第 20 卷 第 5 期 云南农业大学学报 Vol. 20 No. 5
2005 年 10 月 Journal of Yunnan Agricultural University Oct. 2005

遗传算法在通用试题库自动组卷中的应用

张亚静, 杨 毅, 邱 靖, 白春雨
(云南农业大学工程技术学院, 云南 昆明 650201)

摘要: 对组卷问题进行分析, 建立了通用组卷模型, 讨论了运用遗传算法求解在一定约束条件下的多目标参数优化问题, 并针对组卷问题设计了交叉、变异算子和进化模式。结果表明, 改进后的遗传算法性能好、效率高、通用性高, 符合自然界物种微进化的思想, 具有较好的实用性能。

关键词: 遗传算法; 自动组卷; 通用试题库
中图分类号: G 424.79 **文献标识码**: A **文章编号**: 1004 - 390X(2005)05 - 0714 - 06

Application of General Item Bank in Auto-generating Paper Based on GA

ZHANG Ya-jing, YANG Yi, QIU Jing, BAI Chun-yu
(College of Engineering and Technology, Y A U, Kunming 650201, China)

Abstract: This paper analyses the problem of auto generating paper and discusses the multi-object parameter optimizing problem with restrictions which solved with genetic algorithm, It also makes some improvements to the tradition genetic algorithm. The result of the experiment shows that the improved genetic algorithm has high quality and efficiency. It matches the thought of the tiny evolution; It is an applicable method with satisfactory performance.

Key words: genetic algorithm; auto generating paper; exercises store

随着科技的快速发展, 计算机辅助教学(CAI)也得到快速发展, 建立一个好的计算机题库管理系统尤为重要, 而良好的组卷算法却是计算机题库管理系统的核心。

目前, 大约有 4 种组卷方式[1~5]: 包括人工组卷法, 随机选取法, 回溯法, 遗传算法等。前 3 种方法都存在组卷速度慢、质量低、成功率也较低等缺陷。而遗传算法是模拟生物在自然环境中的遗传和进化过程而形成的一种自适应全局优化概率搜索算法, 具有并行性、通用性、全局优化性、稳健性、操作性与简单性等特点。

本文在对组卷问题模型化的基础上, 提出了利用遗传算法求解自动组卷的新思路, 并在模拟环境下实现。对遗传算法在自动组卷应用中的若干问题进行了全面研究, 讨论了各种参数及试题分布对算法的影响及同一参数组卷的试卷重复率等。试验结果表明, 所设计的组卷方法具有收敛速度快、性能好、效率高、适于并行处理等特点, 是一种实用、有效的组卷方法。

1 自动组卷问题

自动组卷即在题库系统中, 由计算机按照用户设定的试卷参数和要求, 从试题库中搜索组织出一份符合条件的试卷。

1.1 与组卷有关的试题参数

一套试卷的构成涉及很多因素, 通过对大量不同学科试卷特点和教师实际出卷方式的分析, 笔者设定了自动组卷试卷构成的 3 组参数: 试题类型, 难度和知识点。其中, 知识点的设置可不以章节为依据, 从而可以淡化教材版本对组卷的影响。知识点的划分采用树形分层的方法, 每个叶子结点(知识点)深度不一定都相同。

收稿日期: 2005 - 04 - 26
作者简介: 张亚静(1981 -), 女, 河北石家庄人, 在读硕士研究生, 主要从事计算机应用方面的研究。

1.2　组卷模型

题库中每道试题包含 4 个属性:题号、题型、难度、知识点。为简单起见,设题库试题按照题型排列,并按序编号,题型即构成了试题库的分段,题型为 5 种,难度为 5 级,知识点为 8 个(实际应用中很容易推广到一般情况)。用户组卷时选定各种题型的数量及分值、各知识点分值及各种难度的分值。当抽取 n 道题时,就决定了一个 $n \times 3$ 的参数矩阵 S。矩阵 S 的每一列代表一个参数,每一行代表一道试题。a_{ij} 表示第 i 套题在第 j 个参数上的每小题累计分值。

$$S = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & a_{n3} \end{bmatrix}$$

2　遗传算法在自动组卷中的应用

遗传算法(GA)[6~9] 最早是由 HOLLAND 于 20 世纪 60 年代提出,是一种模拟生物在自然选择和自然遗传机制的随机优化算法,它是从一个初始种群出发,不断重复执行选择、交叉和变异的过程,使种群进化越来越接近某一目标。

2.1　遗传算法主要运算过程

(1) 随机产生初始群体。设置进化代数计数器 $t \leftarrow 1$;设置最大进化代数 T;随机生成 M 个个体作为初始群体 $P(0)$。

(2) 个体评价,利用适应度函数对个体计算函数值。

(3) 将群体 $P(t)$ 经过选择交叉,变异运算之后得到下一代群体 $P(t+1)$。

(4) 终止条件判断,$t \leqslant T$,则 $t \leftarrow t+1$,转到(2),若 $t > T$,则以进化过程中所得到的具有最大适应度的个体,作为最佳解,终止计算。

2.2　遗传算法求解自动组卷的关键问题

自动组卷问题是一个多目标优化问题,此类问题的复杂度随目标的复杂而增加。如果用传统遗传算法以题型(5 种)、难度(5 级)、知识点(8 个)、总分(100) 为优化目标进行优化,不但进化搜索效率极低,需要的平均过程较长,且易过早收敛,难以得到用户能接受的满意解。因为总分、题型、题量应该是精确达到的,任何近似的满意解都是不可行的。在传统的遗传算法中初始种群个体中“1” 的数目等于用户所需试题的数量,但进行了交叉、变异后,很有可能使串中“1” 的个数大于或小于 n,从而不符合用户的需求,成为非法解。对非法解的修正,会使运算变得复杂,大大影响搜索效率。所以在组卷模型中我们以题型、题量和分值为基础,在形成初始种群和进化的过程中始终保持题型、题量和分值不变,简化了优化目标。另外,如何保护群体多样性,同时又要考虑加快算法的收敛速度,也是进化能否成功的关键。因此,设计一个较简化的模型和较好的算法至关重要。

3　遗传算法在组卷问题中的改进[10~16]

对遗传算法的改进遵循自然界物种微进化论的思想,即进化保持在种内而非跨种进行,它的子代比父代更能适应环境,始终比父代较优。据此,在改进后的遗传算法中,使用了功能块的思想,把每类题型放在同一块,在进行选择、交叉、变异时,保持在功能块内部,这样个体就不可能进化到其他种类,保证了组卷时优化目标中题型、题量、分值的正确匹配。对个体的编码方法及各种算子设计如下:

(1) 编码方法

为了在交叉和变异时不影响每种题型的题目数量,所以用功能块的思想来进行编码。即将整个种群中的个体按照题目类型划分不同的功能块,每个功能块对应一种题型。初始化种群时,随机产生 n 个个体(n 套试卷),每套试卷不是以传统的 1 与 0 来进行编码,而是以题号来编码,传统的二进制编码对问题的表现不直观,且减慢了算法的运行速度。以题号编码的方法所表达的变量意义清楚、明确、不需解码,不会存在编码变换而导致的优化搜索空间急剧增大,从而降低算法性能的问题,能有效改善遗传算法的复杂性,提高运算效率。

(2) 选择算子

适应度函数是度量个体适应度的函数,在遗传算法中是以适应度大小来区分群体中个体的优劣,一般情况下,适应度越大,说明个体越好。本文采用线性代数中欧氏空间距离[2] 的思想来对种群中的个体进行优胜劣汰。在该组卷问题中,把用户需求的参数定为 n 维向量 $\alpha(\alpha_1,\alpha_2,\cdots,\alpha)$,把种群中的个体的各属性看为 n 维向量 $\beta(\beta_1,\beta_2,\cdots,\beta)$,则第 i 个个体向量 β 与向量 α 的欧氏距离:$d_i = \sqrt{(\alpha_1-\beta_1)^2-(\alpha_2-\beta_2)^2-\cdots-(\alpha_n-\beta_n)^2}$。如果欧氏距离越短,适应度越大,则该个体与用户需求越接近。

在算法实现时,把模型中的矩阵 S 增加一行向量,来表示每个个体与用户需求的适应度大小,即矩阵 S 的第 4 列 $a_{i4}(a_{14},a_{24}\cdots a_{n4})(1\leqslant i\leqslant n)$。矩阵 S 就变为 $n\times 4$ 的矩阵 $S(a_{ij})(1\leqslant i\leqslant n,\ 1\leqslant j\leqslant 4)$

$$S=\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & a_{n3} & a_{n4} \end{bmatrix}$$

两向量距离越短则该个体适应度越高,说明越满足用户的需求。对适应度进行排序(从大到小),把前两个个体(距离最短) 直接进入下一代,而把最后两个个体(距离最大) 淘汰。每次交叉变异后,都对种群中的每个个体重新排序,选择前 n 个个体进行交叉变异。这样按照优胜劣汰的规则将适应度较高的个体更多地遗传到下一代。最终达到或接近于最优解。

(3) 交叉算子

在交叉中采用了功能块的思想,即按题型进行划分。题库中有 x 种题型,就把每个个体(每组试卷) 划分为 x 个功能块。把所有个体进行随机配对,用双重循环使每个个体都有机会面对面(即交叉)。但交叉是否进行还是依赖于交叉概率 Pc(Pc 一般为0.4 ~ 0.99),如果随机产生的0 ~ 1之间的随机数小于 Pc,则进行交叉,再随机产生 0 ~ x[random(x)] 的随机数 i(i 为整数),那么交叉就在第 i 个功能块进行,即把两个个体的第 i 个功能块的试题全部交换(见图 1)。这就保证了题型数量的不变,从而达到了局部搜索的目的。

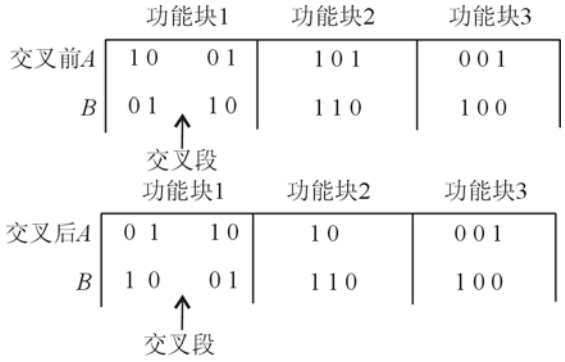

图 1 功能块内交叉

Fig. 1 Crossover in Function block

(4) 变异算子

遗传算法中使用变异算子主要是维持种群体的多样性,防止出现早熟现象。在组卷时,选择了一定的变异概率(Pm),随机产生 0 ~ 1 的随机数,如果该随机数小于 Pm,则进行变异,再随机产生 0 ~ x[random(x)] 的随机数 i(i 为整数),变异就在第 i 个功能块进行。因为每个功能块中的题号的范围已确定,产生第 i 个功能块题号范围内的随机数,如果该随机数已被该个体选中,就重新再产生随机数,直到该随机数没被该个体选中,此时就用该随机数代替该功能块中的任意一道题。在进行变异后,如子代的适应度大于父代的适应度,则用子代替换父代,否则,则保留父代。在进行交叉和变异后,种群中的个体会增加,这就保证了个体的多样性,达到了全局搜索的目的。

(5) 进化完成条件

由于对不同的种群与用户的要求不同,所以,在遗传算法中并不能规定种群进化到了第几代就得到了最优个体(试卷的最优组合)。因此,笔者在遗传算法中设一个标志 f,用来标志群体中是否仍然有个体在进行交叉和变异,也即种群中的个体是否仍然在进化,是否向更优解的方向接近。如 $f=1$,则还有更优个体;否则,就已找到了最优解,种群进化到这一代就结束了,程序满足条件退出。

4 结果与分析

4.1 算法流程图

算法流程如图 2 所示。

4.2 试验初始情况及运行条件统计

为验证该遗传算法的可行性与有效性,以《数据库系统概论》课程为例,在试验中题库共有 1 000 道,单选、填空、简答、编程、设计 5 种题型,8 个知识点。设置试卷中各题型所占的分值比例为:1∶2∶3∶5∶5;题目数量的比例为:20∶10∶10∶4∶2;8 个知识点所占的比例为:10∶20∶10∶20∶10∶10∶10∶10;难∶较难∶中∶较易∶易为:5∶15∶50∶20∶10。

4.3 组卷试验

试验 1:研究改进后的遗传算法的组卷程序的收敛性及不同输入参数对程序收敛性的影响。在试验中,每次固定两种参数,改变另一个参数,用该组卷算法对题库中的试题进行了模拟组卷,并对每种情况进行了 10 组模拟运算,算法满足终止条件($f=0$) 即退出。记录每次的进化代数及选出的最佳组卷个体的适应度。最后对这 10 组结果求平均相异度及平均进化代数。平均相异度越低,说明平均适应度越高。设用户允许的相异度为小于 0.06(大

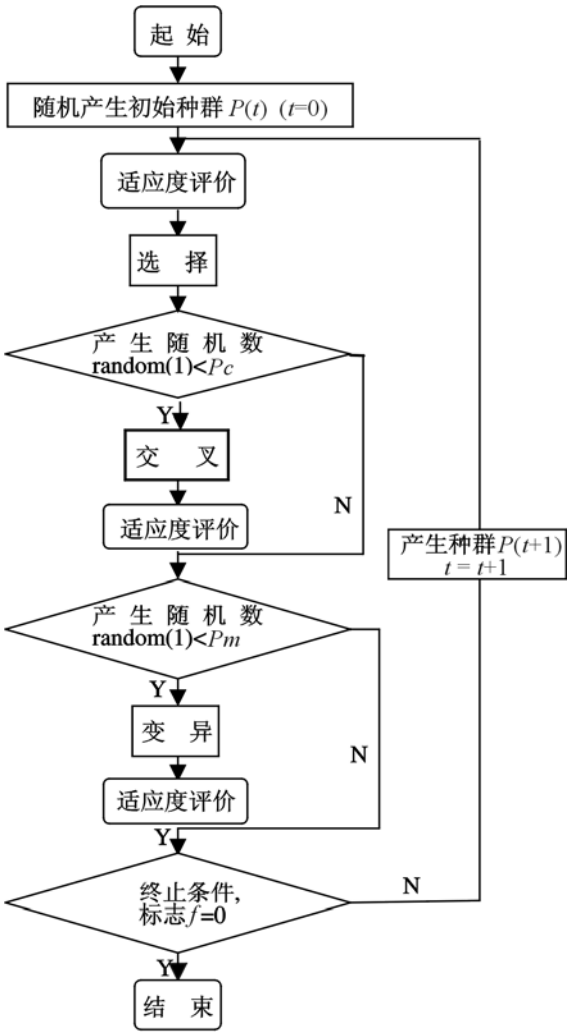

图 2　改进后的遗传算法过程示意图

Fig. 2　Genetic algorithms calculation of improving process sketch map

约相当于符合率最低的指标达到 90%),即如相异度小于 0.06 时,说明该次组卷成功。

(1) 初始种群为 200, $Pm = 0.1$, $Pc = 0.6$

表 1　参数 *q* 对算法收敛性的影响

Tab. 1　The table of parameter q affecting the contraction on the algorithm

交叉和变异种群(q)	种群平均相异度	平均进化代数	成功次数
30	0.060 385 329	7.8	5
50	0.056 301 989	9.1	8
100	0.045 700 114	7.8	10
150	0.042 030 49	7	10
200	0.038 300 647	8.2	10
300	0.0238 994 12	8.4	10
400	0.0349 480 36	9.2	10

(2)初始种群为 200,进行交叉的个体为 200, $Pm = 0.1$

表 2　参数 *Pc* 对算法收敛性的影响

Tab. 2　The table of parameter Pc affecting the contraction on the algorithm

交叉概率(Pc)	种群平均相异度	平均进化代数	成功次数
0	0.050 682 794	7.4	7
0.2	0.037 252 758	8.6	10
0.4	0.035 106 058	9	10
0.6	0.032 009 598	9.16	10
0.8	0.037 075 15	9.6667	10

(3) 初始种群为 200,进行交叉的个体为 200, $Pc = 0.6$

表 3　参数 *Pm* 对算法收敛性的影响

Tab. 3　The table of parameter Pm affecting the contraction on the lgorithm

变异概率(Pm)	种群平均相异度	平均进化代数	成功次数
0.001	0.038 223 32	8.4	10
0.01	0.039 245 746	7.6	10
0.05	0.039 250 372	7.8	10
0.1	0.035 865 660 8	8	10

试验 2:研究改进后的遗传算法对组卷是否有效。

对该算法设置不同的参数来进行模拟组卷,统计每次最佳组卷所抽题的重复率。在每种情况下运行 10 次,求这 10 次试题的平均重复率(平均重复率等于 10 次所选试题相同总数除以总的试题数)。统计表如下:

初始种群为 200, $Pm = 0.1$, $Pc = 0.6$

表 4　统计该算法的所选试题的重复率

Tab. 4　Count the repetitive rate of the examination question selected of this algorithm

交叉变异的种群(q)	30	50	100	150	200
平均重复率	0.216 6	0.158 88	0.083 07	0.052 17	0.048 875

试验 3:分析题库中试题分布对该算法的影响

为了进一步研究该遗传算法的可行性,设置了分布不同的试题库,试题库 1,试题库库中每种题型分布均匀,试题库 2 中每种题型分布不均匀。在相同的运行参数下,分别对两个试题库进行了 10

次模拟组卷,求这10次平均相异度以及成功次数。

初始种群为200, $Pm=0.1$, $Pc=0.6$

表5 不同题库对算法的影响表

Tab.5 The effectible table of different item bank

试题库	平均相异度	成功次数	平均进化代数
试题库1	0.038 300 647	10	8.2
试题库2	0.075 320 388	8	9.2

4.4 分析讨论

试验1 参数对算法的影响

由表1可知,该算法对初始种群的选择和进行交叉和变异的种群(q)敏感性较强。因为种群的不同很大的影响了种群的平均相异度。初始种群为200, $Pm=0.8$, $Pc=0.1$。取不同的进行交叉和变异的种群(q),试验结果有很大的差别。进行交叉和变异的种群(q)越小,种群的平均适应度越小。这是因为种群个体越小,失去了种群的多样性,收敛的速度越快,造成了早熟现象。当进行交叉和变异的种群(q)=200和300时,尽管$q=300$比$q=200$的平均相异度(平均适应度高),但$q=300$时组卷时间慢,效率低。权衡两个方面,我们选择$q=200$。

由表2可知,当初始种群为200,进行交叉和变异的个体为200, $Pm=0.1$,取不同的交叉概率(Pc)对组卷有一定的影响。交叉概率($Pc=0$)时,种群的平均适应度明显比其它大,是因为该试验没有进行交叉运算,减小了个体的多样性,缩小了搜索空间,减少了获得最优解的机会。因此得到的组卷与用户要求有很大的差异。由表2可看出,对该算法$Pc=0.6$较好。

由表3可知,初始种群为200,进行交叉和变异的个体为200, $Pc=0.6$。取不同的变异概率(Pm)对组卷也有一定的影响。由表3可看出,对该算法$Pm=0.1$较好。因为这时种群的平均误差比其它低(越能满足用户的需求),且平均进化代数也不高。

试验2:由表4可知,该算法随着进行交叉与变异算子个体增加而抽题的重复率减少,因此用该算法进行组卷时,为了有效的避免重复率,最好进行交叉与变异的个体较多,但并不是越多越好,多了会影响组卷效率,种群最好为200。此外,我们还可以看出每次抽题的重复率都较低,说明这是一种比较好的遗传算法,可适合于在线考试与各种类型考试,实用性很高。如果每次抽题的重复率高,也即相同的题目数多,说明这个算法实用性低,缺乏有效性、可考性。

试验3:由表5可知,用试题库1组卷的平均相异度比用试题库2组卷的平均相异度低,且成功次数也高。说明用试题库1组卷的效果优于用试题库2组卷的效果,有效性更优。这可能与算法中在产生随机初始种群的随机函数(random)有关,随机函数(random)产生数的概率是平均的。因此,在建试题库时,最好是每种题型分布较均匀,至少每种题型分布基本符合用户组题要求。

4.5 结论

由上述3个试验结果可知,采用功能块的思想以及欧式空间距离求最优解的遗传算法性能优于传统的遗传算法,满意度较高,速度也较快,大概组一套试卷最多只有10 s左右。由于它实行了全局并行搜索,并且在搜索过程中不断向可能包含最优解的方向调整搜索空间,从而能较快的找到满足条件的解。对初始种群的选择和进行交叉和变异的种群(q)敏感性较强。因此,该算法对进行交叉和变异的种群(q)取值相当重要。在建试题库时,最好每种题型分布较均匀。

尽管该算法与用户需求有一定的相异度,但总体的相异度并不大,且该算法组卷的效率高,速度较快。对于该算法,交叉概率$Pc=0.6$,变异概率$Pm=0.1$,进行交叉和变异个体为200较好。变异概率Pm之所以为0.1,是由于在进行交叉时,整个功能块全部互换,交叉较为粗糙,影响了种群的多样性。因此在进行变异时,必须加大变异的概率,以恢复种群的多样性,这就起到了全局收敛的目的。

5 结束语

本文对传统遗传算法的改进,以较快的进化速度求得了多目标组卷问题的满意解,稳定性好,可信度高,是一个比较好的算法,在通用试题库系统建设中具有较高的实用价值。同时,我们在研究中也发现,要进一步提高结果精度,需要花费较大的代价。目前,正继续进行受控变异、诱导进化、二次进化等方面的研究工作。

[参考文献]

[1] 文中林,蔡清万,李元香. 试题库智能组卷的遗传算法[J]. 湖北民族学院学报(自然科学版),2000,18(3):53-55.

[2] 周红晓. 遗传算法在试题库智能组卷中的应用[J]. 浙江师范大学学报(自然科学版),2003,26(4):374

-378.
[3] 毕应洲,苏德富,陈宁江. 基于矩阵编码的遗传算法及其在自动组卷中的应用[J]. 计算机工程,2003,(7):73-75.
[4] 徐一峰. 计算机考试系统抽题算法的哈希函数描述[J]. 佳木斯大学学报(自然科学版), 2004,22(2):239-241.
[5] 张文祥,马银花. 试题库智能组卷算法的设计与实现[J]. 华北科技学院学报,2003,5(1):71-73.
[6] 周明,孙树栋. 遗传算法原理及应用[M]. 北京:国防工业出版社, 1999.
[7] 王小平,曹立明. 遗传算法——理论、软件实现[M]. 西安:西安交通大学出版社,2002.
[8] BERTONI A, DORIGO M. Implicit parallelism in genetic algorithms[J]. Artificial Intelligence, 1993, 61(2):307-314.
[9] HOLLAND F H. Outline for logical theory of adaptive systems[J]. Journal of the Association for Computing Machinery, 1962, 9(3):297-314.
[10] 全惠云. 进化算法[M]. 北京:冶金工业出版社, 2000.
[11] 杨青. 基于遗传算法的试题库自动组卷问题的研究[J]. 济南大学学报(自然科学版), 2004,18(3):228-231.
[12] 樊恽,郑延履. 线性代数与几何引论[M]. 北京:科学出版社,2004.
[13] 李敏强,寇纪淞,林丹,等. 遗传算法的基本理论与应用[M]. 北京:科学出版社, 2004.
[14] 徐丽娜. 神经网络控制[M]. 北京:电子工业出版社,2003.
[15] 孟志青. 通用试题库管理系统的一种优化命题模型[J]. 计算机工程与设计,1998, 19(3):55-58.
[16] 玄光男,程润传. 遗传算法与工程设计[M]. 北京:科学出版社,2000.

(上接第 713 页)

[参考文献]

[1] HENNING S. Antibiotic resistance in aquaculture[J]. Acta Vet, suppl,1999,92:29-36
[2] ROBERT R MUDER, CAROLE Brennen, ANGELLA M GOETZ, et al.. Association with prior fluoroquinolone therapy of widespread ciprofloxacin resistance among gram-negative isolates in a veterans affairs medical center[J]. Antimicrob. Agents Chemother,1999,35: 256-258.
[3] JORDI V, JOAQUIM RUIZ, FERRAN SANCHEZ, et al.. Increase in quinolone resistance in a Haemophilus influenzae strain isolate from a patient with recurrent respiratory infections treated with ofloxacin[J]. Antimicrob. Agents Chemother, 1999,43:161-162.
[4] 陆承平. 嗜水气单胞菌及其所致鱼病[J]. 水产学报, 1992,(9): 282.
[5] 韩文瑜. 现代分子细菌学[M]. 北京:中国解放军农牧大学出版社,1999.
[6] WRETLIND B, MOLLBY R, WADSTROM T. Separation of two hemolysins from Aeromonas hydrophila by isoelectric focusing[J]. Infect Immun, 1971,(4):503-505.
[7] CHOPRA A K, HOUSTON C W, PETERSON J W, et al.. Cloning, expression, and sequence analysis of a cytolytic enterotoxin gene from Aeromonas hydrophila [J]. Can J Microbiol, 1993,39:513-523.
[8] DONTA S T, HADDOW A D. Cytotoxic activity of Aeromonas hydrophila[J]. Infect Immun, 1978,21:989-993.
[9] WILMSEN H U. Aerolysin from Aeromonas hydrophiiila, forms voltage-gated channel in planar lipid bilayers[J]. S Member Biol ,1990,115(1):71-81.
[10] 郭尧君. 蛋白质电泳技术(第 1 版)[M]. 北京:科学出版社,2001.
[11] 涂小林,陆承平. 嗜水气单胞菌毒素的提纯及特性分析[J]. 微生物学报, 1992, 32(6):432-438.
[12] 王世若,王兴龙,韩文瑜. 现代动物免疫学[M]. 长春:吉林科学出版社,2001.
[13] 王世若. 兽医微生物学及免疫学[M]. 长春: 吉林科学技术出版社, 1989.
[14] 朱平,冯书章. 抗体制备技术(第 1 版.)[M]. 长春:长春出版社,1994.
[15] 汪玉松,邹思湘,张玉静. 现代动物生物化学[M]. 北京:中国农业科学出版社,1999.
[16] RM 康普. 蛋白质结构分析、制备、鉴定与微量测定[M]. 北京:科学出版社,2000.