

引用格式:曹一鸣,胡曼,徐永利. 基于蚁群算法的基因位点组合与眼压数据的相关性分析[J]. 北京化工大学学报(自然科学版), 2020, 47(4): 94–100.

CAO YiMing, HU Man, XU YongLi. Correlation analysis between gene locus combinations and intraocular pressure data based on ant colony algorithms[J]. Journal of Beijing University of Chemical Technology (Natural Science), 2020, 47(4): 94–100.

基于蚁群算法的基因位点组合与眼压数据的相关性分析

曹一鸣¹ 胡 曼² 徐永利^{1*}

(1. 北京化工大学 数理学院, 北京 100029; 2. 北京儿童医院 眼科, 北京 100045)

摘 要:设计了一种蚁群算法,用于进行基因多位点组合与连续型表型的关联分析。在欧洲生物信息学研究所发布的一项基因及眼压的数据集上,评估了本文设计方法的性能,实验结果表明,利用本文设计的蚁群算法能够发现与眼压显著相关的基因双位点组合,这为研究青光眼的发病机制提供了新的线索。另外,对于研究其他疾病的多位点组合和连续型表型的关联,本文设计的蚁群算法提供了一种新的思路。

关键词:基因与性状回归分析; 蚁群算法; 基因间相互作用; 青光眼

中图分类号: O29 **DOI:** 10.13543/j.bhxbzr.2020.04.014

引 言

基因的多位点组合与表型的关联分析是生物信息学的研究热点,近年来的研究表明,在对复杂性遗传疾病的分析上,相比于单基因位点,多基因的多位点组合与表型的关联更为显著。然而,现有的方法仅适用于分析多位点组合与离散型表型的关联,不能分析其与连续型表型的关联。

原发性青光眼属于复杂遗传病^[1],是全球第二位的不可逆性致盲眼病^[2–3],在我国,原发性青光眼(不包括先天性青光眼)的致盲率约为 10%^[4]。原发性开角型青光眼(primary open-angle glaucoma, POAG)的单眼致盲率在北京市农村和市区分别为 15.4% 和 10.90%^[5]。在我国,患青光眼的人群中原发性以及先天性青光眼患者的比例占 80% 以上^[6]。针对青光眼进行基因层面的进一步研究有助于对该疾病的早期诊断。病理性眼压升高是大部

分青光眼患者都会出现的临床症状,在一项大型多队列研究中,发现眼压(intraocular pressure, IOP)与原发性开角型青光眼之间存在着强烈的关联^[7]。因此研究与眼压相关的变异有助于发现更多影响青光眼发病的基因。

测序技术的进步使研究人员可对大规模样本数据进行测序,同时可纵向比较一大批样本的基因组。全基因组关联分析(genome-wide association study, GWAS)是一种在全基因组层面上的无假设分析(即关于特定基因或基因座位无现有的假设,但也不存在这些基因之间关联的零假设),其对跨基因组和疾病的标记(单核苷酸多态性,即 SNP)之间的关联进行测试,通常涉及 300 000 个或更多合理多态且在基因组中平均分布的标记^[8]。近年来, GWAS 试图进一步研究多个 SNP 对个体表型差异的综合作用,从而衡量它们对疾病发病的影响大小。

传统 GWAS 一般通过对 Case-Control 数据的分析来筛选出能够影响复杂遗传病的 SNP 位点。2012 年,一项基于单位点对青光眼作用的 GWAS 研究^[9]给出了结果,即经过 T 检验后具有较高显著性的两个单位点:位于 17 号染色体基因 GAS7

收稿日期: 2019-12-09

基金项目: 国家自然科学基金(11571031)

第一作者: 男, 1995 年生, 硕士生

* 通信联系人

E-mail: xuyi@mail.buct.edu.cn

上的 rs11656696 和位于 1 号染色体基因 TMC01 上的 rs7555523。然而对于复杂遗传病,它在基因层面的致病机理不能仅通过几个 SNP 位点单独的作用得到解释,考虑多个位点的共同作用和上位性相互作用等一系列基因间的相互作用关系非常必要^[10]。

近年来,一些学者通过对 Case-Control 数据的分析,在多个 SNP 位点组合与疾病关联性的研究方面已经取得了一些成果。Sapin 等^[10]利用蚁群算法对 I、II 型糖尿病,肠炎和类风湿性关节炎 4 种疾病进行了研究,在 4 个数据集上均发现了对疾病影响统计显著性较高的双 SNP 位点组合。Liu 等^[11]利用 Hi-seeker 模型在对乳腺癌和小肠吸收不良症的 Case-Control 数据研究中,发现了对疾病影响统计显著性较高的三 SNP 位点组合。以上两种方法都是针对 Case-Control 数据(每个病例的状态是患病或健康)的分析,而与 POAG 之间存在强烈关联的眼压数据是连续变化型数据,每个病例的眼压值可以是确定范围内的任意实数,而不是有限的几个状态。据我们所知,针对连续型数据和 SNP 位点的关联性分析目前还没有相关研究。

本文设计了一种蚁群算法,用来从数以亿计(约 1.767×10^{11} 对)的双位点组合里筛选与眼压最为相关的组合。利用本文设计的蚁群算法,对欧洲生物信息学研究所(European Bioinformatics Institute, EBI)所辖数据库中一项记录了样本人群眼压和基因的原发性青光眼数据进行分析,并从中发现了一些具有较高统计显著性的双位点组合;特别是在 2 号染色体上发现了多个 SNP 位点与同一个基因具有较高的显著性关联,这意味着在此数据集上可能存在一些特定位点,通过上位性相互作用影响其他基因的表达式。

1 算法部分

1.1 蚁群算法

蚁群算法是一种启发式算法,能够在节省计算资源的前提下从大量数据中发掘出有效信息。相比于遍历方法的计算复杂度 $O(n^2)$ (n 为总位点数),蚁群算法的计算复杂度为 $O(m)$,其中

$$m = \text{gen}_{\max} \times n_{\text{ant}} \quad (1)$$

式中, gen_{\max} 表示最大迭代次数, n_{ant} 表示蚂蚁个数。相比于遍历计算平方阶的时间复杂度 $O(n^2)$,

蚁群算法的计算复杂度为线性阶 $O(m)$ 。

本文借鉴了 Sapin 等^[10]用来分析 Case-Control 数据的算法,设计出一种通过研究连续型变量和双 SNP 位点组合之间的关联,来寻找可能具有致病性双位点组合的蚁群算法。通常的蚁群算法是一种探索优化路径的概率型算法,蚂蚁个数表示每一次算法迭代中可行解个数的参数。蚂蚁从随机起点开始,移动到与当前节点有连接的下一个节点上。在选择下个节点的过程中,蚂蚁会根据满足条件的节点上信息素值的大小,对较大信息素值有偏向性地随机决定下一步,某节点信息素值越高则蚂蚁选择该节点的概率越高。节点被选择后,蚂蚁在选择的节点上按一定规则留下一一定量的信息素。在本文设计的算法中,信息素用来刻画每个 SNP 位点对于预测眼压的重要程度。

提前为算法设置初始信息素值、蚂蚁个数、竞赛规模、禁忌频率这 4 个参数。另外,由于没有明确终止条件,需要设定最大迭代次数 gen_{\max} 。设置参数后进入算法的循环部分。在算法的每一次循环中,首先采用竞赛选择为每一只蚂蚁找到一个初始的双位点组合,即让每只蚂蚁随机且具有偏向性地从数据集子集中选择 SNP 组合;然后评估每只蚂蚁携带的双位点组合对受试者眼压的划分情况,并计算适应度函数;循环的最后,对本次循环最优的双位点组合进行信息素更新,相应信息素值 $P(\text{SNP}_1) + P(\text{SNP}_2)$ 叠加在双位点组合的每个 SNP 上;最后,在所有 SNP 位点上乘一个小于 1 的常数 r_{loss} 作为信息素值的损失率,以此来蒸发一定量的信息素值,本文中使用了 0.99 这个乘数值。

又考虑到某些位点的信息素值可能过高,从而频繁出现在迭代中,因此本文算法结合了禁忌表每隔一定的子循环数(1000 次)禁忌信息素最高的一个位点。从生物学意义上来说,由于基因间相互作用的影响,两个单体作用都不大的 SNP 位点也可能组成对青光眼发病影响很大的位点组合,这样的组合较难被发掘,未被解释的青光眼遗传机制极有可能在这些位点组合中。结合禁忌表可以找到更多诸如此类对青光眼发病综合作用大、而组合中单个位点影响都很小的双位点组合。

传统的蚁群算法使用轮盘赌法进行路径选择,轮盘赌法又称为比例选择法,其基本思想是各位点被选中的概率与其信息素大小成正比。该方法能选

择既随机又偏向于具有最大信息素的路径。由于本文基因数据包含数量巨大的 SNP(单位点个数接近 600 000 个),当变量数很高时,直接使用轮盘赌法对计算量消耗会成倍增长。因此,本文在轮盘赌法之前加上竞赛选择步骤。竞赛选择即先随机选择一个指定大小的位点子集,再在这个子集上用轮盘赌法选择 2 个 SNP。

1.2 算法流程

基于 1.1 节分析设计的算法,设计伪代码如下。

- 1) 初始化参数(蚂蚁个数 n_{ant} 和最大迭代次数 gen_{max}),将每个 SNP 位点上的信息素初始化为 0;
- 2) repeat 以下步骤:
- 3) for 所有蚂蚁 do:
- 4) 通过竞赛选择为每一只蚂蚁找到两个 SNP 位点的双位点组合;
- 5) 计算每一只蚂蚁的双位点组合的适合度函数值;
- 6) end for;
- 7) 更新具有最高适应度函数值的两个 SNP 位点的信息素;
- 8) for 所有 SNP 位点 do:
- 9) 乘以损失率 r_{loss} 来蒸发一定量的信息素;
- 10) end for;
- 11) 每隔一定迭代次数禁忌信息素最高的一个位点;
- 12) until 最大迭代次数后终止循环。

上面描述的算法伪代码中,主要包括 3 个步骤:

步骤 1 初始化参数(设置蚂蚁个数,最大迭代次数,竞赛规模和禁忌频率),将每个 SNP 上的信息素初始化为 0,然后对步骤 2 迭代 gen_{max} 次;

步骤 2 通过竞赛选择为每一只蚂蚁找到包含两个 SNP 位点的组合,计算每一只蚂蚁位点组合的适合度函数,更新具有最高适应度函数值的两个 SNP 的信息素,最后对所有 SNP 位点的信息素值乘以损失率 r_{loss} 来蒸发一定量的信息素;

步骤 3 每隔一定迭代次数(本文为每 1 000 次)禁忌信息素最高的一个 SNP 位点,当迭代次数达到最大迭代次数 gen_{max} 后,终止算法循环。

1.3 基因型数据的数字表示,双位点组合的逻辑关系及适应度函数的计算

在 1.2 节所述的算法流程中,适应度函数用来刻画基于某一双位点组合状态将全体病例分组的组

间眼压差异。双位点组合把所有对象划分为阳性组(即该组眼压各项统计参数偏高)和阴性组,这两组的眼压平均值之差越大表明这个双位点组合对眼压预测的意义越显著。

对于确定的每个双位点组合都有 9 种基因型及 4 种布尔逻辑运算关系。每一个 SNP 位点的碱基构成可以分为 3 类,即常见纯合子(AA)、罕见纯合子(aa)以及杂合子(Aa),故两个位点的组合有 9 种可能的基因型。考虑了两个位点之间的 4 种布尔逻辑运算^[10],即:

AND 当且仅当第一个 SNP 取特定值并且第二个 SNP 取特定值时,个体是阳性的;

OR 当且仅当第一个 SNP 或第二个 SNP 取特定值时,个体是阳性的;

AND NOT 当且仅当第一个 SNP 取特定值,且第二个 SNP 不取特定值时,个体是阳性的;

XOR 当且仅当两个 SNP 中的一个具有特定值而另一个 SNP 不采用特定值时,个体是阳性的。

当给定了基因型组合和逻辑关系后,可以把样本眼压划分为两组,从而计算适应度函数值。9 种基因型组成和 4 种逻辑关系形成 36 种组合方式,故可得到 36 个适应度函数值,取这 36 个值中的最大值作为该双位点组合最终的适应度函数值。本文算法使用的适应度函数为

$$F = \frac{(P_p - P_m)^2 M_p + (P_n - P_m)^2 M_n}{M_p + M_n} \quad (2)$$

式中, M_p 表示阳性一组的人数; P_p 表示阳性一组的眼压平均值; M_n 表示阴性一组的人数; P_n 表示阴性一组的眼压平均值; P_m 表示所有受测试者的眼压平均值。

2 实验部分

2.1 实验数据及处理

利用本文的蚁群算法分析了来自 EBI 的澳洲人群青光眼患病情况调查数据。该数据包含 2 627 个合格受测试者样本(未测到眼压值的不合格样本有 134 个),在 594 398 个 SNP 位点上的基因数据和眼压值。基因数据按其所在不同染色体位置划分为 26 个子数据(1~22、XY、X、Y、MT,XY 表示 X 染色体和 Y 染色体的重叠部分,MT 表示线粒体),数据表每行记录一个 SNP 位点的信息,前 5 列记录该位点编号、所在位置和碱基成分,之后每 3 列对应一个

受测试者是显性纯合子、杂合子或隐性纯合子的概率。最终的输入数据是一个 $594\,398 \times 2\,627$ 的基因数据矩阵和一个 $2\,627 \times 1$ 的眼压数据矩阵。

2.2 蚁群算法的参数设置

首先用包含 8713 个位点的 21 号染色体进行预实验,来选取算法的核心参数(蚂蚁数)。计算 21 号染色体上任意双位点组合的适应度函数值,并按该值降序排列这些组合,前 750 个组合具有统计显著性。接下来,在其他参数相同的条件下只改变蚂蚁数,把计算结果中属于这 750 个位点对的个数作为评价算法计算效率的标准。在预实验中,将蚂蚁数分别设置为 10、20 和 30,竞赛规模都设置为 20,迭代次数分别设置为 300 000、150 000、100 000 次(确保算法总的计算次数相同)。图 1 展示了在不同蚂蚁数下,随着计算次数增加蚁群算法在预实验子集上能搜索到的前 750 位组合个数。

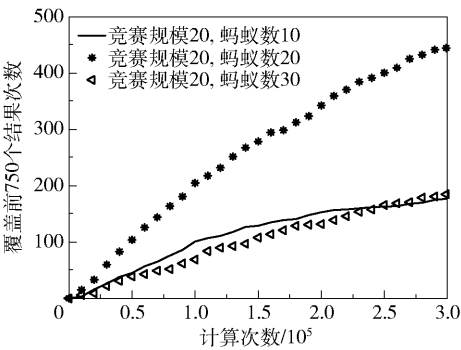


图 1 覆盖前 750 个组合次数随迭代次数增加的变化
Fig. 1 Changes of coverage of the best 750 results as the number of iterations increases

从图 1 中可以看到,蚂蚁数目设置为 20 时,蚁群算法搜索到的关键位点对个数最多。所以在接下来的全位点实验中将蚂蚁数恒定地取为 20,同时将竞赛规模设置为 20。另外,每 1 000 次迭代禁忌一次位点。

2.3 全部位点蚁群实验结果

在全位点实验中使用的计算机包含参数为 Intel i7-i7-6850K 的 CPU,2 400 MHz 的 32G 内存以及 Ubuntu 14.04 的操作系统。根据计算机性能将算法的最大迭代次数设置为 800 000 次。完成本文设计的蚁群算法的 800 000 次迭代总共用时 220 h。表 1 列出了秩和检验 P 值最小的 10 个双位点组合。在所有结果中,有 6 对双位点组合的秩和检验 P 值小于 1×10^{-4} ,秩和检验 P 值小于 1×10^{-2} 的一共有 137 对结果。

表 1 秩和检验 P 值最优的前十组双位点组合

Table 1 Top ten optimal combinations of SNPs of rank sum test P -value

排序	双位点组成	阳性个体		P 值
		总人数	眼压 \geq 2.793 kPa	
I	rs661221 & rs2654703	11	6 (54.5%)	1.84×10^{-5}
II	rs11930632 & rs11702020	8	5 (62.5%)	2.01×10^{-5}
III	rs4876968 & rs10823835	10	6 (60.0%)	2.16×10^{-5}
IV	rs1030891 & rs2822784	7	4 (57.1%)	4.67×10^{-5}
V	rs10265980 & rs17203725	6	5 (83.3%)	7.47×10^{-5}
VI	rs6716551 & rs7201630	6	4 (66.7%)	9.16×10^{-5}
VII	rs2541076 & rs2278997	8	6 (75.0%)	1.05×10^{-4}
VIII	rs7703593 & rs7993513	10	4 (40.0%)	1.23×10^{-4}
IX	rs12126125 & rs8035957	5	4 (80.0%)	3.67×10^{-4}
X	rs969910 & rs7993513	5	3 (60.0%)	4.05×10^{-4}

括号内数字为满足条件的人数占总人数的比例。

在之前关于青光眼数据的眼压与单位点研究中,得到与该疾病显著相关的一个单位点是位于 17 号染色体基因 GAS7 上的 rs11656696^[9],在本文所分析的 EBI 数据集上该位点的秩和检验 P 值为 5.03×10^{-2} 。本文所发现的双位点组合对于眼压预测的秩和检验 P 值低于该单位点的秩和检验 P 值,也即我们发现的双位点组合对眼压的影响更具有统计显著性。

本文实验中,秩和检验 P 值最小的组合是位点 rs661221 与 rs2654703,该组合的秩和检验 P 值为 1.84×10^{-5} ,且这个位点对划分得到的阳性组和阴性组的眼压样本均满足正态分布。

进一步计算了位点 rs661221 与位点 rs2654703 的 T 检验 P 值,为 6.9×10^{-12} ,显著优于文献[9]。位点 rs661221 位于 1 号染色体的基因 RIMS3 上,位点 rs2654703 位于 3 号染色体的基因 FGF12 上。另外,从表 1 可以看到, P 值最小的 10 组双位点组合划分出的阳性组中,眼压高于 2.793 kPa 的病例比例明显高于普通人群中的比例,而眼压高于 2.793 kPa 在临床上是青光眼发病的重要指征。这表明本文算法发现的双位点组合与青光眼发病因素显著相关。

图 2(a)展示了秩和检验 P 值小于 1×10^{-2} 的结果在染色体上分布的热图,图中色块的灰度深浅代表位点对的多少。横纵坐标为位点所在染色体位置编号(1~21),组成位点分别位于横纵坐标标示

的染色体上。图 2(b)是将图 2(a)的内容反映在点和边上的分布图,图中节点度数代表两位点在同一染色体的结果个数,有 2 对双位点组合的组成位点都在 2 号染色体上;3 号和 11 号染色体各有 1 对双位点组合的组成位点都在同一染色体上;其余染色体不包含双位点组合的组成位点都位于同一染色体的情况。图 2(b)中连接各节点的边的粗细代表组成双位点组合的各位点分别位于两不同染色体的组合个数。如图 2(b)所示,最粗的一条边(染色体之间双位点组合最多)在 1 号染色体和 3 号染色体之间,包含 8 对双位点组合;其次为 2 号染色体和 6 号染色体之间的边,包含 6 对双位点组合。

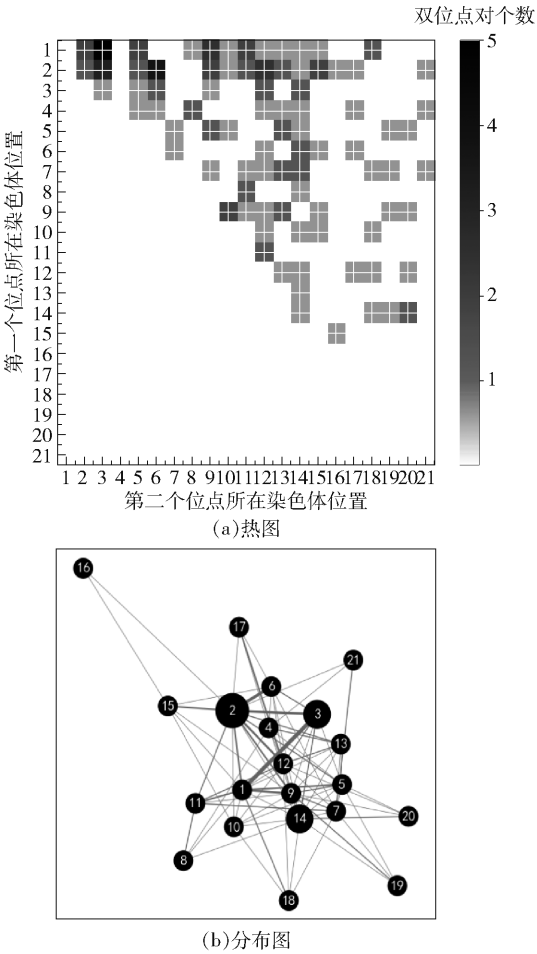


图2 137 个较显著结果在染色体上分布的热图和分布图

Fig. 2 Heat map and the distribution graph of 137 significant results on chromosomes

图 2 表明,137 对显著双位点组合涉及 2 号染色体的个数最多,因此对涉及到 2 号染色体位点的所有结果进行了分析,发现 2 号染色体基因 LOC107985962 上的位点与其他 11 个染色体上的位

点组成了显著性较高的组合。其中包括位点 rs7742590,该位点与基因 FOXC1 都位于 6 号染色体,且距离基因 FOXC1 大约 1 600 万个碱基。据文献[12]报道,6 号染色体上基因 FOXC1 被怀疑与 POAG 发病高度相关。

如表 2 所示,11 个双位点组合都包含 2 号染色体基因 LOC107985962 上的位点。以上位点中有 5 个位于明确的基因上,在这 5 个基因中,有两个基因的功能已被确认(分别为 ERICH6-AS1 和 PDE8B),其余 3 个基因的功能尚未被确认。ERICH6-AS1 属于 RNA 反转录基因,PDE8B 属于磷酸二酯酶调控基因。根据美国国家生物技术信息中心(National Center for Biotechnology Information,NCBI)数据库上收录的常规组织表达量数据,以上两个基因都在胎盘组织具有一定表达量。因此可以猜测该基因通过基因间相互作用影响其他基因的表达,从而引起胎儿的先天性青光眼。本文所提蚁群算法选取的双位点组合与生物学已经证实的结果具有较高的相合性。本文结果的生物学意义有待通过进一步的生物实验来验证。

表 2 与位点 rs13414786 相关的双位点组合

Table 2 Site pair combinations involving SNP rs13414786

另一位点 所在基因 ^{a)}	另一位点信息		<i>P</i> 值 ^{d)}
	位置 ^{b)}	rs 编号 ^{c)}	
ERICH6-AS1	1	rs6678820	1.12×10^{-3}
	3	rs13074670	7.89×10^{-4}
	3	rs9837001	5.21×10^{-4}
PDE8B	5	rs335614	7.38×10^{-3}
	6	rs7742590	8.11×10^{-4}
	9	rs13285154	5.46×10^{-3}
	10	rs2138551	7.44×10^{-4}
LOC105378485	11	rs585166	8.87×10^{-4}
	12	rs6539798	1.17×10^{-3}
LOC105370924	15	rs1948815	8.51×10^{-4}
	17	rs17573984	7.07×10^{-3}

a—与 2 号染色体基因 LOC107985962 上的位点 rs13414786 和 rs1731419 所组成的双位点组合(其中结果 1,3,4,11 包含位点 rs1731419,其他结果包含位点 rs13414786)中,另一位点所在基因的位置(列出了已被确认基因的名称);b—组合中另一位点所在的染色体位置;c—组合中另一位点的 rs 编号;d—对该双位点组合进行秩和检验得到的 *P* 值。

3 结论

本文的研究提供了一种通过探索连续型变量与多位点组合的关联性来发现与疾病相关的 SNP 组合的新思路。通过 10 000 位点子集的实验找到了合理的参数,提高了所设计算法的计算效率,同时也增强了算法对边际效应较高双位点组合的探索能力。另外,通过全部位点的实验,在 2 号染色体上找到了一系列未被报道的显著性水平较高的位点组合,这些位点组合很有可能通过医学实验得到进一步证实,从而为探索与眼压相关的基因功能和青光眼的发病原理提供新的线索。

参考文献:

- [1] 陈宇虹, 孙兴怀. 原发性先天性青光眼的遗传学研究进展[J]. 国际眼科纵览, 2006, 30(2): 104 - 108.
CHEN Y H, SUN X H. Progress in genetics of primary congenital glaucoma[J]. International Journal of Ophthalmology, 2006, 30(2): 104 - 108. (in Chinese)
- [2] 葛坚. 青光眼防治工作中面临的问题与挑战[J]. 中华眼科杂志, 2002, 38(6): 321 - 324.
GE J. Problems and challenges in the prevention and treatment of glaucoma[J]. Chinese Journal of Ophthalmology, 2002, 38(6): 321 - 324. (in Chinese)
- [3] FOSTER P J, JOHNSON G J. Glaucoma in China: how big is the problem? [J]. British Journal of Ophthalmology, 2001, 85 (11): 1277 - 1282.
- [4] 葛坚. 青光眼的研究进展与发展趋势[J]. 中华眼科杂志, 2000, 36(3): 192 - 196.
GE J. Research progress and development trend of glaucoma [J]. Chinese Journal of Ophthalmology, 2000, 36 (3): 192 - 196. (in Chinese)
- [5] 徐亮, 陈建华, 李建军, 等. 北京农村及城市特定人群原发性开角型青光眼的患病率调查及其筛查方法评价[J]. 中华眼科杂志, 2004, 40(11): 726 - 732.
XU L, CHEN J H, LI J J, et al. Prevalence of primary open angle glaucoma in rural and urban specific populations in Beijing and evaluation of screening methods [J]. Chinese Journal of Ophthalmology, 2004, 40 (11): 726 - 732. (in Chinese)
- [6] 林明楷, 葛坚, 陈慧怡, 等. 青光眼住院病人的构成及变化[J]. 中国实用眼科杂志, 2003, 21 (12): 937 - 939.
LIN M K, GE J, CHEN H Y, et al. The analysis in proportion of hospitalized patients with glaucoma [J]. Chinese Journal of Practical Ophthalmology, 2003, 21 (12): 937 - 939. (in Chinese)
- [7] ASCHARD H, KANG J H, IGLESIAS A I, et al. Genetic correlations between intraocular pressure, blood pressure and primary open-angle glaucoma: a multi-cohort analysis [J]. European Journal of Human Genetics, 2017, 25(11): 1261 - 1267.
- [8] HARDY J, SINGLETON A. Genomewide association studies and human disease [J]. New England Journal of Medicine, 2009, 360(17): 1759 - 1768.
- [9] VAN KOOLWIJK L M E, RAMDAS W D, IKRAM M K, et al. Common genetic determinants of intraocular pressure and primary open-angle glaucoma[J]. PLoS Genetics (Online), 2012, 8(5): e1002611.
- [10] SAPIN E, KEEDWELL E, FRAYLING T. An ant colony optimization and tabu list approach to the detection of gene-gene interactions in genome-wide association studies [J]. IEEE Computational Intelligence Magazine, 2015, 10(4): 54 - 65.
- [11] LIU J, YU G X, JIANG Y, et al. HiSeeker: detecting high-order SNP interactions based on pairwise SNP combinations[J]. Genes, 2017, 8(6): 153 - 171.
- [12] 戎晗, 顾珊珊, 张国伟, 等. 非编码区单核苷酸多态性与眼部疾病关系的研究进展[J]. 眼科新进展, 2016, 36(4): 380 - 385.
RONG H, GU S S, ZHANG G W, et al. Recent advances in relationship between SNPs in non-coding regions and ocular diseases [J]. Advances in Ophthalmology, 2016, 36(4): 380 - 385. (in Chinese)

Correlation analysis between gene locus combinations and intraocular pressure data based on ant colony algorithms

CAO YiMing¹ HU Man² XU YongLi^{1*}

(1. College of Mathematics and Science, Beijing University of Chemical Technology, Beijing 100029;

2. Ophthalmology Department, Beijing Children Hospital, Beijing 100045, China)

Abstract: The association analysis of multi-site combinations and phenotype of genes is a research hotspot in bioinformatics studies. Recent studies have shown that for complex genetic diseases, multi-site combinations are more significantly associated with phenotypes than single gene-specific points. However, the existing methods are only suitable for analyzing the association of multi-site combinations with discrete phenotypes, and cannot analyze the association with continuous phenotypes. Glaucoma is the second most common cause of irreversible blindness. High intraocular pressure is the most important indication of the onset of glaucoma. Therefore, studying the correlation between multi-site combination and intraocular pressure is of great significance for studying the pathogenesis of glaucoma. In this paper, an ant colony algorithm has been designed to study the association analysis between gene multi-site combinations and continuous phenotypes. In this work, we evaluated the performance of a method designed by ourselves on a gene and intraocular pressure dataset published by the European Institute of Bioinformatics. The experimental results show that the ant colony algorithm developed in this work can find the two-point combinations of genes significantly related to intraocular pressure, which might provide new clues for studying the pathogenesis of glaucoma. In addition, the ant colony algorithm developed in this work provides new ideas for studying the association of multi-site combinations and continuous phenotypes in other diseases.

Key words: genetic trait regression analysis; ant colony algorithm; intergenic interactions; glaucoma