遗传 HEREDITAS (Beijing) 2011 年 8 月, 33(8): 820－828
ISSN 0253-9772 www.chinagene.cn

综 述

DOI: 10.3724/SP.J.1005.2011.00820

全基因组基因-基因相互作用研究现状

沈佳薇, 胡晓菡, 师咏勇

上海交通大学 Bio-X 研究院, 上海 200230

摘要: 复杂疾病目前正在全球范围流行, 极大地影响人类的健康。研究发现, 复杂疾病的性状受到多个位点的相互作用影响。目前的全基因组关联分析(Genome-wide association study, GWAS)仅仅解析单个 SNP 位点对疾病易感性的贡献, 单纯依靠这一种策略并不能在寻找复杂疾病的病因上得到根本性的突破。基因-基因相互作用可能是复杂疾病致病的主要因素之一。针对这一点, 科学家已经提出了一些检验基因相互作用的算法, 包括惩罚 logistic 回归模型、多因子降维(Multifactor dimensional reduction)、集合关联法(Set-association approach)、贝叶斯网络(Bayesian networks)、随机森林法等。文章首先对目前这些方法做了综述, 并指出了其中的不足, 包括计算复杂度太高、假设驱动、数据会过度拟合、对低维数据不敏感等, 进而简述了一种由笔者所在实验室开发的基于 GPU 的研究基因相互作用的算法, 该算法复杂度低, 不需要任何假设, 没有边际效应, 有很好的稳定性, 速度快, 适用于进行全基因组范围内的基因-基因相互作用计算。

关键词: 复杂疾病; 基因-基因相互作用; GPU; 性状; 易感位点

Current status of studies on genome-wide gene-gene interactions

SHEN Jia-Wei, HU Xiao-Han, SHI Yong-Yong

Bio-X Institutes of Shanghai Jiao Tong University, *Shanghai* 200230, *China*

Abstract: Complex diseases have affected human's health throughout the world. Hundreds of studies show that complex diseases are caused by multiple loci. Currently, genome-wide association studies(GWAS) only focus on the single locus that contributes to the susceptibility of a certain disease. However, the interaction between genes could be one of the main factors that lead to complex traits. This fact has initiated scientists to propose some algorithms to detect these interactions, such as the penalized logistic regression model, multifactor dimensionality reduction method, set association analysis method, Bayesian networks analysis method and random forest. However, these algorithms are of high complexity, hypothesis-driven, causing over fitting of data, or not sensible of data at low dimensions. In this paper, we reviewed these algorithms, and then demonstrated a new algorithm based on GPU to provide a powerful strategy to analyze gene-gene interaction in genome-wide association datasets. This algorithm is of low computing complexity, free of hypothesis, not affected by single locus marginal effect, and also of high stability and speed.

Keywords: complex disease; gene-gene interaction; GPU; traits; susceptibility locus

收稿日期: 2011–05–12; 修回日期: 2011–07–23
基金项目: 国家自然科学基金项目(编号：31000553), 国家高技术研究发展计划项目(863 计划)项目(编号：2009AA022701)资助
作者简介: 沈佳薇, 本科, 专业方向：遗传学、生物信息学。E-mail: celaoforever@foxmail.com
通讯作者: 师咏勇, 博士, 研究员, 研究方向：遗传学、生物信息学。E-mail: shiyongyong@gmail.com

复杂疾病是由多对微效基因与环境因素共同作用所致, 具有明显的遗传异质性、表型复杂性及种族差异性等特征。复杂疾病在人群中的发病率很高, 严重影响人类的身心健康, 如糖尿病、癌症、中风、心脏病、抑郁症、哮喘、自身免疫疾病等。目前最为流行的复杂疾病研究手段是基于单个位点的全基因组关联分析。由于复杂疾病的发病机制十分复杂, 往往是由多个位点共同作用而导致的, 所以相当部分单个位点的作用可能很微弱以致很难被全基因组关联分析所发现。因此, 对于基因-基因相互作用的分析是一种潜在的解决方法。目前, 几种计算基因相互作用的方法被报导, 并取得了一定的研究成果。本文首先对目前现有的算法进行了总结, 并提出了其中的不足之处, 进而陈述了一种由笔者所在课题组开发的基于 GPU(Graphic Processing Unit)的算法, 该算法能有效克服现有算法的不足, 并能以一个十分可观的效率运行。

1　基因相互作用

GWAS(Genome-wide association study, GWAS)解析的只是单个 SNP 位点对疾病易感性的贡献, 而在大部分情况下, 复杂疾病的致病位点不止一个。大量证据及经验说明, 复杂疾病的性状受到多个位点之间的相互作用的共同影响[1~3]。单纯依靠关联研究一种策略并不能在寻找复杂疾病的病因上得到根本性的突破。Moore 等[4]认为上位效应(Epistasis, 基因-基因相互作用)是常见疾病发病的普遍原因。这种基因-基因相互作用可以分为以下几类[5]:

显性上位作用(Dominant epistasis): 两对基因控制同一性状, 其中一对基因的显性效应对另一对基因的表现有遮盖作用;

隐性上位作用(Recessive epistasis): 两对基因控制同一性状, 当其中一对基因是隐性时, 对另一对基因起遮盖作用;

双基因累加效应(Duplicate genes with cumulative effect): 控制同一性状的两对非等位基因的显性表现型效应相同, 双显同时存在时呈现比各自单独存在时更突出的新表现型, 两者的效应可以累加;

双显性基因(Duplicate dominant genes): 控制同一性状的两对非等位基因, 只要其中一个基因是显性, 那么表现型将都相同。只有两对基因都是隐性纯合时, 才会有不同的表现型;

双隐性基因(Duplicate recessive genes): 控制同一性状的两对非等位基因, 只有当两个基因中都存在显性时, 才有一种表现型。反之, 若两个基因其中一个存在隐性纯合时, 表现型将相同;

显性隐性相互作用(Dominant&recessive interaction): 控制同一性状的等位基因当其中一个存在显性, 或者存在两个都是隐性纯合时, 表现型相同。

表 1 列出了以上 6 种不同的基因相互作用的基因型与表现型的关系[6]。

表 1　不同相互作用类型的基因型与表现型

相互作用类型	A-B-	A-bb	aaB-	Aabb
孟德尔比率	9	3	3	1
显性上位作用	12		3	1
隐性上位作用	9	3	4	
双基因累加效应	6	6		1
双显性基因	15			1
双隐性基因	9		7	
显性隐性相互作用	13		3	

2　现有分析基因-基因相互作用的算法

2.1　多因子降维法

多因子降维法(Multifactor-dimensionality reduction, MDR)能够减少多位点信息的维数, 并以相对小的计算复杂度计算与某一种疾病相关的基因的相互作用。该方法是非参数的、无模型的, 并且对于病例-对照研究是可以直接应用的[7]。

多因子降维法中的“因子”是所关注的交互作用的变量, 可以是环境因素或者是基因型等等; “维”是指这些多因子组合中因子的数目。根据疾病易感性将这些多因子组合分类, 分成低危或者高危。然后把这些相互作用的变量看作一个多因子组合。

MDR 算法的基本步骤如图 1 所示[8,9]: ①将样本平均分割成 10 份, 将其中的 9 份作为训练样本, 另外一份作为检验样本, 以便之后做交叉验证。②选择 n 个变量作为研究对象, 可以是环境因素或者是 SNP 位点等等。③计算病例/对照样本中的基因型频率。如图所示, 每个单元格中左边的条带表示病例, 右边条带表示对照。④计算每个变量组合的病例/对照的比值。并将低于阈值的基因型组合标记为

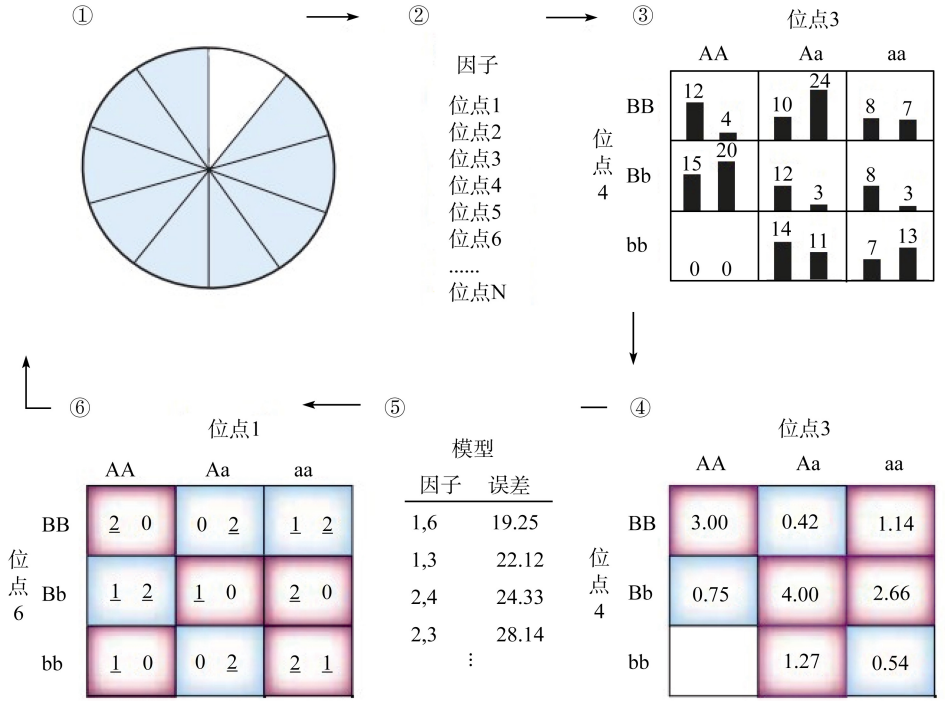

图 1 MDR 算法示意图[8,9]

低危，而高于阈值的基因型组合标记为高危。⑤将剩下的 1/10 样本作为检验样本，来评估每个位点组合的预测误差。并在所有的两因子组合中选择错分最小的 MDR 模型，这两个位点模型在所有可能的模型中将具有最小的预测误差。⑥进行 10 重交叉检验。取 10 次检验的预测误差的平均值，并将这个平均值作为模型相关预测误差的无偏估计，以此来评估单元格分配时的相关误差以及该模型的预测误差。

Tsai 等[10]利用 MDR 方法发现了房颤中相互作用的基因对(*RAS-ACE*)。MDR 显示最佳模型由 3 个 SNP 组成，其中 2 个 SNP 来自 *RAS* 基因，1 个 SNP 来自 *ACE* 基因。这 3 个 SNP 的 10 重交叉检验显示有很好的一致性，100 次的 permutation test 得到的 *P* 值在 0.001 水平上；Cho 等[11]对 II 型糖尿病的 15 个候选基因进行评估，涉及 23 个 SNP，这 23 个 SNP 的频率在病例和对照之间没有很大的不同。对于任何可能的模型的 MDR 分析显示位于基因 *UCP2* 和 *PPARγ* 上的 2 个 SNP 位点有相互作用。

MDR 方法选取合适的基因型组合，检验所有可能的多位点基因型的组合，并且报告最佳的分类组合。由于 MDR 方法采用了 10 重交叉检验，在样本量较小的时候，依然可以达到比较高的准确度。然而，MDR 算法也存在一些不足。首先，随着阶数的增加，可能会导致数据的过度拟合。并且随着模型大小增加，预测误差也会随之增加。其次，当主效应或已知的协同作用存在时，MDR 方法灵敏度并不高[12]。而且，当交互作用存在且是低维度的时候，MDR 算法往往很难发现其中的相互作用[13]。

2.2 惩罚 logistic 回归

传统的 logistic 回归不能分析高阶相互作用，惩罚 logistic 回归通过对经典的 logistic 回归模型做了一些修正，将修正系数 λ 和经典的 logistic 回归模型相结合，有效解决了这一问题。

惩罚 logistic 回归模型的目标函数则表示为:

$$L(\beta) = -\log(y \mid \beta) + \frac{1}{2}\beta^T \Lambda \beta =$$

$$-\sum_{i=1}^{n}(y_i \log p_i + (1-y)\log(1-p_i)) + \frac{1}{2}\beta^T \Lambda \beta$$

其中，L 表示二项式分布的对数似然函数，$\frac{1}{2}\beta^T \Lambda \beta$ 为二次方惩罚项，Λ表示对角元素 $\{0, k \ldots k\}$ 的已知对角矩阵。

Newton-Raphson 迭代表示为:

$$\beta_{new}=\beta-(\frac{\partial^2 L}{\partial\beta\partial\beta^T})^{-1}\frac{\partial L}{\partial\beta};$$

$$\frac{\partial L}{\partial\beta}=X^T(p-y)+\Lambda\beta;\frac{\partial^2 L}{\partial\beta\partial\beta^T}=X^tWX+\Lambda$$

其中, X 是预测因子的$(p+1)\times n$ 阶矩阵(p 和 n 分别表示样本量和预测因子数), y 为二分类(0/1)反应变量的向量。

$$p=(p_1,p_2,\ldots,p_n)^T;$$
$$W=diag(p_1(1-p_1),\ldots,p_n(1-p_n));$$
$$\beta_{new}=(X^TWX+\Lambda)^{-1}X^T(WX\beta+(y-p))$$

模型的有效自由度和系数的方差可以通过下面的公式来得到[14,15]。

有效自由度:

$$df(k)=tr[(X^TWX+\Lambda)^{-1}X^TWX]$$

系数方差:

$$Var(\hat{\beta})=Var[(X^TWX+\Lambda)^{-1}X^TWZ]$$
$$=(\frac{\partial^2 L}{\partial\beta\partial\beta^T})^{-1}I(\beta)(\frac{\partial^2 L}{\partial\beta\partial\beta^T})^{-1}$$

Park 等[13]利用惩罚 logistic 回归对膀胱癌数据进行分析, 并与 MDR 方法进行比较, 结果表明, 惩罚 logistic 回归相比 MDR 有着更好的特异性, 更低的错误率, 然而灵敏度相比 MDR 稍低。

惩罚 logistic 回归相比传统的 logistic 回归偏差更小, 模型更稳定, 在样本量较低并且阶数较高时, 惩罚 logistic 回归更具有优势。然而, 方程中的参数需要通过迭代来计算, 随着迭代次数的上升, 计算复杂度也将呈指数上升, 这对于扫描全基因组范围内的所有两两组合的位点是不可行的。因此, 很难用于全基因组范围内的研究。

2.3　贝叶斯网络(Bayesian networks)法

贝叶斯网络用有向无环图的形式描述一组随机变量之间的联合概率分布(图 2)[16]。图中的一个节点表示一个离散或连续变量(可以是基因或者是 SNP 位点), 节点之间的连线表示变量之间的关联性。节点之间有连线的表明这些节点之间存在相互作用, 而没有直接连线的两个节点相互条件独立, 即在其他变量存在的情况下相互独立。连线的方向用来区分"父节点"和"子节点", 连线指向的节点为子节点, 反之, 是父节点。

如图所示, 节点表示 SNP 位点。图中, A 的父节点是 B 和 C, A 和 B、C 之间存在相关性, 而与 D、E 相互独立。

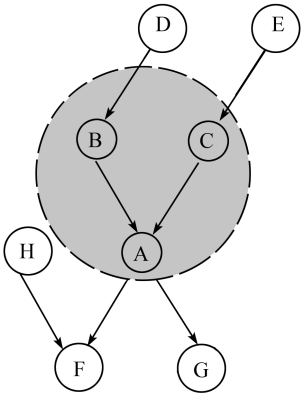

图 2　贝叶斯网络法检测基因基因相互作用示意图[16]

节点和节点之间的关系可以用一个联合概率分布 $P(X_1, \ldots X_n)$来表示:

$$P(X_1,\ldots,X_n)=$$
$$\prod_{i=1}^{N}P(X_i=x_i\mid X_j=x_j,\ldots,X_{j+p}=x_{j+p})$$

条件中的 $p+1$ 个基因($X_i,\ldots, X_{j+p}$)是基因 i 的父节点。联合概率密度可以应用概率和独立性的链式法则, 用条件概率的乘积来表示。这个法则是基于贝叶斯理论的:

$$P(A,B)=P(B\mid A)*P(A)=P(A\mid B)*P(B)$$

为了建立这样一个基因的贝叶斯网络模型, 需要找到一个能够最贴切地描述基因表型的有向无环图。可以选择一个评分方程来评估每个对应不同的基因表型的有向无环图 G, 然后寻找能够使得这个评分最大化的图 G。最著名的评分方程是 Bayesian Information Criteria (BIC) 或 Bayesian Dirichlet equivalence(BDe)[16]。

一个基于贝叶斯网络的方法是 BEAM(Bayesian epistasis association mapping), 该算法结合贝叶斯模型和 Metropolis-Hasting 算法把 SNP 分成 3 组, 第 1 组的 SNP 与疾病无关, 第 2 组的 SNP 对疾病有独立的影响, 第 3 组的 SNP 相互作用对疾病产生作用。然后利用 B statistic 对候选 SNP 进行进一步的分析。对 AMD(Age-related Macular Degeneration)数据的分

析表明, 该方法相比 MDR 和传统 logistic 回归有更好的表现[17]。

贝叶斯网络方法具有一系列良好的性质, 比如可以加入先验知识、避免数据的过度拟合等。然而也存在一定的不足, 如时间复杂度和空间复杂度都很高: 每个节点需要计算的条件概率的次数与该节点的父节点数目是呈指数关系的, 需要很大的空间和时间来计算和存储这些数据。且找出所有可能的有向图 G 是 NP 完全问题, 为了提高效率, 需要一些启发式算法来减少某些可能不合适的有向图 G 的组合, 从而减少计算分支。但是这样可能会丢失一些组合, 并不能进行位点的完全扫描。

2.4 集合关联法(Set-association approach)

集合关联法同时使用了 SNP 的两个信息: 等位基因关联性(Allelic association, AA)和 Hardy-Weinberg disequilibrium (HWD)。这两个值都用卡方检验值来表示。该方法的主要步骤如下[18]:

(1) 删减位点: 由于在病例样本中呈现较高的 HWD 的可能是疾病的易感位点, 但是过高的 HWD 值可能源于错误的基因分型。因此, 在这一步中使用控制样本的 HWD 值作为筛选的依据。例如, 99% 的 SNP 的卡方值都小于 6.6, 那么可以将 6.6 作为一个筛选的阈值, 并记录下所有卡方值大于 6.6 的位点数, 设为 d。

(2) 计算 HWD 值: 分别计算每个 SNP 在病例和对照样本中的卡方值, 并将这两个值相加。将其中 d 个有最大值的 SNP 位点的这个和设为 0。

(3) 加权: 合并 AA 和 HWD 的效益, 计算 $t_i \times u_i$, t_i是 AA 第 i 个 SNP 的统计值, u_i是第 i 个 SNP 的 HWD 的统计值。为了能够将多个 SNP 位点的关联性联合起来考虑, 计算 $S = \sum_i (t_i \times u_i)$。

(4) 分组: 这一步要解决的问题是, 判断将哪些 SNP 的信息包括在统计和中。首先, 将每个位点的 AA 和 HWD 的和, S_i, 按照大小降序排列, 得到 $s_{(1)} \geqslant s_{(2)} \geqslant s_{(3)} \geqslant \cdots\cdots$。然后计算如下项的和:

$$S(n=1) = s_{(1)}, S(n=2) = s_{(1)} + s_{(2)}, \cdots\cdots$$

(5) Permutation test: 对所有数据重复上述步骤进行 permutation test, 得到 P 值。当 S 中所包含的位点数目 n 增加时, P 值也会随之变动, 所要找的是第一次出现 P 极小值时所对应的 n 值。这 n 个 SNP 将是所关注的可能存在相互作用的 SNP 集合。

用该方法研究心血管疾病, 有效地识别出了相互作用的位点。所有的样本在 89 个 SNP 上被基因分型, 共涉及 62 个候选基因, 其中, 当 9 个 SNP 被包含时, 得到最小 P 值(0.021)[19]。

此方法不仅有效地控制 I 类错误, 而且能有效地识别出相互作用的 SNP。其缺点是, 它不是一个全基因组范围内的扫描。另外, 作为这个方法的输入, 需要选出一些候选的位点, 这一步是基于假设驱动的。

2.5 随机森林法

随机森林是一个包含多个决策树的分类器。森林中的每棵树使用 bootstrap 取样法从数据中选取样本, 并从所有的属性中随机选取一个子集。随机森林法通过这些选取的样本根据这个属性的子集来分类, 找出最合适的分类[20]。

随机森林法步骤示意图见图 3[21], 对于一个有 N 个样本 M 个属性的数据集, 个体树可以通过如下步骤建立起来:

(1) 从 N 个样本中可重复地选取 N 个样本;

(2) 对于树中的每个节点, 随机地从所有 M 个属性中选取 m 个属性(m 的大小在树林建立的整个过程中保持恒定);

(3) 找出基于这 m 个属性的最好的样本的分割方式;

(4) 重复第二步和第三步直到这个树完全成长(没有剪枝)。

随机森林法的效能可以通过模拟验证得到。假定有 3 种疾病模型, 在模型 1 中, 两个位点独立的对疾病产生作用; 模型 2 与模型 1 相似, 不同的是, 只有当两个位点至少有一个疾病等位基因时, 疾病的症状才会表现出来; 对于模型 3, 多余的疾病等位基因不会增加疾病的风险。假设在 3 个模型中, 疾病的患病率是 0.1。与其他算法(BEAM, logistic 回归, X2 检验)比较的结果表明, 在模型 1 中, 这些算法的效能相当。而在模型 2 和 3 中, 在 MAF 较小时, 随机森林法的效能不如其他的算法, 在 MAF 较大时, 随机森林法与其他算法效能相当。而当 MAF 在 0.1~0.2 之间时, 随机森林法相比其他算法效能更高[22]。

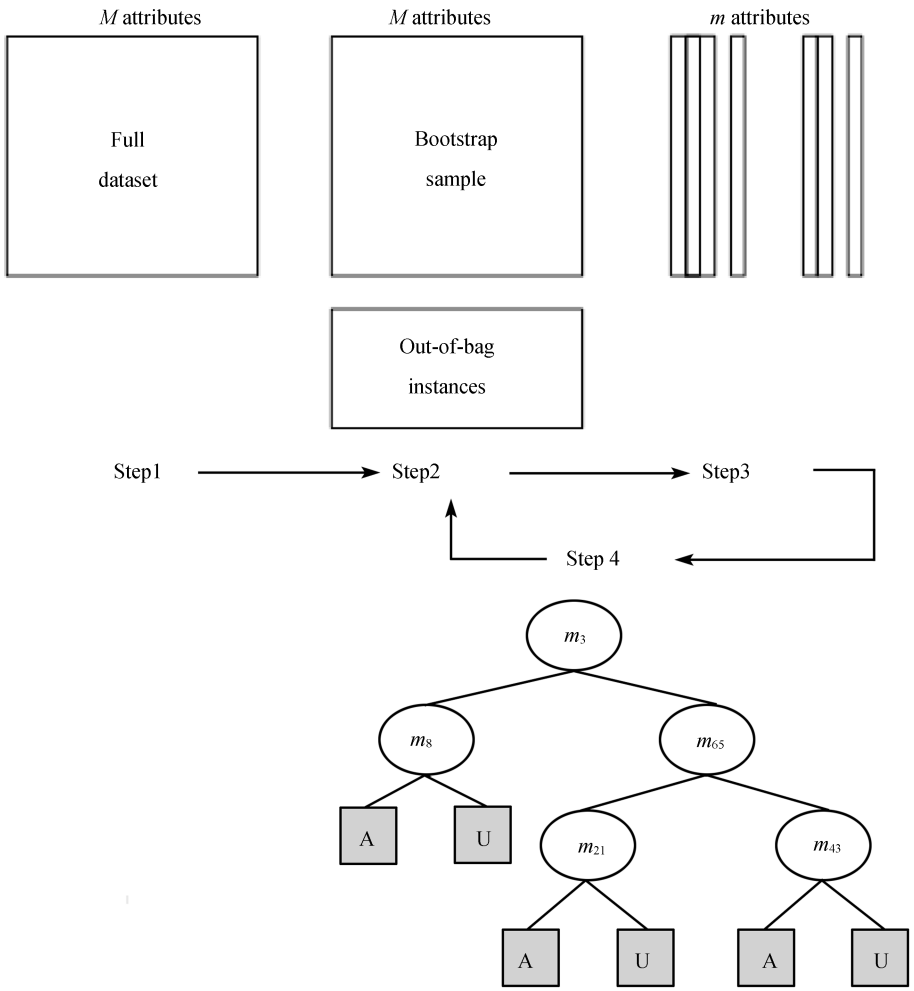

图 3　随机森林法步骤示意图[21]

使用随机森林法生成的决策树揭示了这些属性之间的相互作用[23]，可以应用于研究基因-基因和基因-环境相互作用。这个方法能够揭示那些没有明显边际效应的基因之间或者基因环境之间的相互作用[24]。随机森林法在计算基因-基因相互作用和基因环境相互作用时不需要先验假设[25]。然而，随机森林法会导致过度拟合，而且当被研究的属性在很大程度上不相关时，随机森林法并不能很好地工作。

3　SHEsisEpi

SHEsisEpi 是由我们实验室开发的一个基于 GPU(Graphic Processing Unit)计算基因-基因相互作用的程序(http://analysis.bio-x.cn/SHEsisMain.htm)。它能够在全基因组范围内扫描每对位点，而且速度是 CPU 的数百倍(单个 CPU 与单个 GPU 比较)。

3.1　算法说明

SHEsisEpi 所使用的算法是 Odds Ratio Test(相对危险度测试)的一个衍生。具体描述如下:

假设 A，B 是两个位点，A_i 表示 A 位点上的某个基因型(如 AT)，而 A_i^c 表示非该基因型的其他基因型集合(AA/TT)。同理，B_i 表示 B 位点上的某个基因型，B_i^c 表示非该基因型的其他基因型集合。假定基因型组合 $A_iB_j, A_i^cB_j, A_iB_j^c, A_i^cB_j^c$ 对应的外显率是 $U_{ij}, U_i^{c}{}_j, U_{ij}^c, U_i^{c}{}_j^c$。

$$OR_{Disease:A_i/B_j}=\frac{P(A_iB_j\mid Disease)P(A_i^cB_j^c\mid Disease)}{P(A_iB_j^c\mid Disease)P(A_i^cB_j\mid Disease)}=\frac{D_{ij}D_i^c{}_j^c}{D_{ij}^cD_i^c{}_j}$$

$$OR_{Normal:A_i/B_j}=\frac{P(A_iB_j\mid Normal)P(A_i^cB_j^c\mid Normal)}{P(A_iB_j^c\mid Normal)P(A_{i^c}B_j\mid Normal)}=\frac{N_{ij}N_i^cj^c}{N_{ij}^cN_i^cj}$$

$D_{ij},D_i^cj,D_{ij}^c,D_i^cj^c,N_{ij},N_i^cj,N_{ij}^c,N_i^cj^c$ 分别表示病例组和对照组中基因组合 $A_iB_j,A_i^cB_j,A_iB_j^c,A_i^cB_j^c$ 的观察计数。那么,

$$EOR_{ij}=\frac{OR_{Disease:A_i/B_j}}{OR_{Normal:A_i/B_j}}$$

$$EOR_{ij}=1\Leftrightarrow OR_{disease:A_j/B_j}=OR_{Normal:A_i/B_j}$$

当 A_i 与 B_j 在病例或者对照数据中有相互作用, 并且这种相互作用对外显性有影响, 那么 $EOR_{ij}\neq1$。所以易感性的一个等价数学表述是: 至少有一个基因型组合不满足 $EOR_{ij}=1$。

且该 *OR* 值仅与这两个位点上的基因型的条件概率有关, 而独立于各基因型的边际频率。因此, *EOR* 值可以不受边际效应影响地反映基因型相互作用。

3.2 速度比较

SHEsisEpi 计算同一对文件, GPU(nVidia GTX480)耗时 140 ms, CPU(intel core i7 930)耗时 149386 ms, 该硬件条件下, GPU 的速度是 CPU 的 1000 倍。按照这样的加速比, 使用一块 GTX480 显卡完成一组数据的扫描需要 24 h。如果使用 CPU 的话, 扫描同样的数据需要 3 年时间。

同时, 程序现在已经能够进行 permutation test。如果做 100 次 permutation, 使用一块 GPU 需要 3 个月的时间, 如果使用 5 块显卡, 大约 20 d 就可以完成这 100 次 permutation test, 这是完全可以接受的。而对于 CPU 来说, 一块 CPU 做 100 次 permutation 需要 300 年, 就算使用多个 CPU 同时计算, 总耗时也是天文数字。

3.3 模拟验证结果

我们模拟了 1 000 个 case 和 1 000 个 control 在 1024 个位点的数据, 并假设其中 512 个位点位于一条染色体上, 另外 512 个位点位于另一条染色体上。在这些位点中, 人为设置了一对相互作用的位点。该位点生成的原则如下: 在 case 中, 40%的样本在这两个位点上的基因型不相互独立, 例如, 如果在一个位点上基因型为 A/G 时, 那么在另一位点上基因型为 G/G; 如果在一个位点上基因型为 A/A 时, 那么另一位点上的基因型为 A/G。

然后使用 SHEsisEpi 计算这个模拟数据的位点间的相互作用, 结果表明, SHEsisEpi 能够以 $p=9.103\times10^{-28}$ 识别出该对位点的相互作用。

模拟验证结果表明, SHEsisEpi 的确能够识别出位点间的相互作用。

3.4 SHEsisEpi 产生 I 类错误和 II 类错误的概率

产生 I 类错误的概率由显著水平 α 决定, 该值是可以人为控制的, 一般定为 0.05。程序输出中的 *P* 值即为在 $P<0.05$ 情况下, 产生 I 类错误的概率。

从 SHEsisEpi 的算法描述中可知, *EOR* 是某对位点在正常人群和患病人群中 *OR* 值的比值, 当其偏离 1 时, 表明该对位点存在相互作用。(*EOR*-1)服从 0-1 正态分布, 使用 U 检验法来估计产生 II 类错误的概率 β。

$$\beta=\Phi(z_{\frac{\alpha}{2}}+\lambda)+\Phi(z_{\frac{\alpha}{2}}-\lambda)-1$$

其中, $z_{\frac{\alpha}{2}}=z_{0.025}=1.96,\lambda=\frac{\sqrt{n}}{\sigma}\delta,\delta=|\mu-\mu_0|$

对于 II 类错误的评估, 将从两个方面来看: 样本容量 *n* 一定, δ 变化时, 产生 II 类错误的概率变化情况; δ 一定, 样本容量 *n* 变化时, 产生 II 类错误的概率的变化情况。

取 δ=0.3, 观察 *n* 从 100 到 1 000 变化时, II 类错误产生概率的变化曲线见图 4A。由图可见, 在 δ 一定时, 样本容量越大, 产生 II 类错误的概率越小。当 δ=0.3 时, 样本容量大于 300 时, 产生 II 类错误的概率非常小。

取 *n*=100, δ 在 0 和 1 内变化时, II 类错误产生的概率的变化曲线见图 4B。由图可得, 在样本容量一定时, δ 越大, 产生 II 类错误的概率越小。当在样本容量=100 时, δ>0.4 时, II 类错误的产生概率很小。

3.5 程序使用的成功案例

我们使用 SHEsisEpi 分析了 WTCCC 的双向情感障碍的 GWAS 数据, 耗时 27 h(基于两块 GTX285 显卡), 找出了若干对上位互作的风险 SNP 位点。

将程序的输出信息按照 *P* 值递增排列, 并增加

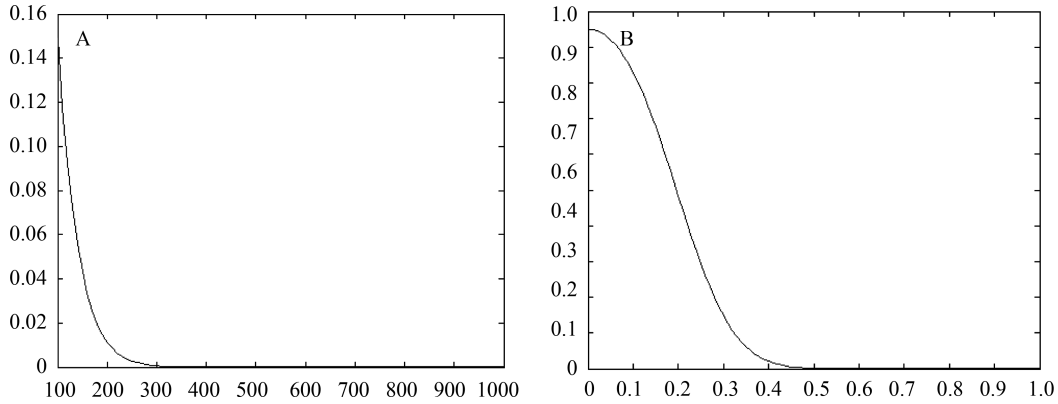

图 4　SHEsisEpi 产生 II 类错误的概率

A: δ 变化时, II 类错误的产生概率与样本容量的关系; B: 样本容量 n 一定时, II 类错误产生的概率与 δ 的关系。

了以下的标准来筛选位点: ①每对中的 SNP 必须位于不同的染色体上, 这样可以避免由于连锁不平衡而导致的相互作用。②每对 SNP 位点 MAF<0.05。③去除那些已经被认定为其他复杂疾病的易感位点的 SNP。因此所要做的就是找出 *P* 值最小的且符合以上 3 条标准的位点对。

根据这些标准, 找出的位点对分别是: ①rs10124883–rs178069; ② rs6004133–rs10124883。为了更进一步地研究这些位点的相互作用, 增加了以下几个位点: rs10123629, rs165596, rs165730 [26] 。

在重复试验中, 我们对这些位点进行了验证, 得到了与程序计算一致的结果。

我们针对公共数据库和本实验室已有的 7 份 GWAS 数据(包括 II 型糖尿病和前列腺癌等), 接近 10 000 个样本, 利用 SHEsisEpi 计算基因的相互作用。结果显示对于同一种疾病的若干个独立样本, SHEsisEpi 的重复性较好,对于从一份数据中候选出来的基因对, 在其他独立样本中都得到了验证, 从而证明 SHEsisEpi 的确有较好的稳定性。

4　结 语

本文对现有的分析基因-基因相互作用的算法作了小结: 在样本量较小的情况下, MDR 算法依然有着较高的正确性, 但对于高阶模型, 其预测能力变差, 且在交互维度较小时, MDR 算法几乎无能为力; 惩罚 logistic 回归在样本量低且阶数高时, 具有一定的优势, 但是随着迭代次数的上升, 其计算复杂度指数上升, 很难应用到全基因组范围内的扫描上; 贝叶斯网络法可以加入先验知识, 能够避免数据过度拟合, 但是它的时间和空间复杂度都很高, 并且需要一些启发式算法来减少分支, 这样会丢失一些位点的组合, 同样也不能应用到全基因组范围内; 集合关联法能够有效控制 I 类错误, 但这个方法是假设驱动的, 在应用前, 需要选出一些候选位点作为这个方法的输入数据。随机森林法虽然不需要任何假设, 但是会导致数据的过度拟合。

我们的算法 SHEsisEpi 能够有效克服上述的局限性: 在速度方面, SHEsisEpi 有着无法比拟的优势, 远远超过了 CPU 的速度, 这也是能够使用这个算法进行 permutation test 的先决条件之一; 在计算范围方面, SHEsisEpi 能够在全基因范围内扫描而不丢失任何位点组合, 在使用这个算法之前不需要任何假设。当然, SHEsisEpi 也是有缺陷的, 目前它只能处理两两位点的相互作用分析。

参考文献**(References)**:

[1] Mackay TFC. Quantitative trait loci in Drosophila. *Nat Rev Genet*, 2001, 2(1): 11–20.

[2] Segre D, DeLuna A, Church GM, Kishony R. Modular epistasis in yeast metabolism. *Nat Genet*, 2004, 37(1): 77–83.

[3] Williams SM, Haines JL, Moore JH. The use of animal models in the study of complex disease: all else is never equal or why do so many human studies fail to replicate animal findings? *BioEssays*, 2004, 26(2): 170–179.

[4] Moore JH. A global view of epistasis. *Nat Genet*, 2005, 37(1): 13–14.

[5] Phillips PC. The language of gene interaction. *Genetics*, 1998, 149(3): 1167–1171.

[6] http://www.microbiologyprocedure.com/genetics/genetic-i

nteraction/dominant-and-recessive-interactions-13-3.htm.

[7] Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet*, 2001, 69(1): 138–147.

[8] Moore JH. Computational analysis of gene-gene interactions using multifactor dimensionality reduction. *Expert Rev Mol Diagn*, 2004, 4(6): 795–803.

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

[10] Tsai CT, Lai LP, Lin JL, Chiang FT, Hwang JJ, Ritchie MD, Moore JH, Hsu KL, Tseng CD, Liau CS, Tseng YZ. Renin-angiotensin system gene polymorphisms and atrial fibrillation. *Circulation*, 2004, 109: 1640–1646.

[11] Cho YM, Ritchie MD, Moore JH, Park JY, Lee KU, Shin HD, Lee HK, Park KS. Multifactor-dimensionality reduction shows a two-locus interaction associated with Type 2 diabetes mellitus. *Diabetologia*, 2004, 47(3): 549–554.

[12] Coffey CS, Hebert PR, Ritchie MD, Krumholz HM, Gaziano JM, Ridker PM, Brown NJ, Vaughan DE, Moore JH. 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(1): 49.

[13] Park MY, Hastie T. Penalized logistic regression for detecting gene interactions. *Biostatistics*, 2008, 9(1): 30–50.

[14] Lee AH, Silvapulle MJ. Ridge estimation in logistic regression. *Comm in Statis-Simulation and Comp*, 1988, 17(4): 1231–1257.

[15] Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Mateo: Morgan Kaufmann, 1988.

[16] Bansal M, Belcastro V, Ambesi-Impiombato A, di Bernardo D. How to infer gene networks from expression profiles. *Mol Syst Biol*, 2007, 3(1): 78.

[17] Zhang Y, Liu JS. Bayesian inference of epistatic interactions in case-control studies. *Nat Genet*, 2007, 39(9): 1167–1173.

[18] Hoh J, Wille A, Ott J. Trimming, weighting, and grouping SNPs in human case-control association studies. *Genome Res*, 2001, 11(12): 2115–2119.

[19] Ott J, Hoh J. Set association analysis of SNP case-control and microarray data. *J Comput Biol*, 2003, 10(3-4): 569–574.

[20] Breiman L. Random forests. *Mach Learn*, 2001, 45(1): 5–32.

[21] McKinney BA, Reif DM, Ritchie MD, Moore JH. Machine learning for detecting gene-gene interactions: a review. *Appl Bioinformatics*, 2006, 5(2): 77–88.

[22] Jiang R, Tang W, Wu X, Fu W. A random forest approach to the detection of epistatic interactions in case-control studies. *BMC Bioinformatics*, 2009, 10(S1): S65.

[23] Breiman L. Classification and Regression Trees. New York: Chapman & Hall/CRC, 1984.

[24] Cook NR, Zee RYL, Ridker PM. Tree and spline based association analysis of gene-gene interaction models for ischemic stroke. *Stat Med*, 2004, 23(9): 1439–1453.

[25] Lunetta KL, Hayward LB, Segal J, van Eerdewegh P. Screening large-scale association study data: exploiting interactions using random forests. *BMC Genetics*, 2004, 5: 32.

[26] Hu XH, Liu Q, Zhang Z, Li ZQ, Wang SL, He L, Shi YY. SHEsisEpi, a GPU-enhanced genome-wide SNP-SNP interaction scanning algorithm, efficiently reveals the risk genetic epistasis in bipolar disorder. *Cell Res*, 2010, 20(7): 854–857.