文章编号 1004-924X(2012)07-1475-10

基于递推遗传的模糊 3-划分熵多阈值 FISH 基因提取

尹诗白*,赵祥模,王卫星
(长安大学 信息工程学院,陕西 西安 710064)

摘要:针对现有寻优算法存在的重复计算问题,提出了基于递推遗传的模糊 3-划分熵多阈值荧光原位杂交(Fluorescence in Situ Hybridization,FISH)基因提取算法来提高用模糊划分熵算法提取多阈值 FISH 基因的效率。采用迭代验证法确定隶属度函数窗宽,并使用附加边界条件及灰度权重的隶属度函数对图像进行模糊 3-划分。为了提高阈值寻优的效率,引入递推算法将模糊熵的计算转化为递推过程,并保存部分不重复的递推结果用于后续的计算,最后采用遗传算法寻优,使得种群个体的计算能使用预存结果快速搜索全局最优阈值。对提取结果与几种常用算法进行了直观比较,并对处理时间、分类概率等性能指标进行了量化分析。对多幅不同类型的仿真人工图像和真实 FISH 图像的测试表明,处理时间仅为常用算法的 1%,错误划分概率小于 6.00×10^{-2}。提出的算法可以准确,高效地提取 FISH 基因目标。

关 键 词:FISH 图像;图像分割;模糊划分熵;递推算法;遗传算法
中图分类号:TP391.4 **文献标识码**:A **doi**:10.3788/OPE.20122007.1475

Fuzzy 3-partition entropy multilevel threshold approach based on recursive genetic algorithm for extracting FISH-labeled genes

YIN Shi-bai*, ZHAO Xiang-mo, WANG Wei-xing

(*School of Information Engineering, Chang'An University, Xi'an* 710064, *China*)
* *Corresponding author, E-mail: shibai. yin@gmail. com*

Abstract: A new fuzzy 3-partition entropy approach based on a fast recursive genetic algorithm was proposed to reduce the repeated computations and to improve the processing efficiency in extraction of FISH-labelled (Fluorescence In Situ Hybridization) genes. An iteration validation method was presented to determine the window width of the membership functions and the membership functions considering the boundary conditions and gray weights were selected to perform the fuzzy 3-partition. To improve the efficiency of selecting optimal thresholds, a recursive algorithm was presented to convert the computation of fuzzy entropy to a recursive process. Then, the no-repetitive results of the processing moments were stored for the succeeding genetic algorithm to compute the fitness of each individual. Finally, the optimal thresholds were searched by the genetic algorithm in a high speed. The result of the proposed algorithm was compared to those of the several common algorithms and the classification

收稿日期:2011-11-23;**修订日期**:2012-02-22.
基金项目:国家自然科学基金资助项目(No. 50978030);长安大学高校助研资助项目(No. CHD2010ZY003);新世纪优秀人才计划资助项目(No. NCET-05-0849);宁夏大学科学研究基金资助项目(No. ZR1122)

probability and run time were analyzed as the test criterion of optimal thresholds. By evaluating various types of simulated images and real FISH images, it shows that the run time of the proposed algorithm is 1% that of other common algorithms and the misclassification error is less than 6.00×10^{-2}. These results demonstrate that the proposed algorithm is effective for improving the precision and efficiency of extracting FISH-labelled genes.

Key words: FISH image; image segmentation; fuzzy partition entropy; recursive algorithm; genetic algorithm

1 引　言

荧光原位杂交（Fluorescence In Situ Hybridization，FISH）技术是对细胞内的目标基因进行荧光染色的常用方法，具有凸显目标，方便统计的优点，尤其适用于肿瘤细胞内癌变基因的统计分析，可为选择合适治疗方案奠定基础。采集该类 FISH 图像并使用图像分割技术高效、准确地提取基因目标，是统计分析的关键。受染色、光照和细胞分泌物的影响，FISH 图像中的目标灰度分布与干扰物灰度分布无明显界限，存在相互交叉及部分重叠的模糊问题，难以客观、准确地分割提取。

目前，分割 FISH 图像的主要方法为阈值法，又可进一步分为单阈值法和多阈值法。常见的人工自定义阈值法[1]为典型的单阈值法，即通过肉眼观察，人工搜索一个全局阈值。该方法因主观性强，精确度低，模糊性目标提取困难等缺陷，而逐渐被大量自动化多阈值分割法取代[2]。目前，模糊划分熵，作为优化复杂模糊性问题的全局最优技术而广受关注。该方法是将同一集合中不同类别的信息进行概率划分，并将最优阈值的选取转化为参数寻优的过程。如 Tao[3-4]选用 S 型，Π 型和 Z 型隶属度函数构建模糊 3-划分熵模型，并结合遗传、蚁群寻优算法确定分割阈值；Nandita[5]采用模糊最大熵准则和细菌觅食寻优算法自适应确定图像的分割阈值；Tenreiro[6]，Mehdi[7]和 Tang[8]等人以香农熵、Rényie 熵、Tsallis 熵作为模糊划分熵的评价标准，采用粒子群寻优算法确定模糊目标提取最优阈值。António[9]，Horng[10]使用基于蜂群寻优的最大模糊划分熵算法对基因染色体进行分割提取。相比于已有的多阈值分割算法，上述方法均能取得较好的分割效果，但计算时间较长，且随着阈值数量的增加而大幅增长。这主要由寻优算法中种群个体的重复计算造成。为了进一步提高 FISH 图像的分割效率，高效的模糊划分熵多阈值分割算法亟待研发。

近年来，递推算法作为数值计算的一类新的寻优算法广泛应用于图像分割等领域[11]。但是递推算法往往需要与穷举算法相结合，这限制了该算法在多阈值分割领域的应用。如 Tang[12]和 Bena[13]提出的基于模糊 2-划分熵的递推穷举算法，虽减少了重复计算，提高了效率；但应用于多阈值分割，不但递推公式难以推出，穷举搜索还会造成计算量的重复叠加。为此，针对 FISH 图像固有的模糊特性，本文提出了一种高效的基于递推遗传的模糊 3-划分熵多阈值 FISH 基因提取算法。该算法的主要贡献在于：将递推算法引入到多阈值模糊划分熵分割中，并可在此基础上，结合遗传算法高效地确定目标阈值。

2 基本思路

肿瘤细胞中染色体携带的原癌基因 erbB-2 及预癌基因 p53 经 DNA 探针 FISH 原位杂交后，在荧光显微镜下分别呈现红色和绿色，如图 1 所示。其中红色的 erbB-2 基因与背景色反差较大，易分离；绿色的 p53 基因受染色、光照、及细胞背景像素中部分绿色成分的干扰，易隐藏在细胞分泌的绿色絮状物中，难以简单、高效地提取。

考虑到绿色基因目标的灰度分布与干扰物的灰度分布存在互相交叉及部分重叠的模糊特性，提出采用递推遗传的模糊 3-划分熵算法提取基

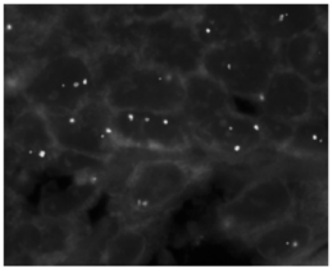


图 1　典型的 FISH 图像

Fig. 1　Typical FISH image

因目标。图 2 为算法流程图，以 FISH 图像为输入，主要包括隶属度函数窗宽确定、图像模糊 3-划分熵分割和基于递推遗传的阈值选择。

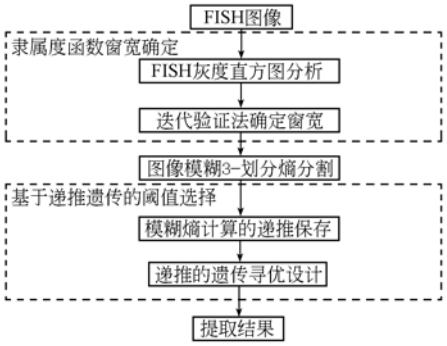


图 2　算法流程

Fig. 2　Algorithm flow

各步骤的概要说明如下：

(1)隶属度函数窗宽确定

Murthy[14] 指出模糊划分熵算法中，隶属度函数窗宽对阈值选取有一定的影响。为了更准确地确定阈值，首先分析了绿色通道下 FISH 灰度直方图的分布特点，特别是目标基因可能存在的灰度级区域，然后使用迭代验证法自适应地确定窗宽。

(2)图像模糊 3-划分熵分割

考虑到目标与干扰物的灰度分布存在相互交叉的模糊特性，选用附带边界条件及灰度权重的 S 函数、Ⅱ函数和 Z 函数作为隶属度函数，进行图像的模糊 3-划分熵分割。

(3)基于递推遗传的阈值选择

为了加快寻找模糊 3-划分熵分割时的最优阈值，将步骤(2)中的图像模糊熵公式作为输入，使用递推算法将其计算转化为递推过程，并保存不重复的递推结果。在此基础上，设计合适的遗传算法以快速选择最优阈值。

3　递推遗传的模糊 3-划分熵多阈值 FISH 基因提取

3.1　隶属度函数窗宽确定

3.1.1　FISH 灰度直方图分析

由上述分析可知，在大量 FISH 图像的绿色通道灰度直方图中找出目标的灰度分布特征是确定窗宽区域的基础。经实验反复测试表明，其直方图均呈非对称的单峰分布，且目标灰度主要分布于峰尾至其后的 ΔL 级灰度区域内，整体平滑毛刺较少，如图 3 所示，峰尾对应的下标为 90，该区域为[90，90＋ΔL]。因峰内高度均高于峰顶右边的峰尾高度(高度即某一灰度级上对应像素点的个数)，峰尾高度又略高于其后的灰度级高度，使得峰尾具有相对唯一性，寻找时可采用迭代验证法来确定。

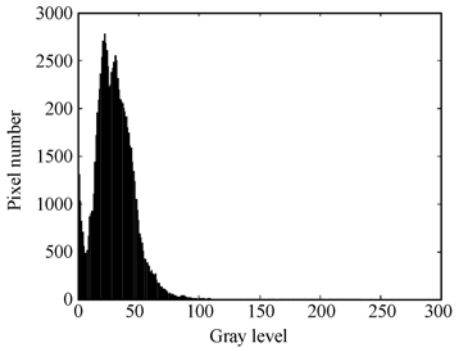


图 3　图 1 的绿色通道灰度直方图

Fig. 3　Grayscale histogram of Fig. 1 in green channel

3.1.2　迭代验证法确定窗宽

为了加快峰尾的搜索，减少不必要的操作，可根据该类 FISH 图像的峰尾高度信息预先估计一个近似的初始值，然后针对每幅 FISH 灰度直方图的分布情况，不断验证改进这一估计值，直到该值趋于稳定为止。具体步骤如下：

(1)假定峰尾初始高度：根据绿色通道下大量 FISH 灰度直方图的分布特性，假定一个全局的峰尾初始高度 h_0。

(2)确定峰尾下标：在灰度直方图上，寻找到灰度高度小于等于 h_0 的最大高度 h_{max} 所对应的灰度级下标，将其设为该图像的峰尾下标 L_{min}。

(3)迭代验证：从峰尾灰度级下标 L_{min} 开始，按下标增长的方向，统计高度非 0 的灰度级下标总数 S_{num}，此灰度级范围内所有不重复的高度 $\{h_1, h_2, \cdots, h_n\}$ 以及各高度所对应的不重复灰度

级下标个数{S_1,S_2,…,S_n}。使用公式(1)计算灰度下标 $L_{min}+1$ 到 255 范围内的高度加权平均值 h',其中权系数 $g(h_i)$为高度 h_i 所对应的不同灰度级下标个数 S_i 占灰度级下标总数的比例。由概率统计理论可知,若 $h'>h_{max}$,则自 L_{min}下标往后的灰度级高度总体高于 h_{max},更新 h_{max}的值为灰度级递增方向上,距离 L_{min}下标最近,对应的灰度级高度大于 h_{max}的 h'_{max}值;否则保存 h_{max}。重复执行(2),(3)操作,直到 h_{max}保持不变。

$$\begin{cases} h' = \sum_{i=1}^{n} h_i \times g(h_i),\ g(h_i) = S_i / S_{sum} \\ S_{sum} = S_1 + S_2 + \cdots + S_n \end{cases}. \tag{1}$$

(4)确定窗宽:将当前的峰尾灰度级下标 L_{min}作为窗宽下限,$L_{max}=L_{min}+\Delta L$ 作为窗宽上限,则该图像的窗宽[L_{min},L_{max}]确定完毕。不同 FISH 图像窗宽区域的确定只需重复执行(2)~(4)操作即可。

3.2 图像模糊 3-划分熵分割

在使用模糊 3-划分熵理论进行 FISH 图像的阈值分割时,首先将图像的灰度直方图信息映射到模糊域中。考虑到基因目标的灰度分布与周围介质的灰度分布存在相互交叉和部分重叠的模糊特性,选择附加边界条件及灰度权重的 S 函数,Π 函数,Z 函数[3][14-15]来定义 FISH 图像模糊域模糊子集的隶属度函数,详细的定义如式(2)~(4),对应的 3 个模糊集合 E_d,E_m 和 E_b 的概率划分见式(5)。

$$u_d=\begin{cases} 1, & k\leqslant a_1 \\ 1-\dfrac{(k-a_1)^2}{(c_1-a_1)\times(b_1-a_1)}\times\dfrac{n_{b_1}}{n_k}, & a_1<k\leqslant b_1 \\ \dfrac{(k-c_1)^2}{(c_1-a_1)\times(c_1-b_1)}\times\dfrac{n_{c_1}}{n_k}, & b_1<k\leqslant c_1 \\ 0, & k>c_1 \end{cases}, \tag{2}$$

$$u_m=\begin{cases} 0, & k\leqslant a_1 \\ \dfrac{(k-a_1)^2}{(c_1-a_1)\times(b_1-a_1)}\times\dfrac{n_{b_1}}{n_k}, & a_1<k\leqslant b_1 \\ 1-\dfrac{(k-c_1)^2}{(c_1-a_1)\times(c_1-b_1)}\times\dfrac{n_{c_1}}{n_k}, & b_1<k\leqslant c_1 \\ 1, & c_1<k\leqslant a_2 \\ 1-\dfrac{(k-a_2)^2}{(c_2-a_2)\times(b_2-a_2)}\times\dfrac{n_{b_2}}{n_k}, & a_2<k\leqslant b_2 \\ \dfrac{(k-c_2)^2}{(c_2-a_2)\times(c_2-b_2)}\times\dfrac{n_{c_2}}{n_k}, & b_2<k\leqslant c_2 \\ 0, & k>c_2 \end{cases}, \tag{3}$$

$$u_b=\begin{cases} 0, & k\leqslant a_2 \\ \dfrac{(k-a_2)^2}{(c_2-a_2)\times(b_2-a_2)}\times\dfrac{n_{b_2}}{n_k}, & a_2<k\leqslant b_2 \\ 1-\dfrac{(k-c_2)^2}{(c_2-a_2)\times(c_2-b_2)}\times\dfrac{n_{c_2}}{n_k}, & b_2<k\leqslant c_2 \\ 1, & k>c_2 \end{cases}, \tag{4}$$

$$\begin{cases} p_d = \sum_{k=1}^{255} p_k \times u_d(k) \\ p_m = \sum_{k=0}^{255} p_k \times u_m(k) \\ p_b = \sum_{k=0}^{255} p_k \times u_b(k) \end{cases}. \tag{5}$$

其中 k 是 FISH 图像的灰度级,a_1,b_1,c_1,a_2,b_2,c_2 是决定隶属度函数形状的参数变量,且满足:$L_{min}\leqslant a_1\leqslant b_1\leqslant c_1\leqslant a_2\leqslant b_2\leqslant c_2\leqslant L_{max}$。$n_{b_1}$,$n_{c_1}$,$n_{b_2}$,$n_{c_2}$为 FISH 直方图中灰度级 b_1,c_1,b_2,c_2 的像素个数。p_k 可由 $p_k=n_k/(M\times N)$($k=0,1,\cdots,255$)计算得到,其中 n_k 为 FISH 灰度直方图中灰度级为 k 的像素个数,$M\times N$ 为 FISH 图像的大小。3 个模糊集合划分的 FISH 图像总模糊熵如式(6):

$$H(a_1,b_1,c_1,a_2,b_2,c_2) = -p_d\log(p_d)-p_m\log(p_m)-p_b\log(p_b). \tag{6}$$

根据最大模糊熵准则,使 H 最大的 a_1,b_1,c_1,a_2,b_2,c_2 最佳组合,即为保留 FISH 图像最大信息量的分割阈值 T_1,T_2。其中 T_1 为隶属度函数 $u_d(k)$与 $u_m(k)$交点处的灰度值,T_2 为 $u_m(k)$与 $u_b(k)$交点处的灰度值。基于公式(1)~(3),T_1,T_2 可用公式(7)计算得到。

$$T_1=\begin{cases} a_1+\sqrt{(c_1-a_1)\times(b_1-a_1)\times n_{a_1}/(n_{a_1}+n_{b_1})}, \\ \qquad (a_1+c_1)/2\leqslant b_1\leqslant c_1 \\ c_1-\sqrt{(c_1-a_1)\times(c_1-b_1)\times n_{b_1}/(n_{b_1}+n_{c_1})}, \\ \qquad a_1\leqslant b_1\leqslant(a_1+c_1)/2 \end{cases}$$

$$T_2=\begin{cases} a_2+\sqrt{(c_2-a_2)\times(b_2-a_2)\times n_{a_2}/(n_{a_2}+n_{b_2})}, \\ \qquad (a_2+c_2)/2\leqslant b_2\leqslant c_2 \\ c_2-\sqrt{(c_2-a_2)\times(c_2-b_2)\times n_{b_2}/(n_{b_2}+n_{c_2})}, \\ \qquad a_2\leqslant b_2\leqslant(a_2+c_2)/2 \end{cases}. \tag{7}$$

3.3 基于递推遗传的阈值选择

3.3.1 模糊熵计算的递推保存

最大模糊熵的参数寻优并不是一个简单的工作,无论采用穷举寻优[12-13],还是群体寻

优[3-4][7-8]，P_d，P_m，P_b 和 u_d，u_m，u_b 均涉及（a_1，b_1，c_1，a_2，b_2，c_2）由（$L_{\min}$，$L_{\min}+1$，$L_{\min}+2$，$L_{\min}+3$，$L_{\min}+4$，$L_{\min}+5$）到（$L_{\max}-5$，$L_{\max}-4$，$L_{\max}-3$，$L_{\max}-2$，$L_{\max}-1$，$L_{\max}$）的重复计算。为提高效率，本文将 FISH 图像总模糊熵的计算转化为递推过程，并预存瞬间的递推结果用于后续操作，具体的递推过程如下：

首先，由公式(2)、(3)、(4)得出：$u_m=1-(u_b+u_d)$，则由式(5)可知：$1-p_m=p_b+p_d$。若将 u_b+u_d 合成一个隶属度函数 u_m'，即将 u_b 与 u_d 的函数曲线合为隶属度函数 u_m'的曲线，则 $1-p_m=p_m'$，且 $p_m'=\sum_{k=0}^{255} p_k * u_m'(k)$，上述基于 u_b，u_d，u_m 的模糊 3-划分简化为基于 u_m，u_m'的模糊 2-划分：

$$\begin{aligned}&H(a_1,b_1,c_1,a_2,b_2,c_2)=\\&-p_m'\log(p_m')-p_m\log(p_m)=\\&-(1-p_m)\log(1-p_m)-p_m\log(p_m)=\\&-\log(1-p_m)+p_m\log\left(\frac{1-p_m}{p_m}\right).\end{aligned}\tag{8}$$

从公式(8)可见，由 p_m 可计算出图像的总模糊熵。将式(3)代入式(5)的 p_m 中，得到：

$$\begin{aligned}p_m=\frac{1}{M\times N}\Bigg[&\frac{n_{b_1}}{(c_1-a_1)(b_1-a_1)}\sum_{k=a_1+1}^{b_1}(k-a_1)^2-\\&\frac{n_{c_1}}{(c_1-a_1)(c_1-b_1)}\sum_{k=b_1+1}^{c_1}(k-c_1)^2-\\&\frac{n_{b_2}}{(c_2-a_2)(b_2-a_2)}\sum_{k=a_2+1}^{b_2}(k-a_2)^2+\\&\frac{n_{c_2}}{(c_2-a_2)(c_2-b_2)}\sum_{k=b_2+1}^{c_2}(k-c_2)^2+\\&\sum_{k=b_1+1}^{b_2}p_k\Bigg].\end{aligned}\tag{9}$$

式(9)由 5 个部分组成，因每个部分均涉及整数操作，故采用递推的方式解决重复计算的问题。设第 1 部分为 P_{a_1,b_1}，则：

$$P_{a_1,b_1}=\sum_{k=a_1+1}^{b_1}(k-a_1)^2,\tag{10}$$

因 P_{a_1,b_1}随 b_1 而变化，故递推公式为：

$$\begin{aligned}P_{a_1,b_1}&=\sum_{k=a_1+1}^{b_1-1}(k-a_1)^2+(b_1-a_1)^2=\\&P_{a_1,b_1-1}+(b_1-a_1)^2.\end{aligned}\tag{11}$$

可进一步写成：

$$\begin{cases}P_{a_1,b_1}=P_{a_1,b_1-1}+(b_1-a_1)^2,b_1=a_1+2,\cdots,254\\P_{a_1,a_1+1}=1\end{cases}.\tag{12}$$

设第 2 部分为 P_{b_1,c_1}：

$$P_{b_1,c_1}=\sum_{k=b_1+1}^{c_1}(k-c_1)^2,\tag{13}$$

同样，公式(13)随 c_1 而变化，递推公式如下：

$$\begin{aligned}&P_{b_1,c_1}=P_{b_1,c_1-1}-2P^*_{b_1,c_1-1}+[(c_1-1)-b_1],\\&c_1=b_1+2,\cdots,255,\end{aligned}\tag{14}$$

其中：

$$\begin{cases}P^*_{b_1,c_1-1}=P^*_{b_1,c_1-2}-[(c_1-2)-b_1],\\\qquad\qquad c_1=b_1+3,\cdots,255\\P^*_{b_1,b_1+1}=0\end{cases}.\tag{15}$$

由于公式(9)的 3，4 部分与 1，2 部分相似，故可采用相同的递推方式。设第 3 部分为 P_{a_2,b_2}，第 4 部分为 P_{b_2,c_2}，递推结果参见 P_{a_1,b_1}，P_{b_1,c_1}。

第 5 部分设为 P_{b_1,b_2}，递推公式如下：

$$P_{b_1,b_2}=\sum_{k=b_1+1}^{b_2}p_k=p_{b_1,b_2-1}+p(b_2).\tag{16}$$

考虑到 3，4 与 1，2 部分递推重复，仅保存 $L_{\min}\leqslant a_1\leqslant b_1\leqslant c_1\leqslant b_2\leqslant L_{\max}$ 范围内 P_{a_1,b_1}，P_{b_1,c_1}，P_{b_1,b_2}的所有瞬间值（因 $L_{\min}\leqslant a_2\leqslant b_2\leqslant c_2\leqslant L_{\max}$，$p_{a_2,b_2}$，$p_{b_2,c_2}$的瞬间值可使用 p_{a_1,b_2}，p_{b_1,c_1}的预存结果），结合后续的遗传算法快速寻优。

3.3.2　递推的遗传寻优设计

由上述分析可知，递推的方式已减少大量的重复计算，但式(9)的 3，4 部分与 1，2 部分仍具有相似的推导过程，采用文献[12-13]中的穷举算法寻优，会有重复操作。遗传算法能自适应地调整搜索速度，获得全局最优阈值，在参数寻优方面，具有比其他算法[4][7-8]更好的搜索能力，故本文采用该法进一步确定最优阈值。

详细的递推遗传寻优设计如下，其中种群个体（a_1，b_1，c_1，a_2，b_2，c_2）采用 48 位的格雷码编码：

(1)随机产生 p 个体的初始种群，并进行适当的数学处理[3]，使种群个体中的参数满足窗宽区域的限制：$L_{\min}\leqslant a_1\leqslant b_1\leqslant c_1\leqslant a_2\leqslant b_2\leqslant c_2\leqslant L_{\max}$。

(2)按公式(8)计算每个个体的适应度值。式中 p_m 的计算可使用已预存的递推结果 P_{a_1,b_1}，P_{b_1,c_1}，P_{a_2,b_2}，P_{b_2,c_2}，P_{b_1,b_2}。

(3)将适应度较大的 X 个体无条件地复制到子代种群。

(4)对父代种群进行选择，交叉和变异等遗传算子操作，繁殖出子代剩余的 $P-X$ 个个体。选择方法采用常见的轮盘赌；交叉算子采用多点交叉，染色体的每 8 位间插一个交叉点；交叉及变异概率采用文献[16]的方法，通过种群中个体适应

度大小自适应的调整初始交叉率 P_{c_1}，P_{c_2} 和变异率 P_{m_1}，P_{m_2}。

(5)将子代个体与父代个体的适应度值相比较，若前者大于后者，用子代替换父代，否则保留父代。重复执行(2)～(5)，直到进化预定的代数。

(6)得到(a_1，b_1，c_1，a_2，b_2，c_2)的最佳组合，并利用公式(7)计算最佳阈值。

4　实验结果及性能分析

为了验证本算法的有效性，选用仿真的人工图像和大量真实的 FISH 图像进行提取测试，算法实现均在 Inter G840 CPU，内存为 2 G 的条件下，使用 VC++6.0 编程完成。由于本算法的主要贡献在于引入递推算法减少重复的计算量，所以更关注算法效率。比较算法除选用常用的人工阈值法外[1]，还选用文献[3][4][8]提到的基于遗传算法(Genetic Algorithm，GA)，蚁群算法(Ant Colony Optimization，ACO)和粒子群算法(Particle Swarm Optimization，PSO)的模糊划分熵多阈值提取方法，进一步验证递推遗传的阈值选择效率。

经大量实验及综合因素考虑，实验相关参数设置如下：峰尾初始高度 $h_0=150$；窗宽区域 $\Delta L=60$；种群数目 $N=100$；交叉和变异概率 $P_{c_1}=0.9$，$P_{c_2}=0.6$，$P_{m_1}=0.1$，$P_{m_2}=0.001$[16]，最大繁衍代数 $G_{\max}=200$。GA 参数同上，其它关键参数设置如下：

PSO：种群大小 $PS=100$，$c_1=c_2=2$，惯性因数从 $w_{\max}=0.95$ 到 $w_{\min}=0.45$ 线性递减。

ACO：种群大小 $PS=100$，其余参数同文献[4]。

量化指标除选用运行时间外，还选用了误分概率(False-positive，FP)，计算背景像素误分为对象的概率，公式见(17)；正分概率(True-positive，TP)，计算真正的对象像素划分为对象的概率，公式见(18)；错误划分概率 $p(\text{error})$，计算背景像素误分为对象的概率与对象像素误分为背景的概率之和，公式见(19)。Backgnd 为真实情况下的所有背景像素，Obj 为真实情况下的所有对象像素，Obj_c 为分割后的所有对象像素，Backgnd_c 为分割后的所有背景像素[17]。

$$FP=p(Backgnd)\,p(Obj_c|Backgnd), \quad (17)$$

$$TP=p(Obj)\,p(Obj_c|Obj), \quad (18)$$

$$p(error)=p(Backgnd)\,p(Obj_c|Backgnd)+p(Obj)\,p(Backgnd_c|Obj), \quad (19)$$

4.1　仿真图像的对比试验

采用仿真的方式构建与 FISH 图像大小一致的人工仿真图像，旨在已知 FISH 图像真实情况下对上述算法进行定量分析。根据实际预癌基因 p53 的形状，大小，单位区域个数创建目标物，并以泊松噪声，背景噪声，非特异染色噪声，点伸展函数来再现 FISH 图像中的主要噪声[17]。由于目标物提取的主要难点在于其灰度级与干扰物的灰度级存在交叉重叠的模糊问题，故采用业内广泛认可的泊松分布作为 FISH 图像中灰度级的噪声分布。

表 1 是 5 种算法在不同程度噪声干扰下的分类概率。其中 STD 表示泊松噪声为 1 倍时的标准噪声，表达式 2×，3× 分别为泊松噪声的强化倍数。背景噪声，非特异染色噪声及点伸展函数均保持常量。从中可见，同等条件下本文算法的划分概率与 GA 一致，FP，$p(\text{error})$ 值均低于人工阈值法，PSO 及 ACO，且运行时间最短，甚至可以忽略不计。这主要是由于引入的递推算法并没有改变后续遗传算法的寻优能力，且在参数寻优的特定场合下，能获得比 PSO，ACO 更好的最优阈值[18]；同时，递推保存的计算方式减少了重复的计算量，大大缩短了处理时间，使本文算法在确保精度的前提下，具有较高的运行效率。

表 1　5 种算法的分类概率及处理时间比较

Tab. 1　Comparison of classification probability and processing time for 5 s

噪声	方法名	FP /%	TP /%	$p(\text{error})$ /%	时间 /s
STD	GA	4.37	91.8	4.39	6.574
	人工阈值	7.01	94.2	7.12	>10
	PSO	7.51	94.5	7.62	7.455
	ACO	7.69	94.9	7.73	7.008
	本文算法	4.37	91.8	4.39	0.052
2×	GA	5.02	92.8	5.08	6.604
	人工阈值	8.31	96.0	8.36	>10
	PSO	8.68	96.4	8.74	7.621
	ACO	8.73	96.8	8.85	7.143
	本文算法	5.02	92.8	5.08	0.054
3×	GA	5.92	93.3	6.99	7.105
	人工阈值	9.03	97.5	9.26	>10
	PSO	9.12	98.1	9.37	7.757
	ACO	9.57	98.8	9.74	7.434
	本文算法	5.92	93.3	6.99	0.056

4.2　真实图像的对比试验

根据目标基因受干扰物影响的模糊程度，从大量测试图像中选择了 3 种不同类型的典型图像进行提取结果的对比说明。分别为少量绿色絮状物的 FISH 图(4(a))，对应的直方图(4(c))单峰区域较窄，模糊度低；适量絮状物的 FISH 图(5(a))，其直方图(5(c))单峰区域适中，模糊度一般；大量絮状物的 FISH 图(6(a))，其直方图(6(c))单峰区域较宽，模糊度高。图(4～6 中的图(b))为对应的绿色通道灰度图，其中箭头显示标出了预癌基因 p53 的位置。红线勾画的目标为专业医师人工判断结果，为衡量各算法的 FP 和 p(error)值提供参考依据。分割后的结果如图 4～6中的图(d)～(h)所示，其中方框内显示标出的是各算法提取出的绿基因目标。相应的分割阈值及处理时间见表 2，划分概率见表 3。从中可见，真实图像的提取结果与仿真图像一致，本文算法仍具有较好的分割精度和最短的运行时间。

表 2　5 种算法的阈值及处理时间比较

Tab. 1　Comparison of thresholds and processing time from 5 algorithms

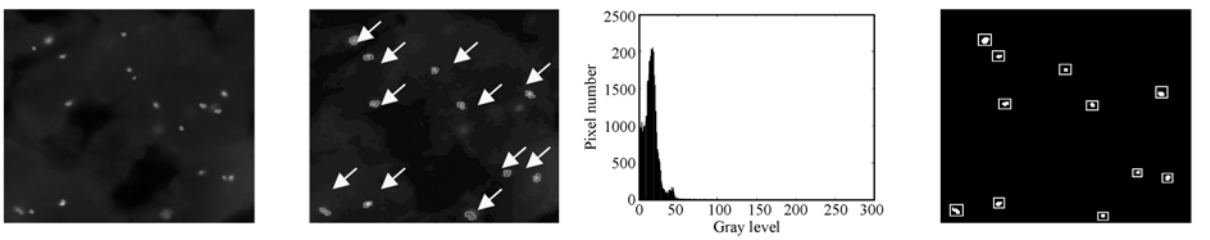
图像	人工阈值法		GA		PSO		ACO		本文算法	
	阈值	时间/s	阈值	时间/s	阈值	时间/s	阈值	时间/s	阈值	时间/s
图 4	84	＞10.0	79,92	6.753	82,95	7.672	83,97	7.135	79,92	0.050
图 5	114	＞10.0	106,134	6.815	109,137	7.817	104,140	7.302	106,134	0.053
图 6	146	＞10.0	141,167	7.291	145,171	7.850	151,174	7.526	141,167	0.055

表 3　5 种算法的分类概率对比

Tab. 3　Comparison of classification probability from 5 algorithms　　(%)

图像	人工阈值法		GA		PSO		ACO		本文算法	
	FP	p(error)	FP	p(error)	FP	p(error)	FP	p(error)	FP	p(error)
图 4	7.06	7.15	4.39	4.42	7.56	7.67	7.73	7.78	4.39	4.42
图 5	8.43	8.49	5.12	5.19	8.75	8.84	8.86	8.97	5.12	5.19
图 6	9.08	9.33	5.96	6.95	9.18	9.43	9.63	9.82	5.96	6.95

使用本算法对 FISH 图 4(a)处理后，确定的窗宽区域为[58,118]，最佳阈值与 GA 选取的结果一致，即 $T_1=79$，$T_2=92$，图像按目标像素的灰度级大小分为黑、白、灰 3 个区域如图 4(h)、(e)所示，即灰度级较高的白色目标、较低的灰色目标及黑色背景区域。人工阈值 $T=84$ 介于 $T_1=79$ 与 $T_2=92$ 之间，故对应相同的目标，本文算法提取的灰色目标面积略大，白色目标面积略小，如图 4(d)、(h)。PSO，ACO 提取结果图 4(f)、(g)与本文算法近似，但划分概率 FP，p(error)值和处理时间均大于本文算法，见表 2，3。

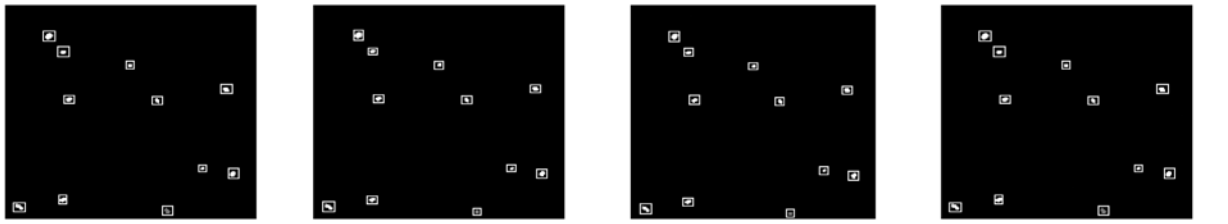


(a)初始图像　(a) Original image

(b)对应的绿色通道灰度图　(b) Corresponding gray image in green channel

(c)对应的绿色通道直方图　(c) Corresponding histogram in green channel

(d)人工阈值法，T=84　(d) Manual segmentation, T=84



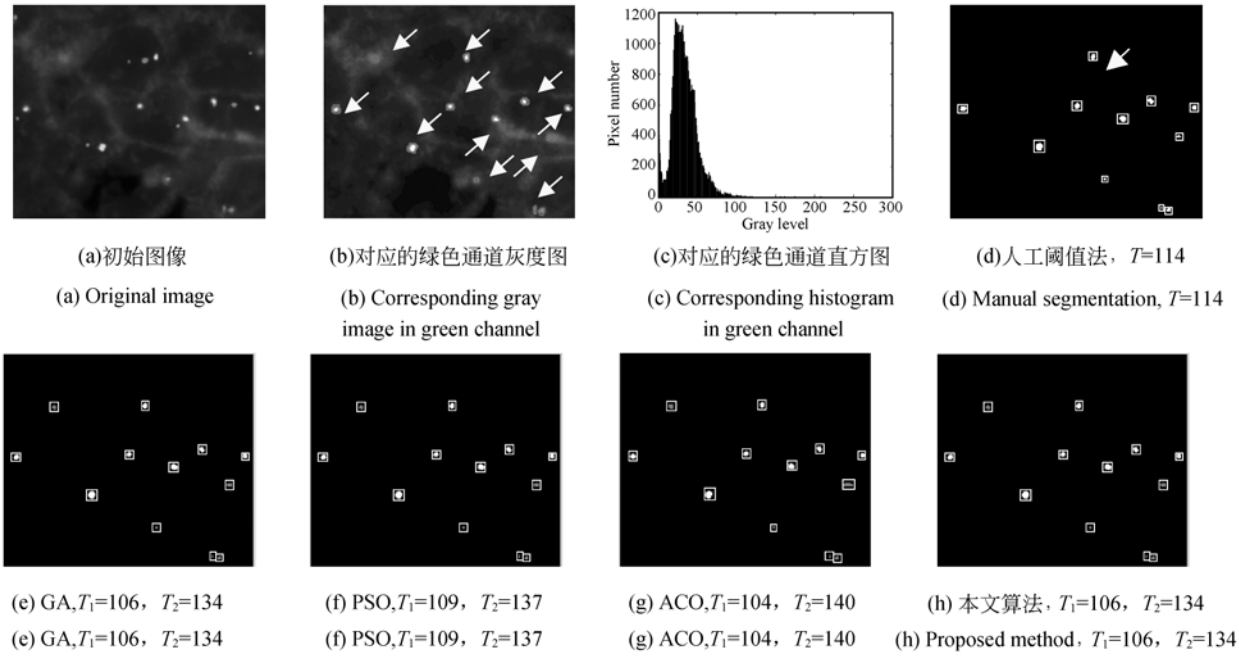
(e) GA,T_1=79 T_2=92　(f) PSO,T_1=82 T_2=95　(g) ACO,T_1=83 T_2=97　(h) 本文算法，T_1=79 T_2=92

(e) GA,T_1=79 T_2=92　(f) PSO,T_1=82 T_2=95　(g) ACO,T_1=83 T_2=97　(h) Proposed method，T_1=79 T_2=92

图 4　FISH 图 4(a)的分割结果对比

Fig. 4　Comparison of segmentation results of FISH image 4(a) from 5 algorithms

FISH 图 5～6 中的图(a)的分割结果与此类似。本文算法确定的窗宽区域分别为[94,154]和[129,189],最优阈值同 GA,提取结果图 5～6 中的图(h)、(e)与人工勾画的结果图 5～6 中的图(b)一致。人工阈值法,PSO,ACO 均存在不同程度的目标丢失,如图 5～6 中的图(d)、(f)、(g)箭头所示。对应的划分概率 FP,p(error)值和处理时间大于本文算法,详见表 2,3。



(a)初始图像　(b)对应的绿色通道灰度图　(c)对应的绿色通道直方图　(d)人工阈值法，T=114

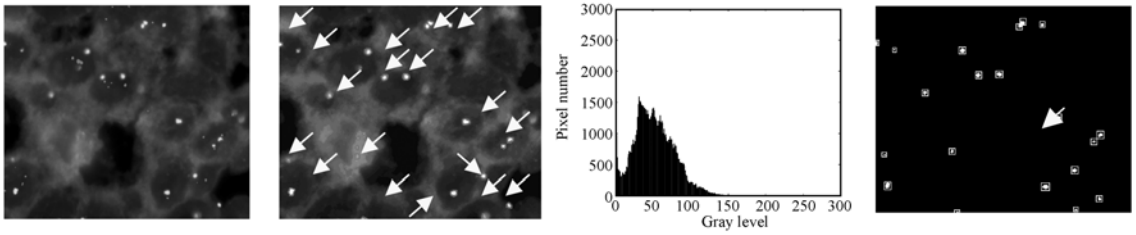
(a) Original image　(b) Corresponding gray image in green channel　(c) Corresponding histogram in green channel　(d) Manual segmentation, T=114

(e) GA,T_1=106，T_2=134　(f) PSO,T_1=109，T_2=137　(g) ACO,T_1=104，T_2=140　(h) 本文算法，T_1=106，T_2=134

(e) GA,T_1=106，T_2=134　(f) PSO,T_1=109，T_2=137　(g) ACO,T_1=104，T_2=140　(h) Proposed method，T_1=106，T_2=134

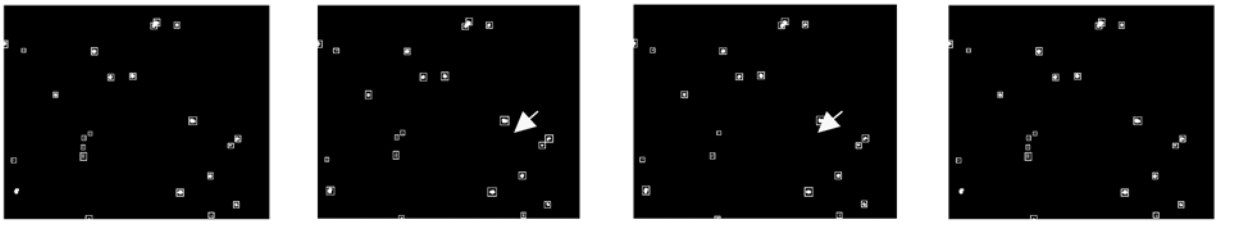
图 5　FISH 图 5(a)的分割结果对比

Fig. 5　Comparison of segmentation results of FISH image 5(a) from 5 algorithms



(a)初始图像　(b)对应的绿色通道灰度图　(c)对应的绿色通道直方图　(d)人工阈值法，T=146

(a) Original image　(b) Corresponding gray image in green channel　(c) Corresponding histogram in green channel　(d) Manual segmentation, T=146



(e) GA, T_1=141, T_2=167　(f) PSO, T_1=145, T_2=171　(g) ACO, T_1=151, T_2=174　(h) 本文算法, T_1=141, T_2=167

(e) GA, T_1=141, T_2=167　(f) PSO, T_1=145, T_2=171　(g) ACO, T_1=151, T_2=174　(h) Proposed method, T_1=141, T_2=167

图 6　FISH 图 6(a)的分割结果对比

Fig. 6　Comparison of segmentation results of FISH image 6(a) from 5 algorithms

4.3　时间复杂度析

从表 1 和表 2 可见，不同噪声环境下，图像的处理时间仅差 10^{-3} 个数量级。这主要是由本文算法的时间复杂度决定，整个执行时间包括 3 个部分，第 1 部分是阈值区域的搜索时间，执行对象为统计好的灰度直方图，搜索范围为 0～255 灰度级，故搜索时间与图像大小无关，仅与灰度分布有关。第 2 部分是瞬间变量的递推保存时间，因递推公式及递推范围不变，运行时间相对独立。第 3 部分是遗传算法寻优时间，因个体适应度的计算可使用预存的瞬间结果，耗时仅与遗传算法的参数有关，如迭代次数，停止条件等。综上所述，本算法具有复杂度低，耗时短及鲁棒性强的特点。

5　结　论

针对现有的模糊划分熵算法在多阈值 FISH 基因提取时存在效率低和计算量重复的问题，提出了一种基于递推遗传的模糊 3-划分熵多阈值 FISH 基因提取算法。该算法对待处理的 FISH 图像采用迭代验证法确定窗宽区域，然后选用附加边界条件及灰度权重的 S，Π，Z 函数对图像进行模糊 3-划分。重点研究了阈值寻优的效率问题，为了确保精度，本文将递推算法引入到多阈值模糊划分熵的计算中，将不同变量的组合计算转化为递推的过程，然后保存不重复的递推结果，用于后续的遗传算法，使得种群个体的计算能使用预存的递推结果，高效准确地确定全局最优阈值。通过多幅人工仿真图像及真实 FISH 图像的测试，得到以下结论：与常用的基因提取算法相比，本文算法所提出的递推遗传寻优策略，延续了遗传算法的寻优能力，提取精度明显优于经典的人工阈值法和一般寻优策略的模糊划分熵多阈值提取算法，错误划分概率小于 6.00×10^{-2}；引入的递推算法有效减少了重复的计算量，结合遗传算法寻优使处理时间大大缩短，同等条件下，处理时间仅为其他常用算法的 1%。综上，本文算法可以更加高效、准确地提取 FISH 基因目标。

参考文献：

[1] TANKE H J. Studies of the human genome fluorescent in situ hybridization and image analysis[J]. *Biology of Cell*, 1991, 93(2): 33-41.

[2] POLETTI E, ZAPPELLI F, RUGGERI A, *et al*.. A review of thresholding strategies applied to human chromosome segmentation [J]. *Computer Methods and Programs in Biomedicine*, 2012, 6(2): 121-132.

[3] TAO W B, TIAN J W, JIAN L. Image segmentation by three-level thresholding based on maximum fuzzy entropy and genetic algorithm[J]. *Pattern Recognition Letters*, 2003, 24(16): 3069-3078.

[4] TAO W B, JIN H, LIU L M. Object segmentation using ant colony optimization algorithm and fuzzy entropy[J]. *Pattern Recognition Letters*, 2007, 28(7): 788-796.

[5] NANDITA S, AMITAVA C, SUGATA M. An adaptive bacterial foraging algorithm for fuzzy entropy based image segmentation[J]. *Expert Systems with Applications*, 2011, 38(12): 15489-15498.

[6] MACHADO J A T, COSTA A C, QUELHAS M D. Shannon, Rényie and Tsallis entropy analysis of DNA

using phase plane[J]. *Nonlinear Analysis: Real World Applications*, 2011, 12(6): 3135-3144.

[7] MEHDI S, HASSAN S, ARIAA. Minimum entropy control of chaos via online particle swarm optimization method[J]. *Applied Mathematical Modelling*, 2011, 21(10): 171-195.

[8] TANG Y G, DI Q Y, GUAN X P, *et al.*. Threshold selection based on fuzzy Tsallis entropy and particle swarm optimization[J]. *Neuro Quantology*, 2008, 6(4): 412-419.

[9] MACHADO J A T, COSTA A C, QUELHAS M D. Analysis and visualization of chromosome information[J]. *Gene*, 2011, 49(1): 81-87.

[10] HORNG M H. Multilevel thresholding selection based on the artificial bee colony algorithm for image segmentation[J]. *Expert Systems with Applications*, 2011, 38(11): 13785-13791.

[11] PENG H J, WU Z G , ZHONG W X. Fourier expansion based recursive algorithms for periodic Riccati and Lyapunov matrix differential equations [J]. *Journal of Computational and Applied Mathematics*, 2011, 235(12): 3571-3588.

[12] TANG Y G, MU W W, YING Z, *et al.*. A fast recursive algorithm based on fuzzy 2-partition entropy approach for threshold selection[J]. *Neurocomputing*, 2011, 74(17): 3072-3078.

[13] BENABDELKADER S, BOULEMDEN B. Recursive algorithm based on fuzzy 2-partition entropy for 2-level image thresholding[J]. *Pattern Recognition*, 2005, 38(8): 1289-1294.

[14] MURTHY C A, PAL S K. Fuzzy thresholding mathematical framework, bound functions and weighted moving average technique[J]. *Pattern Recognition Letter*, 1990, 11(2): 197-206.

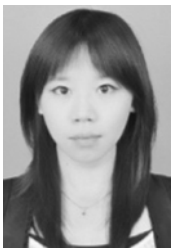
[15] 周学成,罗锡文,严小龙,等.基于遗传算法的原位根系 CT 图像的模糊阈值分割[J].中国图象图形学报,2009,14(4): 682-687.
ZHOU X CH, LUO X W, YAN X L, *et al.*. A fuzzy thresholding segmentation for plant root CT images based on genetic algorithm[J]. *Journal of Image and Graphics*, 2009, 14(4): 682-687. (in Chinese)

[16] 张怀柱,向长波,宋建中,等.改进的遗传算法在实时图像分割中的应用[J].光学 精密工程, 2008,16(2): 334-337.
ZHANG H Z, XIANG CH B, SONG J ZH, *et al.*. Application of improved adaptive genetic algorithm to image segmentation in realtime[J]. *Opt. Precision Eng.*, 2008, 16(2): 334-337. (in Chinese)

[17] RICHARD A. RUSSELL N M A, DAVIA A S, *et al.*. Segmentation of fluorescence microscopy images for quantitative analysis of cell nuclear architecture[J]. *Biophysical Journal*, 2009, 96(8): 3379-3389.

[18] GOZDE B, DERYA B, ALP K. An incremental genetic algorithm for classification and sensitivity analysis of its parameters[J]. *System with Application*, 2011, 38(3): 2609-2620.

作者简介：



尹诗白(1984－)，女，四川成都人，博士研究生，2006 年、2009 年于西安工业大学分别获得学士和硕士学位，主要从事医学图像处理，机器视觉方面的研究。E-mail: shibai.yin@gmail.com

王卫星(1959－)，男，湖南人，博士，教授，博士生导师，1997 年于瑞典皇家工学院获得博士学位，现为长安大学信息工程学院特聘教授，主要从事医学图像处理，目标识别和机器视觉方面的研究。E-mail: wxwang@chd.edu.com

导师简介：

赵祥模(1966－)，男，重庆人，博士，教授，博士生导师，2003 年，2006 年于长安大学分别获得硕士和博士学位，现为长安大学校长助理，主要从事图像处理与目标识别方面的研究。E-mail: xmzhao@chd.edu.cn

（本栏目编辑：曹　金）