

硕士学位论文

基于群体智能的基因表达数据双聚类研究

**RESEARCH ON BICLUSTERING OF
GENE EXPRESSION DATA BASED ON
SWARM INTELLIGENCE**

凡振豪

西南大学

2020 年 2 月

国内图书分类号：TM301.2
国际图书分类号：62-5

学校代码：10635
密级：公开

工学硕士学位论文

基于群体智能的基因表达数据双聚类研究

硕士研究生：凡振豪

导 师：欧灵副教授

申 请 学 位：工学硕士

学 科：计算机软件与理论

所 在 单 位：计算机与信息科学学院

答 辩 日 期：2020 年 2 月

授予学位单位：西南大学

Classified Index: TM301.2

U.D.C: 62-5

Dissertation for the Master's Degree in Engineering

RESEARCH ON BICLUSTERING OF GENE EXPRESSION DATA BASED ON SWARM INTELLIGENCE

Candidate:	FAN Zhenhao
Supervisor:	OU Ling
Academic Degree Applied for:	Master of Engineering
Specialty:	Computer Architecture
Affiliation:	College of Computer and Information Science
Date of Defence:	February, 2020
Degree-Conferring-Institution:	Southwest University

摘 要

高通量基因微阵列技术的出现, 产生了大量的基因表达数据。这些数据在追踪生物过程, 基因规则发现以及病理分析中有着至关重要的作用。基因表达数据的双聚类是指, 找出在某些条件子集下包含一致表达波动的基因子集。双聚类可以看作是一种多目标优化问题。针对表达数据高维度, 高冗余的特点, 许多群智能算法被用于双聚类中来。

本文基于布谷鸟搜索算法、萤火虫算法和细菌觅食算法等群智能优化算法, 从算法结合以及多目标优化等方面进行基因表达数据双聚类的分析研究, 意在解决当前双聚类算法的聚类质量差和生物意义不明显等问题。论文的主要工作包括:

(1) 提出基于布谷鸟搜索算法和萤火虫算法的混合双聚类算法 (Cuckoo Search and Firfly Algorithm hybrid Biclustering, CSFAB)。通过将布谷鸟搜索算法的全局搜索能力与萤火虫算法的快速收敛能力有效地结合起来, CSFAB 算法可以显著地提高搜索速度和范围, 同时能够跳出局部最优解和找到包含不同基因的双聚类, 从而提高双聚类的多样性。与 CC、ISA、CSB、FAB 和 PSOB 等算法比较, 实验表明 CSFAB 算法的双聚类质量和生物意义更优。

(2) 提出基于多目标细菌觅食算法的双聚类算法 (Multi-Object Bacterial Foraging Algorithm Biclustering, MOBFOB)。因为双聚类可以看作多目标优化问题, 该算法使用多目标细菌觅食算法同时优化均方残差和体积等双聚类质量评价指标, 找到占优的双聚类解集。与 CC、CSFAB、CSB 和 FAB 相比, MOBFOB 算法在计算效率、双聚类的质量评价指标和生物意义等方面获得提高。

关键词: 群智能算法; 基因表达数据; 双聚类; 多目标优化

Abstract

The advent of high-throughput gene microarray technology has generated a large amount of gene expression data. These data play a vital role in tracking biological processes, discovering genetic rules, and analyzing pathology. Biclustering of gene expression data refers to finding a subset of genes that contain consistent expression fluctuations under certain conditional subsets. Biclustering can be considered as a multi-objective optimization problem. In view of the high dimensionality and high redundancy of expression data, many swarm intelligence algorithms have been used in biclustering.

Based on swarm intelligence optimization algorithms such as cuckoo search algorithm, firefly algorithm and bacterial foraging algorithm, this paper conducts analysis and research on the biclustering of gene expression data from the aspects of algorithm combination and multi-objective optimization. Problems such as poor quality and insignificant biological significance. The main work of the thesis includes:

(1) A hybrid biclustering algorithm CSFAB(Cuckoo Search and Firfly Algorithm hybrid Biclustering) based on cuckoo search algorithm and firefly algorithm is proposed. By effectively combining the global search ability of the cuckoo search algorithm with the fast convergence ability of the firefly algorithm, the CSFAB algorithm can significantly improve the search speed and range. And at the same time, the algorithm can jump out of the local optimal solution and find biclusters containing different genes, that increasing the diversity of biclusters. Compared with CC, ISA, CSB, FAB, and PSOB algorithms, experiments show that the quality and biological significance of CSFAB algorithm is better.

(2) MOBFOB (Multi-object Bacterial Foraging Algorithm Biclustering) based on a multi-object bacterial foraging algorithm is proposed. Because biclustering can be considered as a multi-objective optimization problem, the algorithm uses a multi-target bacterial foraging algorithm to simultaneously optimize the bicluster's quality evaluation indicators such as mean square residual and volume, and finds the dominant bicluster solution set. Compared with CC, CSFAB, CSB, and FAB, the MOBFOB algorithm has improved the computational efficiency, the quality evaluation index of bi-clustering, and biological significance.

Keywords: Swarm Intelligence Algorithm, Gene Expression Data, Biclustering, Multiple-optimistic

目 录

摘 要	I
ABSTRACT	II
 第 1 章 绪论	 1
1.1 研究背景及意义	1
1.2 相关研究进展	2
1.3 本文的研究内容及组织结构	3
第 2 章 基因表达数据的双聚类相关概述	4
2.1 基因表达数据	4
2.2 双聚类的相关概念	4
2.2.1 双聚类的定义	4
2.2.2 双聚类的类型	4
2.2.3 双聚类的结构	6
2.3 双聚类的验证指标	6
2.3.1 质量验证指标	6
2.3.2 生物验证指标	6
2.4 群智能算法	6
2.4.1 布谷鸟搜索算法	6
2.4.2 萤火虫算法	7
2.4.3 细菌觅食算法	8
2.5 本章小结	8
第 3 章 基于 CS-FA 的双聚类算法	9
3.1 CS-FA 搜索算法	9
3.2 基于 CS-FA 搜索双聚类算法	9
3.3 实验结果及分析	9
3.4 本章小结	9
第 4 章 基于多目标 BFO 优化的双聚类算法	10
4.1 多目标优化问题的基本概念	10
4.2 多目标 BFO 搜索算法	10

4.3 基于多目标 BFO 搜索双聚类算法	10
4.4 实验结果及分析	10
4.5 本章小结	10
第 5 章 总结与展望	11

第 1 章 绪论

1.1 研究背景及意义

随着科技的进步，人类对自然，以及对生命有了更为深刻的认识。从显微镜的发明到 DNA 双链结构的提出，再到如今的 21 世纪，人们越来越多地发现，许多重大疾病跟基因相关，如癌症、一些先天性的心脏病和肥胖症。对于基因的研究成为了解决这些难题的关键。无论是 2003 年的 SARS，还是 2020 年的新型冠状病毒，基因都是研究这些病毒的突破点，搞懂了其基因的组成和作用，将极大地帮助研究人员研制出对应的抗体。每次病毒的大规模扩散，对所在国家的经济和发展都会带来巨大的损失，因此，基因研究是一项任重而道远的，属于全人类的任务。

随着生物实验技术的进步，基因测序变得越来越方便和便宜。在美国，已经有公司推出了亲民的基因测序产品，人们只需花费数百美元就可以测自己的基因，从而可以提前知道将来容易得哪些疾病，并做好预防。在细胞的生物过程中，基因通过转录成信使核糖核酸（mRNA），在不同酶和氨基酸的参与下合成各种各样的蛋白质，这一过程称之为基因的表达。尽管同一生命体中的各个细胞的基因序列是相同的，但不同的细胞仅会在特定的条件下表达特定的极少数基因。所以，研究基因的表达调控是基因组表达分析的重要内容，也有很大的意义，比如，它有助于确认病毒感染和致癌基因，并有助于确认在细胞的各个生命周期内活性基因等。通过高通量基因表达测量技术如微阵列技术（Microarray），可以测得不同 mRNA 的在细胞中的含量。该数据代表着基因的表达水平，因此被成为基因表达数据。

在生物信息学中，对基因表达数据的挖掘是研究热点。通过对基因表达数据的聚类分析，可以得到感兴趣的差异基因。这些基因在不同的实验条件（如样本，时间）下，存在某种一致的表达模式。通过对这些基因的富集或旁路分析，找到这些基因的功能以及相互的调控关系。

然而，基因表达数据有两个主要特点。一，一般而言，基因的数量在几千到几万，而样本或条件的数量只有几十到几百。二，正如前面所说，基因的表达是条件相关的。基因有可能会在多个条件下表达，也有可能所有的条件下都没有表达。这些特点使得常规的聚类方法无法胜任，于是就引入了双聚类（Biccluster）的概念。不同与常规聚类的全局模式（Global Pattern），双聚类专注于寻找局部模式（Local Pattern），它不要求同一类基因只有在所有实验条件下才具有相似表达，而只要求在部分实验条件下具有相似的表达。找到的基因子集和条件子集就构成了一个双

聚类。

1.2 相关研究进展

双聚类这一单词最早由 Hartigan 于 1972 年提出。直到 2000 年, Cheng 和 Church 将双聚类引入到基因表达数据挖掘中, 提出了 CC 算法, 并得到了较好的效果。他们提出了 MSR (Mean Square Residue, 均方残差) 用于评价双聚类相似性。MSR 越小, 则表明行相似性和列相似性越高。该算法先通过不断地删除基因节点和条件节点, 找到小于事先给定的 MSR 阈值 δ 的双聚类, 然后将其作为初始双聚类, 在保证 MSR 不会增大的前提下, 不断向其中添加基因节点和条件节点, 最终得到一个双聚类结果。如果想要多个结果, 算法会把之前找到的双聚类使用随机数覆盖, 再重复上述操作, 直到获得想要数量的双聚类。随机数的引入, 会导致结果不准确, 而且无法找到重叠的双聚类。2002 年, Yang 等对 CC 算法进行改进, 提出了 FLOC 算法。算法从多个初始双聚类出发, 根据最大增益的原则, 来执行基因和条件节点的删除或增加。但并没有解决贪心策略带来的陷入局部最优解的问题。

2002 年, Ihmels 等人提出了 ISA 算法 (Iterative Signature Algorithm, 迭代签名算法)。ISA 并没有使用 MSR, 而是将双聚类视为转录模块, 然后通过对其进行打分, 迭代地修改基因集和条件集, 直到无法再继续修改。同年, Tanny 等提出了 SAMBA (Statistical Algorithm Method For Biclustet Analysis) 算法。该算法使用了图论和统计学的知识, 将基因表达数据看作一个二分图, 一边为基因, 一边为条件。通过寻找最大稠密子图的方式来找到基因表达数据中的最大子矩阵, 即双聚类。

由于双聚类问题是 NP 难问题, 通过贪婪或枚举的方法很难高效地得出结果, 人们开始使用元启发式算法, 如群智能算法来寻找双聚类。2004 年, stanfan 等人设计了一个基因表达数据双聚类分析的进化算法框架。Ken-neth Bryan 等人于 2006 年提出了基于模拟退火算法的双聚类算法, 并通过实验证明所提方法获得了质量较优的双聚类。2006 年 Mitra 提出了多目标进化双聚类的框架, 应用经典的非支配排序遗传算法 (Non-dominated Sorting Genetic Algorithm-II, NSGA-II) 算法, 整合局部搜索策略, 并提出新的定量度量方法估计双聚类的质量。2007 年, Goiraldez 应用最大标准区域 (Maximal Standard Area, MSA) 作为度量双聚类的标准, 并和 MSR 一起应用到多目标进化算法 MOEA 中, 提出多目标连续变化双聚类算法 (Sequential Multi objective Biclustetring, SMOB), 有效解决成比例模式的双聚类问题。2013 年 Carlos A. Brizula 提出了一种改进的多目标遗传双聚类算法 (Enhanced Multi-objective Genetic Biclustetring, EMOGB), 与其它多目标双聚类算法不同的

是他采用了一种基因和实验条件组来代表一个双聚类的编码方式，并且在搜索双聚类的过程中减少了局部搜索环节，从而算法的执行效率有了很大的提高。

1.3 本文的研究内容及组织结构

本文首先对基因表达数据和双聚类分析等相关基础知识进行阐述。然后从单目标优化，混合优化和多目标优化等方面，研究了以群智能算法如布谷鸟搜索算法、萤火虫算法和细菌觅食算法为框架的双聚类算法。最后对本文的工作进行了总结和展望。本文各章节内容安排如下：

第一章从生物背景知识出发，讨论了基因研究的重要性以及双聚类的作用和发展现状。双聚类克服了常规聚类与基因表达数据之间的矛盾，能更好的挖掘出有价值的基因子集和条件子集。群智能算法由于其快速且效果更好，在双聚类分析中得到了很广泛地应用。

第二章先是更具体地讨论的基因表达数据的数学形式，以及双聚类的定义、类型和结构。接着将双聚类算法分为了基于质量评价指标和基于模型两种。然后，对于双聚类结果，讨论了其质量验证指标和生物验证指标。最后分别介绍了本文所关注的群智能算法。

第三章

第四章

第五章对全文工作进行了总结和展望。

第 2 章 基因表达数据的双聚类相关概述

2.1 基因表达数据

数据的好坏在很大程度上决定了结果的好坏。基因表达数据主要通过基因芯片技术和下一代测序技术等高通量基因表达测量技术获得。这些技术可以同时不同的样本或条件下，对成千上万的基因进行高效、精确、定量地测量。这时只是得到了一些原始数据，由于在特定条件下只有很少数的基因会表达，里面会存在很多缺失数据。需要对原始数据进行缺失值和去噪处理，才能得到适合数据挖掘的数据。

基因表达数据极其庞大，以及需要专业的生物知识，导致很多跨领域的研究者很难涉足。为了打破这种学科壁垒，科学家们提出了关于描述和存储基因表达数据的标准，并在标准之上建立了基因表达数据库。最广泛的数据库为 GEO (Gene Expression Omnibus)，由美国国家生物技术信息中心于 2000 年开发。该数据库提供了共享基因表达数据的平台，并且有专业的人员进行审核。公共数据库的出现，极大促进了生物信息学的发展。

基因表达数据一般以一个二维矩阵 E 表示，一行代表一个基因，一列代表一个实验条件。实验条件包括不同时期，不同组织，不同个体，不同外部环境等等。矩阵 E 中的每个元素 e_{ij} 表示基因 g_i 在实验条件 c_j 下的表达水平值，其生物含义是该基因在此条件下，细胞中 mRNA 的含量。

2.2 双聚类的相关概念

2.2.1 双聚类的定义

给定一个大小为 $n \times m$ 基因表达数据 $E(X, Y)$ ，假定集合 $I \subseteq X, (|I| = k \leq n)$ 是 E 的基因集合 X 的子集；集合 $J \subseteq Y, (|J| = l \leq m)$ 是 E 的基因集合 Y 的子集。如图2-1所示，双聚类是指在条件子集 J 下的基因表达模型表现出同源特性的基因子集 I 。因此，双聚类可以定义为一个 $k \times l$ 的子矩阵 $B(I, J)$ ，也简称为 B 。

2.2.2 双聚类的类型

给定一个二维矩阵 $A(I, J)$ ， a_{ij} 为其中第 i 行第 j 列的值， α_i 是一个与行有关而与列无关的变量， β_j 是一个与列有关而与行无关的变量， $0 \leq h, r, t, d \leq |I|$ 或 $0 \leq h, r, t, d \leq |J|$ 。Madeira 和 Oliveira 提出，在双聚类中主要有以下四种结构：

1. 具有相同常量值的双聚类，如图2-2 a)所示，公式如下。

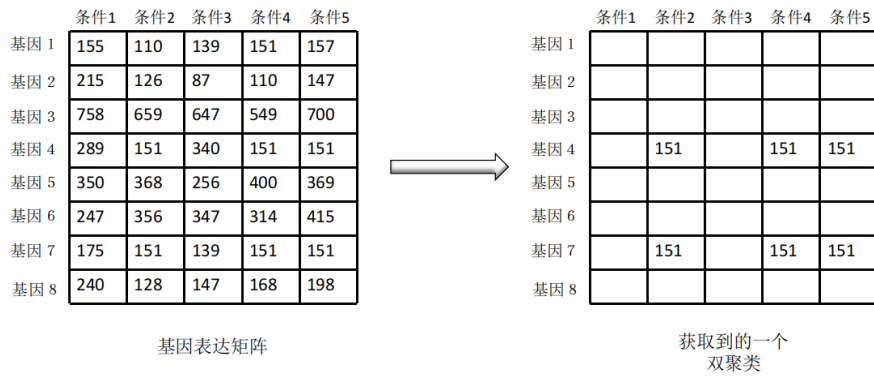


图 2-1 双聚类定义示例

$$a_{ij} = \mu \quad (2-1)$$

2. 列或行具有相同常量值的双聚类，如图2-2 b)和2-2 c)所示，公式如下。

$$(a_{ij} = \mu + \alpha_i \text{ 或 } a_{ij} = \mu * \alpha_i) \text{ 且 } (a_{ij} = \mu + \beta_j \text{ 或 } a_{ij} = \mu * \beta_j) \quad (2-2)$$

3. 具有连贯值的双聚类，如图2-2 d)和2-2 e)所示，公式如下。

$$a_{ij} = \mu + \alpha_i + \beta_j \text{ 或 } a_{ij} = \mu * \alpha_i * \beta_j \quad (2-3)$$

4. 具有连贯演变的双聚类，如图2-2 f)所示，公式如下。

$$a_{ih} \leq a_{ir} \leq a_{it} \leq a_{id} \text{ 或 } a_{hj} \leq a_{rj} \leq a_{tj} \leq a_{dj} \quad (2-4)$$

则双聚类可以根据相对应的结构分类。根据要解决问题的特定属性，需要考虑不同结构的双聚类。

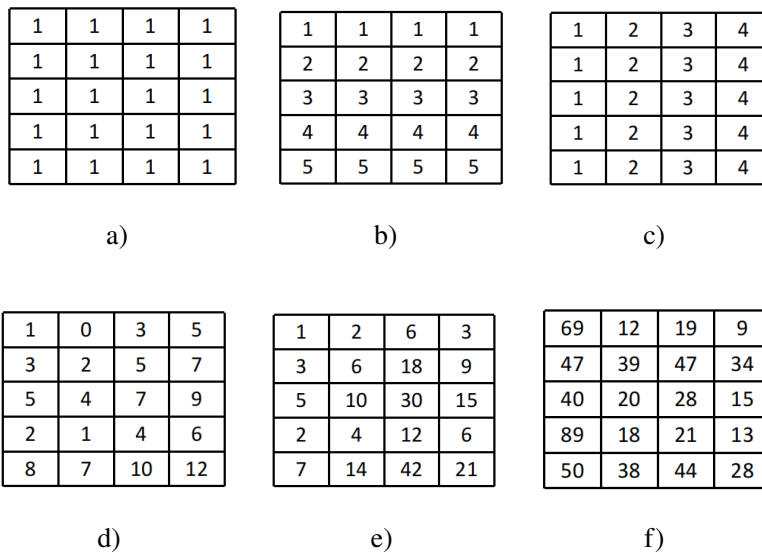


图 2-2 双聚类的类型

2.2.3 双聚类的结构

双聚类的结构是指，通过算法找到的双聚类之间在原始矩阵时间的相对位置。根据结构，大体可以分为 8 类，如下图所示：

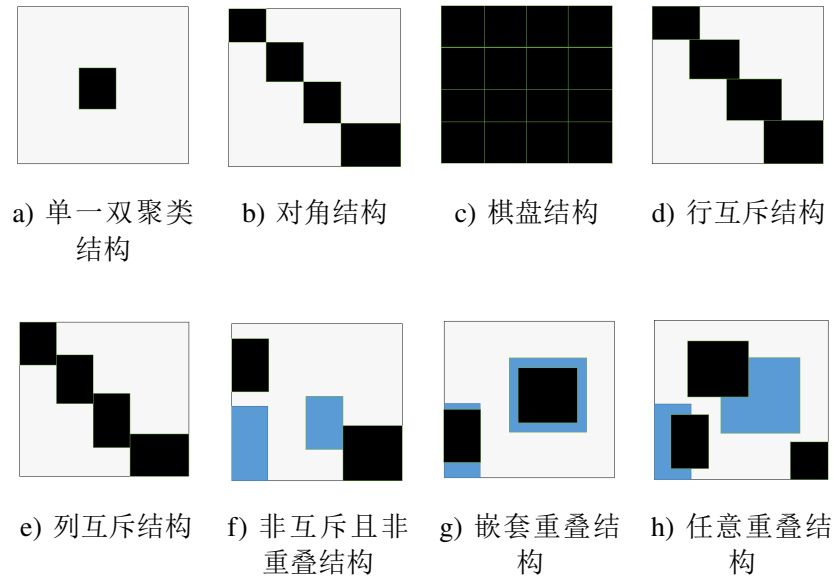


图 2-3 双聚类的结构

2.3 双聚类的验证指标

2.3.1 质量验证指标

2.3.2 生物验证指标

2.4 群智能算法

目前, 有很多优秀的优化算法, 有确定性方法如线性规划、二次规划, 动态规划和梯度下降; 以及随机性方法如群体智能。这些方法让我们能够在一定的时间内解决某些问题。然而, 处理大量高维数据时, 确定性方法太过复杂导致需要大量的计算成本。元启发式的群体智能算法因其高效率越来越受到关注。

2.4.1 布谷鸟搜索算法

布谷鸟搜索算法 (Cuckoo Search, CS) 是 Yang 和 Deb 于 2009 年提出的新兴启发算法。该算法通过模拟布谷鸟寄生育雏行为, 在可行域中通过 Levy 飞行寻找合适的鸟巢, 来找到较优解。该算法有三条理想化的规则:

1. 每只布谷鸟每次下一个蛋, 并将其放入随机选择的巢中。
2. 具有优质蛋的最佳巢会被带到下一代。
3. 可用的寄主巢数量是固定的, 且寄主以概率 $P_a \in (0, 1)$ 发现布谷鸟放的蛋。

在这种情况下，寄主可以消灭该蛋或放弃旧巢另建新巢。

CS 中有两种更新方式，一种是布谷鸟寻找宿主鸟巢的 Levy 飞行：

$$x_{i+1} = x_i + \alpha \otimes Levy(\beta) \quad (2-5)$$

其中， α 是步长缩放因子， $Levy(\beta)$ 是 Levy 飞行路径。

另一种是寄主以概率 P_a 发现外来鸟蛋后，采用随机方式重新建巢：

$$x_{i+1} = x_i + r \otimes Heaviside(P_a - \varepsilon) \otimes (x_t - x_k) \quad (2-6)$$

其中， r, ε 是服从均匀分布的随机数， $Heaviside()$ 是跳跃函数， x_t, x_k 是其他任意的两个鸟巢。

2.4.2 萤火虫算法

萤火虫算法 (Firefly Algorithm, FA) 是 Yang 于 2008 年提出的一种启发算法。把空间各点看成萤火虫，利用发光强的萤火虫会吸引发光弱的萤火虫的特点，在发光弱的萤火虫向发光强的萤火虫移动的过程中，完成位置的迭代，从而找出最优位置。算法有以下三条假设：

1. 萤火虫不分性别，这样一个萤火虫将会吸引到所有其他的萤火虫。
2. 吸引力与它们的亮度成正比，对于任何两个萤火虫，不那么明亮的萤火虫被吸引，因此移动到更亮的一个，然而，亮度又随着其距离的增加而减少。
3. 如果没有比一个给定的萤火虫更亮的萤火虫，它会随机移动。

萤火虫的相对荧光亮度计算方式：

$$I = I_0 e^{-\gamma r_{ij}} \quad (2-7)$$

其中， I_0 表示最亮萤火虫的亮度，即自身 ($r = 0$ 处) 荧光亮度，与目标函数值相关，目标函数值越优，自身亮度越高； γ 表示光吸收系数，因为荧光会随着距离的增加和传播媒介的吸收逐渐减弱，所以设置光强吸收系数以体现此特性，可设置为常数； r_{ij} 表示萤火虫 i 与 j 之间的距离。

当萤火虫 i 的相对亮度小于萤火虫 j 时，向萤火虫 j 靠拢。位置的更新方式为：

$$\beta(r) = \beta_0 e^{-\gamma r_{ij}^2} \quad (2-8)$$

$$x_i = x_i + \beta(x_j - x_i) + \alpha(rand - 1/2) \quad (2-9)$$

其中， β_0 表示最大吸引度，即光源处 ($r = 0$ 处) 的吸引度。 α 为步长因子， $rand$ 为 $[0, 1]$ 上服从均匀分布的随机因子。

2.4.3 细菌觅食算法算法

细菌觅食算法(Bacterial Foraging Optimization, BFO)由 Passino 于 2002 年提出。通过模拟大肠杆菌菌落的觅食行为,不断地使用鞭毛游动和翻转,最终躲开有毒的地方并找到营养度高的位置,如图2-4所示。算法分为趋向性操作、复制操作和迁徙操作。

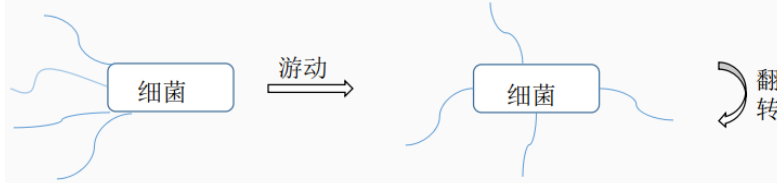


图 2-4 细菌的游动和翻转

1. 趋向性操作。这一操作模拟得是大肠杆菌的游动和翻转。在营养度高的地区,细菌会更多地游动,在营养度低的地区,细菌会更多地翻转,以逃出该地区。设细菌的种群规模为 S , 维度为 n 。细菌的觅食行为可以用以下公式表示:

$$\theta(i, j + 1, k, l) = \theta(i, j, k, l) + C(i) \times \phi(i, j) \quad (2-10)$$

$$\phi(i, j) = \frac{\Delta(i)}{\sqrt{\Delta^T(i)\Delta(i)}} \quad (2-11)$$

其中, $\theta(i, j, k, l)$ 表示细菌在第 j 次趋向性操作, 第 k 次复制操作和第 l 次迁徙操作时的位置。 $C(i)$ 是细菌 i 的趋向性步长。 $\phi(i, j)$ 表示细菌在第 j 次趋向性操作时的随机方向的单位向量。 $\Delta(i)$ 为随机向量。

2. 复制操作。复制操作的目的是将表现不好的细菌淘汰掉。首先,对种群按适应度排序,然后,前一半的细菌会复制一份覆盖后一半的细菌。保持种群数量不变的同时,实现优胜劣汰的机制。

3. 迁徙操作。在生物观察中发现,随着某些条件的改变,可能会使该地区的细菌突然死亡或迁移。算法通过迁徙操作模拟这一现象,提高种群的多样性。细菌会以一定的概率被清除,并随机生成一个新的细菌。

2.5 本章小结

第 3 章 基于 CS-FA 的双聚类算法

3.1 CS-FA 搜索算法

3.2 基于 CS-FA 搜索双聚类算法

3.3 实验结果及分析

3.4 本章小结

第 4 章 基于多目标 BFO 优化的双聚类算法

- 4.1 多目标优化问题的基本概念
- 4.2 多目标 BFO 搜索算法
- 4.3 基于多目标 BFO 搜索双聚类算法
- 4.4 实验结果及分析
- 4.5 本章小结

第 5 章 总结与展望