doi: 10.3969/j.issn.1007-2861.2012.03.010

基于关联规则挖掘的蛋白质相互作用的预测

林合同， 龚云路， 秦殿刚， 冯铁男， 王翼飞

(上海大学 理学院,上海 200444)

摘要: 利用蛋白质的一级结构信息,采用三肽频数方法刻画蛋白质序列,将关联规则(association rule,AR)挖掘应用于蛋白质相互作用(protein-protein interactions, PPIs)的预测. 计算结果表明,提出的方法在半胱氨酸不同分类的情况下都能够准确地预测蛋白质相互作用. 最后,比较半胱氨酸的不同分类对预测结果的影响.

关键词: 关联规则挖掘;蛋白质相互作用;序列编码;氨基酸分类

中图分类号: Q 516 **文献标志码**: A **文章编号**: 1007-2861(2012)03-0265-06

Prediction of Protein-Protein Interactions Based on Association Rule Mining

LIN He-tong, GONG Yun-lu, QIN Dian-gang, FENG Tie-nan, WANG Yi-fei

(College of Sciences, Shanghai University, Shanghai 200444, China)

Abstract: Association rule (AR) mining has been successfully applied to predict protein-protein interactions (PPIs) through protein's primary sequence. A conjoint triad feature is used to describe amino acids. Experimental results show that the proposed method can predict PPIs with high accuracy under different classifications of Cys. The predicted results of two classifications of Cys are compared.

Key words: association rule mining; protein-protein interactions (PPIs); sequential coding; classification of amino acids

蛋白质间的相互作用在生命活动中扮演着重要的角色,研究蛋白质间的相互作用对了解蛋白质的功能和生命活动的规律至关重要. 目前研究蛋白质间相互作用的方法有很多[1-8]. 蛋白质的一级结构包含了蛋白质的很多信息,因此,依赖蛋白质的一级结构预测蛋白质的相互作用具有巨大的挑战性.

本工作从蛋白质的一级结构信息出发,采用三肽频数方法编码蛋白质序列,将关联规则挖掘应用于蛋白质相互作用预测中. 二硫键是某些蛋白质维持结构稳定的重要因素,它由2个半胱氨酸残基的侧链—SH氧化形成[7]. 鉴于半胱氨酸在蛋白质研究中的重要作用,本工作在不同的半胱氨酸分类情况下,对蛋白质的相互作用分别进行了预测.

1 材料与方法

1.1 数据来源

本工作采用小鼠和人的蛋白质相互作用数据作为样本数据,它们均来源于IntAct数据库(网址为ftp://ftp.ebi.ac.uk/pub/databases/intact/current/psimitab),下载得到的文件每一行均是一个蛋白质

收稿日期:2011-03-10

基金项目:国家自然科学基金资助项目(30871341);上海市重点学科建设资助项目(S30104);上海市教委重点学科建设资助项目(J50101)

通信作者:王翼飞(1948~),男,教授,博士生导师,研究方向为计算分子生物学. E-mail:yifei_wang@staff.shu.edu.cn

相互作用对. 除去自相互作用数据和冗余数据,选取了1 151对小鼠蛋白质相互作用数据和4 149对人蛋白质相互作用数据作为阳性数据.

在预测蛋白质相互作用的过程中,阴性数据的选取非常重要,因为阴性数据的选择直接决定了挖掘出规则的数目以及关联性的强弱,从而决定了模型的预测能力. 本工作采用亚细胞定位的方法选取阴性数据,认为处于细胞不同位置的蛋白质不会发生相互作用. 例如,由 Gene Ontology 标注结果可知,O43609 位于细胞质或质膜上,而 Q6PEV8 位于高尔基体或细胞核上,因此,认为 O43609 与 Q6PEV8 不会发生相互作用[9]. 为了避免数据不平衡对预测结果的影响,本工作根据 Gene Ontology 标注结果,选取1 151对小鼠蛋白质不相互作用数据和4 149对人蛋白质不相互作用数据作为阴性数据.

1.2 关联规则挖掘

关联规则挖掘(association rule mining)问题于1993年由 Agrawal 等提出,并已成为了数据挖掘领域中最活跃的一个分支,它能发现大量数据中项集之间有趣的关联. 最初的研究主要应用于市场分析、决策制定、商业管理等领域,但随着研究的不断深入,关联规则挖掘研究的应用越来越广泛,已扩展到了网络分析、天文学和生物信息学等领域. 近年来,关联规则挖掘已成为数据挖掘领域一个非常重要的研究课题[10].

1.2.1 关联规则的基本概念

在众多的数据挖掘教材和文献中,对关联规则挖掘涉及到的基本概念定义如下[10-12].

定义1 项为数据库中不可分割的最小信息单位,一般用 i 表示,项的集合称为项集. 设集合 $I=\{i_1,i_2,\cdots,i_k\}$ 为项集,如果 I 中项的数目为 k,则称 I 为 k-项集.

定义2 关联规则是形如 $X\Rightarrow Y$ 的蕴含式,其中 $X\subseteq I, Y\subseteq I$,并且 $X\cap Y=\varnothing$. X 称为前项,Y 称为后项. 关联规则反映了当 X 中的项出现时,Y 中的项也跟着出现的规律.

定义3 如果项集 $U=\{u_1,u_2,\cdots,u_k\}$ 出现的频率大于或等于最小支持度计数,即满足最小支持度阈值,则称它为频繁项集(frequent itemset),频繁 k-项集的集合通常记为 L_k.

定义4 关联规则的支持度(support)是指交易集中同时包含 X 和 Y 的交易数与所有交易数的比值,记为 $\sup(X\Rightarrow Y)$,即 $\sup(X\Rightarrow Y)=\sup(X\cup Y)=p(X\cup Y)$. 支持度表示 X 和 Y 中所含的项在事务集中同时出现的概率.

定义5 关联规则的置信度(confidence)是指交易集中同时包含 X 和 Y 的交易数与包含 X 的交易数的比值,记为 $\mathrm{conf}(X\Rightarrow Y)$,即

$$\mathrm{conf}(X\Rightarrow Y)=\frac{\sup(X\cup Y)}{\sup(X)}=p\left(\frac{Y}{X}\right).$$

置信度表示在包含 X 的事务中,出现 Y 的条件概率.

一般地,用户可以定义两个阈值,分别为最小支持度阈值和最小置信度阈值. 当挖掘出的关联规则的支持度和置信度都满足这两个阈值时,就认为这个规则是有效的,否则,就是无效的. 这两个阈值一般可由领域专家或用户设定.

定义6 同时满足最小支持度阈值(minsupp)和最小置信度阈值(minconf)的关联规则称为强关联规则. 强关联规则可由频繁项集产生.

1.2.2 关联规则挖掘算法

Agrawal 等在1993年设计了一个基本算法,称为 Apriori 算法. Apriori 算法是一个基于两阶段频繁项集思想的方法,此算法的设计可以分解为两个子问题:① 找到所有支持度大于最小支持度的项集,这些项集称为频繁项集;② 用第①步找到的频繁项集生成期望的规则.

Apriori 算法是一种宽度优先算法,它是通过对数据库 D 的多次扫描来发现所有的频繁项集,在每一次扫描中只考虑具有同一长度 k 的所有 k-项集. 在第一次扫描中,Apriori 算法计算数据库 D 中所有单个项目的支持度,生成所有长度为1的频繁项集(记为 L_1);然后利用 L_1 来挖掘 L_2,即频繁2-项集,如此不断循环,直至不能找到频繁 k-项集为止,其中在发现每个 L_k 的过程中需要对整个事务数据库进行扫描. 本工作利用 Apriori 算法实现了关联规则挖掘,调用微软公司开发的 Microsoft SQL Server 2005 中的关联规则挖掘程序模块[13].

1.3 蛋白质序列的向量化

首先根据理化性质将氨基酸分成不同的类,蛋白质序列根据氨基酸的类别转化为数值序列. 由于半胱氨酸具有比较特殊的理化性质,因此,本工作考虑将半胱氨酸划分为不同的类别,并对预测结果进行了讨论. 基于本工作提出的方法比较了半胱氨酸在不同分类下对预测结果的影响,将氨基酸划分为7类[14]和6类[9],分别记为A类和B类,其中A类是将B类第3类中的半胱氨酸单独作为一类(见表1).

表 1　氨基酸分类

Table 1　Classification of amino acids

分类	氨基酸																			
	C	A	G	V	I	L	F	P	Y	M	T	S	H	N	Q	W	R	K	D	E
A	1	1	1	2	2	2	2	3	3	3	3	4	4	4	4	5	5	6	6	7
B	1	1	1	2	2	2	2	3	3	3	3	4	4	4	4	5	5	6	6	3

为了能够准确预测蛋白质相互作用,必须采用能充分刻画蛋白质序列信息的编码方法. 本工作采用三肽频数方法描述蛋白质序列[14]. 该方法是将蛋白质序列的氨基酸残基赋值后,统计连续 3 个数值的残基,计算它们在序列中出现的数目,作为向量对应位置的坐标. 如果将蛋白质分成 7 类,则向量维数是 7 ×7 ×7;若分成 6 类,则向量维数为 6 ×6 ×6. 假定分成 7 类,则空间维数是 343,d_i 为第 i 个元素,i =1,2,…,343. 蛋白质相互作用的两条序列长度一般不同,在长的蛋白质序列中,某一个 d_i 有可能过大. 为了解决这个问题,利用下式将 d_i 正规化为 f_i,即

$$f_i = \frac{(d_i - \min(d_1, d_2, \cdots, d_{343}))}{\max(d_1, d_2, \cdots, d_{343})},$$

$$i = 1, 2, \cdots, 343. \qquad (1)$$

根据这种编码方法,一条蛋白质序列在 A 分类下,转化成一个 343 维的向量;在 B 分类下,转化成一个 216 维的向量[14].

假设 D_M, D_N 为两个蛋白质的数值序列,D_{MN} 表示它们之间的相互作用,则有

$$\{D_{MN}\} = \{D_M\} \oplus \{D_N\}. \qquad (2)$$

因此,当氨基酸分为 7 类时,686 维[343(一个蛋白质) +343(另一个蛋白质)]的向量表示一对蛋白质的相互作用;当氨基酸分为 6 类时,432 维的向量表示一对蛋白质的相互作用.

1.4　训练集和测试集的选取

分类关联规则挖掘是在已知数据的基础上产生强关联规则,然后用这些强关联规则来预测新的数据. 首先将样本数据划分为训练集和测试集. 小鼠的样本数据训练集由 921 对阳性数据和 921 对阴性数据组成,人的样本数据训练集由 3 319 对阳性数据和 3 319 对阴性数据组成,剩余的阳性数据和阴性数据分别组成小鼠和人的样本数据测试集. 训练集和测试集中的样本数据均为随机选取的.

1.5　数据结果的评价标准

本工作使用准确率、敏感度和特异性 3 项指标来评价蛋白质相互作用预测方法的性能. 先作如下定义:P_T(true positive)表示正确预测蛋白质相互作用对的数目;N_T(true negative)表示正确预测蛋白质非相互作用对的数目;P_F(false positive)表示将蛋白质非相互作用对预测为相互作用对的数目;N_F(false negative)表示将蛋白质相互作用对预测为非相互作用对的数目;N 表示样本集总数,$N = P_T + N_T + P_F + N_F$;准确率 $R_A = (P_T + N_T)/N$;敏感度 $E_S = P_T/(P_T + P_F)$;特异性 $P_S = N_T/(N_T + N_F)$.

1.6　应用分类关联规则预测蛋白质相互作用的流程

本工作应用分类关联规则预测蛋白质相互作用的具体步骤如下(见图 1).

(1) 将训练集中的数据导入数据库.

(2) 对训练集中的数据进行相关性分析,剔除相关性差的变量.

(3) 分别将阳性数据和阴性数据随机分为 5 份,应用 5 次交叉验证.

(4) 调用关联规则挖掘程序模块[12]生成关联规则.

(5) 通过最小支持度阈值和最小置信度阈值对规则进行筛选,生成强关联规则.

(6) 对本次循环生成的强关联规则进行检测,并将检测后的关联规则归入规则集 A 中.

(7) 5 份样本数据依次循环重复步骤(4) ~步骤(6),5 次循环结束后转至步骤(8).

(8) 删除规则集 A 中重复出现的规则.

(9) 应用规则集 A 中的强关联规则对测试集中的数据进行预测.

2　结果与讨论

本工作首先对训练集中的蛋白质序列数据进行相关性分析,剔除相关性差的变量. 小鼠样本数据在 A 分类下通过相关性分析剔除了 108 列,剩余 578 列;在B分类下剔除了69列,剩余363列. 人的样本

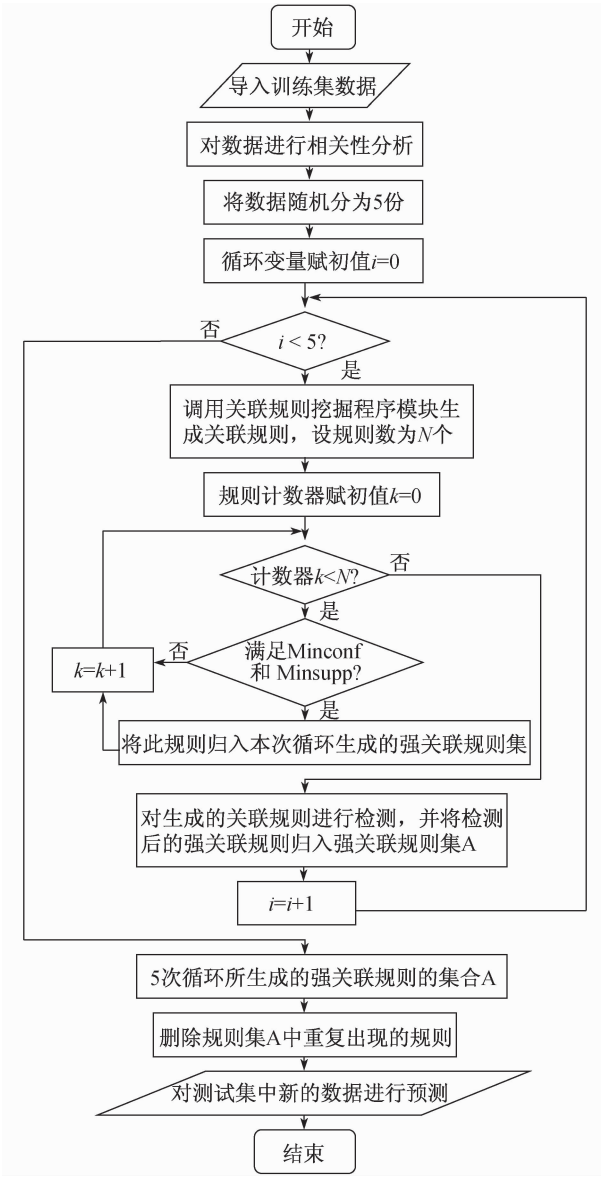

图1 应用关联规则预测蛋白质相互作用的流程图

Fig. 1 Flowchart of the predicting PPIs used association rule

数据在A分类下通过相关性分析剔除了135列,剩余551列;在B分类下剔除了74列,剩余358列. 为了验证模型的有效性,本工作采取了5次交叉验证的方法,即把训练集中的样本数据随机划分为5份,应用其中4份数据生成频繁项集和规则,通过设定最小支持度阈值和最小置信度阈值对规则进行筛选. 选择其中的强关联规则(即满足最小支持度和最小置信度阈值的关联规则),然后由另一份数据对生成的强关联规则进行检测,5份数据依次循环实验5次,并保留每次所生成的频繁项集和强关联规则. 小鼠样本数据氨基酸在A分类下的最小支持度为1 300,共产生1 613个频繁项集;B分类下的最小支持度为500,共产生114个频繁项集. 人的样本数据氨基酸在A分类下的最小支持度为2 300,共产生1 994个频繁项集;B分类下的最小支持度为1 600,共产生571个频繁项集. 表2为小鼠样本数据在A分类下产生的前10个频繁项集.

表2 小鼠样本数据氨基酸在A分类下的频繁项集

Table 2 Frequent itemset under A classifications of mouse data

序号	支持度	项的数目	项集
1	1 873	1	列 566 < 8.203 125E - 02
2	1 871	1	列 290 < 8.203 125E - 02
3	1 813	1	列 287 < 2.175 098E - 03
4	1 792	1	列 337 < 1.917 213E - 03
5	1 773	2	列 317 < 4.516 234E - 03, 列 290 < 8.203 125E - 02
6	1 762	1	列 316 < 5.022 638E - 03
7	1 756	1	列 143 < 3.189 308E - 03
8	1 748	2	列 316 < 5.022 638E - 03, 列 290 < 8.203 125E - 02
9	1 748	2	列 681 < 4.470 794E - 03, 列 587 < 4.101 563E - 02
10	1 747	2	列 319 < 2.883 595E - 03, 列 305 < 0.093 75

表2中的频繁项集按支持度由大到小进行排序,其中支持度代表频繁项集在事务数据库中出现的次数. 表中的项集1为1-项集,表示列566 < 8.203 125E - 02在事务数据库中出现了1 873次;项集5为2-项集,表示列317 < 4.516 234E - 03且列290 < 8.203 125E - 02在事物数据库中出现了1 773次,其余项集依次类推.

由5次交叉验证对挖掘模型的训练可知,在氨基酸两种分类下最小,当最小支持度阈值为0.35,最小置信度阈值为0.60时,准确率达到最大. 小鼠样本数据在A分类下5次验证的平均检测结果为89.45%,共产生了52条强关联规则;B分类下5次验证的平均检测结果为86.37%,共产生了47条强关联规则. 人的样本数据在A分类下5次验证的平均检测结果为83.55%,共产生了57条强关联规则;B分类下5次验证的平均检测结果为80.89%,共产生了54条强关联规则. 小鼠样本数据在A分类下生成的前10条强关联规则如表3所示.

表 3　小鼠样本数据氨基酸在 A 分类下的关联规则

Table 3　Association rules under A classifications of mouse data

序号	概率	重要性	规则
1	1.000	0.402	列 222 = 2.825 183E - 02 ~ 0.137 466→Lebel = yes
2	1.000	0.378	列 129 = 2.623 247E - 02 ~ 0.170 677 1→Lebel = yes
3	1.000	0.402	列 213 = 4.639 892E - 02 ~ 0.186 536 3→Lebel = yes
4	1.000	0.354	列 42 = 1.929 555E - 02 ~ 9.248 666E - 02→Lebel = yes
5	0.995	0.383	列 545 = 0.234 266 6 ~ 0.250 688 1→Lebel = yes
6	0.985	0.471	列 227 = 0.327 075 7 ~ 0.382 406 8→Lebel = no
7	0.975	0.465	列 278 = 0.303 242 1 ~ 0.380 248→Lebel = no
8	0.959	0.365	列 387 = 0.248 763 8 ~ 0.386 202→Lebel = yes
9	0.857	0.394	列 180 = 0.228 412 2 ~ 0.363 971 3→Lebel = no
10	0.805	0.359	列 270 = 0.258 367 ~ 0.402 513 6→Lebel = no

表 3 中概率,即规则的可能性,定义为在给定左侧的情况下右侧的项出现的条件概率;重要性用于度量规则的有用性,值越大,则意味着规则越有用. 表 3 中的规则按概率从大到小进行排序. 表中的规则 1 表示列 222 介于 2.825 183E - 02 ~ 0.137 466 之间,则这个行向量所代表的蛋白质对是相互作用的,且这条规则出现的概率为 1.000,重要性为 0.402;规则 6 表示列 227 介于 0.327 075 7 ~ 0.382 406 8之间,则这个行向量所代表的蛋白质对是不相互作用的,且这条规则出现的概率为 0.985,重要性为 0.471,其余规则依次类推.

利用上述挖掘出的强关联规则对测试集中新的数据进行预测. 小鼠样本数据在氨基酸 A 分类下的预测准确率为 86.96%,B 分类下为84.57%. 人的样本数据在氨基酸 A 分类下的预测准确率为85.78%,B 分类下为 82.95%. 为了进一步说明本方法在预测蛋白质相互作用上的有效性,分别在 A 分类和 B 分类下与 SVM 进行了比较. 应用 Libsvm 软件包,由训练集中的蛋白质序列数据通过支持向量机学习算法建立预测模型,然后对测试集中新的数据进行预测. 关于支持向量机的详细计算过程可参见文献[3-5]. 小鼠样本数据在氨基酸不同分类下两种方法的预测结果比较如表 4 所示. 人样本数据的预测结果比较如表 5 所示.

表 4　小鼠样本数据在氨基酸不同分类下两种方法预测结果的比较

Table 4　Mouse data prediction results comparison for two methods for both A and B classifications　%

分类	方法	R_A	E_s	P_s
A 分类	AR	86.96	87.39	86.52
	SVM	83.48	84.35	82.62
B 分类	AR	84.57	85.22	83.91
	SVM	82.61	83.46	81.74

表 5　人的样本数据在氨基酸不同分类下两种方法预测结果的比较

Table 5　Human data prediction results comparison for two methods for both A and B classifications　%

分类	方法	R_A	E_s	P_s
A 分类	AR	85.78	86.14	85.42
	SVM	85.43	85.90	84.94
B 分类	AR	82.95	83.13	82.77
	SVM	83.07	83.73	82.41

从表 4 可以看出,本方法对小鼠样本数据在氨基酸两种分类下预测的准确率、敏感度和特异性都较 SVM 有了一定的提高,而对人样本数据在氨基酸两种分类下预测的准确率、敏感度和特异性与 SVM 相差不大(见表 5). 由于对蛋白质间相互作用的计算机预测受到多种因素的影响,预测准确率虽然是判断一种方法好坏非常重要的标准,但是其他的一些标准,如算法的稳定性、计算的效率等也可作为评价的指标. 究竟哪一种方法是最优的,还要通过大量的工作进行综合的评价.

尽管如此,从表 4 和表 5 中可以看出,本方法在预测蛋白质相互作用上是有效的,其敏感度、特异性和准确率在两类物种下均达到了 80% 以上. 另外,在两类物种下氨基酸在 A 分类下的预测效果总是好于 B 分类下. 因肽链折叠而处在特定部位的半胱氨酸侧链—SH 氧化形成的二硫键,是某些蛋白质维持结构稳定的重要因素[7],将半胱氨酸单独作为一类是合理的,这可能是在 A 分类下的预测性能优于 B 分类下的一个重要原因.

3 结束语

本工作从蛋白质的一级结构信息出发,利用关联规则挖掘模型预测了蛋白质的相互作用,比较了半胱氨酸在不同分类下模型的预测性能.从蛋白质的一级结构到蛋白质相互作用的高级结构,均受到了多种因素的影响,其中包括物理因素、化学因素以及生物因素.氨基酸的分类对蛋白质相互作用预测有着一定的影响,哪一种分类方法最优,还需要进行大量研究.

尽管目前研究蛋白质相互作用的方法很多,但把关联规则挖掘应用到生物信息学方面将会产生非常好的发展前景.随着对算法的不断改进,蛋白质相互作用实验数据的增加以及对蛋白质相互作用机理的深入研究,一定会对蛋白质相互作用的研究带来更多的信息,使预测的准确率有更大的提高.蛋白质相互作用计算机预测的研究非常具有挑战性,还需要更多工具以及方法的综合与应用.

参考文献:

[1] MCDOWALL M D, SCOTT M S, BARTON G L. Human protein-protein interactions prediction database [J]. Nucleic Acid Research, 2009, 37:651-656.

[2] HAN J W, KAMBER M. Data mining concepts and techniques [M]. San Francosco: Morgan Kaufmann, 2006:1-40.

[3] JANSEN R, YU H, GREENBAUM D, et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data [J]. Science, 2003, 302 (5644):449-53.

[4] WANG J, LI C H, WANG E K, et al. Uncovering the rules for protein-protein interactions from yeast genomic data [J]. PNAS, 2009, 106:3752-3757.

[5] SOONG T T, WRZESZCZYNSKI K O, ROST B. Physical protein-protein interactions predicted from microarrays [J]. Bioinformatics, 2008, 24:2608-2614.

[6] FAWCETT T. An introduction to ROC analysis [J]. Pattern Recognition Letters, 2006, 27:861-874.

[7] 王翼飞,史定华.生物信息学——智能化算法及其应用[M].北京:化学工业出版社,2006:11-18.

[8] 冯铁男,江浩,王翼飞.基于信噪比的蛋白质相互作用的预测[J].上海大学学报:自然科学版,2008,14(6):604-610.

[9] 秦殿刚,高松,冯铁男,等.通过序列编码预测蛋白质相互作用[J].应用科学学报,2009,27(6):601-605.

[10] GIONIS A, MANNILS H, MIELIKAINEN T, et al. Assessing data mining results via swap randomization [C] // Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Date Mining. 2006:167-176.

[11] VAPIK V. The nature of statistical learning theory [M]. 2nd ed. New York: Springer, 2000:96-99.

[12] HAN J, PEI J. Mining frequent patterns without candidate generation: a frequent-patter tree approach [J]. Data Mining Knowledge Discovery, 2004, 8:53-87.

[13] 谢邦昌.商务智能与数据挖掘 Microsoft SQL Server 应用[M].北京:机械工业出版社,2008.

[14] SHEN J W, ZHANG J, LUO X M, et al. Predicting protein-protein interactions based only on sequences information [J]. PNAS, 2007, 104(11):4337-4341.