DOI:10.3969/j.issn.1005-202X.2017.05.004

医学影像物理

相似度检测法在遗传性疾病和遗传位点上的应用

何世钧,程小龙,张婷,周媛媛,朱吉光
上海海洋大学信息学院, 上海 201306

【摘 要】在生物学中,寻找致病位点、表现型性状差异及疾病的易感性等都属于分类问题,目的是把等待处理的样本进行分类,找出异常数据。相似度的优点在于类与类之间寻找异常数据,将相似度检测运用于致病位点的查询,结合卡方检验和支持向量机,能更精确地找到致病位点。实验中相似度检测得到的致病位点和卡方检验得到的致病位点都有可能是该病的致病位点,致病位点可能存在于两者的集合之中,同时根据支持向量机的小样本、非线性问题中表现出许多特有的优势,在实验中使用支持向量机进行建模,在模型精度较高的基础上确定各个位点的权重,权重大的位点对结果的影响明显,达到筛选出最有可能致病的位点的目的。

【关键词】遗传性疾病;遗传位点;相似度检测法;支持向量分类算法;全基因组关联性分析
【中图分类号】R394.8 【文献标志码】A 【文章编号】1005-202X(2017)05-0450-06

Similarity detection method in the determination of disease-causing gene and genetic loci

HE Shijun, CHENG Xiaolong, ZHANG Ting, ZHOU Yuanyuan, ZHU Jiguang
College of Information, Shanghai Ocean University, Shanghai 201306, China

Abstract: In biology, searching for the pathogenic loci, the difference of phenotype and the susceptibility of disease is a kind of classification problem, aiming to classify the samples and to find the abnormal data. The similarity can find the abnormal data between classes. Therefore, combining similarity detection method with chi square test and support vector machine can find the pathogenic loci more accurately. Both the pathogenic loci detected by similarity detection method and chi square test would be the disease loci of the disease, and the disease loci may exist in the collection of both. With the advantages in the small sample and nonlinear problems, support vector machine is used for the modeling. Based on the established model of high precision, we select the most likely loci of disease by determining the weights of loci because the effect of the loci with large weight on the results is significant.

Keywords: genetic disease; genetic loci; similarity detection method; support vector classification; genome-wide association analysis

前 言

人体的每条染色体携带一个DNA分子,人的遗传密码由人体中的DNA携带。DNA是由分别带有A、T、C、G四种碱基的脱氧核苷酸链接组成的双螺旋长链分子。基因是DNA长链中有遗传效应的一些片段。在组成DNA的数量浩瀚的碱基对(或对应的脱氧核苷酸)中,有一些特定位置的单个核苷酸经常发生变异引起DNA的多态性,称之为位点。染色体、基因和位点的结构关系见图1。

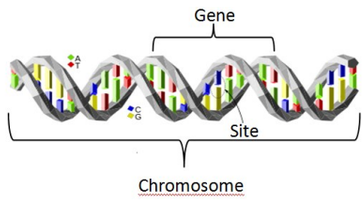

图1 染色体、基因和位点的结构关系

Fig.1 Structural relationships of chromosomes, genes and loci

大量研究表明,人体的许多表型性状差异以及对药物和疾病的易感性等都可能与某些位点相关联,或与包含有多个位点的基因相关联[1]。因此,定位与性状或疾病相关联的位点在染色体或基因中的位置,能帮助研究人员了解性状和一些疾病的遗传机理,也能使人们对致病位点加以干预,防止某些遗

【收稿日期】2017-01-08
【基金项目】上海市科委科研计划项目(10510502800)
【作者简介】何世钧,男,教授,主要从事海洋信息处理与预测模型、自动控制与检测技术、传感器与数字技术等研究,E-mail: heshijun6@163.com

传病的发生。

近年来，研究人员大都采用全基因组的方法来确定致病位点或致病基因，全基因组关联性分析研究最早应用于人类流行病学的研究，是在种群水平上用来寻找与某一表型相关的基因位点，即单核苷酸多态性(SNP)的强大技术手段，能够客观地筛选出与某个特定的表型相关度较高的新位点[2]。并已发展成为研究真核生物遗传疾病基因多态性的一项重要研究手段[3]。但是采用全基因组的方法确定致病位点或致病基因需要招募大量的志愿者，耗费大量的人力物力。房雅楠等[4]报道使用卡方检验比较位点基因型频率和等位基因频率分布，采用Logistic回归分析疾病危险因素，判断群体中位点的多态性与疾病的发生是否有关。目前也有研究人员使用一些基于生物信息学分析的位点致病性预测软件对致病位点进行预测[5]。但是因为疾病的类型不同，预测软件所预测的结果差异也较大，同时，不同位点致病性预测软件的预测敏感度和特异性也有差异，从而导致不同软件预测结果的假阳性率和假阴性率也不尽相同，其精确性有待提高。

相似度检测法在寻找致病位点的过程中不需要耗费大量的人力物力，由于相似度检测模型是对位点的碱基对进行处理，因此不会因为疾病类型的不同而导致预测结果的差异，考虑到相似度检测法的这些特点，结合碱基和位点的关系(两个碱基的组合表示1个位点的信息)建立模型，采用相似度分析法确定致病位点或致病基因，从而发现遗传病或性状的遗传机理。

1 位点的分析与处理

1.1 碱基的编码方式

本文实验数据由重庆大学提供，数据分别为针对某种遗传疾病(简称疾病A)提供1 000个样本信息，这些信息包括这1 000个样本的疾病信息、样本的9 445个位点编码信息，以及包含这些位点的基因信息。为便于进行数据分析、找出致病位点，把数据中每个位点的碱基(A,T,C,G)编码方式转化为数值编码方式。将数据进行数字化不但便于存储和查询，而且便于数值计算和数据处理。数字编码方式有按分子量大小顺序排列的碱基数字编码、按π电子能级顺序编码的碱基数字编码[6]及其他编码方式，如格雷码、ASCII、Unicode码等。

按分子量大小顺序排列的碱基数字编码方式简介如下：

使用0(00)、1(01)、2(10)和3(11)这4个数字对(A,T,C,G)4种碱基进行编码。这种编码方式共有4!=24种编码方式组合。基于二进制数字中0与1的互补关系和碱基互补法则，筛选出的数字编码格式共有8种，即0123/CTAG，0123/CATG，0123/GTAC，0123/GATC，0123/TCGA，0123/TGCA，0123/ACGT，0123/AGCT。其余16种数字编码格式，因不满足互补法则而被摒弃。采用0123/CTAG这种编码格式，首先，其参考了化学上以原子量大小顺序排列元素周期表，4种碱基的分子量按大小顺序排列是C=111.10，T=126.12，A=135.13，G=151.13；其次，按数码计算，两对互补数字之和相等。而CTAG两对互补碱基的分子量之和亦呈相等关系，二者绝对误差为0.98<1，其相对误差为0.3%。故可以认为两对互补碱基对的分子量之和几乎相等，而且两对互补的碱基对呈对角互补关系。另外0123/CTAG编码格式可以反映出4种碱基的化学性质[7]。

1.2 碱基的数值化编码

按分子量大小顺序排列的碱基数字编码方式具有可以对碱基结构、功能基团、碱基互补、氢键强弱等性质进行编码，方便求出任意碱基重复单元的重复系列的数字编码法则，压缩信息冗余度，提高编码效率、方便进行各种数学运算和逻辑运算，对促进DNA生物计算机的发展具有重大推动作用等优点[8]，故采用按分子量大小顺序的编码方式对位点数据进行预处理。

用0(00)表示C，1(01)表示T，2(10)表示A，3(11)表示G，16组多聚双核苷酸链的数值编码如下：
CC(0000)=0 CT(0001)=1 CA(0010)=2 CG(0011)=3
TC(0100)=4 TT(0101)=5 TA(0110)=6 TG(0111)=7
AC(1000)=8 AT(1001)=9 AA(1010)=10 AG(1011)=11
GC(1100)=12 GT(1101)=13 GA(1110)=14 GG(1111)=15

由此将数据文件中的碱基编码方式转化成数值编码方式，下面给出部分位点的碱基序列(表1)和对应的数值序列(表2)对比。

将每个位点的碱基数值化之后，运用相似度检测模型和支持向量机对数值化之后的碱基进行数据分析，找出致病位点。

2 模型的建立及分析

寻找致病位点的问题归为一个高纬度数据分类问题[9]。在对高维度数据进行分类时，数据包含大量特征，过高的特征维数使分类的正确率下降，因此进行有效的降维是一个关键问题。使用一些统计数据模型进行分类降维，对其进行特征提取和特征选择，找出这些数据中最能够代表该特征的一部分数据，

表1 部分位点的碱基编码方式

Tab.1 Base coding of partial loci

rs4040617	rs2980300	rs4970383	rs4475691	rs1806509
AA	CC	GG	CC	AA
AA	CC	GG	CC	AA
GG	CC	GG	CC	AA
AG	CT	GT	CT	AC

表2 部分位点的数字化编码

Tab.2 Digital coding of partial loci

rs4040617	rs2980300	rs4970383	rs4475691	rs1806509
10	0	15	0	10
10	0	15	0	10
15	0	15	0	10
11	1	13	1	8

从而达到简化阵列数据的目的[10]。相似度检测法不但能保留数据特征，还能对高维数据进行降维，从而简化数据，找到致病位点。

2.1 相似度检测法模型

计算系统相似度数值的大小可表示为数学表达式。其中，和分别表示两个系统的各自组成要素的数量，表示相似要素的数量，表示相似要素的相似程度。由相似要素的数量确定的相似度为数量相似度，记为Sr。相似度的计算方法分为以下几种：

（1）谱图相似度（Spectral Similarity，SS）法[11]，按下式计算：

$$Sr = SS = \frac{\sum_{i=1}^{p}\sqrt{e_i f_i}}{\sqrt{\sum_{i=1}^{p} e_i \sum_{i=1}^{p} f_i}} \tag{1}$$

e_i 为系统对应的第 i 个对象的量化值，即特征值。Sr 值越大，表示A变量与B变量越接近。

（2）相关系数（Correlation coefficient，CC）法：

$$Sr = CC = \frac{\sum_{i=1}^{p}(e_i - \bar{f})(f_i - \bar{f})}{\sqrt{\sum_{i=1}^{p}(f_i - \bar{f})^2 \sum_{i=1}^{p}(e_i - \bar{e})^2}} \tag{2}$$

（3）夹角余弦（Cosine Distance，CD）法[12]：

$$Sr = CD = \frac{\sum_{i=1}^{p} e_i f_i}{\sqrt{\sum_{i=1}^{p} e^2{}_i \sum_{i=1}^{p} f_i^2}} \tag{3}$$

e_i 和 f_i 是第 i 个对象A指标和B指标的值。

（4）相似度D按下式计算：

$$Sr = D = 1 - \frac{\sum_{i=1}^{p} abs(e_i - f_i)}{\sum_{i=1}^{p} abs(e_i + f_i)} \tag{4}$$

e_i 为系统对应的第 i 个对象的量化值，即特征值。

考虑到相关系数法在比较两个矩阵的相似度方面更加方便、易操作，在下面的实验中采用相关系数法建立相似度模型。

2.2 非线性支持向量分类模型

支持向量机有很强的非线性建模能力，而且具有全局最优、结构简单、小样本推广能力强等优点，很适合解决分类问题[13]。用以下线性函数对给定的训练样本集进行回归拟合：

$$f(x) = w \cdot x + b \tag{5}$$

其中 w 为法向量，b 是偏移量。为使式(5)具有平滑性，需要求出上式成立的最小 w 。假设在精度 ε 下，训练样本集 (x_i, y_i) 可以用线性函数进行回归，以上问题的求解可以表示为以下的约束问题的求解：

$$\min \ \frac{1}{2}\|w\|^2 \quad s.t. \ \begin{cases} y_i - w \cdot x_i - b \le \varepsilon \\ w \cdot x_i + b - y_i \le \varepsilon \end{cases} \tag{6}$$

为保证回归拟合的泛化性，利用松弛因子 $\xi_i \ge 0$ 和 $\xi_i^* \ge 0$ 及惩罚参数 C 对 ε -带进行软化，于是将式(6)升级为以下形式：

$$\min \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \quad s.t. \ \begin{cases} y_i - w \cdot x_i - b \le \varepsilon + \xi_i \\ w \cdot x_i + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i \ge 0 \\ \xi_i^* \ge 0 \end{cases} \quad i = 1, \ldots, n \tag{7}$$

其中回归中的 ε 不敏感损失函数形式如下：

$$|\xi|_\varepsilon = \begin{cases} 0 & if \ |\xi| \le \varepsilon \\ |\xi| - \varepsilon & otherwise \end{cases} \tag{8}$$

损失函数的图像如图2所示。当样本点的实际值与预测值之差小于事先给定的 ε 时，则认为该样本点的预测值是无损失的，即使二者之间存在一定误差。

为得到式(7)的解，通过引入Lagrange乘子将其转化为对偶问题进行求解：

$$\min \ \frac{1}{2}\sum_{i,j=1}^{n}(a_i - a_i^*)(a_j - a_j^*) < x_i, x_j > + \sum_{i=1}^{n} a_i(\varepsilon - y_i) + \sum_{i=1}^{n} a_i^*(\varepsilon + y_i) \quad s.t. \ \begin{cases} \sum_{i=1}^{n}(a_i - a_i^*) = 0 \\ a_i, a_i^* \in [0, C] \end{cases} \tag{9}$$

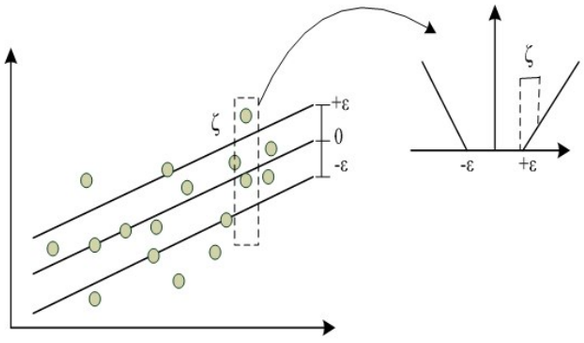

图2　ε –不敏感损失函数图

Fig.2 Graph of ε –insensitive loss function

求解以上对偶问题，可得到最优的分类函数为 $f(x)=\mathrm{sgn}(\sum_i^n(a_i-a_i^*)(x_i\cdot x)+b)$，$w$ 即为决策函数的权重向量，$w=\sum_i^n(a_i-a_i^*)x_i$，其各分量绝对值大小分别代表各分量 x_i 在决策函数中的权重。

对于非线性分类，利用核函数把数据从输入空间 R^n 映射到一个特征空间[14]：

$$\Phi:\ \begin{matrix}\mathrm{X}\subset R^n\to\mathrm{X}\subset\mathrm{H}\\ x\to\phi(x)\end{matrix}\qquad(10)$$

输入空间中的训练样本 $T=\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$，映射到特征空间的样本集为：$\bar{T}=\{(\phi(x_1),y_1),(\phi(x_2),y_2),\ldots,(\phi(x_n),y_n)\}$，接着将 $\bar{T}$ 在特征空间进行回归拟合。这种利用核函数 $K(x_i,x_j)$ 进行空间转换并得到决策函数的过程中计算复杂度并没有增加[15]。

样本 X 通过 $\phi(X)$ 映射到特征空间后，非线性分类问题转化以下形式：

$$\begin{aligned}&\min\ \frac{1}{2}\sum_{i,j=1}^n(a_i-a_i^*)(a_j-a_j^*)<\phi(x_i),\phi(x_j)>\\&+\sum_{i=1}^n a_i(\varepsilon-y_i)+\sum_{i=1}^n a_i^*(\varepsilon+y_i)\\&s.t.\quad\begin{cases}\sum_{i=1}^n(a_i-a_i^*)=0\\ a_i,a_i^*\in[0,C]\end{cases}\end{aligned}\qquad(11)$$

解式(11)可得最优的分类函数为 $f(x)=\mathrm{sgn}(\sum_i^n(a_i-a_i^*)K(x_i\cdot x)+b)$，法向量 $w=\sum_i^n(a_i-a_i^*)\phi(x_i)$。

3 结　果

3.1 相似度检测模型的求解

把数据中1 000×9 445个碱基以位点为单位进行处理，先对第一列的前500个碱基进行处理，把一列中的最小值用0标记，最大值用2标记，其余的用1标记，依据位置填入到矩阵中，同理处理后500个碱基对，这两个矩阵反映了样本的信息，求解前500个健康样本形成的矩阵和后500个患疾病A的样本组成矩阵的相似度。若两个矩阵的相似度较低，则该位点为致病位点的可能性大，若相似度较高，则该位点为致病位点的可能性较小。

以第一个位点rs3094315为例，表3、表4分别为rs3094315碱基对的相对位置统计信息。处理后得到2个矩阵，$\boldsymbol{E}_{mn}$ 为前500个健康样本形成的矩阵，$\boldsymbol{F}_{mn}$ 为后500个患疾病A的样本组成矩阵。

表3　前500个碱基对的相对位置统计信息

Tab.3 Relative position statistics of the first 500 base pairs

	CC	TC	TT
CC	7	31	24
TC	22	103	104
TT	33	94	81

表4　后500个碱基对的相对位置统计信息

Tab.4 Relative position statistics of the last 500 base pairs

	CC	TC	TT
CC	6	25	25
TC	27	98	98
TT	24	99	97

$$\boldsymbol{E}_{mn}=\begin{bmatrix}7&31&24\\22&103&104\\33&94&81\end{bmatrix}\quad\boldsymbol{F}_{mn}=\begin{bmatrix}6&25&25\\27&98&98\\24&99&97\end{bmatrix}$$

使用 $Sr=CC=\dfrac{\sum_m\sum_n(\boldsymbol{E}_{mn}-\bar{E})(\boldsymbol{F}_{mn}-\bar{F})}{\sqrt{(\sum_m\sum_n(\boldsymbol{E}_{mn}-\bar{E})^2)(\sum_m\sum_n\boldsymbol{F}_{mn}-\bar{F})^2}}$

解得两个矩阵的相似度为0.997 9，其中 $\bar{E}$ 和 $\bar{F}$ 分别表示矩阵 $\boldsymbol{E}$ 和矩阵 $\boldsymbol{F}$ 的均值，m=1,2,3；n=1,2,3。表明此位点为致病位点的可能性极小。图3是9 445个位点患病情况的碱基对的相似度，最后找出图中相似度低的位点作为与疾病有显著相关的位点。

$SNPs_{corr}(Q<0.15)=\{rs10915668, rs938962, rs950601, rs2273298, rs12124123, rs6671790, rs12408946\}$，$|SNPs_{corr}|=7$。此方法不仅表征了1个位点的各种碱基对的数量，还反映了碱基对间的相对位置，能较好地反映健康和患病样本的碱基特征。

3.2 非线性支持向量分类模型的建立与求解

记得到的卡方检测结果中的 $SNPs_{Chi}$ 和相似度检验结果中 $SNPs_{corr}$ 为可能是致病位点的集合

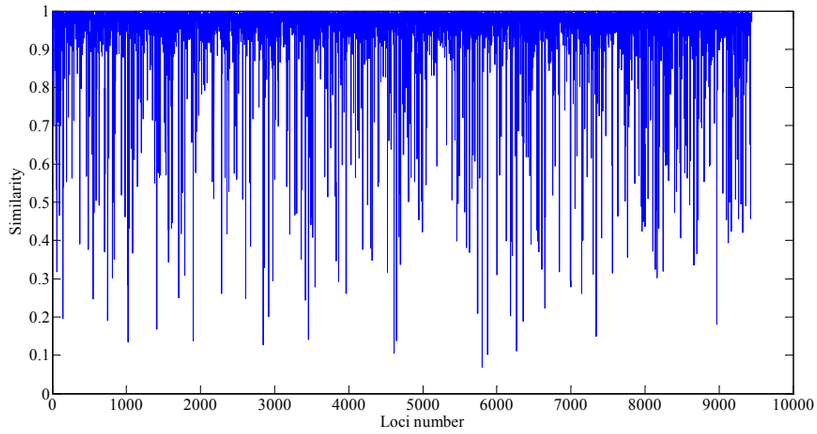

图3 9 445个位点患病情况的碱基对的相似度

Fig.3 Similarity of the base pairs of 9 445 loci related to disease

$S=\{SNPs_{Chi},SNPs_{corr}\}$，其中 $SNP_{Chi}=\{rs17262293,rs2273298,rs12742921\}$，$|S|=10$。由此将高维数据分类问题转化成非线性的分类问题。根据支持向量机的小样本、非线性问题中表现出许多特有的优势，我们在该问题中使用SVM进行建模，包括对于惩罚参数、核函数的选择等，然后将集合 S 中位点对应的碱基对数据作为输入样本，是否患病 Y 作为输出数据进行SVM建模处理。使用事先准备好的测试数据进行对比，以确定训练之后的模型是否有较好效果。在模型精度较高的基础上确定各个位点的权重，权重大的位点对结果的影响明显，达到筛选出最有可能致病的位点的目的。

考虑到不同位点之间存在的相互关系对疾病的检测会造成不同程度的影响，同时为了在集合 S 的基础上寻找其中最有可能的致病位点的组合，故对 S 进行组合排序，并使用SVM对不同组合的数据进行建模，对比所有组合的模型分类精度来判断致病位点与疾病之间的关系。通过不同的位点组合，发现编号为9的位点序列（$illness=\{SNPs_{Chi},SNPs_{corr}\}-\{rs12408946\}$）具有较高的精度，其中 acc_{train}=84.1%，acc_{test}=68%，即除位点（rs12408946）外的9个位点更能反应位点与疾病之间的关系。

图4是位点集合illness对应碱基对信息作为输入数据所得计算模型中每个位点的权重信息。由图5可知，每个位点都对疾病有影响，即都可能是致病位点，其中位点rs2273298对应权重明显高于其他位点权重，说明位点rs2273298是致病位点可能性最大。

4 结 论

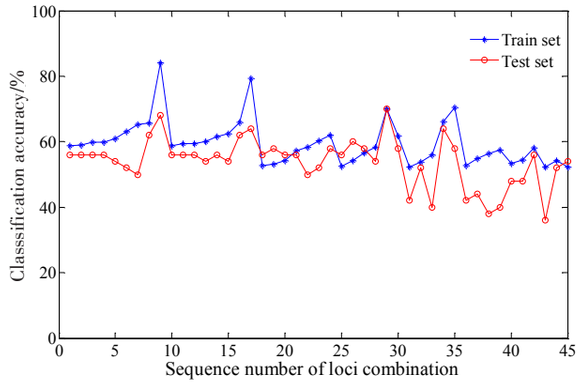

图4 不同位点组合的分类精度

Fig.4 Classification accuracy of different combinations of loci

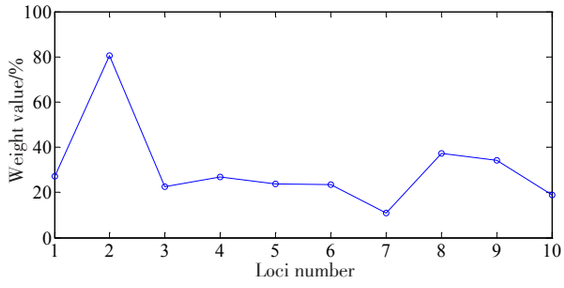

图5 最优模型中位点的权重

Fig.5 Weight value of the sites in the optimal model

相似度结合卡方检验和支持向量机模型能在寻找致病位点、表现型性状差异及疾病的易感性等问题中精确找到异常数据，该方法可简化致病位点的样本数量，找到可能的致病位点，定位与性状或疾病相关联的位点在染色体或基因中的位置，帮助研究人员了解性状和一些疾病的遗传机理，也能使人们

对致病位点加以干预,防止一些遗传病的发生。该模型还可用于建立相关遗传疾病的预测模型,可直接通过样本位点上的信息对样本的健康状况进行判断。同时,该模型还可以用于分析性状与位点之间的关系,实现分析多种性状之间的相关性。

致谢:感谢第十三届全国研究生数学建模专家组提供的数据以及对本文部分内容做出荣获全国研究生数学建模三等奖的评审,衷心感谢他们公平的审议和帮助。

【参考文献】

[1] REN P, PENG W, YOU W, et al. Genetic mapping and quantitative trait loci analysis of growth-related traits in the small abalone Haliotis diversicolor, using restriction-site-associated DNA sequencing[J]. Aquaculture, 2016, 454: 163-170.

[2] BUSH W S, MOORE J H. Chapter11: Genome-wide association studies[J]. PLoS Comput Biol, 2012, 8(12): e1002822.

[3] PATRICIA A, SONIA D M, RAUL G, et al. Genome-wide survey of yeast mutations leading to activation of the yeast cell integrity MAPK pathway: novel insights into diverse MAPK outcomes[J]. BMC Genomics, 2011, 12: 390.

[4] 房雅楠, 隋汝波. EDNRA和EDNRB基因单核苷酸位点多态性与缺血性脑卒中的相关性[J]. 中国老年学杂志, 2016, 18: 4446-4449.
FANG Y N, SUI R B. Association between single nucleotide polymorphism of EDNRA and EDNRB gene and ischemic stroke[J]. Chinese Journal of Geriatrics, 2016, 18: 4446-4449.

[5] 章亮, 苏志熙. 位点致病性预测软件对错义突变的预测效用评估[J]. 基因组学与应用生物学, 2016, 35(8): 1916-1925.
ZHANG L, SU Z X. Evaluation utility of predicting the missense mutation by locus predicting software[J]. Genomics and Application Biology, 2016, 35(8): 1916-1925.

[6] 罗辽复. 物理学家看生命[M]. 长沙: 湖南教育出版社, 1994.
LUO L F. Physicist to see life[M]. Changsha: Hunan Education Press, 1994.

[7] 陈惟昌, 陈志华, 陈志义, 等. 遗传密码和DNA序列的高维空间数字编码[J]. 生物物理学报, 2000, 16(4): 760-768.
CHEN W C, CHEN Z H, CHEN Z Y, et al. Genetic dimension coding of genetic codes and DNA sequences[J]. Acta Biophysica Sinica, 2000, 16(4): 760-768.

[8] 李书超, 许进, 刘光武. 基于DNA计算的RNA序列多点"突变"与首段碱基的数字编码及其运算法则[J]. 计算机工程与应用, 2004, 40(8): 15-18.
LI S C, XU J, LIU G W. Multi-point "Mutation" of RNA sequences based on DNA computing and digital codingof the first base and algorithms of computation [J]. Computer Engineering and Applications, 2004, 40(8): 15-18.

[9] 朱扬勇, 熊赟. DNA序列数据挖掘技术[J]. 软件学报, 2007, 11: 2766-2781.
ZHU Y Y, XIONG Y. Data mining technology of DNA sequence[J]. Journal of Software, 2007, 11: 2766-2781.

[10] 朱坤. 基于改进的Relief算法与支持向量机的高通量基因数据分析[D]. 福州: 福建农林大学, 2012.
ZHU K. Analysis of high throughput gene data based on improved relief algorithm and support vector machine[D]. Fuzhou: Fujian Agriculture and Forestry University, 2012.

[11] 苏越, 刘素红, 王呈仲, 等. 谱图相似度分析结合保留指数对单萜烯同分异构体的GC-MS定性分析[J]. 分析测试学报, 2009, 28(5): 525-528.
SU Y, LIU S H, WANG C Z, et al. Spectral similarity analysis combined with qualitative GC-MS retention index of monoterpene isomers analysis[J]. Journal of Instrumental Analysis, 2009, 28(5): 525-528.

[12] 相秉仁. 计算药学[M]. 北京: 中国医药科技出版社, 1990: 160-163.
XIANG B R. Calculation of pharmacy[M]. Beijing: China Medical Science and Technology Press, 1990: 160-163.

[13] 唐莹莉, 何世钧, 李煜, 等. 基于小波支持向量机的小鼠脑电波与呼吸的关系[J]. 中国医学物理学杂志, 2015, 32(3): 365-369.
TANG Y L, HE S J, LI Y, et al. Relationship between brain waves and respiration based on wavelet support vector machines[J]. Chinese Journal of Medical Physics, 2015, 32(3): 365-369.

[14] 王英奇. 基于核学习方法的聚类算法研究[D]. 兰州: 兰州交通大学, 2009.
WANG Y Q. Research on clustering algorithm based on kernel learning method[D]. Lanzhou: Lanzhou Jiaotong University, 2009.

[15] PLATT J C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods[C]//SMOLA A J, BARTLETT P L. Advances in large margin classifiers. Cambridge: MIT Press, 1999.

(编辑:黄开颜)