

基因识别及其算法研究

古毅伟¹, 王松², 张旭¹, 全辉¹

(1. 北京邮电大学 计算学院, 北京 100876)

(2. 中国科学院软件研究所, 北京 100190)

摘要: 针对基因识别问题, 基于 DNA 序列的 3 周期这一性质, 首先给出了 DNA 序列功率和信噪比的快速算法并讨论了不同物种基因类型的阈值确定方法; 在此基础上, 本文建立了基于背景噪声抑制和频谱平滑的 SNR 频谱预处理模型, 经过预处理后的频谱不仅大幅度抑制了背景噪声, 同时保留了 SNR 频谱的模式特征. 在编码序列识别上, 本文对经典的 EPND 预测算法进行了改进, 使用改进的 EPND 算法对经过预处理后频谱进行基因识别, 实验结果显示这种基因识别模型具有优异的基因识别性能, 比传统直接使用基于滑动窗口 DFT 的 EPND 识别算法在敏感度、特异性等评价指标上提高了 2%-12% 左右.

关键词: 基因识别; 频谱分析; 噪声抑制; EPND; 滤波; 信噪比

1 问题描述

详见第九届全国研究生数学建模竞赛 A 题

2 模型分析

问题一 a) 针对使用 DFT 定义计算 DNA 序列的功率谱或信噪比计算量较大的问题, 本文探求分别针对 Voss 和实数映射的功率谱与信噪比的快速计算方法; b) 探讨 *Z-curve* 映射的频谱和 Voss 映射下的频谱之间的关系.

问题二 对于不同物种基因序列, 研究其阈值确定方法和阈值结果. 并分析按照频谱或信噪比特征将编码与非编码区间分类的有效性以及分类错误问题.

问题三 针对现有基因识别算法的不足设计新的基因识别算法探测、预报尚未被注释的 DNA 序列的外显子, 并评估算法有效性.

3 模型建立与求解

3.1 问题一求解

3.1.1 功率谱与信噪比的快速算法

使用 Voss 或实数映射方法数值化基因序列后利用 DFT 定义进行功率谱和信噪比的计算, 其时间复杂度为 $O(N^2)$, 对于很长的 DNA 序列使用 DFT 的计算量将会很大, 一种可能

收稿日期: 20 -00-00

资助项目:

的降低计算复杂度的方法是使用时间复杂度为 $O(N \log N)$ 的快速 Fourier 变换 (FFT)[Ref] 来代替 DFT 定义进行计算. 特别地, 对于 $N/3$ 频率点处的功率和信噪比可以找到时间复杂度更优的为 $O(N)$ 的快速算法.

1) VOSS 映射下 $N/3$ 频率点处的功率快速算法

将核苷酸符 $b \in I = \{A, T, G, C\}$ 出现在该序列的 $0, 3, 6, \dots, N-3$ 与 $1, 4, 7, \dots, N-2$ 以及 $2, 5, 8, \dots, N-1$ 等位置上的频数分别记为 x_b, y_b 和 z_b , 则 $\frac{N}{3}$ 处的总功率谱值为:

$$\begin{aligned} P[\frac{N}{3}] &= \sum_{b \in I} \left| U_b[\frac{N}{3}] \right|^2 = \sum_{b \in I} \left| \sum_{n=0}^{N-1} u_b[n] \cdot e^{-j \frac{2\pi n \cdot \frac{N}{3}}{N}} \right|^2 \\ &= \sum_{b \in I} (x_b^2 + y_b^2 + z_b^2 - x_b y_b - x_b z_b - y_b z_b) \\ &= \sum_{b \in I} (x_b, y_b, z_b) \begin{pmatrix} 1 & -1/2 & -1/2 \\ -1/2 & 1 & -1/2 \\ -1/2 & -1/2 & 1 \end{pmatrix} \begin{pmatrix} x_b \\ y_b \\ z_b \end{pmatrix} = \sum_{b \in I} X_b^T M X_b \end{aligned}$$

2) VOSS 映射下 $N/3$ 频率点处的信噪比快速算法

假设 DNA 序列的长度为 N , DNA 序列的四种核苷酸 A, C, T, G 的出现次数为 N_A, N_C, N_G 和 N_T . 因此指示序列 $u_b(n)$ 的功率谱为:

$$|U_b|^2 = \sum_{k=0}^{N-1} |U_b[k]|^2 = N \cdot N_b \quad (1)$$

所以全部序列的总功率 E 为:

$$E = \sum_{b \in I} |U_b|^2 = N \cdot N_A + N \cdot N_C + N \cdot N_G + N \cdot N_T = N^2 \quad (2)$$

根据 Parseval 定理, 可以得到

$$\sum_{n=0}^{N-1} |u_b[n]|^2 = \frac{1}{N} \sum_{k=0}^{N-1} |U_b[k]|^2 \quad (3)$$

观察上式, 易得出此等式的左边等于 N_b , 即:

$$|U_b|^2 = \sum_{k=0}^{N-1} |U_b[k]|^2 = N \cdot N_b \quad (4)$$

因此, DNA 序列的总功率谱 E 能由以下式子得到:

$$E = \sum_{b \in I} |U_b|^2 = N \cdot N_A + N \cdot N_C + N \cdot N_G + N \cdot N_T = N^2 \quad (5)$$

由式 3-8 和题目中给出的信噪比定义式可得:

$$R = \frac{P[\frac{N}{3}]}{E} = \frac{P[\frac{N}{3}]}{E/N} = \frac{\sum_{b \in I} X_b^T M X_b}{N} \quad (6)$$

3) 实数映射下 $N/3$ 频率点处的功率快速算法 (类似于 VOSS 映射)

$$P[\frac{N}{3}] = (x_b, y_b, z_b) \begin{pmatrix} 1 & -1/2 & -1/2 \\ -1/2 & 1 & -1/2 \\ -1/2 & -1/2 & 1 \end{pmatrix} \begin{pmatrix} x_b \\ y_b \\ z_b \end{pmatrix} = X_b^T M X_b$$

其中 $x_b = \sum_{i \in \text{position1}}^{u(i)}$, $y_b = \sum_{i \in \text{position2}}^{u(i)}$, $z_b = \sum_{i \in \text{position3}}^{u(i)}$ (Position1: $0, 3, 6, \dots$ 与 Position2: $1, 4, 7, \dots$ 及 Position3: $2, 5, 8, \dots$).

4) 实数映射下 $N/3$ 频率点处的信噪比快速算法 (类似于 Voss 映射)

$$R = \frac{P[\frac{N}{3}]}{\bar{E}} = \frac{P[\frac{N}{3}]}{E/N} = \frac{\sum_{b \in I} X_b^T M X_b}{N_C + 4N_G + 9N_T}$$

N_C, N_G, N_T 分别代表 C, G, T 三种核苷酸在 DNA 序列中出现的频数.

3.1.2 不同映射 (Voss 映射与 Z-curve 映射) 下功率 (信噪比) 之间关系

1) 信噪比关系

对于两种线性映射关系

$$\begin{pmatrix} x[n] \\ y[n] \\ z[n] \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{pmatrix} \begin{pmatrix} u_A[n] \\ u_C[n] \\ u_G[n] \\ u_T[n] \end{pmatrix} = \begin{pmatrix} \beta_1 & \beta_2 & \beta_3 & \beta_4 \end{pmatrix} \begin{pmatrix} u_A[n] \\ u_C[n] \\ u_G[n] \\ u_T[n] \end{pmatrix}$$

β_i 为列向量根据文献 [2] 中证明了: (引理) 如果列向量 β_i 满足以下两个条 (i) $\forall i (1 \leq i \leq 4), \|\beta_i\|^2 \equiv c_1$, c_1 为常量; (ii) $\forall i, j (1 \leq i, j \leq 4, i \neq j)$, 内积 $\langle \beta_i, \beta_j \rangle \equiv c_2$ c_2 为常量. 则 $x[n], y[n], z[n]$ 序列信噪比 R_Z 与 $u_i[n]$ 序列信噪比 R_I 满足关系 $R_S = \frac{c_1 - c_2}{c_1} R_I$

由 Voss 映射序列与 Z-curve 映射序列的线性关系

$$\begin{pmatrix} \Delta x[n] \\ \Delta y[n] \\ \Delta z[n] \end{pmatrix} = \begin{pmatrix} 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} u_A[n] \\ u_C[n] \\ u_G[n] \\ u_T[n] \end{pmatrix} \quad \begin{cases} \beta_1 = \{1, 1, 1\} \\ \beta_2 = \{-1, 1, -1\} \\ \beta_3 = \{1, -1, -1\} \\ \beta_4 = \{-1, -1, 1\} \end{cases}$$

易见 (i) $\forall i (1 \leq i \leq 4) \quad \|\beta_i\|^2 \equiv 3 = c_1$, 即满足上述假设 (i); (ii) $\forall i, j (1 \leq i, j \leq 4, i \neq j)$, $\langle \beta_i, \beta_j \rangle \equiv -1 = c_2$ 即满足上述假设 (ii). 从而由引理可得 Z-curve 映射序列信噪比 R_Z 与 Voss 映射序列信噪比 R_I 之间存在比例关系:

$$R_Z = \frac{4}{3} R_I \quad (7)$$

2) 频谱关系

假设一个 DNA 序列 $u(n)$ 长度为 N , $u(n)$ 的滑动窗口, M 点 DFT 定义为:

$$U(m, k) \triangleq \sum_{n=0}^{M-1} u(n+m) e^{-j2\pi nk/M} \quad (8)$$

窗口开始的点设 $m = 0, P, \dots, (N-1)/P$ (若 $(N-1)/P$ 整数, $u(n)$ 取 0), P 为窗口滑动的次数, 如果 $P = 1$, 则窗口每次滑动一个核苷酸的距离, 而如果 $P = 3$, 则窗口的位移是密码子的基数. 为了造成 $u(n)$ 周期 3 的情况, 我们设 $M = 3L$ (L 为正整数) 然后设频率 k 为 $L = M/3$, 则等式 (3-14) 可变换为:

$$U(m) \triangleq U(m, L) = \sum_{n=0}^{M-1} u(n+m) e^{-j2\pi n/3} \quad (9)$$

继续对该公式进行 $P=3$ 的多项展开, 使 $u(m+n) \equiv u_m(n)$ 我们能重新得到 $U(m)$ 的表达式

$$U(m) = \sum_{r=0}^2 \sum_{n=r, r+3, \dots}^{\lfloor \frac{N-1}{3} \rfloor} u_m(3n+r) e^{-j2\pi r/3} \triangleq \sum_{r=0}^2 U_{m,r} e^{-j2\pi r/3} \quad (10)$$

接着计算 Voss 映射的 DNA 频谱 P_S , 首先找到 4 个 Voss DFT 序列的多相组成, 如下所示:

$$U_{bm} \triangleq U_{bm0} + U_{bm1}e^{-j2\pi/3} + U_{bm2}e^{-j4\pi/3}, b \in I \quad (11)$$

根据上文的假设条件, DNA 总频谱 P_S 表达式为:

$$P_S = |U_{Am}|^2 + |U_{Tm}|^2 + |U_{Gm}|^2 + |U_{Cm}|^2 \quad (12)$$

由于 $|U_{bm}|^2 = U_{bm}U_{bm}^*$, * 表示复数. 可以推出:

$$|U_{bm}|^2 = 1/2 \sum_{r=0}^2 [U_{bm_r} - U_{bm_q}]^2$$

将上述公式带入 (12), P_S 的计算公式为:

$$P_S = 1/2 \sum_{b \in I} \sum_{r=0}^2 [U_{bm_r} - U_{bm_q}]^2 \quad (q = (r+1) \bmod 3) \quad (13)$$

对于 Z-curve 映射, 首先分别找到 $x(n)$, $y(n)$ 以及 $z(n)$ 对应的多相表达式, 以 $x(n)$ 为例, 我们能写成:

$$U_m = \sum_{n=0}^{M-1} x(n)e^{-j2\pi n/3} = 2 \sum_{n=0}^{M-1} [x_A(n) + x_G(n)]e^{-j2\pi n/3} - \sum_{n=0}^{M-1} e^{-j2\pi n/3} \quad (14)$$

由于 $M = 3L$, 公式 (3-20) 中的第二个总和为 0, 通过多相表达式, 我们能得到

$$U_m = 2 \sum_{r=0}^2 (U_{Am_r} + U_{Gm_r})e^{-j2\pi r/3} \quad (15)$$

由于 $|U_{bm}|^2 = U_{bm}U_{bm}^*$, 能得到

$$|U(m)|^2 = 2 \sum_{r=0}^2 [U_{Am_r} + U_{Gm_r} - U_{Am_q} - U_{Gm_q}]^2$$

其中 $q = (r+1) \bmod 3$, 同理可得 $y(n)$ 和 $z(n)$ 的频谱. 则 Z-curve DNA 频谱表达式为:

$$P_Z = 2 \sum_{r=0}^2 [U_{Am_r} + U_{Gm_r} - U_{Am_q} - U_{Gm_q}]^2 \quad (16)$$

其中 $q = (r+1) \bmod 3$, I^1 是核苷酸集合 I 的子集, 即 $I^1 = \{C, G, T\} \subset I$, 重新整理有:

$$P_Z = 4P_S + 4 \sum_{r=0}^2 (U_{Am_r} - U_{Am_q}) \sum_{b \in I^1} (U_{Am_r} - U_{Am_q}) \quad (17)$$

通过观察, 可以得到对于 $r \in \{0, 1, 2\}$, $[U_{Ar} + U_{Gr} + U_{Cr} + U_{Tr}]$ 等于处于窗口中的第 r 个密码子位置所包含的基因数的总和. 这是因为处于窗口中的每个密码子, 第 r 个位置永远都是被一个核苷酸占据的, 这个数值是个常量而且等于窗口长度的 $1/3$, 由此可得:

$$P_Z = 4P_S \quad (18)$$

3.2 问题二对不同物种类型基因的阈值确定

1) 判别效果的评价指标

假设所选定的信噪比分类阈值为 R_0 , 将 $R \geq R_0$ 作为外显子的判别, $R < R_0$ 作为内含子的判别. 通过阈值判别外显子与内含子的效果可用敏感度和专一性两种指标来表示:

敏感度: $S_n = \frac{T_P}{T_P + F_N}$ 专一性: $S_p = \frac{T_N}{T_N + F_P}$ 总正确率 $AC = (S_n + S_p)/2$ T_P 表示被正确判为外显子的个数; T_N 表示被正确判为内含子的个数; F_N 表示被错误地判为内含子的个数; F_P 表示被错误地判为外显子的个数.

2) 信噪比阈值的确定方法

对于给定的基因数据, 设所有外显子的信噪比均值为 m_1 , 所有内含子的信噪比均值为 m_2 , 设所有外显子的信噪比标准差为 σ_1 , 所有内含子的信噪比标准差为 σ_2 .

A. 均值平均法: $R_0 = (m_1 + m_2) / 2$

B. 概率平均法: 本方法基于对给定基因数据外显子和内含子信噪比统计量均服从正态分布的假设. 阈值 R_0 的估计值可由 (19) 式给出:

$$R_0 = \frac{(m_1 \cdot \sigma_2 + m_2 \cdot \sigma_1)}{(\sigma_1 + \sigma_2)} \quad (19)$$

C. 线性判别分类器方法: 本文采用 Ho-Kashyap 算法.

3) 阈值确定方法的评价与分析

表 3.1 不同算法对应的阈值的正确率对比 (数据来自于 A 题附件 Genes100, genes200)

基因种类	均值平均法		概率平均法		线性判别分类器方法		R0=2
	R0	AC	R0	AC	R0	AC	AC
Homosapiens	1.8858	0.7375	1.1577	0.7750	1.0039	0.7750	0.7250
Musmusculus	1.6065	0.7712	1.0884	0.8058	1.0402	0.8136	0.7199
Mammal	1.6957	0.7789	0.8161	0.7778	1.2582	0.8166	0.7481

对表 3-1 中的数据分析可知:

两种均值算法和线性分类器方法的判别比常值 2 为阈值进行判别时的正确率高.

概率平均法的正确率比均值平均法的效果好但均比线性分类器方法的判别差.

对于测试数据, 三种算法的信噪比阈值均比题目中给出的常数阈值 2 小, 因此以 2 为信噪比阈值来判别外显子和内含子并不总是合适的.

3) 外显子判别正确率与外显子序列长度的关系讨论

由表 3-2 可见: 对于 Mammal 和 Musmusculus 基因序列, 随着外显子长度的增加, 外显子序列的信噪比和判别正确率也随之增加, 短外显子的判别正确率比较低.

表 3-2 外显子长度与正确性关系 (数据来自 A 题附件 gene200 和 gene100)

外显子类型	Mammal		Musmusculus	
	R0	AC	R0	AC
短外显子 $i \leq 100$	1.0532	0.7037	1.0231	0.7046
中等长度外显子 $i=100 \ \&\& \ i \leq 300$	1.6141	0.9078	1.1914	0.8735
较长外显子 $i=300 \ \&\& \ i \leq 500$	1.7544	0.9472	2.0834	0.9465
长外显子 $i=500$	2.8271	0.9953	2.7664	0.9907

3) 信噪比阈值与物种类型关系讨论

由表 3-3 可以得出以下结论:

不同物种生物, 信噪比阈值是不同的, 因此信噪比阈值应该至少基于物种类型;

DNA 序列相近的生物, 信噪比阈值往往也接近, 譬如人类、老鼠和哺乳动物;

高等生物信噪比阈值通常趋于高于低等生物;

表 3.3 不同物种基因的阈值及其正确率

基因种类	均值平均法	
	R0	AC
genes100, Homo sapiens	1.8858	0.7375
genes100, Musmusculus	1.6065	0.7712
genes200, Mammal	1.6957	0.7789
拟南芥	2.0765	0.7553
酵母菌	4.6758	0.7843
线粒体	12.0872	0.7555

数据来自 A 题附件 genes200 以及酵母菌, 拟南芥和人类线粒体的 DNA 序列

3.3 问题三基因识别算法的实现

DNA 序列映射方法的选择

DNA 映射方法有多种, 对不同映射方法进行 DFT, 产生的频谱可能会对 DNA 编码序列的预测结果造成不同的影响, 因此第一步我们需要确定对 DNA 序列进行映射的方法, 本文中, 主要对比 Voss 映射, Z-curve 映射, 以及复数映射对基因预测的效果.

文献 [3] 中指出复数映射中噪声干扰严重, 无法直接判别外显子的位置. 其效果明显不如 Voss 和 Z-curve 映射的效果. 由第一问求解可知 Z-curve 和 Voss 具有相同的频谱. 因此, 在本文中采取的数值映射方式是 Voss 映射.

滑动窗口的选择

由图 3-6 可以看出, 使用短窗口让频谱振荡加剧, 这会使很多内含子序列被错误判定为外显子序列. 而使用长窗口, 短的外显子将无法捕获. 因此一个合适的滑动窗口长度将对基因识别产生比较大的影响, 结合实验的数据, 本文提出两个经验性的滑动窗口长度选择策略.

策略一: 设定滑动窗口长度 M 为 $L/2-L$ 之间的一个值, L 为平均外显子长度.

策略二: 使用不同窗口长度进行滑动 DFT 频谱, 然后对频谱进行加权平均.

滤波处理

平滑滤波

滑动窗口频谱曲线图 (以信噪比 SNR 为纵轴) 呈现高频震荡, 直观上表现为频谱的“毛边”, 这将会影响之后基因识别算法的判断, 这里使用平滑滤波预处理. 平滑预处理算法可以简单描述成对于滑动窗口频谱曲线的每一点使用其相邻 K 点 (包括该点) 的信噪比 (R) 的均值替代为该点的信噪比.

滑动窗口频谱为: $R(n; \frac{M}{3}) \quad n = 1, \dots, N$ M 为滑动窗口长度. 使用平滑处理的滑动窗口的频谱曲为: $R(n; \frac{M}{3}) = \sum_{i=n-\frac{k-1}{2}}^{n+\frac{k-1}{2}} R(i; \frac{M}{3})/K \quad n = 1, \dots, N$

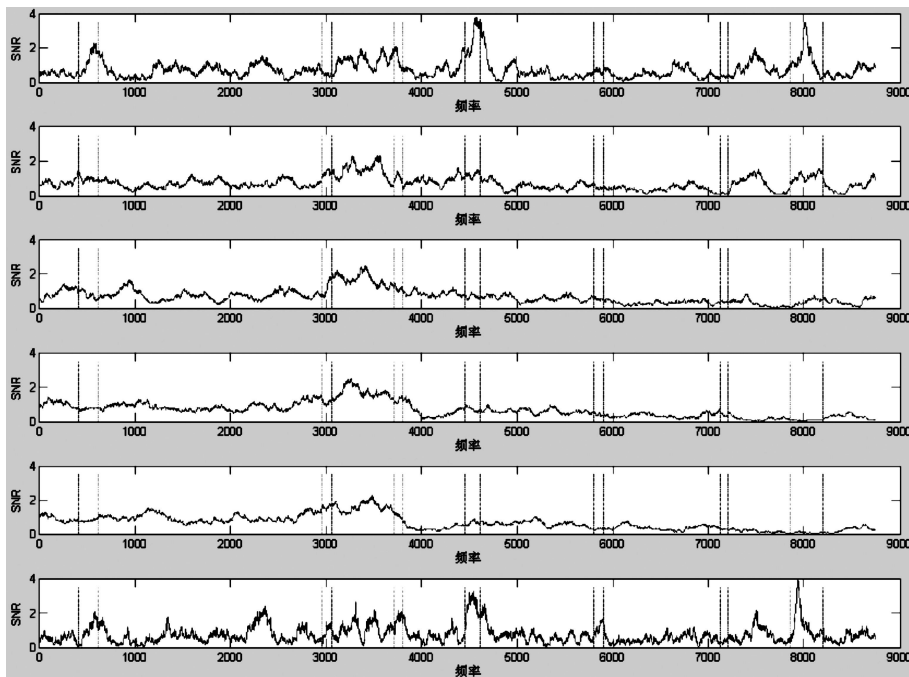


图 3-6 不同滑动窗口长度的 DNA 片段的 DFT 频谱图

(数据来自于 A 题附件 Gene100, 自上而下滑窗长度分别为 $L/10$, $L/6$, $L/2$, $4L$, $8L$ 和 L , L 为平均外显子长度)

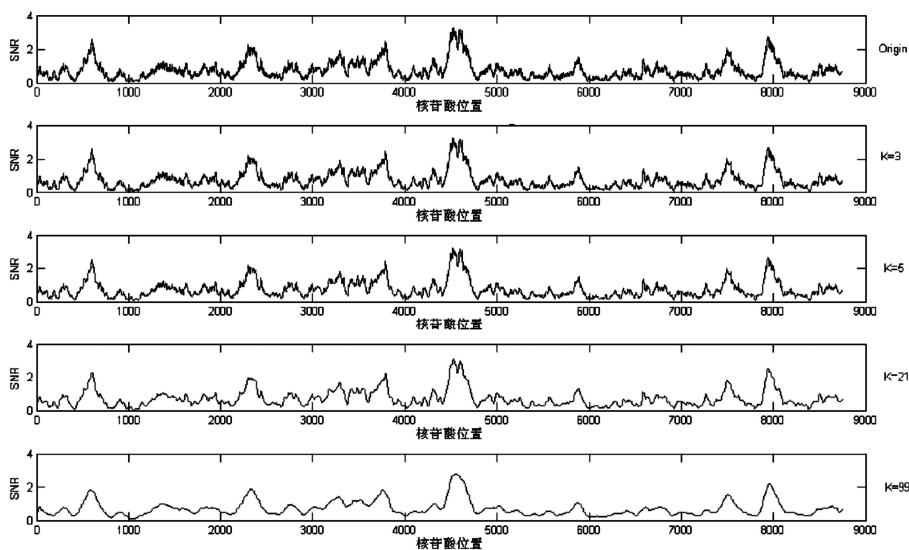


图 3-7 DNA 片段滑动窗口频谱曲线平滑处理结果

图 3-7 是针对附件 Genes100 第一个 DNA 片段数据的滑动窗口频谱曲线平滑处理结果, K 取不同值的结果对比. 可见滑动平滑滤波很好地平滑了频谱曲线, 消除了频谱毛边, 合适的 K 值 (譬如 $K=21$) 不但很好地消除了频谱的毛边, 也很好的保持了频谱的低频轮廓. 这将对

基因识别算法的准确性产生积极意义.

随机噪声抑制

由于 DNA 序列随机噪声的影响, 频谱的特征会不同程度上被削弱或淹没. 因而抑制背景噪声可以成为提高基因算法准确率的一种有效手段, 本文提出一种滤波算法抑制序列随机噪声以增强频谱特征.

注意到外显子在 $N/3$ 出现频谱峰值而内含子没有这样的特征, 因而在滑动窗口频谱曲线图滑动到外显子区域时往往会出现 SNR 峰值. 基于这样的事实, 如果对整个 DNA 的 $N/3$ 中心频率进行陷波处理, 将可以剥离出噪声序列. 再对 DNA 序列和噪声序列进行滑动窗口 DFT, 所得 SNR 结果之差可以认为是抑制噪声后的信号功率的信噪比. 算法流程如下.

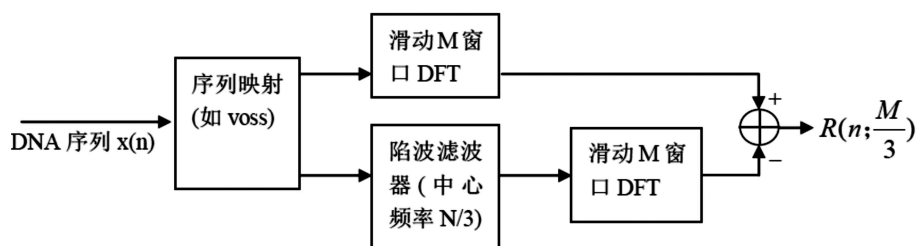


图 3-8 背景噪声抑制算法流程

算法中的陷波滤波器可以使用二阶 Notch 滤波器, 针对附件 Genes100 第一个 DNA 片段数据的随机噪声抑制结果如下.

下图 3-9 是 Genes100 第一个 DNA 片段数据频谱结果, 虚线框为外显子位置, 图 3-9 中图 1 为原始滑动窗口 DFT 频谱图 (窗口大小设置为 200), 图 2 为抑制噪声算法处理后的滑动窗口 DFT 频谱图, 图 3 为图 2 经平滑处理后的频谱图.

由图 3-9 可以看出, 结合滑动窗口长度的合理选取以及平滑处理, 该算法有效抑制了背景噪声, 并保留了 SNR 峰值特征, 这将为后面的基因识别算法提供了易于进行准确识别的频谱.

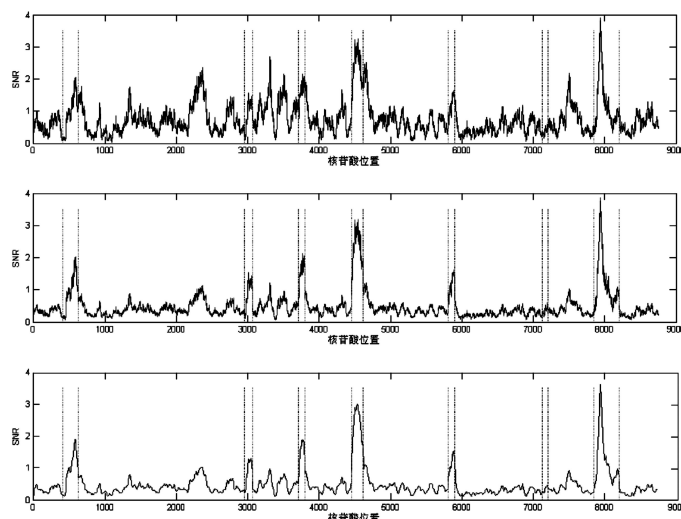


图 3-9 DNA 片段数据频谱结果 (图 1, 图 2, 图 3 为图中依次从上到下的 3 个子图)

基因识别算法

选择合适的序列映射、合适的滑动窗口长度并进行噪声抑制滤波和滑动平滑之后将得到比较好的可供进行基因识别的算法的 SNR 频谱. 基因识别算法将根据得到的频谱识别出对应的外显子和内含子区域. 通常使用 EPND 算法 [7](由 Yin 和 Yau 于 2007 年提出).

由于 EPND 算法对长 DNA 序列预测正确率不高, 尤其是当一个外显子跟着一段很长外显子的时候, 外显子序列往往不能被很好地预测. 鉴于此本文对 EPND 算法进行了一点改进, 改进后的 EPND 算法被实验证明可以更准确地预测外显子和内含子.

改进的 EPND 算法

将 DNA 序列分割成 2000bp 的子序列.

将每一个 2000bp 的子序列等分成 k 段, 这将产生 $k+1$ 个端点 $P_1, P_2, \dots, P_{K+1}$.

对子序列每一段 $[P_i, P_{K+1}] i = 1, \dots, K$ 使用 EPND 算法; 单个核苷酸将被多次计算, 如果在大多数的计算中, 某一个位置核苷酸被预测为属于外显子区域则将该核苷酸判定为外显子核苷酸, 反之判定为内含子核苷酸.

为比较不同 k 值情况下改进的 EPND 的效果, 使用本文提出的算法对 Genes100 中包含的 92 个 *Mus musculus* 的 DNA 序列进行预测, 并对预测结果求平均 S_n , 平均 S_p 和平均 AC .

表 3-4 改进的 EPND 与 EPND 的比较

基因识别算法	平均 S_n	平均 SP	平均 AC
DFT ¹ + EPND	69.37%	95.06%	82.21%
DFT+ 改进的 EPND($k=2$)	68.56%	94.17%	79.36%
DFT+ 改进的 EPND($k=4$)	72.54%	95.22%	83.88%
DFT+ 改进的 EPND($k=6$)	69.60%	94.54%	82.07%
DFT+ 改进的 EPND($k=8$)	67.80%	94.20%	81.00%

由表 3-4 数据可以得到以下结论:

改进的 EPND 比 EPND 的效果要好.

在 $k=4$ 的情况下, 改进的 EPND 的效果最好.

算法评估

表 3-5 Genes100 中 92 个 *Mus musculus* 的 DNA 序列预测结果指标的平均值

基因识别算法	平均 S_n	平均 SP	平均 AC
DFT+EPND	69.37%	95.06%	82.21%
DFT+ 改进的 EPND	72.54%	95.22%	83.88%
DFT+ 滤波 + 改进的 EPND	81.25%	96.30%	88.77%

使用本文提出的基因预测算法对 Genes100 中包含的 92 个 *Mus musculus* 的 DNA 序列进行预测, 并对预测结果求平均 S_n , 平均 S_p , 平均 AC , 如表 3-5 所示:

由表 3-5 中的数据可以看出在 DFT 上增加抑制背景噪声滤波使用改进的 EPND 的预测算法能够显著的提升预测结果的敏感度, 同时也能够提供预测结果的特异性.

基因识别算法流程

使用本文提出的基因识别算法预测 DNA 序列编码区的流程如下图 3-10 所示:

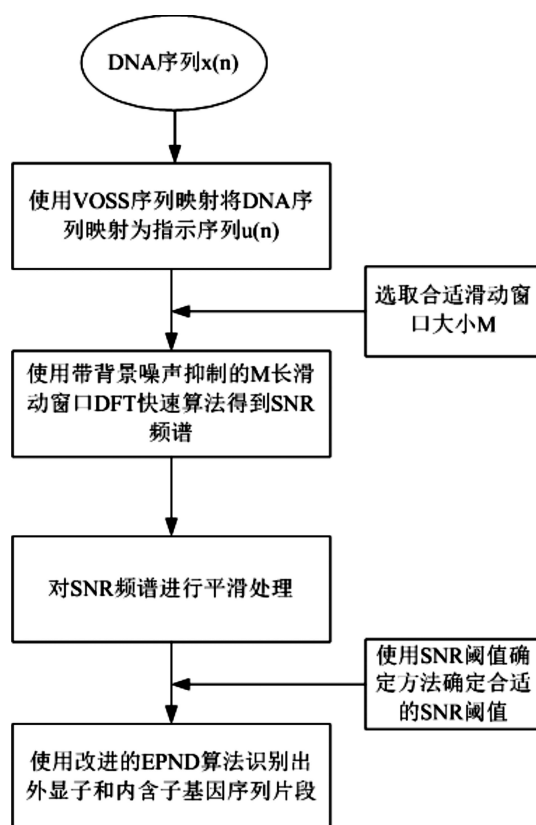


图 3-10 本文基因识别算法识别基因流程

5 模型总结

由于随机背景噪声等原因的影响,使得很多基因识别算法准确率偏低,本文建立的基于背景噪声抑制和频谱平滑的 SNR 频谱预处理模型,经过预处理后的频谱不仅大幅度抑制了背景噪声,同时保留了 SNR 频谱的模式特征.在编码序列识别上,本文对经典的 EPND 预测算法进行了改进,使用改进的 EPND 算法对经过预处理后频谱进行基因识别,实验结果显示这种基因识别模型具有优异的基因识别性能.

参考文献

- [1] 马宝山, 朱义胜. 用于基因预测的自适应滤波器的仿真研究 [J]. 系统仿真学报, 2007, 9(24): 5620-5623.
- [2] Shao J, Yan X, Shao S. 2012 SNR of DNA sequences mapped by general affine transformations of the indicator sequences Journal of Mathematical Biology
- [3] 饶妮妮, 邱丽君, DNA 序列数值映射方法的研究 [J]. 生物医学工程学杂志, 2005, 22(4): 681-685.
- [4] Burge, C., Karlin, S.. Prediction of complete gene structures in human genomic DNA[J]. J. Mol. Biol, 1997, 268, 78-94.
- [5] 马宝山. 基于信号处理理论和方法的基因预测研究 [D]. 大连海事大学博士论文, 2008,08,01.

-
- [6] 马宝山, 朱义胜, 陈玉珍. 用多种统计特征识别基因序列 [J]. 计算机工程与应用, 2009,10,11.
- [7] Yin, C., Yau, S.S.-T. Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. *Journal of Theoretical Biology*, 2007, 247, 687-694.

请作者补上英文题目, 作单, 单位, 英文摘要, 英文关键词

Abstract:

Keywords: