第43卷 第6A期
2016年6月

计 算 机 科 学
Computer Science

Vol. 43 No. 6A
June 2016

单核苷酸多态性在疾病相关性分析中的编码问题研究

赵 婧[1] 魏 彬[2] 张 瑾[1]

(西京学院控制工程学院 西安 710123)[1] (武警工程大学电子技术系 西安 710086)[2]

摘 要 作为第三代遗传标记的单核苷酸多态性(SNP)具有数量众多、分布广泛且遗传稳定性等特点,其是疾病-基因相关性以及药物设计等研究的基础所在。这类研究多采用基于计算的方法,因此如何对SNP进行适当的编码进而提升算法的性能是其中十分关键的一个环节,然而目前专门针对SNP编码问题的研究还相对较少。在常用SNP表示方式的基础上,根据疾病易感性研究的特点,并结合SNP之间的关联性,提出了几种新的编码方法。大量实验表明,编码方式对疾病易感性分析算法的性能有着较大的影响,基于分布信息的编码方法能获得更好的结果,即其能更好地对SNP序列进行描述,在最大程度上保留原有生物序列所携带的丰富信息,更适合于疾病易感性研究。

关键词 单核苷酸多态性,编码,疾病易感性

中图法分类号 TP183 **文献标识码** A

Research on Single Nucleotide Polymorphism Encoding in Disease Association Studies

ZHAO Jing[1] WEI Bin[2] ZHANG Jin[1]

(School of Control Engineering, Xijing University, Xi'an 710123, China)[1]

(Department of Electronic Technique, Armed Police Engineering University, Xi'an 710086, China)[2]

Abstract Due to the SNP has some characteristics (such as high abundance and low mutation rate), they are suitable for disease association studies. Lots of those studies were based on calculated methods, so encoding the SNP to enhance the performance of disease associated analysis algorithm was critical aspect. However, few of studies were dedicated to that issue. Therefore, based on common SNP encoding method and association between them, we proposed several new encoding methods. The experiments results show that encoding methods has a greater impact on algorithm performance, and the methods described herein are better than others. Namely, the encoding methods proposed in this paper are better to describe the SNP sequence and retain the original biological sequence information, and are more suitable for disease susceptibility research.

Keywords Single nucleotide polymorphism, Encoding, Disease susceptibility

1 引言

快速发展的基因分型技术使得研究疾病与基因之间的相关关系成为可能,单核苷酸多态性(Single Nucleotide Polymorphism,SNP)由于具有高密度以及高保守性等特点,已经成为疾病相关性问题研究的重要基础[1]。研究表明,大多数SNP位点都是中性的[2],即不会对表现型产生任何影响,然而位于编码区内的SNP可能会改变氨基酸序列,从而影响蛋白质的功能,进而对疾病易感性产生影响[3]。因此,研究疾病与SNP之间的相关关系对于疾病的预防、治疗以及药物的开发等都具有十分重要的意义[4]。

目前针对单核苷酸多态性与疾病之间的关系已经进行了大量的研究[5],并发现了部分与疾病紧密相关的SNP,其中大多数研究都是基于数据挖掘[6]、统计等计算方法进行的[7],而部分算法不能直接对字符型数据进行处理,这就要求在算法初始化阶段对SNP以一定的形式进行编码,以适应算法的需求。因此,编码问题就成为进行分析及各种算法设计的基础所在,有效的编码方式可以使得数据的压缩率接近于信息论的极限,而在同时还能接近完整地保存数据所携带的信息[8]。目前,专门针对SNP编码方式的研究相对来说还比较少,系统性的研究几乎没有,而现在常用的SNP表示方法仅为某种信息的客观描述(例如大多数研究都将SNP编码为类似于0、1和2的形式(0和1分别表示纯合位点中常见的和不常见的等位基因,2表示杂合位点)[9]),其中含有较大的资源浪费和冗余性。此外,随着测序技术的飞速发展,大量低频或稀有变异被发现[10],而这些变异可能与疾病的关联性更为紧密,所以如何在编码的过程中突出这些变异,从计算角度出发,探索特征表示的新途径,并从中发现有价值的信息,将为本领域研究提供重要的参考。

综上所述,本文以疾病易感性分析为模版,并结合SNP之间的关联性对SNP的编码问题进行了初探,提出了几种新的编码方式,并在真实疾病数据集上对其进行了测试,实验结

本文受陕西省教育厅科研计划项目(15JK2187),西京学院科研基金项目(XJ140115),武警工程大学基础研究基金项目(WJY201518)资助。

赵 婧(1983—),女,博士,讲师,主要研究方向为模式识别、生物信息处理,E-mail:zhaojing.83@163.com(通信作者);魏 彬(1982—),男,博士,讲师,主要研究方向为生物信息处理。

果表明不同的编码方式所携带的信息是不可同日而语的，本文所设计的方法在很大程度上提升了疾病易感性分析算法的性能，即其能更好地对 SNP 序列进行描述，在最大程度上保留原有生物序列所携带信息，更适合于疾病易感性研究。

2 疾病易感性分析模型

疾病主要可以分为两个大类：(1)符合孟德尔遗传规律的称为单基因疾病；(2)不符合孟德尔遗传规律的称为复杂疾病(或称多基因疾病)。疾病易感性研究主要关注的是复杂疾病，它是由多个因素相互作用引起的，这类疾病在人类疾病中占 80% 以上，其基因与表型之间没有简单的对应关系[11]。

疾病易感性研究是选择有特定疾病的人群组(case)与未患这种病的对照组(control)，通过分析确定与疾病相关的 SNP 组合，用这些 SNP 构建相应疾病的分析模型，最后对新样本进行评估。这种研究方法是由“果”探“因”，其具体过程可描述如下。

给定一组样本 $G=\{g_1,g_2,\cdots,g_m\}$，每个样本包含有 n 个 SNP 和其疾病状态。$g_i=(g_{i1},g_{i2},\cdots,g_{in},y_i)$，其中 $g_{ij}\in\{0,1,2\}$，$y_i\in\{-1,+1\}$，-1 和 $+1$ 分别用于表示 case 和 control。设 $R=(r_1,r_2,\cdots,r_n)$ 与 n 个 SNP 相对应，研究目的就是从 R 中选择一个子集 R_s，并满足：

$$\max\ p_e=f(R_s)$$

$$\text{s.t. } R_s=\{r_{s_1},r_{s_2},\cdots,r_{s_d}\},d<n$$

该子集的选择可以被看成是一个优化问题，因此如何定义 R_s 的最优性是解该问题的基础。本文定义 p_e 为疾病状态预测精度，所以该优化问题就可以被描述为寻找对于疾病状态有着最高预测精度的 R_s。

3 SNP 编码问题研究

3.1 基于频率的编码方法

许多已有疾病易感性研究都是基于 SNP 在 case 和 control 样本中统计信息差值进行的，结果表明这种差别还是明显存在的。因此，本文基于位置相关统计特征，利用不同位置上各类基因型的使用偏好性，挖掘 SNP 的保守特异性及不同位置上 SNP 间的关联关系。

本文主要用到的频率信息为基因型频率(记为 M)和二联体频率(记为 B)，针对这两类频率特征，每一类又可以用以下 3 种具体的方式将 SNP 映射到新的特征空间：(1)仅利用 case 的频率矩阵进行映射(T-mapping)；(2)利用 case 和 control 数据频率差值矩阵进行映射(D-mapping)；(3)同时利用 case 和 control 数据的频率矩阵进行映射(B-mapping)。两类频率信息在 3 种特征映射方式下，将产生 6 种具体的特征表示法。其中，频率计算方法如下。

假设作为训练数据的样本数为 N(case 或 control)，长度为 L，X_i 表示在位置 i 处的基因型的类型($X_i\in\{0,1,2\}$，$1\leqslant i\leqslant L$)，X_iX_{i+1} 表示位置 i 和 $i+1$ 上的二联体类型($1\leqslant i\leqslant L-1$)，$\#X_i$ 表示在训练数据中的位置 i 处特定类型的碱基出现的次数，$\#X_iX_{i+1}$ 表示训练数据集中位置 i 和 $i+1$ 上特定类型二联体出现的次数。

如果把两类频率矩阵分别记为：$P^M=\{p_i^M(X_i)\}_{3\times L}$，$P^B=\{p_i^B(X_iX_{i+1})\}_{9\times(L-1)}$，则有 $p_i^M(X_i)=\dfrac{\#X_i}{N}(i=1,\cdots,L)$，$p_i^B(X_iX_{i+1})=\dfrac{\#X_iX_{i+1}}{N}(i=1,\cdots,L-1)$。

3.2 基于分布信息的编码方法

基于频率差值的编码方法可能会出现“同一化”编码现象，例如表 1 中第 i 列的基因型 0 与第 j 列的基因型 2 频率差值相同，因此都用 -0.1 进行编码，然而这样编码却无法体现这两种基因型分布的差异性以及其所携带的不同信息，而这样的差异很可能会与疾病有着某种关联性。

表 1 D-mapping 编码示例

基因型	case 频率		control 频率		差值	
	i	j	i	j	i	j
0	0.3	0.6	0.4	0.5	−0.1	0.1
1	0.4	0.2	0.3	0.2	0.1	0
2	0.3	0.2	0.3	0.3	0	−0.1

因此，引入了分布信息(Distribution Information，DI)对此现象进行描述并对上述方法进行校正，具体方法如下：

$$M_{ij}^{*}=\log(p_j(S_{kj}\mid c_t))\times P^{*D} \tag{1}$$

其中，$p_j(S_{kj}\mid c_t)$ 表示 S_{kj} 在 c_t 类中的分布信息，$c_t\in\{case,control\}$。

4 实验结果及分析

4.1 实验数据

本文用表 2 所列的 3 个真实疾病数据集(Crohn's disease (CD)，Autoimmune disorder (AD) 和 Tick-borne encephalitis (TE))对所设计的编码方法的性能进行测试，所有数据由 Brinza 博士提供[12]。其中，CD 取自人类染色体 5q31 区域，该区域内的变异可能对 Crohn's 疾病有一定的影响，该数据集共包含 144 个 case 样本和 243 个 control 样本，每个样本含有 103 个 SNP 位点；AD 从对 DNA 中已被证明和 Autoimmune disorder 相关的 3 个基因(CD28，CTLA4 和 ICONS)分型获得，该数据集共有 384 个 case 样本和 652 个 control 样本，每个样本含有 108 个 SNP 位点；TE 数据集由对 21 个 case 和 54 个 control 分型得到的 41 个 SNP 位点构成。

表 2 实验数据集详细信息

数据集	SNPs 个数	cases 样本个数	controls 样本个数
CD	103	144	243
AD	108	384	652
TE	41	21	54

4.2 评价指标

本文用灵敏度(Sn)、专一度(Sp)和预测精度(Acc)对方法性能进行评估，并用五倍交叉验证对算法进行测试。

$$Sn=\frac{TP}{TP+FN} \tag{2}$$

$$Sp=\frac{TN}{TN+FP} \tag{3}$$

$$Acc=\frac{TP+TN}{TP+FP+FN+TN} \tag{4}$$

其中，TP 为正确预测的 case 样本数目；TN 为正确预测的 control 样本数目；FN 为错误预测的 case 数目；FP 为错误预测的 control 数目。

4.3 实验结果及分析

首先，为了对比传统编码方式(分别用 0、1、2 表示 3 种基因型，记为 MS)与基于频率编码法的区别及哪种方式更能有效地描述 SNP 所提供的丰富信息，利用笔者发表的疾病易感

性分析算法 SI[13] 对几种编码方式在 3 种数据集上的识别性能进行了测试，结果如图 1 所示（其中纵轴表示预测精度，横轴表示不同的编码方法）。

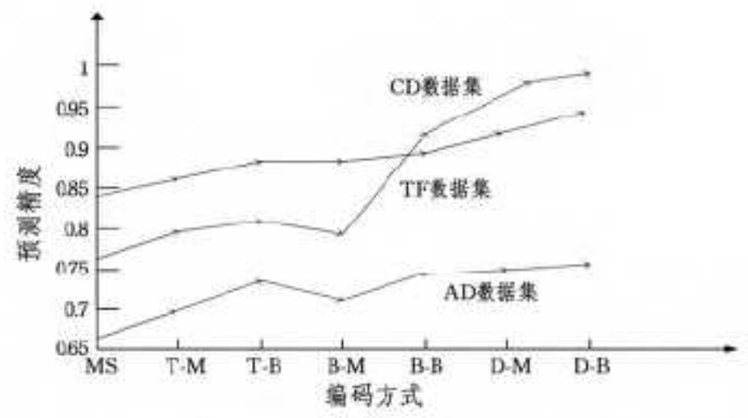


图 1　不同编码方式在 3 种数据集上的性能比较

从图 1 中可以看出，在 3 种实验数据集上，各种编码方式的识别性能有如下规律：(1)基于二联体频率的方法均取得了相对较高的识别精度；(2)基于 D-mapping 的方法相比于其他两种方式来说效果更好。因此我们可以得出以下一些结论：(1)有效地利用不同类别的数据信息可以进一步提高疾病预测精度，且实验结果表明相比于其他方法，采用 D-mapping 方式可在两类特征差异显著的条件下取得更好的识别效果，这是由于该方式考虑到了两类数据频率特征的差异性，频率差值绝对值的大小能直观地反映该 SNP 的出现所提供的类别信息的大小，正负号则表明了在其出现情况下的类别倾向性；(2)二联体频率特征能够取得较好的识别精度，表明 SNP 之间存在着一定关联关系；(3)不同数据集上，各种编码方法所表现出来的差别并不完全一致，这表明各种疾病与 SNP 间的关联关系存在着差异性，即致病机理存在较大差异。

其次，由于在上面的实验中基于二联体频率特征的 D-mapping 方式取得了最好的效果，因此将基于分布信息的编码方法在此基础上进行进一步的实验，结果如表 3 所列。AD 由于致病机理上的复杂性原因，Sn 出现了一定程度的下降而 Sp 则大幅上升，但总体评价指标 Acc 则在 D-mapping-B-DI 方式下得到了提升，即其效果更加好。因此，可以得出结论，带分布信息的编码方式取得了更好的结果，这是由于其不仅保留了原有方式的优点，而且因为对数操作的引入更加深了对于稀有元素的权重，这也符合之前大多数研究的结果，即稀有元素更有可能对疾病的发生造成影响。所以说基于 DI 编码方式较其他方法更为合理，更能准确地反映 case 和 control 样本间的差别。

表 3　带分布信息编码方式的测试结果

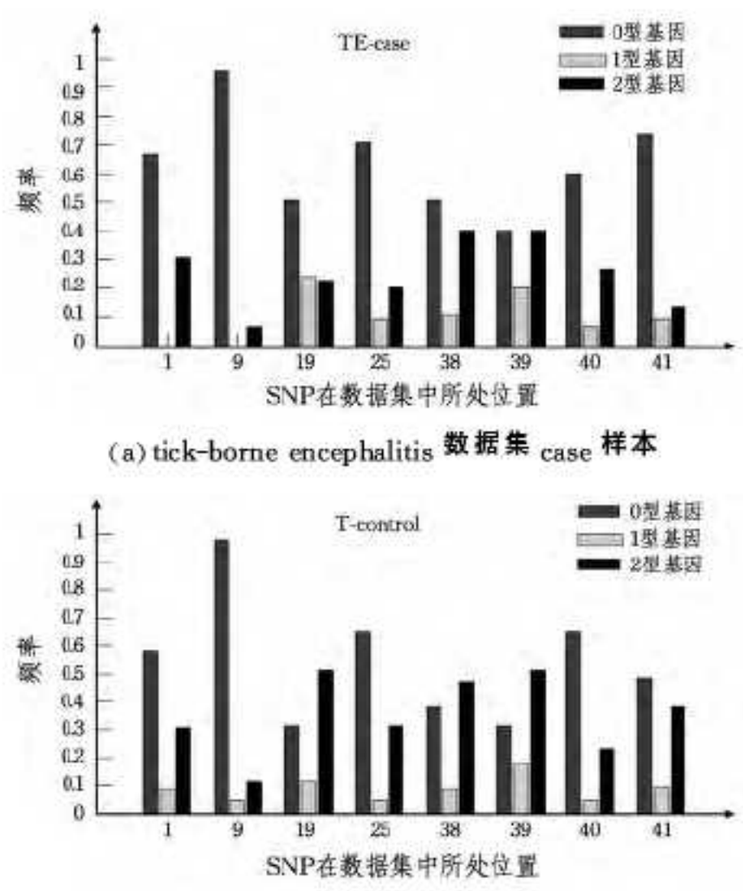
数据集	Encoding	Sn	Sp	Acc
AD	D-mapping-B	0.6854	0.7926	0.7530
	D-mapping-B-DI	0.5449	0.9304	0.7881
CD	D-mapping-B	0.8961	0.9717	0.9509
	D-mapping-B-DI	0.8963	0.9876	0.9535
TE	D-mapping-B	0.9500	1.0000	0.9857
	D-mapping-B-DI	1.0000	1.0000	1.0000

再次，为了进一步说明编码方法的重要性、本文方法的性能以及不同编码方法之间性能的区别，我们在 4 种已发表算法（USVM[14]，MultiBLUP[15]，BPSO-EDA[16]，HCAF[17]）上进行了测试，结果见表 4。从表中结果可以看出，尽管 HCAF 所取得的结果并不十分理想，但是两种编码方式还是表现出了明显的不同。因此，可以得出结论，不论在哪种算法下，本文设计的编码方法所取得的结果都要明显优于传统编码法，即文中编码方式与具体分析算法无关。

表 4　不同编码方式预测精度比较

数据集	Encoding	USVM	MultiBLUP	BPSO-EDA	HCAF
AD	MS	0.6732	0.6365	0.6821	0.6309
	D-mapping-B-DI	0.7821	0.7732	0.7698	0.7527
CD	MS	0.8532	0.8643	0.8598	0.7943
	D-mapping-B-DI	0.9743	0.9787	0.9678	0.9587
TE	MS	0.7862	0.7743	0.7784	0.6943
	D-mapping-B-DI	0.9926	1.0000	1.0000	0.9788

最后，给出了本文编码方法结合 SI 算法所选子集（子集的大小也从侧面反映了疾病的复杂度）中基因型在两类样本中的分布（如图 2 所示，横轴为 SNP 在数据集中所处位置，纵轴表示频率）。从图中可以看出，被选中的 SNP 在两类样本中的分布有着明显的区别，这种差别在 TE 数据集上尤为明显，例如基因型 1 在 case 样本的 1 和 9 号位置上是不存在的，而 control 样本却含有少量的基因型 1，说明该位置上的基因型 1 起保护作用。



(a) tick-borne encephalitis 数据集 case 样本

(b) tick-borne encephalitis 数据集 control 样本

图 2　基因型在两类样本中的分布

结束语　分析与疾病密切相关的 SNP 对于人类对疾病的认识和治疗有着十分深远的影响，除了对算法模型的设计外，SNP 的表示方式（即编码方法）对分析算法的性能也有着较大的影响，然而相比于对算法模型设计的研究，针对编码方式的研究相对较少。本文基于疾病易感性分析算法的特点以及 SNP 间的相关关系等，设计了 6 种基于频率的编码方式以及一种带分布信息的编码法，最后通过系统的实验测试对几种方法的性能和特点进行了详细的对比分析，证明了编码方法研究的重要性以及文中所设计方法的有效性。实验结果一方面使我们更加清楚地认识了各种映射方式的特点和性能，另外部分结果也为认识 SNP 致病机理提供了一些线索。

参 考 文 献

[1] Zhang Han, Shi Jian-xin, Liang Fa-ming, et al. A fast multilocus test with adaptive SNP selection for large-scale genetic-association studies[J]. European Journal of Human Genetics, 2014, 22(5): 696-702

（下转第 235 页）

[25] 刘俊涛,刘文予,吴彩华,等. 一种提取物体线形骨架的新方法[J]. 自动化学报,2008,34(6):617-622
[26] 瞿鑫,丁天怀. 皮棉中异性纤维骨架快速提取算法[J]. 农业机械学报,2010,41(6):177-181
[27] 陈雪松,王乘,徐学军,等. 一种运用势能理论的骨架特征提取方法[J]. 小型微型计算机系统,2011,32(1):151-155
[28] 贾挺猛,荀一,鲍官军,等. 基于机器视觉的葡萄树枝骨架提取算法研究[J]. 机电工程,2013,30(4):501-504
[29] 王松伟,李言俊,张科,等. 一种快速的目标骨架提取算法[J]. 红外与激光工程,2009,38(4):731-736
[30] Sarkar N,Chaudhuri B B. An efficient differential box-counting approach to compute fractal dimension of image[J]. IEEE Transactions on Systems Man and Cybernetics,1994,24(1):115-120
[31] 赵巨波,孙华燕,杜巍. 一种图像边缘特征提取算法[J]. 光学精密工程,2000,8(4):325-327
[32] 张国立,杨瑾,李晶,等. 基于小波包和数学形态学结合的图像特征提取方法[J]. 仪器仪表学报,2010,31(10):2285-2290
[33] 陈雪松,徐学军. 一种二值图像特征提取的新理论[J]. 计算机工程与科学,2011,33(6):31-36
[34] 陈桂芬,曾广伟,陈航,等. 基于纹理特征和神经网络算法的遥感影像分类方法研究[J]. 中国农机化学报,2014,35(1):270-274
[35] Sun Zheng, Yan Qi. Motion estimation of 3D coronary vessel skeletons from X-ray angiographic sequences[J]. Computerized Medical Imaging and Graphics,2011,35:353-364
[36] Wang Sen, Wu Jianhuang, Wei Mingqiang, et al. Robust curve skeleton extraction for vascular structures[J]. Graphical Models,2012,74:109-120
[37] 吴健,崔志明,徐倩,等. 基于 Level Set 模型的彩色脑血管图像骨架提取算法[J]. 计算机科学,2009,36(2):278-281
[38] 陈国栋,李建微,潘林,等. 基于人体特征三维人体模型的骨架提取算法[J]. 计算机科学,2009,36(7):295-297
[39] 耿欢,杨金柱,赵大哲,等. 一种面向人体管状组织的三维骨架提取算法[J]. 仪器仪表学报,2014,35(4):754-761
[40] 朱文博,李彬,田联房,等. 新型基于分层多假设跟踪的冠脉骨架提取算法[J]. 自动化学报,2014,40(8):1783-1792
[41] 张曼婷,闫文耀,王庆. 基于图像处理的车牌定位及字符识别算法研究[J]. 现代雷达,2014,36(8):26-40
[42] 徐辉. 基于 MATLAB 实现汽车车牌自动识别系统[J]. 电脑知识与技术,2010,6(17):4752-4754
[43] 张广清. 基于图像识别技术的运输车辆识别与定位方法及实现[J]. 冶金自动化,2014,38(2):28-33
[44] 任其亮. 复杂背景下运动车辆车标定位与识别方法[J]. 数学的实践与认识,2009,39(2):57-63
[45] 宁彬. 图像处理技术在机动车车牌自动识别技术中的应用[J]. 科学技术与工程,2013,13(2):366-371
[46] 廉宁,徐艳蕾. 基于数学形态学和颜色特征的车牌定位方法[J]. 图学学报,2014,35(5):774-779
[47] 厉荣宣,沈希忠,张树行,等. 基于图像处理的轴类零件表面裂纹检测[J]. 图学学报,2015,36(1):62-67
[48] 韩丽,楚秉智,高小山. 高斯曲率约束的 MRG 骨架提取优化算法[J]. 计算机辅助设计与图形学学报,2009,21(9):1227-1231
[49] 马艳娥,张波涛,高磊,等. 基于图像处理的零件尺寸测量研究[J]. 电子测试,2011,(8):39-41,95
[50] 王玉槐,王琦晖,寿周翔,等. 应用改进 Canny 法检测工业零件含噪图像边缘[J]. 轻工机械,2012,30(4):77-80

(上接第 221 页)

[2] Liu Xin-yu,Wang Yu-peng,Sriram T N. Determination of sample size for a multi-class classifier based on single-nucleotide polymorphisms:a volume under the surface approach[J]. Bmc Bioinformatics,2014,15:190-198
[3] Vogler C,Gschwind L,Coynel D,et al. Substantial SNP-based heritability estimates for working memory performance[J]. Translational Psychiatry,2014,4(9):438-438
[4] Schierding W,Cutfield W S,O'Sullivan J M. The missing story behind Genome Wide Association Studies:single nucleotide polymorphisms in gene deserts have a story to tell[J]. Frontiers in Genetics,2014,5:39
[5] Wei Bin,Peng Qin-ke,Kang Xue-jiao. A Hybrid Feature Selection Algorithm used in Disease Association Study[C]// the 8th World Congress on Intelligent Control and Automation. 2010,5:2931-2935
[6] Talluri R, Wang Jian, Shete S. Calculation of exact p-values when SNPs are tested using multiple genetic models[J]. Bmc Genetics,2014,15:75
[7] Roshyara N R,Kirsten H,Horn K,et al. Impact of pre-imputation SNP-filtering on genotype imputation results[J]. Bmc Genetics,2014,15:88-99
[8] Richardson A M, Lidbury B A. Infection status outcome, machine learning method and virus type interact to affect the optimised prediction of hepatitis virus immunoassay results from routine pathology laboratory assays in unbalanced data[J]. Bmc Bioinformatics,2013,14:206-221
[9] Duan L,Thomas D C. A Bayesian Hierarchical Model for Relating Multiple SNPs within Multiple Genes to Disease Risk[J]. International Journal of Genomics,2013,15:406217
[10] 周家蓬,裴智勇,陈禹保,等. 基于高通量测序的全基因组关联研究策略[J]. 遗传,2014,10(5):1-22
[11] Thieme S,Groth P. Genome Fusion Detection:a novel method to detect fusion genes from SNP-array data[J]. Bioinformatics,2013,29(6):671-677
[12] Brinza D, Zelikovsky A. Design and validation of methods searching for risk factors in genotype case-control studies[J]. Journal of Computational Biology,2008,15(1):81-90
[13] Wei Bin,Peng Qin-ke,Zhang Quan-wei. Identification of Combination of SNPs Associated with Graves' Disease using Swarm Intelligence[J]. Science China Life Sciences,2011,2(2):139-145
[14] Wei Bin,Peng Qin-ke,Li Jing, et al. USVM:Selection of SNPs in diseases association study using UMDA and SVM[C]// 4th International Conference on Bioinformatics and Biomedical Engineering. 2010
[15] Speed D,Balding D J. MultiBLUP:improved SNP-based prediction for complex traits[J]. Genome Research,2014,15,24(9):1550-1557
[16] Wei Bin,Peng Qin-ke,Li Chen-yao. A Hybrid of Binary Particle Swarm Optimization and Estimation Distribution Algorithm for Feature Selection[C]//6th International Conference on Natural Computation. 2010:2510-2514
[17] Sluga D,Curk T,Zupan B,et al. Heterogeneous computing architecture for fast detection of SNP-SNP interactions[J]. Bmc Bioinformatics,2014,15:216-222