(19) 107-112

1998 年 1 月 重庆大学学报 (自然科学版) Vol. 21, №. 1
第 21 卷第 1 期 Journal of Chongqing University (Natural Science Edition) Jan. 1998

·研究简报·

紫外光谱与遗传算法用于多组分氨基酸同时测定*

0623.736
0657.32

夏之宁① 胡 芳② 邱细敏② 石乐明② 李志良①
(① 重庆大学应用化学系,重庆,400044;② 湖南大学化学化工系,第一作者 36 岁,男,副教授,博士)

摘 要 将遗传算法用于紫外光谱数据处理,实现了多元分辨与校正。并对多组分分析体系进行同时定量测量。

关键词 遗传算法;紫外光谱;多元校正;分辨;氨基酸
中国图书资料分类法分类号 O657.32

0 引 言

遗传算法(GA, Genetic Algorithm)[1]是由美国 Michigan 大学 J. H. Holland 教授提出来的,早在 1962 年就已形成了 GA 基本思想并于 60 年代末期提出了 GA 数学框架,后于 1975 年在其专著[1]中介绍了 GA 完整算法。随后特别是近年来,GA 作为一种思想上和算法上新颖的全局优化搜索方法吸引了广大研究者,获得了迅速发展和广泛应用[2-6]。GA 以其极强的解决问题能力和对使用对象广泛的适应性,在许多科学研究与工程技术等领域取得了优良的成效,已成为国际学术界跨学科领域的热门研究课题之一。GA 遗传算法的基本思想来源于自然遗传学和生物进化论。它可以实现全局搜索最优求解,已用来解决许多问题包括化学中的构象问题[2-6]。笔者研究了 GA 遗传算法并用于多元校正及光谱分辨。将所提出的方法对多组分体系进行同时定量测定,获得良好的分析结果。对四种氨基酸多元分析,回收率为 95.9%～104.3%.

1 理论与方法

1) 基本理论 根据生物遗传学及自然进化论机理:"自然进化,物竞天演,适者生存,优胜劣汰",每一物种进化过程中总是朝着适应所处的环境方向演化。这种自然选择性和物种适应性构成了 GA 遗传算法的主旋律。GA 结合了达尔文"适者生存"进化理论和信息交换随机过程,前者消除了解域中的不适应因素,后者则利用了原解中的已知信息,从而有效的加

* 收文日期 1997-07-22
国家自然科学基金,编号:29405036,29775035;优秀年轻教师基金(教人司[1996]486),回国留学人员基金(教外司留[1996]644)资助项目

快了优化搜索过程。此相当于在一种"自调节"或"自适应"的变量空间中，搜索出与环境最相适、最匹配的最优解。故 GA 属于自适应方法。每一物种(解)在一代接一代的进化过程中不断吸收"信息"和"知识"，并均在其成员的染色体(chromosome)组成上得到反映。一定数量的物种(又称个体或可能解)组成一自然群体(解集)。在给定环境下，由于物种之间存在某种差异故而具有不同的生存能力或适应能力。选择某些物种作为父母体使之进行生殖繁衍即经过遗传操作改变染色体(串)组成，而产生新的物种。对这些后代物种的生存能力进行评价并同该群体中其它物种比较，按照"适者生存，优胜劣汰"的进化规则淘汰其中适应能力最差的物种，使群体的规模保持不变，物种数目既不增加亦不减少。随着"新陈代谢"的进行，物种不断更新换代，群体的结构也不断更新换代，群体的结构也不断进化或逐步调整，促使整个生物体系(群体)向全局最优解逐步演化、逐渐逼近，最终获得全部最优解，完成整个优化演进过程。

2) 基本操作　GA 的遗传操作是自然界生物进化过程的数学抽象。现有多种类型的遗传操作，如交叉(crossover)和变异(mutation)等简单基本操作，删除与复制等相对低级操作和婚配与迁居等相对高级操作。此处主要介绍几种重要的基本操作：

① 择种操作(selection)　依据染色体串的适应值(fitness)在当前串集(bit strings set)中随机地选择若干对物种(个体)作为双亲(父母体)用于繁殖后代(子女体)，所产生的新物种(个体)加入下一代群体。通常个体(物种)可用矢量描述，其构成元素在 GA 中称为染色体(或工作串)。其中适应值较大串被选中的概率较大而适应性 较小者被选中的概率较小，或者说适于生存环境的优良个体将有更多的繁殖机会，从而使优良特性得以遗传。故选择是遗传算法 GA 的关键之一，它体现了适者生存的自然界进化思想。

② 交叉操作　对于选中用于繁殖的每一(对)个体，先随机地把当前串集中的(染色体)串配对，后按一定概率部分地交换双亲(父母体)的基因信息或染色体(串)。若串长为 m，则部分交换长度在[1，m-1]之间随机确定。例如有两个串(其串长度均为 $m=60$：$A1=001010$ 和 $A2=110101$)相互配成对，如在位置 3 处经交叉产生新个体(两新串：$B1=A1'=000010$ 和 $B2=A2'=111101$)。交叉体现了自然界中信息交换的思想。

③ 变异操作　对于以一定概率随机地选取的若干个体(物种)，随机地在染色体(串)中取某一位改变其值，常为取反运算即由 1→0 或由 0→1。例如将串集 $A1$ 与 $A2$ 分别改变为 $C1=A1''=101010$ 和 / 或 $A2''=C2=010101$(前者由 1→0，后者由 0→1)。变异操作又称为突 变操作；同自然界一样，每一位均发生变异的可能性是很小的；因此变异操作模拟了生物进化过程中偶然的基因突变现象。上述几种 GA 基本操作(即选择，交叉和变异)涉及的仅是对染色体串进行复制或交换，因而实现相当简易，但由它们所构成的 GA 其功能却很强[1]。GA 的搜索优化能力主要是由选择与交叉赋予的；而变异算子则保证了算法能搜索到问题解域空间的每一点，因此使 GA 算法具有全局最优。故变异进一步增强了 GA 的能力。可以用 GA 解决许多复杂优化问题。

3) 基础算法　在遗传算法 GA 中常将个体(物种或解)以矢量形式来表述，相应构成矢量的元素即为位串(bit string)或染色体。染色体由基因组成，基因则相当于串的位。它们的进化是经遗传操作实现的。设欲求问题的全局最优解(s_0 矢量)为：

$$S_0=(s_{01},s_{02},s_{02},\cdots,s_{0k},\cdots,s_{0n}) \tag{1}$$

假设群体(解域)中存在如下两种个体(单解)，S_1 和 S_2

$$S_1 = (s_{01}, s_{02}, s_{03}, \cdots, s_{0k}, \cdots, s_{0n}) \tag{2}$$

$$S_2 = (s_{r1}, s_{r2}, s_{r3}, \cdots, s_{rk}, \cdots, s_{rn}) \tag{3}$$

其中解 S_1 为除第 k 元素外($s_{jk} \neq s_{0k}$)，其余各染色体(基因位)都为最优解 S_0 的元素。而另一解 S_2 则仅第 k 元素($S_{rk} = S_{0k}$)为最优解 S_0 的第 k 元素，其余各染色体(基因位)都不是最优解 S_0 的元素。故若将也只需将 S_1 和 S_2 的第 k 基因位相互交换，便可使 S_1 变成为全局最优解 S_0。一般地假设最优解 S_0 的 n 个元素各分别分布于 n 种个体(物种或解)中，则任取其中一个体，经(n-1)次上述有效变换后便找到全局最优解 S_0。在 GA 中将任意两个体(物种)中染色体予以交换的遗传操作，即上述交叉。它可使父母体优良染色体及其基因遗传给子女体。但交叉不产生新染色体。若最优解元素不存在于群体，则无论如何交叉均不产生最优解及其解元素。因此需要进行其它操作，如变异操作：对于 S_1 中的第 k 位基因 $s_{jk}(\neq s_{0k})$，若取常数 r 使

$$s_{jk} \cdot (1 + r) = s_{0k} \tag{4}$$

则可以构造一单位基矢(向)量

$$R = (0,0,0,\cdots 1 \cdots 0) \tag{5}$$

即其中第 k 个元素为一，而其余元素均为零。则

$$s_0 = s_1 + s_{jk} \cdot r \cdot R \tag{6}$$

可求得最优解。此种变异运算遗传操作实际上相当于改变了个体中某条染色体的某些基因。若构成不同的单位矢量则能对不同基因(位置)的染色体作变异遗传操作。由于参数 r 实际上无法预先确定，通常在 ±1 范围内选择随机数。随着生物进化或遗传的进行，逐步减小优化区域，得到全局最优的精确解。

在遗传或进化过程，不断对新产生的一代群体进行重新评价、选择、交叉、变异等，如此循环往复，使群体中最优个体的适应度和平均适应度不再提高，则迭代过程收敛，遗传算法中止。通常对于任意给定的初始矢量集(解群集)，经上述交叉与变异等遗传操作，总能使群体或其中某些个体演化成为或非常接近全局最优解 S_0，GA 即告完成。

2　实验部分

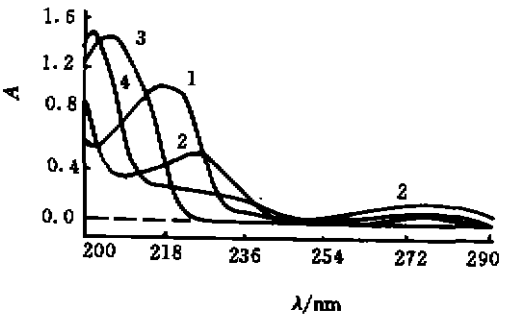
1) 仪器与试剂　日本岛津 UV-2600 型，日立 Hitachi-557 型，美国 Beckman DU-7HS 型及上海 740 型紫外可见分光光度计；IBM-PC/AT486 微机与 VAX-DMS 工作站用于编制运算 GA 及其它算法程序；上海雷磁 pHS-2 型酸度计用于调节溶液 pH 值或标定酸碱浓度。色氨酸 Trp，珞氨酸 Tyr，苯丙氨酸 Phe 及 2,4-二羟基苯丙氨酸 Dhp 等均为色谱纯试剂(Merck)。

用常规法配成 0.1mol/L 的 HCl 储备液并稀释作为工作液。其它试剂均为 AR 以上。所用水为二次(去离子蒸馏)水。

2) 实验方法 分别取适量 Trp,Tyr,Phe,Dhp 配成系列纯组分标准液和多组分混合液,在分光光谱仪上扫描记录紫外光谱,取波长范围为 190～310 nm,狭缝宽度为 1.0 nm,在 200～300 nm 范围内间隔 1 nm 读取数据并输入计算机处理。

3 结果与讨论

1) 紫外光谱 四组分氨基酸(1. Trp,2. Tyr,3. Phe,4. Dhp)在 0.1 mol/L 的 HCl 介质中的紫外吸收光谱如附图所示,其重叠十分严重,须作多元分辩或予分离定量。



附 图 氨基酸的紫外光谱
1) Trp,2) Tyr,3) Phe, 4) Dhp

2) 试验条件 依据文献报导并经实验考察氨基酸溶液介质,发现用 0.1 mol/L 的 HCl 作测试介质则光谱数据稳定,分析结果较好。而采用 0.1 mol/L 的 NaOH 介质,则紫外吸收光谱略有移动,但量测数据不够稳定。故本文采用 0.1mol/L 的 HCl 为介质,其线性加合性良好,相对偏差一般小于 2.5%,最大不超过 5%.

3) 模拟结果 以对称函数

$$f(x_1,x_2)=y=x_1^2+2(1-\cos 2\pi x_1)+3x_2^2+4(1-\cos 4\pi x_2)$$

为目标函数进行最优化(此处为最小化)为例,考察 GA 算法的性能。此函数 $y=f(x_1,x_2)$ 有若干局优点和零值最优点($x_1=0$, $x_2=0$). 设定优化约束区域为 $-1\leqslant(x_1$ 或 $x_2)\leqslant+1$ 即 $x_1\in[-1,+1]$, $x_2\in[-1,+1]$. 最始群体(解集)依据 $z_i=(b_i-a_i)c+a_i$ 作初值赋值,其中 z_i 为第 i 个待求解参数,c 为一其值在[0,1]之间的随机数,即 $0\leqslant c\leqslant 1$, a_i 与 b_i 则为每一参数的优化区域,通常是已知的或予先设定的,此例中 $a_i=-1$ 而 $b_i=+1$ 为上下界。

GA 算法参数分别取择如下:1) 优化求解精度为 $\varepsilon=0.002=2\cdot10^{-3}$ 或遗传操作总次数限取 2 千次,2) 交叉与变异操作的概率均为 50%,即各为 1 千次。于是经较小次数的遗传操作便能很快收敛接近全局最优解;如经 200 次操作得优化目标函数为 $f(x_1,x_2)=y=3.59\times10^{-4}$ 和 4.66×10^{-4},取均值则为 $f(x_1,x_2)=y=4.12\times10^{-4}$,其优化求解值分别为 $x_1,x_2=-0.001\sim0.001$. 确实非常接近全局最优解($x_1=0$, $x_2=0$, $y=0$). 由模拟结果可知,GA 确能跨越局部极优而达到全局最优。

4) 适应函数 依据"适者生存,优胜劣汰"自然进化原则,需要选择合适的适应值函数或其它评价函数,它是 GA 算法的一个重要问题。选择了一个优良的评价函数不仅有利加速收敛,而且可获正确求解。对于不同问题可选择不同的性能(或适应)评价函数,包括最小一

乘，最小二乘，均方误差等。总的原则是选择性能评价，或者是适应性函数要能有效地指导搜索沿着面向参数最优组合方向逐步逼近，即向全局最优参数组合(解)方向逐步逼近，而不致使优化搜索不能收敛或陷入局部极优。经考察本文采用了简单而实用的均方根误差类函数作适应性能评价函数：$f(x) = 1/\mathrm{MSE}$.

5) 多元分辨　多元分辨与校正通常以满足线性加和性原理的多组分分析体系来实现，例如满足 Beer 定律矩阵形式 $A = K'C + E$ 的多元重叠光谱分析作多组分同时定量。对于本文所考察的四组分氨基酸混合体系的光谱数据处理，则实现氨基酸多组分分析。考虑到各组分浓度均应为正值或至少为零而不能出现负值，可以定出优化区域的下界 $a_i = 0$，故有约束条件 $x_i > 0$；又依据测试光谱数据在其最大峰值处，各组分浓度均不应超过仅由单组分存在时的对应，从而可以设置可能上界，即 $b_i = x_{i\max}$. 于是可设定优化空间的约束区域 $a_i = 0 \leqslant x_i \leqslant x_{i\max} = b_i (i = 1,2,3,4)$ 即 $x_i \in [0, x_{i\max}] = [a_i, b_i]$. 依据最小二乘准则，以估计均方误差(MSE)作评价函数，其倒数作适应值，用 GA 对混合试样中各氨基酸组分进行定量(同时测定)。与卡尔漫滤波(KF)、目标因子分析(FA)、主成分回归、偏最小二乘、人工神经网络及通用模拟退火等方法结果比较，相互一致。其优化精度与预测效果均属良好，回收率为 95.9 ～ 104.3%.

6) 同时测定　以正交设计[依表 $L9(3^4)$]及随机设计配制混合已知表液及混合未知试样，以 GA 对所测定的光谱数据进行多元分辨可预测未知浓度即实现多组分同时测定(见附表)，结果良好。

附 表　四组分氨基酸混合试样的同时多组分分析结果　10^{-6}

方 法	氨基酸\试样	No. 1	2	3	4	5	6	7	8	9	10
Taken	Tyr	4.670	3.736	3.736	8.406	7.472	6.538	5.604	2.802	2.802	4.670
	Trp	5.180	6.216	4.144	2.072	1.036	2.072	5.180	4.144	4.144	3.108
	Phe	14.33	23.89	33.45	23.89	9.556	4.778	14.33	23.89	23.89	23.89
	Dhp	8.208	4.104	12.31	6.156	16.42	10.26	2.052	12.31	12.31	2.052
Found	Tyr	4.570	3.734	3.561	8.315	7.319	6.530	5.646	2.774	2.871	4.890
	Trp	5.127	6.137	4.159	2.072	1.045	2.026	5.060	4.075	4.058	3.021
by KF	Phe	14.01	23.43	32.71	24.74	9.998	4.516	13.72	23.62	23.58	24.43
	Dhp	8.274	4.116	12.43	6.428	16.43	10.34	2.146	12.40	12.41	1.995
Found	Tyr	4.501	3.716	3.422	8.161	7.086	6.500	5.689	2.783	2.893	4.941
	Trp	5.214	6.225	4.226	2.161	1.131	2.125	5.167	4.202	4.187	3.005
by FA	Phe	14.57	23.79	32.60	23.95	9.732	5.014	14.65	23.95	24.11	24.70
	Dhp	8.532	4.306	12.87	6.626	16.96	10.34	2.164	12.61	12.44	1.986
Found	Tyr	4.678	3.126	3.572	8.321	7.234	6.531	5.626	2.769	2.866	4.793
	Trp	5.132	2.231	4.157	2.072	1.048	2.044	5.137	4.098	4.106	3.032
by GA	Phe	14.12	23.54	32.87	24.26	9.708	4.652	13.96	23.67	23.86	24.45
	Dhp	8.267	4.118	12.45	6.284	16.43	10.32	2.142	12.43	12.34	1.998

7) 初步结论　GA 系新颖的实现全局最优求解的化学计量学新方法。为氨基酸分析提供了一种优良的和有效的新方法，也可望用于其它方面。

参 考 文 献

1 Holland J A. Adaptation in Natural and Artificial Systems. Ann Arbor MI: University of Michigan Press, 1975, 7～56

2 Kirkpatrick S, Gelatt Jr C D, Vecchi M P. Optimization by simulated annealing. Science, 1983, 220: 671～680

3 Metropolis N, Rosenbluth A, Rosenbluth M et al. Equation of state calculations by fast computing machines. J Chem Phys, 1953, 21: 1087～1092

4 McGarrah D B, Judson R S. Analysis of the genetic algorithm method of molecular conformation determination. J Comp Chem, 1993, 14(11): 1385～1395

5 Judson R S, Jaeeger E P, Treasurywala A M, Perterson M L. Conformational searching methods for small molecules II. genetic algorithm approach. J Comp Chem, 1993, 14: 1407～1414

6 Lucasius C B, Blommers M J J, Buydens L M C, Kateman G. A genetic algorithm for conformational analysis of DNA. In: Davis L(Ed). Handbook of Genetic Algorithms, New York NY: Van Nostrand Reinhold, 1991, Chapt 18, 251

7 Goldberg D. Genetic Algorithms in Search, Optimization, and Learning, Reading MA: Addison- Wesley, 1989, 1～276

8 Davis L, Steenstrup M. Genetic Algorithms and Simulated Annealing, London: Pitman, 1987, 1st Ed: 1～11

9 Li Z L, Shi L M, Li M L et al. A highly sensitive and selective method for trace multicomponent analysis. Acta Chim Sin, 1990, 48(11): 1101～1107

10 Shi L M, Xu Z H, Li Z L et al. Application of the Kalman filter algorithm to the simultaneous determination four amino acids by direct uv spectrophotometry. J Micronutr Anal, 1990, 8: 1～12

Genetic Algorithms and Ultravoilet Spectroscopy as Applied to Multicomponent Analysis of Amino Acids

*Xia Zhining*① *Hu Fang*② *Qu Ximin*② *Shi Leming*② *Li Zhiliang*①

(① Dept. of Applied Chemistry, Chongqing University, Chongqing, 400044;
② Department of Chemistry and Chemical Engineering;
Hunan University; First Author, male, age 36, Ph. D.)

ABSTRACT Modified Genetic algorithm (GA) was proposed for simultaneous multicomponent analysis of ultraviolet spectroscopy is of four amino acids Tyr, Trp, Phe and Dhp.

KEYWORDS genetic algorithms (GA); ultraviolet spectroscopy; multivariate calibration, resolution; amino acids