用的方法之一，以核心家系为基础的设计可以有效避免传统病例对照研究中由于人群分层(population stratification)导致的虚假关联结论问题，因而备受遗传流行病学研究者的青睐[12-14]。因此，为了扩展 MDR 的应用范围，2006 年 Martin 等[15]在原始 MDR 的基础上提出了一种基于基因型传递不平衡检验(genotype-pedigree disequilibrium test, geno-PDT)的多因子降维法，即 MDR-PDT。该方法整合了 MDR 和 PDT 的优点，可用于核心家系研究中分析复杂性状疾病多个因素间交互作用，并具有良好的效能。

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

基本原理

MDR-PDT 是对原始 MDR 方法的扩展，主要是

基于 geno-PDT 的原理并采用降维的策略来分析多个候选基因位点间交互作用对复杂性状疾病的影响。其涉及到的基本理论方法主要有 geno-PDT 和排列检验(permutation test)。

1. geno-PDT 的原理[16]：首先，核心家系是指由双亲及其后代组成的两代人的家庭，为实际工作中最常见的家系类型。本文所谈到的核心家系(也称可提供信息的家系)主要有两种基本类型：患者及其父母(父母中至少一人的基因型为杂合型)组成的“三联体”(trio)，患病不一致同胞对(discordant sibpair, DSP)(有或无父母亲基因型信息)。那些具有多个家庭成员的核心家系则可认为是上述 trios 和/或 DSPs 的不同形式的组合。geno-PDT 就是充分利用这些家系中多个家庭成员的信息，以某候选基因位点的基因型信息为观察指标，通过构造出一个家系的综合统计量来检验某位点基因型与疾病是否存在关联，其零检验假设是基因型与疾病既不存在连锁也不存在关联。

如果某基因座只有两个等位基因 M_1、M_2，则其基因型为 M_1M_2；如果某基因座有多个等位基因，其基因型可记为 M_iM_j。设定 $|M_iM_j| = g$，则对于单个 trio 和 DSP 家系资料，理论上可分别构造出如下随机统计量：

$$X_{T(g)} = (g \text{ 传递的次数}) - (g \text{ 未传递的次数}) \quad (1)$$

$$X_{S(g)} = (\text{患病同胞中 } g \text{ 出现的次数}) - (\text{未患病同胞中 } g \text{ 出现的次数}) \quad (2)$$

对于某个包含有 n_T 个 Trio 和 n_S 个 DSP 的核心家系，可得到如下的综合随机变量：

$$D(g) = [\sum_{j=1}^{n_T} X_{Tj}(g) + \sum_{j=1}^{n_S} X_{Sj}(g)] \quad (3)$$

在零假设成立的条件下，所有 trio 的预测值 $E[X_T(g)] = 0$，所有的 DSP 的预测值 $E[X_S(g)] = 0$；且对于整个家系的综合随机变量的预测值，则有 $E[D(g)] = 0$。

假设有 N 个这样的核心家系，$D_i(g)$ 是第 i 个核心家系的综合随机变量，并且每个家系的 $D_i(g)$ 是相互独立的，则研究中总的检验统计量可按如下表示：

$$T(g) = \frac{\sum_{i=1}^{N} D_i(g)}{\sqrt{\sum_{i=1}^{N} D_i(g)^2}} \quad (4)$$

在零假设成立的条件下，$T(g)$ 近似服从均值为 0，方差为 1 的标准正态分布。

MDR-PDT 就是在 geno-PDT 原理的基础上，通过构造家系的综合随机统计量来充分利用家庭中多个成员的基因型信息，从而将 MDR 的方法扩展应用至核心家系研究中来分析基因-基因交互作用的。

2. 排列检验的原理[17,18]：“permutation”在数学上是“排列、置换”的含义。考虑到顺序时，“permutation”意为“排列”，不考虑顺序时即为“组合”。其具体的含义需根据实际情况而定：如原始资料为成组设计的类型，则取“组合”之意；如资料为配对设计的资料，则理解为“置换”更合适。其基本思想是：根据所研究的问题构造一个检验统计量；利用现有样本，按排列组合的原理，导出检验统计量的理论分布，当无法由此得到确切的理论分布时，可采用抽样模拟的方法估计其近似分布；然后求出从该分布中获得现有样本或更极端样本的概率(P 值)，并作出推论。

3. MDR-PDT 的分析步骤：MDR-PDT 的分析过程与 MDR 基本相似[19,20]，也是采用降维的策略来分析多个基因型间是否存在交互作用。如图 1 所示。

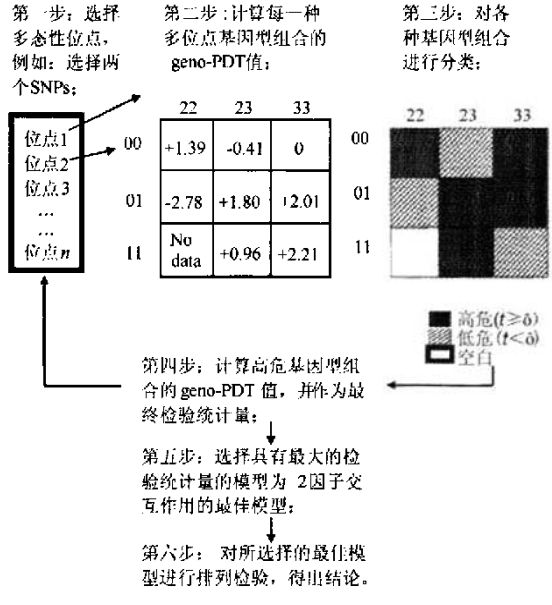

图1 两位点模型的 MDR-PDT 分析步骤

第 1 步，从多个候选基因标记位点中选择 n 个感兴趣的基因标记位点，n 个基因位点即可代表 n 个空间维度。例如，图 1 中所示的就是选择了 2 个候选基因位点，对 2 维基因-基因交互作用进行分析的过程。

第 2 步，对多个基因型进行组合。如果 2 个候选基因位点各有 3 种不同的基因型，就会产生 9 种两基因位点的基因型组合方式。对所有的基因型进行组合，并根据公式(4)计算每一种基因型组合的 geno-PDT 统计量值。

第 3 步，根据阈值对各种基因型组合进行分类。若某种基因型组合的 geno-PDT 统计量小于阈值 τ，则定义该组合(单元格)为低危，反之，则定义为高危。这样，就可以把 n 维结构的交互作用降低至只有一个维度两个水平的形式(高危和低危)。需要说明的是，分析过程中通常设定阈值 $\tau=0$，这是因为根据 geno-PDT 的运算原理，该值可以作为区分各种多位点基因型组合与疾病是否存在关联及关联类型(正关联和负关联)的分界点。

第 4 步，选择高危基因型组合的 geno-PDT 检验统计量作为 MDR-PDT 分析的最终检验统计量。因为上一步中已明确定义了基因型只有两种：高危和低危，并且根据 geno-PDT 的原理，高危基因型的 $D_i(g)$ 值通常与低危基因型相反，只需计算其中之一即可。因此 MDR-PDT 仅选择了高危基因型组合的 geno-PDT 作为最终的检验统计量。

第 5 步，对所有可能的2-位点基因型组合重复上述过程，选择最大的 MDR-PDT 统计值作为 2 位点基因交互作用的检验统计量，并将此模型定义为 2-位点交互作用的最佳模型。

第 6 步，对所选择的模型进行排列检验，得出最终结论。首先确定零假设，即假设2-位点的高危基因型组合与某疾病状态无关；在零假设成立的条件下，基于现有核心家系资料，通过计算机模拟(通常重复 1000 次以上)得到 MDR-PDT 检验统计量的“经验抽样分布”；然后计算 P 值，在零假设成立的前提下，P 值为“经验抽样分布”中 MDR-PDT 值等于及大于(或等于及小于)现有样本检验统计量值的概率；最后根据小概率事件的原理作出统计推断。

上述步骤是以 2 个位点的基因-基因交互作用为例进行说明的，对于 2 个以上位点的基因-基因交互作用分析过程与此类似。

实例分析

以 Martin 等[15]在美国白人中进行的一项研究为例，说明MDR-PDT的分析过程。该研究从阿尔茨海默病（AD）协作项目（the Collaborative Alzheimer Projec, CAP)中选取了 533 个家系，共组成 2018 对 DSPs，探讨 *VR22*、*LRRTM3* 和 *APOE* 3 个基因多个位点间交互作用对 AD 的影响。研究中选择的基因单核甘酸多态性(SNPs)信息具体见表 1。其中 *VR22* 也称 *CTNNA3*，位于 10q21.3，*VR22* 可以编码一种α-连接蛋白，与细胞粘附作用有关，并且 *VR22* 所在区域也是 AD 连锁定位重现率最高的区域之一，因此有研究者认为该基因可能是 AD 的候选基因；*LRRTM3* 基因嵌套于 *VR22* 基因的第 7 内含子区域，该基因可在海马等神经系统特异地表达，可能与 AD 发病有关[21]，Martin 等选取了其中的 3 个 SNPs(SNP1-SNP3)作为候选基因位点；*APOE* 是 AD 关联研究中惟一重现性较高的候选基因，为目前公认的 AD 的易感基因。

表1 研究中选择的等位基因位点特征

标记位点	SNP	等位基因频率
LRRTM3		
SNP1	rs1925617	0.43/0.57
SNP2	rs942780	0.20/0.80
SNP3	rs1925583	0.45/0.55
VR22		
SNP4	rs997225	0.21/0.79
SNP5	rs6480140	0.39/0.61
SNP6	rs7074454	0.37/0.63
SNP7	rs7070570	0.28/0.72
SNP8	rs12357560	0.23/0.77
SNP9	rs7911820	0.37/0.63
SNP10	rs2126750	0.35/0.65
SNP11	rs1786927	0.41/0.59
APOE[a]		
SNP + 3937	rs429358	0.10～0.42/0.58～0.90
SNP + 4075	rs7412	

注：[a] *APOE* 基因共有 3 种等位基因 ε2、ε3 和 ε4，为了与其他位点统一，将 ε2、ε3 合并，故表中等位基因频率分别为 ε4 和 ε2 + ε3 的等位基因频率

采用 MDR-PDT 的方法对这些 SNPs 间的交互作用进行分析，结果表明：对于 1 个位点的模型，只有 *APOE*-4 模型具有统计学意义（$t=8.24$，$P<0.05$）；对于两个位点模型，所有包括 *APOE*-4 的模型均具有统计学意义（$t=6.60\sim8.75$，P 均<0.05），其中 *APOE*-SNP2模型的检验统计量值最大（$t=8.75$），另外一个未包括 *APOE*-4 的SNP1-SNP2 模型也具有统计学意义（$t=5.08$，$P<0.05$）；对于 3 个位点的模型，所有包括 *APOE*-4 的模型均具有统计学意义（$t=6.21\sim8.94$，P 均<0.05），其中 *APOE*-SNP1-SNP9 模型的检验统计量值最大（$t=8.94$）。见表 2。

为了验证 MDR-PDT 分析结果，研究者进一步采用条件 logistic 回归的方法对两位点数据进行分

析，结果发现：以 *APOE*、SNP2 为自变量的两因素模型中仅 *APOE*-4主效应项具有统计学意义（*OR* = 5.99，95% *CI*：3.22～11.12），SNP2 主效应项和 *APOE*-SNP2交互作用项均无统计学意义，从而提示 *APOE*-SNP2 存在交互作用的原因可能是 *APOE*-4存在的强主效应影响的结果。以 SNP1、SNP2 为自变量的模型中，尽管 SNP1、SNP2 的无显著主效应，但两者的交互作用项却具有统计学意义（*OR* = 2.14，95% *CI*：1.30～3.52）。

表2 基于 AD 患病不一致同胞对数据用 MDR-PDT 分析产生的最佳模型

位点	最佳模型	*t* 值	*P* 值
1	*APOE*	8.24	<0.05
2	*APOE*-SNP2	8.75	<0.05
3	*APOE*-SNP1-SNP9	8.94	<0.05

注：上述检验是采用 Permutation test 重复 1000 次的结果；1～3 个位点模型 1000 次重复的界值分别为3.08、4.26和5.50

因此，对于两位点交互作用分析的结果，条件 logistic 回归分析与MDR-PDT均提示 *APOE*-4与 AD 存在关联，另一方面提示无显著主效应的两个 SNPs 之间存在交互作用共同影响 AD 的发生。两者的区别是MDR-PDT识别出的 *APOE*-SNP2交互作用模型在条件 logistic 回归分析中没有重现，这可能是由于 *APOE* 存在强主效应作用的缘故。不过在分析复杂疾病多个因素间的高阶交互作用时，MDR-PDT比传统的 logistic 回归法更有效。

讨　论

MDR-PDT 与原始 MDR 的原理相似，都采用了降维的策略来评价多因子间交互作用。因此MDR-PDT同样具有非参数、不依赖于遗传模式的优点，适用于复杂疾病病因研究中的基因-基因交互作用分析。另外，MDR-PDT还具有geno-PDT的优点，可充分利用了家系中多个家庭成员的信息，在基因型水平剖析复杂性状疾病病因模式，并能用于多种结构类型的核心家系资料。Martin 等[15]在不同的等位基因频率、不同外显率以及存在拟表型的条件下进行了模拟研究，结果表明：对于由 200 个 trios 和 200 个复合 DSPs（家系中包括两个患病同胞和一个非患病同胞）组成的家系资料，分析 1 个位点、2 个位点和 3 个位点交互作用的假阳性率（Ⅰ类错误）均在0.042～0.054之间。在包含有 200 个 trios 的家系研究中分析两位点间交互作用时：无拟表型条件下，MDR-PDT与原始 MDR 均具有较高的效能，所有模型的把握度均在88%～100%；存在拟表型时，两者的效能都会明显降低；但是在等位基因频率较低时，如果不存在拟表型，MDR-PDT的效能则高于原始 MDR。

与原始 MDR 相比，MDR-PDT主要有两个明显的特征。首先是MDR-PDT采用geno-PDT的方法，通过构造家系的综合统计量充分利用了核心家系中多个家庭成员的信息，既扩展了应用范围，也提高了分析效能。Martin 等[15]比较了从一个家系中仅选择一个 DSP 和选择复合 DSPs 时MDR-PDT的效能，结果发现当所有的同胞都纳入分析时，MDR-PDT具有较大的检验统计量值，并具有更高的效能。

其次，MDR-PDT 仅采用排列检验选择最佳交互作用模型，简化了原始 MDR 的分析过程。原始 MDR 则同时采用了交叉验证（cross-validation）和排列检验两种方法选择最佳模型，通常选择交叉验证一致性（consisitency，CV）最大、平均预测误差（prediction error，PE）最小的模型作为最佳交互作用模型。但是实际工作中，常会出现 CV 和 PE 矛盾的情况，即具有最大 CV 值的模型却同时具有较高的 PE 水平，或者具有较低 PE 水平的模型 CV 值也较小。Mei 等[22]曾讨论过不同排列检验（固定、非固定和混合排列检验）以及有无使用交叉验证对 MDR 分析结果的影响，结果发现：有无使用交叉验证对最佳模型的识别过程没有影响；固定排列检验可能具有增大Ⅰ类错误的风险，混合排列检验相对比较保守，识别低维度多因子交互作用的能力稍低，而无交叉验证的非固定排列检验则与使用了交叉验证的原始 MDR 分析结果一致，并具有适宜的把握度和假阳性率。因此MDR-PDT在模型选择时，仅采用了非固定的排列检验，对各种k-位点组合分别进行独立的检验。这样，MDR-PDT既可用于单个位点的主效应分析，也可用于多个位点的交互作用分析，并且其把握度不受进入待评价模型中的位点个数的影响。但也有研究者认为，采用降维的策略选择多因子交互作用模型时最好同时运用交叉验证的方法[23]。因此，在 MDR-PDT 分析过程中究竟是否应该采用交叉验证的方法，有待进一步的研究。另外，为了避免仅采用排列检验可能会降低MDR-PDT识别最佳模型特异度的问题，一方面，尽量不要选择具有显著主效应的候选基因位点来进行分析，除非研究者已确切把握了该候选基因位点的

有关信息；其次，可以采用多元 logistic 回归的方法对MDR-PDT的结果进行验证，即采用“两步走”的分析策略：第一步，对所有的候选基因位点组合进行MDR-PDT分析，筛出具有统计学意义的多因子交互作用模型；第二步，采用多元 logistic 回归对这些具有统计学意义的交互作用模型进行验证。如本文列举的研究实例中，研究者运用MDR-PDT的方法对多个位点基因型的交互作用进行分析后，又进一步采用多元 logistic 回归的方法对前面的分析结果进行了验证。

MDR-PDT 是原始 MDR 的丰富和扩展，可用于多种结构类型的核心家系研究。模拟研究和实例分析均表明该方法具有良好的效能[11,15]，不过其也存在一定的局限性，例如该方法还不适用于具有两代以上家庭成员的扩展家系研究，这主要是因为排列检验不太适用于有多代家庭成员的扩展家系资料，并且把一个大的扩展家系分成几个独立的核心家系进行分析，也不利于同时采用连锁分析对候选基因进行区域定位。尽管如此，MDR-PDT仍是应用于核心家系研究分析基因-基因交互作用的一种有效方法，是将多因子降维策略拓展应用至以家系为基础的遗传流行病学研究的一次崭新尝试。无论从该方法本身的形成过程来看，还是考虑到其应用价值，都对复杂疾病的遗传流行病学研究有着重要意义，这将有助于今后更好地剖析其遗传本质，更全面地认识其复杂的病因基础。

参 考 文 献

[1] 严卫丽，顾东风. 复杂疾病关联研究中的若干问题. 遗传学报，2004，31：533-537.

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

[3] Elston RC, Spence MA. Advances in statistical human genetics over the last 25 years. Statistics in Medicine, 2006, 25: 3049-3080.

[4] 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 Human Genetics, 2001, 69: 138-147.

[5] Cho YM, Ritchie MD, Moore JH, et al. Multifactor-dimensionality reduction shows a two-locus interaction associated with Type 2 diabetes mellitus. Diabetologia, 2004, 47: 549-554.

[6] Moore JH, Williams SM. New strategies for identifying gene-gene interactions in hypertension. Ann Med, 2002, 34: 88-95.

[7] 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 Bioninformatics, 2004, 5: 49.

[8] Chan IHS, Leung TF, Tang NLS, et al. Gene-gene interactions for asthma and plasma total 19E concentration in Chinese children. J Allergy Clin Immunol, 2006, 117: 127-133.

[9] Ma DQ, Whitehead PL, Menold MM, et al. Identification of significant association and gene-gene interaction of GABA receptor subunit genes in autism. Am J Human Genetics, 2005, 77: 377-388.

[10] Brassat D, Motsinger AA, Caillier SJ, et al. Multifactor dimensionality reduction reveals gene-gene interactions associated with multiple sclerosis susceptibility in African Americans. Genes and Immunity, 2006, 7: 310-315.

[11] 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. Genetic Epidemiol, 2003, 24: 150-157.

[12] Cordell HJ, Clayton DG. Genetic epidemiology 3-Genetic association studies. Lancet, 2005, 366: 1121-1131.

[13] 易洪刚，陈峰. 病例父母亲对照研究. 中华流行病学杂志，2004，25(5)：439-444.

[14] 易洪刚，陈峰，于浩，等. 病例同胞对照设计. 中华流行病学杂志，2006，27(2)：170-173.

[15] Martin ER, Ritchie MD, Hahn L, et al. A novel method to identify gene-gene effects in nuclear families: the MDR-PDT. Genetic Epidemiol, 2006, 30: 111-123.

[16] Martin ER, Bass MP, Hauser ER. A genotype-based association test for general pedigrees: The geno-PDT. Am J Human Genetics, 2002, 71: S2365.

[17] Klebanov L, Gordon A, Xiao Y, et al. A permutation test motivated by microarray data analysis. Computational Statistics & Data Analysis, 2006, 50: 3619-3628.

[18] 荀鹏程，赵杨，易洪刚，等. Permutation Test 在假设检验中的应用. 数理统计与管理，2006，25：616-621.

[19] 唐迅，李娜，胡永华. 应用多因子降维法分析基因-基因交互作用. 中华流行病学杂志，2006，27(5)：437-441.

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

[21] Majercak J, Ray WJ, Espeseth A, et al. Promotes processing of amyloid-precursor protein by BACE1 and is a positional candidate gene for late-onset Alzheimer's disease. Proceeding of the National Academy of Sciences of the United States of America, 2006, 103(47): 17967-17972.

[22] Mei H, Ma DQ, Ashley-Koch A, et al. Extension of multifactor dimensionality reduction for identifying multilocus effects in the GAW14 simulated data. BMC Genetics, 2005, 6: S145.

[23] Motsinger AA, Ritchie MD. The effect of reduction in cross-validation intervals on the performance of multifactor dimensionality reduction. Genetic Epidemiol, 2006, 30: 546-555.

（收稿日期：2007-02-12）

（本文编辑：张林东）