文章编号:1674-2974(2016)10-0155-06

一种分析全基因组上位性的新方法*

李泽军[1,2],陈 敏[1,2†],曾利军[2]

(1.湖南大学 信息科学与工程学院,湖南 长沙 410082;2.湖南工学院 计算机科学与信息学院,湖南 衡阳 421002)

摘 要:传统基于单位点的全基因组关联研究存在重复性低、难以解释性等缺陷,而采用基于机器学习的上位性分析中面临计算复杂度高、预测准确度不足等问题.本文提出一种分析全基因组上位性的新方法,该方法采用二阶段框架的上位性分析方法,它包含特征过滤阶段以及上位性组合优化阶段,在特征过滤阶段提出了多准则融合策略,从多个不同角度评价遗传变异位点,以保证易感的弱效位点能被保留,然后采用多准测排序融合策略剔除与疾病状态关联程度低的遗传变异,进一步在上位性组合优化阶段采用贪婪算法启发式地搜索组合空间,以降低时间复杂度,最后采用支持向量机作为上位性评价模型.实验中采用不同的连锁不平衡参数与经典算法 SNPruler 与 ACO 的性能进行对比,实验结果表明:本文方法能有效保留弱效位点,一定程度上提高了疾病预测的正确度.

关键词:全基因组关联研究;上位性;复杂疾病;智能计算

中图分类号:TP39 **文献标识码**:A

A Genome-wide Epistasis Analysis Method Based on Multiple Criteria Fusion

LI Ze-jun[1,2],CHEN Min[1,2†],ZENG Li-jun[2]

(1. College of Computer Science and Electronic Engineering, Hunan Univ, Changsha, Hunan 410082, China;
2. School of Computer and Information Science, Hunan Institute of Technology, Hengyang , Hunan 412002, China)

Abstract: Traditional units of genome-wide association studies have serious defects such as low repeatability, difficulty to interpret, and epistasis analysis based on machine learning has troubles such as high computational complexity and insufficient prediction accuracy. This paper presented a new approach for the analysis of genome-wide epistatic. This method uses the framework of two-phase epistatic analysis method. It includes a filtering stage and an epistatic combinatorial optimization stage. The characteristics of the filtering stage presents a multicriteria fusion strategy for the evaluation of genetic loci from multiple perspectives to ensure that the weak effect of susceptibility loci can be retained, and then, this method uses the multiple criteria sorting fusion strategy to eliminate the low degree of genetic variation associated with disease states. Epistatic combinatorial optimization phase uses the greedy algorithm combination of heuristic search space in order to reduce the time complexity. Finally, a support vector machine was used as the epistatic evaluation model. Experiments with different parameters of linkage disequilibrium SNPruler with

* **收稿日期**:2016-03-26

基金项目:国家自然科学基金资助项目(61672223),National Natural Science Foundation of China(61672223);湖南省自然科学基金资助项目(2016JJ4029)

作者简介:李泽军(1972—),男,湖南常宁人,湖南工学院副教授,湖南大学博士生

† 通讯联系人,E-mail:9918428@qq.com

classical algorithms were compared with the performance of the ACO, and the experiment results show that the method can effectively keep weak effect locus and improve disease forecasting accuracy considerably.

Key words: GWAS (Genome-Wide Association Study); epistasis; complex diseases; intelligent computing

复杂疾病如癌症等严重威胁着人类的健康,它的形成和发展通常是由多种基因变异所导致,因此对不同患者采用相同的治疗措施可能产生不同的治疗反应. 目前,肿瘤等复杂疾病通常主要采用两种高通量的数据,一种是基因表达谱数据,另一种是全基因组单核苷酸多态性数据(SNP). 为提高疾病分析的效率,数据挖掘、机器学习等方法被广泛用于复杂疾病分析中. 令人惊讶地是机器学习方法在基因表达谱分析中应用非常广泛,而在 SNP 数据中的应用则较为少见[1]. 研究证实,利用机器学习方法提取基于样本状态密切相关的特征基因,然后以此构建预测模型,其性能优于仅使用先验信息证实的候选位点或者通过显著性检验的候选位点[2-4]. 然而从高维的全基因组 SNP 数据或者下一代测序数据中提取具有预测效果的信号仍然是一种挑战,所面临的实验或者计算难题与传统的基因表达谱数据也有差别[5-6]. 同时,传统的基于单位点的复杂疾病分析方法忽略了位点之间的上位性相互作用,导致关联研究难以解释且预测准确度低等不足. 为了构建更准确可靠的基于全基因组遗传变异的复杂表达预测模型,有必要将位点之间上位性效应融入疾病预测模型中[7-8].

当前已经有一些基于机器学习的上位性分析方法,比如将复杂性状(如血压、身高等)的预测看作是回归问题或者是疾病状态预测问题,根据学习模型的准确率以评价所选择的特征属性(易感位点组合)与疾病状态相关的程度[9-10]. 以较早的机器学习应用为例,Wei 等人[11]在 2009 年利用支持向量机及 L2 范数正则逻辑回归构造了一个具有较高的预测性能的风险模型,并选择 500 个左右的 SNP 用以预测 1 型糖尿病(type 1 diabetes T1D). 与之相比,仅考虑少数已知的 T1D 易感位点的上位性,使得预测性能大大降低. 该方法采用了 5 折交叉验证,实验结果显示该方法 AUC 值在 0.9 左右. 但是,简单的交叉验证使得两个特征选择阶段之间存在信息泄露,从而导致过于乐观的验证结果. 2013 年 Wei 等人再次从 15 个欧洲国家获取更大规模的样本数据集(大于 10,000 个体),然后对克罗恩病(CD)和溃疡性结肠炎(UC)进行风险预测[12]. 该研究利用了定制的具有更高分辨率的 SNP 分型芯片,可以获取常见变异以及第一个阶段 GWAS 研究中所忽略的稀罕变异,然后采用相对宽松的阈值($P < 10^{-4}$)预测选定的 10 000 个左右 SNP,紧接着采用 L1 范数正则化逻辑回归对稀疏遗传风险建模. 该研究的结果与以往一些重要研究结果[13-15]相互印证,从而说明风险预测性能是与样本规模、稀罕变异以及机器学习模型密切相关的.

为了解释身高性状的遗传缺失性,Yang 等人[16]改进了两阶段框架并且采用简单线性回归模型对 294 831 个变异位点进行分析. 该研究采用的是全基因组预测方法,并没有采用任何特征选择方法,实验中对 4 000 个欧洲后裔人群的身高表型变异进行分析,结果表明,所识别的易感位点能解释其中 45%的表型变异. Makowsky[17]等在欧洲人群中进行全基因组预测模型训练,然而在完全独立数据集上采用 10 折交叉验证 R^2 值大幅下降. 这两个研究进一步阐释了缺乏特征选择或者有效模型验证将产生过度拟合现象. 吴蓉晖等[18]利用多种统计准则综合评价每个遗传标记,然后利用蚁群算法搜索上位性组合空间,尽管该方法一定程度提高了疾病分类的准确度,但是它存在两点不足:1)多种准则之间的互补性低,准则之间存在重叠,因此可能导致准则偏差,并且这些准则仅考虑致病等位基因在不同类样本中的分布差异,而忽略了类内的稳定性;2)蚁群算法具有一定随机性,在不同的运行环境下得到一致的病组合可能存在差别,因此降低了算法鲁棒性.

对以上研究中存在的不足,本文提出一种基于二阶段框架的上位性识别方法,它包含特征过滤阶段以及组合优化阶段. 在特征过滤阶段中,本文提出了利用多准则融合策略从多个互补角度更准确地评价每个位点,以避免有效的弱效位点被剔除;在组合优化阶段,为了寻找与疾病状态关联程度最大的多位点上位性组合,本文采用贪婪算法启发式地搜索高阶组合空间. 为验证易感上位性组合的分类准确度,本文采用留一交叉法评估学习模型的性能. 多组数据集上实验结果表明了本文方法的优越性.

1　二阶段框架的上位性分析

1.1　特征过滤阶段

特征过滤阶段的主要目的是删除大量位点中的噪声位点以及冗余位点,以显著减少 SNP 的数量,从而使得后续上位性分析阶段的组合爆炸现象得到一定程度的缓解.传统的过滤准则都是采用某一种度量方法如统计检验法从单个位点在患病对照组中的分布差异进行评价,导致一些真实易感的弱效位点被剔除,从而使得后续上位性分析的准确度不足.本文从多个角度同时考察每个位点,以更准确地保留弱效位点.

1.1.1　多个单位点评价准则

1)互信息准则

信息熵理论被广泛用于 SNP 数据分析中[19-20],熵作为一种平均自信息度量方法,可以定量度量信息源中信息量,可以描述随机变量中不确定性程度.在数据集噪声测量中,熵可以直接测量数据集的冗余程度或者是噪声信息含量,其中冗余程度是指数据集中各个元素之间的相互依赖程度.假设 X 表示一个随机变量空间,可以表示为 $X=\{x_1, x_2, x_3 \cdots x_n\}$,$p(X=x)$ 表示值 x 出现的频率,因此随机变量 X 信息熵可以表示为:

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log\ p(x_i). \tag{1}$$

本文利用互信息公式度量遗传变异与疾病之间的依赖程度,如式(2)所示.

$$I(X,Y) = H(X) + H(Y) - H(X,Y). \tag{2}$$

式中:$H(X)$为变量 X 的熵,从该式看它描述变量取值不同时对于熵的影响,也即不确定性程度.通常某种研究背景下,变量值的分布呈现某种概率分布如正态分布等.信息熵作为一种不确定性度量方法,其值越小则表明随机变量的不确定越小,即确定性越大.X 为不同遗传位点,Y 为疾病状态标签变量;$p(xi)$ 为位点上不同等位基因的分布频率;$H(X)$为位点的熵,因此式(2)表示单个位点与疾病状态之间的关联程度,如果关联程度越大则表明该弱效位点是易感位点可能性越大.

2)频率差异准则

易感遗传位点上不同等位基因与疾病状态之间存在明显的关联,当某个样本携带有易感位点上的致病等位基因时,那么该样本则患病的风险更大.因此,某个等位基因在患病组和对照组中频率分布差异非常显著,那么该等位基因更有可能与疾病相关,那么该位点即为易感位点.频率差异准则如式(3)所示:

$$\mathrm{FS} = \frac{|\ \mathrm{Fcontrol} - \mathrm{Fcase}\ |}{m}. \tag{3}$$

式中:Fcontrol 表示等位基因在对照组中的频率;Fcase 表示在患病组中的频率;m 表示样本的数量.可以发现 $FS \in [0,1]$.

3)类间类内平方和比准则

类间差异表示保证位点在不同类样本中的分布差异最大,而类内一致表示位点上等位基因在同组样本中的分布差异应该保持最小化.当某个遗传位点上等位基因在不同类中的差异越大,并且在相同类的变化越小,那么表明该位点越可能与疾病分类相关,因此打分越高,如式(4)所示:

$$\mathrm{BW}(j) = \frac{\sum_i \sum_k I(y_i = k)\ (\overline{x_{kj}} - \overline{x_j})^2}{\sum_i \sum_k I(y_i = k)\ (x_{ij} - \overline{x_{kj}})^2}. \tag{4}$$

式中:BW(j)为遗传位点 j 的打分;分子和分母分别代表类间差异值与类内差异值;y_i 表示第 i 个样本的疾病状态.

1.1.2　多准则融合策略

每个遗传位点在经过多个准则评价度量后,需要将不同尺度的打分值进行融合,以确定遗传位点的不同性能.本文首先利用不同准则对每个位点进行打分排序,然后将不同准则的打分排名融合,如图 1 所示.

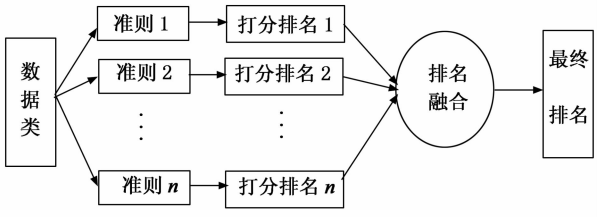

图 1　多准则融合

Fig. 1　Multi-criteria evaluation

该阶段分析中,首先各准则基于各自度量的含义基础上,按每个位点成为致病位点的可能性程度进行排序,比如互信息准则中互信息值越大,则该位点对于疾病状态的影响越大,因此其排名越高.多准则融合过程中,可以对不同准则的排名顺序作加权融合,然后选择总排名靠前的位点构成候选易感位点.由于本文中,对各种准则没有任何先验信息,因此将不同准则看作是等权重的.

1.2 上位性分析阶段

基于单个位点的遗传特征评价方法的计算成本低，因此更适合用于全基因组上的 SNP 数据分析过滤阶段. 同时，该过滤准则不与学习模型相嵌套，因此进一步降低时间复杂度. 该过滤准则的最大弊端在于忽略了位点之间的相互作用即上位性，因此，本文在第二阶段采用学习模型进一步考虑多个位点的相互作用.

1.2.1 基于支持向量机的疾病状态预测

支持向量机(Support vector machine, SVM)是一种基于有监督的学习预测模型，该模型最大优势在于维度无关性，即属性的维度不影响模型的复杂度，因此特别适用于高维的全基因组 SNP 数据. 概括而言，SVM 目标是构造最优超平面保证不同类样本间隔最大化，同时保证泛化误差尽可能小.

将候选的上位性位点作为特征属性组合，利用留一交叉法验证候选上位性组合的疾病预测性能. 留一交叉法将每个样本单独作为一份测试集，其余样本作为训练样本，利用训练样本上特征属性组合训练支持向量机，然后将所预测的测试样本类别号与其真实的类别号进行比较，反复迭代训练、测试 N 次，然后计算平均错误率.

1.2.2 贪婪算法

从候选易感 SNP 集合中选择一个最优的上位性 SNP 组合，使其数量最少并且疾病预测准确度最高，该问题是一个 NP 问题，尤其当候选易感 SNP 集合中 SNP 数量仍然较大时，难以搜索到最优解. 因此，本文采用贪婪策略寻找近优解. 贪婪算法见表 1(又称贪心算法)，首先将待求解的问题划分为若干个子问题，然后分别得到子问题的最优解，最后将所有子问题的解合成原问题的解. 在决策过程中并不是从全局考虑每个阶段的策略，而仅仅从当前情况下选取最有利的策略，可以看出，该算法并不能保证求解的全局最优，但是当解集空间过大时，采用该算法寻找近优解也是一种较好的替代策略.

假设候选易感 SNP 集合 I_0 中含有 s 个 SNP，那么在第 1 次迭代情况下，要从 s 个 SNP 中选择一个位点剔除，找出一个包含有 $s-1$ 个位点的子集 I_1，该子集的疾病预测准确度增加的最多，该次迭代的计算复杂度为 $O(s)$，然后在 I_1 的子集中选择一个位点满足式(5)，依次迭代直到满足退出条件. 该优化表达式如下：

$$\mathrm{Max}:\mathrm{Acct}-1-\mathrm{Acct}. \tag{5}$$

Acct 表示第 t 次迭代过程中所对应的子集具有的最大预测准确度. 贪婪算法的结束条件是，从当前子集中删除任何一个位点，预测准确度都会降低，即子集中已经没有冗余位点.

表 1 基于贪婪算法的上位性分析

Tab. 1 Epistatic analysis based on greedy algorithm

```
基于贪婪算法的上位性分析：
输入：候选易感 SNP 集合 S
输出：最佳易感上位性组合 I
Begin
初始化候选易感 SNP 集合 I 等于集合 S；
  Do
    计算集合 I 的预测准确度 Acc0；
    For i=1:|I|
       从集合 I 中选择一个位点删除，得到 Acci；
    End For
       选择最大的 Acci，并将所对应的集合覆盖集合 I；
  while(Acci> Acc0)
  输出集合 I；
End
```

从算法流程来看，当最终上位性组合中包含的 SNP 数目为 n 时，该贪婪算法的时间复杂度为 $O(n*s*T)$，其中 T 表示预测过程所消耗的时间. 贪婪算法的流程图如图 2 所示.

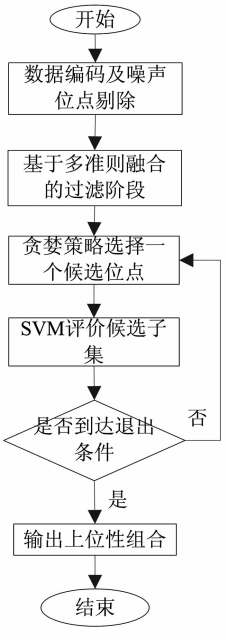

图 2 贪婪算法流程图

Fig. 2 Flow chart of greedy algorithm

2 数据集及评价指标

为验证本文多准则策略的有效性，将利用模拟数据集评价本文方法. 模拟数据集中包含 2 000 个

样本(1 000 个病例样本和 1 000 个对照样本),每个数据集包含 100 个标记(其中 2 个致病标记和 98 个非致病标记).首先,生成 2 个致病标记的基因型分布,要设置 4 个参数:疾病外显率 $p(D)$,边际效应 λ,连锁不平衡 r^2,次等位基因频率 MAF.通过这 4 个参数的值计算得到基因效应 θ 和基线效应 α 的值.再根据基因效应 θ 和基线效应 α 的值生成相应的仿真数据集.本文将这 4 个参数分别设置为 $p(D)=0.1$;$\lambda=0.3$;$r^2=0.5$, 0.7, 1;MAF=0.05, 0.1, 0.2 和 0.5 构造模拟数据集,数据集分别标记为 sim1,sim2 和 sim3.

本文将采用预测准确度指标.如果某个上位性组合是真实的致病上位性组合,那么利用它理论上可以很好地预测个体的患病状态.预测准确度如式(6)所示:

$$\mathrm{Acc}=\frac{\sum_{i=1}^{m}Pi}{m}. \tag{6}$$

3　实验结果及分析

图 3～图 5 比较了 ACO 算法[18]和 SNPRuler[21]这几种方法的预测准确度.

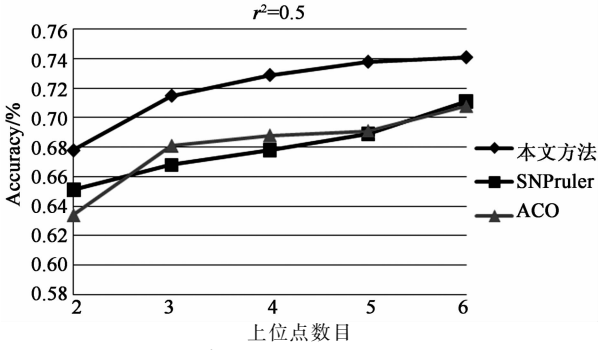

图 3　$r^2=0.5$ 时预测准确度比较

Fig. 3　The prediction accuracy on $r^2=0.5$

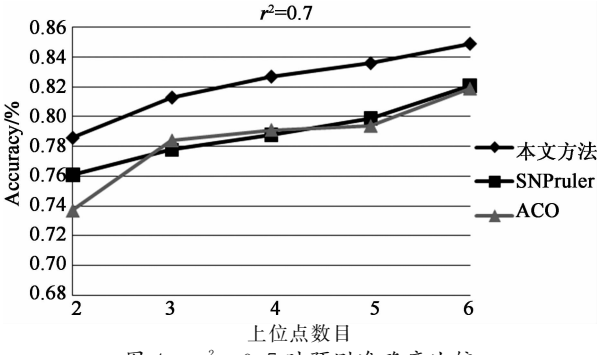

图 4　$r^2=0.7$ 时预测准确度比较

Fig. 4　The prediction accuracy on $r^2=0.7$

从图 3～图 5 可以发现,随着上位性位点数目的增加,本文方法的预测准确度也在增加,这是因为上位性位点增加,能引入更多与疾病分类状态相关的易感位点.同时,由于采用了贪婪策略以保证每轮迭代中预测准确度至少保持不变,因此,以上图中本文方法所对应的曲线基本上是递增的.图中也可以看出,其余两种方法的准确度不断增加,这是因为两种方法都能从候选子集中选择与疾病状态相关的易感位点.

上位性分析的时间复杂度为 C_n^m,其中 n 表示数据集中包含的位点数量,m 则表示上位性的阶数,因此搜索全局最优解的成本过于昂贵.尽管本方法中贪婪算法难以保证搜索到全局最优解,但该方法适用于在较大数据集上搜索高阶上位性,一定程度提高了上位性分析策略的适用性.

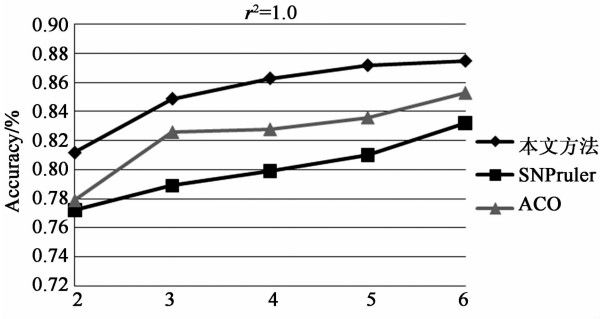

图 5　$r^2=1$ 时预测准确度比较

Fig. 5　The Prediction accuracy on $r^2=1$

从图 3～图 5 中发现,本文方法的预测准确度总体上看来是高于其它两个方法的.这表明本文方法能选出与疾病更为相关的易感上位性组合.其它两种方法 SNPruler 与 ACO 的性能存在一定波动性,在图 3 和图 4 中,后两种方法的准确度比较接近,但是在图 5 中,ACO 方法优于 SNPruler 方法.将图 3,图 4 和图 5 进行比较发现,随着连锁平衡值的增加.3 种方法的准确度都有所增加,以本文方法为例,图 3 中本文方法的评价准确度为 0.72,图 4 的平均准确度为 0.83,而图 5 中的平均准确度为 0.86.进一步分析该现象可以推测,当致病位点与其它标记之间的连锁不平衡性增加,那么间接表明该致病位点更可能被其它标记所描述,因此,致病位点所包含的信息也能被学习模型所利用,从而提高了预测准确度.

4　结　论

传统基于单位点的全基因组关联研究具有计算

简便性,但是过度简化了复杂疾病的致病模型,从而导致研究结果的重复性低.本文首次提出了一种多准则的上位性分析方法,它从多个角度互补地评价了遗传位点,从而避免了真实的弱效位点被剔除.在上位性组合优化阶段,本文采用贪婪算法作为优化策略,尽管该策略仍然不能全局最优性,但是该算法的鲁棒性高,避免了研究结果的随机性,从而更适用于临床应用.

参考文献

[1] KRUPPA J, ZIEGLER A, KONIG I R. Risk estimation and risk prediction using machine-learning methods[J]. Human Genetics, 2012, 131(10): 1639－1654.

[2] PAHIKALA T, OKSER S, Airola A, *et al*. Wrapper-based selection of genetic features in genome-wide association studies through fast matrix operations[J]. Algorithm Mol Biol,2012, 7(1):11.

[3] OKSER S, LEHTIMAKI T, ELO L L, *et al*. Genetic variants and their interactions in the prediction of increased Pre-clinical carotid atherosclerosis[J]. The Cardiovascular Risk in Young Finns Study, PLoS Genet,2010, 6(9):e1001146.

[4] KOOPERBERG C, LEBLANC M, OBENCHAIN V. Risk prediction using genome-wide association studies[J]. Genet Epidemiol, 2010, 34(7):643－652.

[5] KRAFT P, WACHOLDER S, CORNELIS M C, *et al*. Beyond odds ratios: communicating disease risk based on genetic profiles[J]. Nat Rev Genet ,2009,10(4):264－269.

[6] ASHLEY E A, BUTTE A J, WHEELER M T, *et al*. Clinical assessment incorporating a personal genome[J]. Lancet, 2010,375(9725):1525－1535.

[7] MANOLIO T A. Bringing genome-wide association findings into clinical use[J]. Nat Rev Genet, 2013,14(8):549－558.

[8] GIBSON G. Hints of hidden heritability in GWAS[J]. Nat Genet,2010, 42(8):558－560.

[9] YANG J, BENYAMIN B, MCE VOY B P, *et al*. Common SNPs explain a large proportion of the heritability for human height[J]. Nat Genet,2010,42(11):565－569.

[10] MAKOWSKY R, PAJEWSKI N M, KLIMENTIDIS YC, *et al*. Beyond missing heritability: prediction of complex traits [J]. PLoS Genet ,2011,7:e1002051.

[11] WEI Z, WANG K, QU H Q, *et al*. From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes[J]. PLoS Genet, 2009,5: e1000678.

[12] WEI Z, WANG W, BRADFIELD J, *et al*. Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease[J]. Am J Hum Genetics,2013, 92(6):1008－1012.

[13] CHATTERJEE N, WHEELER B, SAMPSONJ, *et al*. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies s[J]. Nat Genet, 2013, 45(4):400－405.

[14] DUDBRIDGE F. Power and predictive accuracy of polygenic risk scores[J]. PLoS Genet ,2013,9: e1003348.

[15] DO C B, HINDS D A, FRANCKE U, *et al*. Comparison of family history and SNPs for predicting risk of complex disease [J]. PLoS Genet,8: e,2012,1002973.

[16] YANG J, BENYAMIN B, MCEVOY B P, *et al*. Common SNPs explain a large proportion of the heritability for human height[J]. Nat Genet, 2010,42(7):565－569.

[17] MAKOWSKY R, PAJEWSKI N M, KLIMENTIDIS Y C, *et al*. Beyond missing heritability: prediction of complex traits [J]. PLoS Genet ,2011,7:e1002051.

[18] 吴蓉晖, 卢友敏. 基于蚁群算法的复杂疾病上位性分析方法[J]. 湖南大学学报:自然科学版, 2014,42(8): 125－131.
WU Rong-hui , LU You-min. An epistasis analysis method of complex diseasesBased on ant colony algorithm[J]. Journal of Hunan University: Natural Sciences, 2014,42(8): 125－131. (In Chinese)

[19] LIU Z, LIN S. Multiocus LD measure and tagging SNP selection with generalized mutual information[J]. Genetic Epidemiology, 2005, 29(4): 353－364.

[20] LI X, LIAO B, ZHU W, *et al*. Informative SNPs selection based on two-locus and multilocus linkage disequilibrium: criteria of max-correlation and min-redundancy[J]. IEEE/ACM Trans Comput Biol Bioinform, 2013, 10(3): 688－695.

[21] XIANG Wan, CAN Yang, QIANG Yang, *et al*. Predictive rule inference for epistatic interaction detection in genome-wide association studies[J]. Bioinformatics, 2010,26 (1):30－37.