doi:10.3969/j.issn.1672-5565.2014.01.08

基于共词分析的国内生物信息学热点领域研究

宋茂海[1,2],李东方[2*]
(1.第二军医大学基础部生物信息学教研室,上海 200433;
2.第二军医大学基础部计算机教研室,上海 200433)

摘　要:利用共词分析和可视化方法对生物信息学的关键词进行聚类分析,探讨该研究领域的学科分类和热点内容。以中国知网、中华医学会数据库中期刊论文为统计来源,对1998~2013年间的5 707篇生物信息学相关文献进行计量分析,提取出40个高频关键词。利用ROST软件得到关键词共词矩阵,在此基础上利用SPSS进行因子分析、聚类分析和多维尺度分析。结合因子分析和聚类分析将生物信息学领域主要研究内容分为7类,结合多维尺度分析对研究热点及变化趋势进行了初步探讨。研究结果较为客观地反映了当前生物信息学领域的学科分类和研究热点,为科研人员进行生物信息学研究提供一些思路。

关键词:文献计量学;共词分析;聚类分析;数据挖掘
中图分类号:G350; R857　　**文献标志码**:A　　**文章编号**:1672-5565(2014)-01-046-07

Hot spots analysis of China's bioinformatics based on co-word analysis method

SONG Maohai[1,2], LI Dongfang[*]
(1. *Department of bioinformatics*, *the Second Military Medical University*, *Shanghai* 200433;
2. *Department of computer Science*, *the Second Military Medical University*, *Shanghai* 200433)

Abstract: Using co-word analysis and visual method, we clustered the key words of bioinformatics, and revealed the subject classification and hotspots of this subject. We got 5 707 bioinformatics related articles in the database of China National Knowledge Infrastructure (CNKI) and the Chinese medical association from 1998 to 2013. From these, we extracted 40 high frequency keywords. Based on these keywords, we established the co-word matrix using ROST software, and processed these data by factor analysis, clustering dendrogram, Multidimensional Scaling diagram using SPSS software. We divided the topics of bioinformatics research content into seven categories using factor analysis and cluster dendrogram. We also discussed the hotspots and tendency preliminarily using multidimensional scaling diagram. The results reflected the current subject classification and research hotspots objectively in the field of bioinformatics; our conclusions could provide a certain reference to scientific research community for bioinformatics.

Keywords: Bibliometrics; Co-word analysis; Cluster analysis; Data mining

利用信息计量学对某一领域的论文进行统计分析,归纳出该学科的研究分类、结构与范式,对于规划学科布局,促进学科发展,调整科研方向具有重要的参考价值[1]。共词分析作为信息计量方法的一种,通过主题分析能直观地揭示学科微观结构,其原理是当两个学科领域内的关键词在一篇文献中同时出现时,表明这两个词之间具有一定的内在关系,出现的次数越多,表明它们的关系越密切[2-3]。在此基础上,利用因子分析、聚类分析和多维尺度分析等多元分析方法,按照关键词之间的"距离"将某一领域内关键词加以分类,从而揭示学科领域的发展与演进趋势、课题研究的扩散与传播关系[4-6]。本文采用共词分析方

收稿日期:2013-10-31;**修回日期**:2013-11-13.
作者简介:宋茂海,男,讲师,研究方向:生物信息学与数据挖掘;E-mail:mhsong@smmu.edu.cn.
*通信作者:李东方,男,教授,研究方向:计算机教育与生物信息学;E-mail:dfli@smmu.edu.cn.

法,通过分析期刊论文的关键词,考察近十年来我国生物信息学的研究分类和发展趋势[7]。

1　数据来源

本文选择中国知网学术期刊网络出版总库、中国重要会议论文全文数据库、国际会议论文全文数据库和中华医学会/中国医师协会全文期刊库为数据源,以“关键词”为检索途径,以“生物信息学”为检索词,采用“精确”检索方式,共检索到 1998~2013 年 3 月相关期刊论文 5 707 篇(去除无关键词的论文及会议通知、征稿启示等文献),论文的年份分布见表 1。

表 1　1998~2013 年 3 月生物信息学文献年份分布

Table 1　Distribution of bioinformatics articles between 1998 and 2013

年份	论文数量	年份	论文数量
1998	39	2006	409
1999	55	2007	463
2000	135	2008	489
2001	155	2009	647
2002	242	2010	650
2003	318	2011	624
2004	342	2012	670
2005	327	2013.1~3	81

2　数据处理和分析

2.1　高频关键词确定

关键词作为一篇论文的元数据,是文章核心内容的浓缩和提炼。对 5 707 篇期刊论文进行数据统计,共提取关键词 27 402 个。去除不参与后期分析的“生物信息”、“生物信息学”关键词,合并“蛋白质组”、“蛋白质组学”,“miRNA”、“microRNA”等同义关键词,按词频由高到低排序,选择前 40 个关键词作为分析对象(见表 2)。这 40 个高频关键词共累计出现 3 891 次,占论文总数的 68.2%,在一定程度上能体现国内生物信息学的研究现状。

2.2　共词矩阵与相关矩阵

利用 ROST 数据挖掘软件对 40 个关键词进行两两共词检索,统计其在所有论文中同时出现的次数,形成一个 40×40 的共词矩阵,对角线上的数值为该关键词在所有论文中出现的次数,非对角线上的数值表示两个关键词共同出现在同一篇论文中的次数(见表 3、表 4)。

表 2　1998~2013 年生物信息学文献高频关键词表

Table 2　High frequency keywords sheet of bioinformatics between 1998 and 2013

序号	关键词	频次	序号	关键词	频次
1	蛋白质组学	498	21	水稻	59
2	基因克隆	360	22	质谱	58
3	基因组学	345	23	日本血吸虫	48
4	人类基因组计划	229	24	双向凝胶电泳	47
5	数据库	187	25	双向电泳	46
6	序列分析	153	26	计算生物学	46
7	基因芯片	131	27	肿瘤	45
8	基因表达	129	28	抗原表位	45
9	生命科学	124	29	预测	44
10	microRNA	100	30	新基因	43
11	电子克隆	97	31	转录因子	43
12	序列比对	96	32	生物芯片	42
13	功能基因组学	83	33	原核表达	41
14	基因表达谱	76	34	靶基因	40
15	数据挖掘	72	35	比较基因组学	39
16	生物技术	72	36	蛋白质相互作用	39
17	表达序列标签	72	37	支持向量机	38
18	启动子	66	38	基因组研究	37
19	系统生物学	65	39	单核苷酸多态性	36
20	分子生物学	62	40	乳腺癌	36

表 3　生物信息学文献高频关键词共词矩阵(部分)

Table 3　Co-word matrix of bioinformatics high frequency keywords

序号	1	2	3	4	5	6	7	8	9	10
1	498	1	59	33	20	2	10	7	14	0
2	1	360	11	2	2	28	0	26	1	0
3	59	11	345	31	28	7	10	5	17	1
4	33	2	31	229	29	2	10	7	42	0
5	20	2	28	29	187	2	4	4	8	0
6	2	28	7	2	2	153	1	3	2	0
7	10	0	10	10	4	1	131	9	2	5
8	7	26	5	7	4	3	9	129	2	1
9	14	1	17	42	8	2	2	2	124	0
10	0	0	1	0	0	0	5	1	0	100

为了消除频次悬殊造成的影响,用 Ochiia 相似系数将共词矩阵转换成相关矩阵[8]。即将共词矩阵中的每个数值都除以与之相对行列的两个词频总数乘积的平方根。

其计算公式为:$\text{Ochiia}=\frac{\overline{AB}}{\sqrt{\sum A\times\sum B}}$

表 4 生物信息学文献高频关键词相关矩阵(部分)

Table 4 Correlation matrix of bioinformatics high frequency keywords

序号	1	2	3	4	5	6	7	8	9	10
1	1.000	0.002	0.142	0.098	0.066	0.007	0.039	0.028	0.056	0.000
2	0.002	1.000	0.031	0.007	0.008	0.119	0.000	0.121	0.005	0.000
3	0.142	0.031	1.000	0.110	0.110	0.030	0.047	0.024	0.082	0.005
4	0.098	0.007	0.110	1.000	0.140	0.011	0.058	0.041	0.249	0.000
5	0.066	0.008	0.110	0.140	1.000	0.012	0.026	0.026	0.053	0.000
6	0.007	0.119	0.030	0.011	0.012	1.000	0.007	0.021	0.015	0.000
7	0.039	0.000	0.047	0.058	0.026	0.007	1.000	0.069	0.016	0.044
8	0.028	0.121	0.024	0.041	0.026	0.021	0.069	1.000	0.016	0.009
9	0.056	0.005	0.082	0.249	0.053	0.015	0.016	0.016	1.000	0.000
10	0.000	0.000	0.005	0.000	0.000	0.000	0.044	0.009	0.000	1.000

2.3 多元统计分析

将相关矩阵的数据导入 SPSS 19.0,进行多元统计分析,包括因子分析、聚类分析和多维尺度分析。

2.3.1 因子分析

因子分析通过研究众多变量之间的内部依赖关系,探求观测数据中的基本结构,并以最少的信息丢失将多个变量化为少数几个综合变量,原始的变量是可观测的显在变量,而假想变量是不可观测的潜在变量,称为因子。将表 4 的相关矩阵的数据导入 SPSS,选择主成分法(Principal components)进行因子分析得到各行的特征根、方差(见表 5)和碎石图(见图 1)。

通过因子矩阵的总方差表,可见有 18 个主成分被提取,这些主成分累积解释全部信息的 61.17%。从载荷因子分布情况来看,因子分析结果中的关键词分布比较离散,若严格按照载荷因子大于 1 的条件分类,则类别将多达 18 个,不利于分析讨论;若按图 1 曲线的拐点位置来分类,则类别只有 4 个,也不便于展开讨论。因此,综合因子矩阵和碎石图分析结果[9-10],结合其他高频关键词的特点,选取因子载荷大于 1.3 的主成分进行分类,可将 40 个关键词归为 7 类。

表 5 生物信息学文献相关矩阵的因子分析

Table 5 Factor analysis of correlation matrix of bioinformatics

Component	Total Variance Explained					
	Initial Eigenvalues			Extraction Sums ofSquared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.474	6.186	6.186	2.474	6.186	6.186
2	2.386	5.964	12.150	2.386	5.964	12.150
3	1.684	4.209	16.359	1.684	4.209	16.359
4	1.484	3.710	20.069	1.484	3.710	20.069
5	1.391	3.478	23.547	1.391	3.478	23.547
6	1.337	3.342	26.889	1.337	3.342	26.889
7	1.297	3.242	30.131	1.297	3.242	30.131
8	1.266	3.165	33.297	1.266	3.165	33.297
9	1.254	3.136	36.432	1.254	3.136	36.432
10	1.224	3.060	39.493	1.224	3.060	39.493
11	1.174	2.936	42.429	1.174	2.936	42.429
12	1.145	2.864	45.292	1.145	2.864	45.292
13	1.104	2.759	48.051	1.104	2.759	48.051
14	1.083	2.708	50.759	1.083	2.708	50.759
15	1.070	2.676	53.435	1.070	2.676	53.435
16	1.060	2.650	56.085	1.060	2.650	56.085
17	1.028	2.570	58.655	1.028	2.570	58.655
18	1.008	2.519	61.174	1.008	2.519	61.174
19	.990	2.475	63.648			

Extraction Method: Principal Component Analysis.

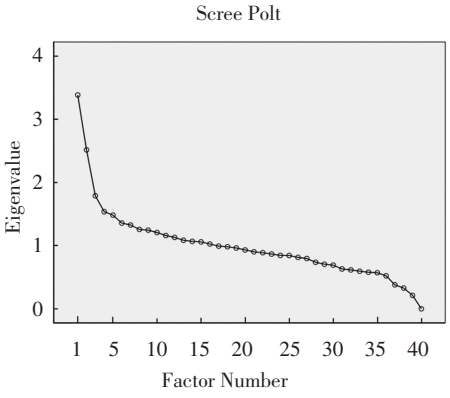

图 1　生物信息学文献高频关键词碎石图

Fig.1　Scree plot of bioinformatics high frequency keywords

2.3.2　聚类分析

聚类分析是一组将研究对象分为相对同质的群组的统计分析技术,其基本思想是把相似程度较大的变量聚合为一类,把另外一些相似的变量聚合为另一类,关系密切的聚合到一个小的分类,关系疏远的聚合到一个大的分类,直到把所有的变量都聚合完毕,最后再把整个分类系统画成一张谱系图,用它把所有变量间的亲疏关系表示出来[11]。图 2 是生物信息学高频关键词聚类分析树形图,显示了各关键词之间的关联程度,上端 0~25 的代表各类之间的距离,越早被聚为一类的关键词之间的距离越近,关联越紧密。

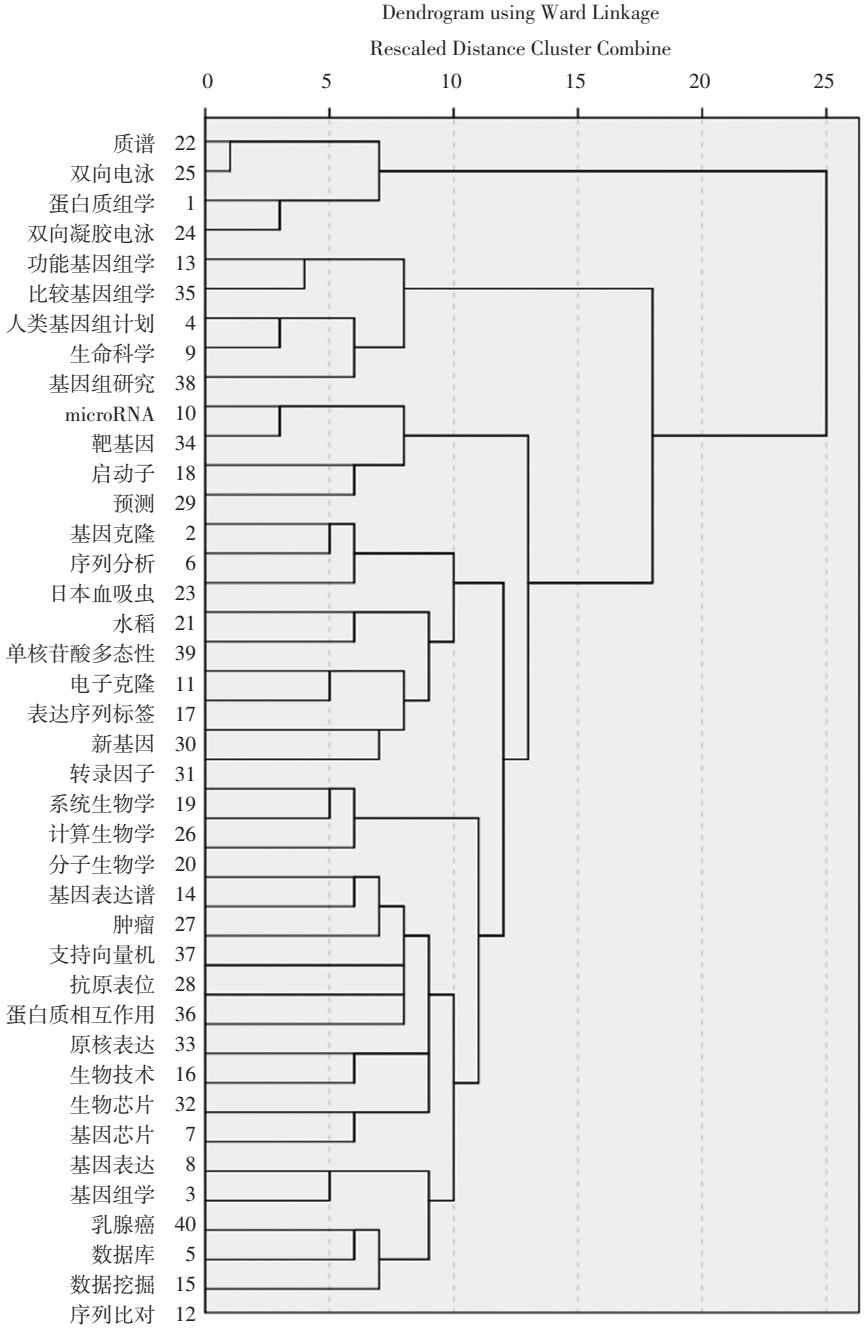

图 2　生物信息学文献高频关键词聚类分析树形图

Fig.2　Cluster dendrogram of bioinformatics high frequency keywords

依据聚类过程同时参考因子分析结果,本研究所用的高频关键词可分为以下 7 类:

(1)蛋白质组学分析。蛋白质组学直接研究编码基因翻译出的蛋白质产物,比转录组学注释基因组获得的结果更直接。蛋白质特有的翻译后处理现象使得蛋白质组学在提供基因表达产物、确认和校正编码基因、解析翻译后处理现象,以及发现新的编码基因及其规律上拥有先天的优势[12]。

(2)系统生物学分析。系统生物学是研究基因和蛋白质的一种新方法,和传统生物科学研究单个基因或者蛋白质不同,系统生物学研究的是生物信息(DNA、mRNA、蛋白质、功能蛋白、生物信息途径、生物信息网络)在所有水平上复杂的相互作用,重点考察这些生物信息是如何一起工作的[13]。

(3)功能基因组学分析。基因组学的研究已从建立高分辨遗传、物理和转录图谱为主的结构基因组学转向功能基因组学。功能基因组学主要研究 DNA 序列变异性、基因组表达调控、模式生物体和生物信息平台与数据库构建[14]。

(4)microRNA 研究分析。microRNA 主要与靶 mRNA 分子的 3' 非编码区的不完全互补序列结合,通过靶向降解 mRNA 或抑制 mRNA 翻译,达到基因沉默的调控效果[15]。近年来,随着测序技术的发展和多种分子生物学实验手段的结合,越来越多的 microRNA 相继被发现,相应的表达变化、作用机制等后续研究正在迅速兴起。

(5)基因克隆表达分析。基因克隆技术把来自不同生物的基因同有自主复制能力的载体 DNA 在体外人工连接,构建成新的重组 DNA,然后送入受体生物中去表达,从而产生遗传物质和状态的转移和重新组合,再进行基因相关结构、功能的研究。

(6)电子克隆研究。电子克隆是利用生物信息学手段进行基因克隆的新方法,它借助计算机的高速运算能力,通过 EST 或基因组的序列组装和拼接,利用 RT-PCR 方法快速获得新基因,具有投入低、速度快、针对性强等优点[16]。电子克隆技术成为基因工程中获得新基因的重要手段,对开展人类基因功能的研究,在基因水平上预防疾病具有重要的意义和价值。

(7)基因的数据挖掘分析。高通量测序带来了海量的核酸及蛋白质序列数据,人们很难直观地解读这些高维数据中的信息[17-18]。利用计算机科学及应用数学知识,通过降维、关联分析、分类和识别等数据处理方法,更好地理解基因表达谱、预测基因功能、分子结构和优化先导分子等。

2.3.3 多维尺度分析

多维尺度分析是一种通过二维空间展现关键词之间的联系,利用平面距离来反映关键词之间的相似程度,同时又保留数据对象间原始关系的数据分析方法[19]。根据因子矩阵,利用 SPSS 进行多维尺度分析并加以整理得出多维尺度图,如图 3 所示。图中,有高度相似性的点聚集到一起形成一类,并且越居中的关键词与其他关键词的联系越多,在该领域中的地位越核心。

分析生物信息学高频关键词在多维尺度图上的分布情况。其中,"蛋白质相互作用"关键词靠近图形中心,说明蛋白质组学是生物信息学研究的热点方向。另外,系统生物学和比较基因组学、基因芯片、计算生物学研究仍将是今后的热点和方向。

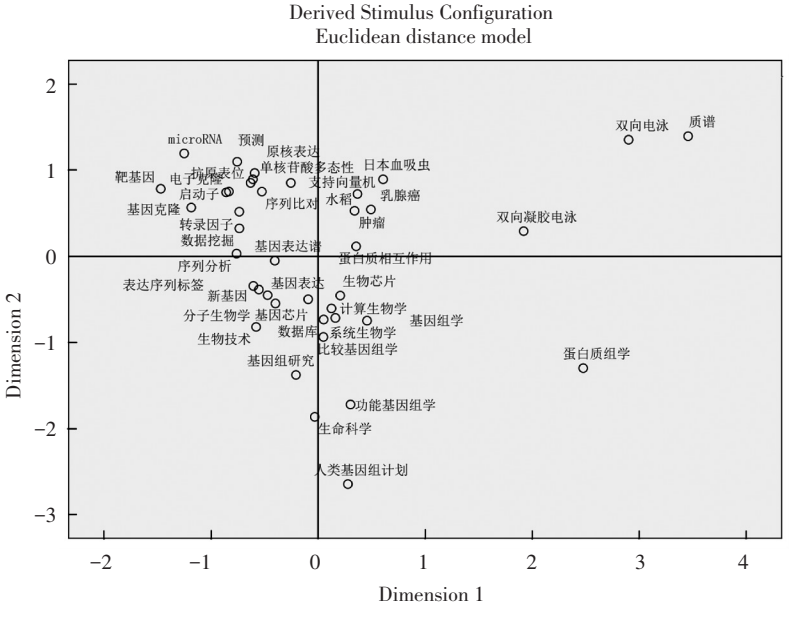

图 3 生物信息学文献高频关键词多维尺度图

Fig.3 Multidimensional scale diagram of bioinformatics high frequency keywords

3 结　论

本文在提炼生物信息学期刊论文 40 个高频关键词的基础上,运用共词分析方法,通过因子分析,聚类分析和多维尺度分析,探讨了生物信息学研究的结构、关注的热点和研究趋势,得出该领域研究颇受关注的 7 个类别。由于论文发表的时滞性,特别是国内和国外研究热点的时滞性,单纯通过关键词列表进行统计分析存在一定的偏差。另外,有些新出现的关键词,因出现频次较低,未能引起共词分析方法的"注意",所以分析时还要结合时间序列,才能更精确地预测未来的研究热点。

参考文献(References)

[1] 邱均平.信息计量学(九):第九讲 文献信息引证规律和引文分析法[J].情报理论与实践,2001,24(3):236-240.
QIU Junping. Bibliometrics (IX): Document Information Law Citations and Citation Analysis [J]. Information Studies: Theory & Application, 2001,24(3):236-240.

[2] 郭文姣,欧阳昭连,李阳,等. 应用共词分析法揭示生物医学工程领域的研究主题[J]. 中国生物医学工程学报, 2012, 31(4): 545-551.
GUO Wenjiao, OUYANG Zhaolian, LI Yang, et al. Revealing Theme Structure of Biomedical Engineering Using Co-Word Analysis [J]. Chinese Journal of Biomedical Engineering, 2012, 31(4): 545-551.

[3] 朱安青,周金元. 我国科技查新研究热点及趋势分析——共词分析视角[J]. 图书情报研究, 2009, 2(4): 45-49.
ZHU Anqing, ZHOU Jinyuan. Co-Word Analysis of Sci-Tech Novelty Retrieval Research in China [J]. Library & Information Studies, 2009, 2(4): 45-49.

[4] LIN S M, MCCONNELL P, JOHNSON K F, et al. MedlineR: an open source library in R for Medline literature data mining [J]. Bioinformatics, 2004, 20(18): 3659-3661.

[5] KRALLINGER M, ERHARDT R A A, VALENCIA A. Text-mining approaches in molecular biology and biomedicine [J]. Drug discovery today, 2005, 10(6): 439-445.

[6] ZHANG J, JASTRAM I. A study of metadata element co-occurrence [J]. Online Information Review, 2006, 30(4): 428-453.

[7] 朱杰.生物信息学的研究现状及其发展问题的探讨[J],生物信息学,2005,3(4):185-188.
ZHU Jie. Bioinformatics' Status in Quo and Its Development in the Future [J]. China journal of Bioinformatics, 2005,3(4):185-188.

[8] 许梅华.基于共词分析的近年国内发展心理学研究热点分析[J].现代情报,2010,30(8):171-175.
XU Meihua. Hot Spots Analysis of China's Developmental Psychology Based on Co-Words Analysis Method [J]. Journal of Modern Information, 2010,30(8):171-175.

[9] 张晗,韩爽,白星,等.利用遗传算法确定医学文献的研究热点[J].现代图书情报技术,2011,(3):57-61.
ZHANG Han, HAN Shuang, BAI Xing, et al. Application of Genetic Algorithm to Identify Hot Topics from Medical Literature [J]. New Technology of Library and Information Service, 2011,(3):57-61.

[10] 刁雪涛,张小芳,宋洁,等.生物信息学研究进展[J].安徽农学通报,2008,14(22):160-162.
DIAO Xuetao, ZHANG Xiaofang, SONG Jie, et al. Advances in Bioinformatics Research [J]. Anhui Agriculture Science Bulletin, 2008,14(22):160-162.

[11] 曹利霞, 葛淼, 何进伟. 主成分分析法评估地理分布对成年人肺顺应性参考值的影响[J]. 第二军医大学学报, 2009, 30(1): 35-39.
CAO Lixia, GE Miao, HE Jinwei. Principal Component Analysis of Geographic Influence on Adult Lung Compliance [J]. Academic Journal of Second Military Medical University, 2009, 30(1): 35-39.

[12] 张昆,王乐珩,迟浩,等. 蛋白质基因组学:运用蛋白质组技术注释基因组[J].生物化学与生物物理进展,2013,40(4):297-308.
ZHANG Kun, WANG Leheng, CHI Hao, et al. Proteogenomics: Improving Genomes Annotation by Proteomics [J]. Progress in Biochemistry and Biophysics, 2013,40(4):297-308.

[13] 资治科,孙之荣.系统生物学: 面向系统的生物学研究[J].系统工程理论与实践,2005,(2):47-55.
ZI Zhike, SUN Zhirong. Systems Biology: System-oriented Biological Research [J]. Systems Engineering-Theory & Practice, 2005,(2):47-55.

[14] STEIN L. Genome annotation: from sequence tobiology [J]. Nat Rev Genet, 2001, 2(7): 493-503.

[15] 赵海苹,罗玉敏.微波 RNA-144 的研究进展[J].首都医科大学学报,2013,34(1):80-85.
ZHAO Haiping, LUO Yumin. Progress in Studies of MicroRNA-144-Associated Diseases and Related Mechanism [J]. Journal of Capital Medical University, 2013, 34(1):80-85.

[16] 王冬冬,朱延明,李勇,等.电子克隆技术及其在植物基

因工程中的应用[J].东北农业大学学报,2006,37(3):403-408.
WANG Dongdong, ZHU Yanming, LI Yong, et al. Application of in Silico Cloning Technique in Plant Gene Engineering [J]. Journal of Northeast Agricultural University, 2006,37(3):403-408.

[17] 黄子夏,柯才焕,陈军.大规模 GO 注释的生物信息学流程[J].厦门大学学报(自然科学版),2012,51(1):139-143.
HUANG Zixia, KE Caihuan, CHEN Jun. Bioinformatics Procedure of Large-Scale GO Annotation [J]. Journal of Xiamen University (Natural Science), 2012,51(1):139-143.

[18] BRENT M R. Genome annotation past, present and future: how to define an ORF at each locus. Genome Research. 2005,15 (12): 1777-1786.

[19] 赵守盈,吕红云.多维尺度分析技术的特点及几个基础问题[J].中国考试,2010,(4):13-19.
ZHAO Shouying, LÜ Hongyun. The Characteristic and Several Basic Problem of Multidimensional Scaling Analysis [J]. China Examinations, 2010,(4):13-19.