文章编号:1004-9037(2013)05-0679-06

真核生物多基因预测结果整合算法

刘金定 朱毅华 黄水清

(南京农业大学信息科学技术学院,南京,210095)

摘要:针对独立基因预测算法可靠性较差的缺点,提出了真核生物多基因预测结果整合算法(Algorithm for integration of multiple eukaryotic gene prediction results, AIMEGPR)。该算法在综合分析各种预测算法结果的基础上,首先用极大似然法估计各种预测算法的性能参数,然后利用这些性能参数计算基因证据区间上各个基因片段归属于各种基因元件类型的后验概率,最后采用动态规划法在基因证据区间上确定最优的一致基因结构。AIMEGPR 既不需要人工定制整合规则,也不需要复杂的训练学习,因此 AIMEGPR 尤其在基因组新测序物种上进行编码基因注释时具有十分显著的优越性。实验结果表明,利用 AIMEGPR 算法对多基因预测结果整合可以明显提高一致基因的可靠性。

关键词:多基因整合;一致基因;基因预测;动态规划

中图分类号:TN713;Q751 **文献标志码**:A

Algorithm for Integration of Multiple Eukaryotic Gene Prediction Results

Liu Jinding, *Zhu Yihua*, *Huang Shuiqing*

(College of Information and Technology, Nanjing Agricultural University, Nanjing, 210095, China)

Abstract: Considering the disadvantage of poor reliability of the single prediction algorithm, the paper presents an algorithm for integration of multiple eukaryotic gene prediction results (AIMEGPR). AIMEGPR took account of the results of multiple prediction software and used the maximum likelihood method to estimate the performance parameters of various prediction algorithms. Then AIMEGPR used these performance parameters to calculate the posterior probability which class the gene segment in gene evidence region belongs to. Finally, AIMEGPR determined an optimal consensus gene structure for gene evidence region using dynamic programming method. AIMEGPR does not require integration of the rules offered by experts, and also does not require complex training study. Therefore, AIMEGPR has a significant superiority on annotation for the new sequencing genome. The experimental results show that AIMEGPR can improve the reliability of consensus genes.

Key words: integration of multiple genes; consensus gene; gene prediction; dynamic programming

引 言

随着测序成本下降,越来越多真核生物基因组项目得以启动,不少物种的基因组测序工作已经完成。基因预测作为基因组研究的首要任务,其结果的可靠性直接影响后续生物学研究。目前独立基因预测方法可分为从头预测算法和同源预测算法两大类[1]。从头预测法用各种基因信号和序列统计特征进行基因预测[2-4],同源预测方法则用已知蛋白序列或者核酸序列和基因组序列进行比对,通过解析比对结果获得基因结构[5]。从头预测算法大多基于统计学原理进行基因预测,通常获得的基

基金项目:国家自然科学基金 (31171843)资助项目;国家高技术研究发展计划("八六三"计划) (2012AA101505)资助项目。

收稿日期:2013-05-15;**修订日期**:2013-07-26

因数量较多,但假阳性偏高。同源预测算法依赖于物种亲缘关系和搜索序列的数量及完整性,通常预测的基因数量偏少、基因完整性较差,但识别出来的基因片段的可靠性较高[6]。

由于各种预测方法采用的算法模型和侧重点不同,因此各种方法的预测结果必然会产生差异。领域专家根据自身经验对不一致的预测结果进行整合是目前获得一致基因的最可靠方法[7]。由于人工整合一致基因成本高、耗时长,为此人们提出了多基因预测结果整合算法,如 ExonHunter[8],JIGSAW[9]等。这些整合算法总体上可分为投票法和规则学习法两大类[10]。投票法对各种基因结构进行投票统计,以得票的多少来确定最终基因结构。该类方法虽然简单直接,但由于投票总数偏少不能达到统计学效果,所以一致基因可靠性改善并不明显。规则学习法则需提供各种预测算法结果样本以及对应的专家整合结果样本,然后在两部分数据集上利用机器学习方法进行规则学习。虽然规则学习法可以取得较好的整合性能,但必须具备两个前提:(1)领域专家具有足够整合经验,能够整合出可靠的一致基因结构;(2)选择的训练样本集能够保证机器学习到整合过程中的所需的整合规则。目前规则学习法在模式生物基因组上得到了很好的应用,主要原因是模式生物可以提供很好的训练集。对于基因组新测序的物种而言,因为缺少足够可靠基因用于学习,所以规则学习法的优越性受到了限制。

针对上述基因整合算法的不足,本文提出了一种真核生物多基因预测结果整合算法(Algorithm for integration of multiple eukaryotic gene prediction results, AIMEGPR)。该算法利用"诊断性能估计模型"[11]估计各种独立预测算法的性能参数,然后计算基因证据区间上各种基因片段类型的后验概率,最后采用动态规划算法获得一致基因结构。该算法避免了投票法过于简单、缺少统计学支持的不足,而且解决了规则学习法难以获得可靠训练集的困难。实验结果表明,该方法明显提高了基因整合结果的可靠性。

1 真核多基因预测结果整合算法

1.1 真核基因结构

真核生物基因在基因组上以外显子(Exon)和内含子(Intron)相互交替的形式出现。基因表达时以基因组序列为模板,拷贝整个基因序列(包括外显子和内含子),形成信使 RNA 前体(Pre-mRNA),然后在相关酶的作用下,从 mRNA 前体中剪去内含子序列形成成熟的 mRNA 序列(见图 1),最后成熟的 mRNA 根据密码子和氨基酸对应关系翻译成对应的氨基酸序列,在细胞内发挥生物学功能[11]。一些基因可能存在几种剪接方式,因此一个基因会转录出几种不同的mRNA,这些mRNA 被称为这个基因转录出来的不同转录本。

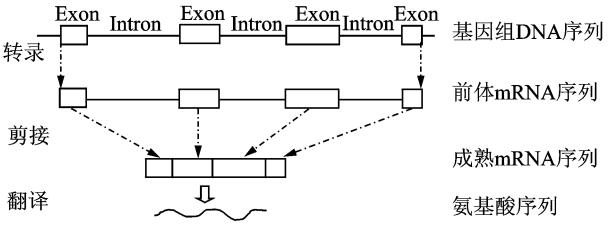


图 1 真核基因结构以及转录、剪接过程

1.2 聚类计算基因证据区间

对所有预测方法结果进行聚类处理,把在基因组位置上有重叠的基因聚类在一起,构成基因证据区间(Gene evidence region, GER)。比如,方法 A 预测到两个基因(只列出了基因中外显子坐标,相邻两个外显子间为内含子):a1(600.. 800, 1 200..1 500, 3 000..4 500), a2(8 000..9 000, 12 000..14 000);方法 B 预测到了一个基因:b1(2 800..4 300,8 000..9 500)。由于 a1 和 b1 之间有重叠,a2 和 b1 有重叠,因此聚类后得到的一个基因证据区间为[600,14 000]。该基因证据区间上存在 3 个基因证据,分别来自于 a1,a2,b1。后续的基因整合将以基因证据区间为基因范围整合一致基因。

1.3 基因证据区间片段化处理及类型标志

当不同预测方法在基因证据区间上预测的基因结构存在冲突时,外显子和内含子作为整合的基本单位显得粒度太粗,为此对基因证据区间进行片段化处理,使基因证据区间的基本单位粒度变细,从而有利于一致基因的整合。如图 2 所示,在某个基因证据区间上方法 1,方法 2,方法 4 识别了 3 个外显子,2 个内含子;方法 3 识别了 2 个外显子,1 个内含子;方法 5 识别了 4 个外显子,3 个内含子。按照图 2 表示法,基因证据区间经片段化处理后得到了 11 个基因片段(Gene segment, GS)。

根据基因片段在不同基因元件(外显子或内含子)上的位置,定义 8 种基因片段类型(Gene segment type, GST):完整外显子片段 EC、外显子上游片段 EU、外显子下游片段 ED、外显子内部片段

EI、完整内含子片段 IC、内含子上游片段 IU、内含子下游片段 ID、内含子内部片段 II。根据 5 种预测方法的预测结果，5 种预测方法对 11 个基因片段的类型标志结果如图 2 所示。

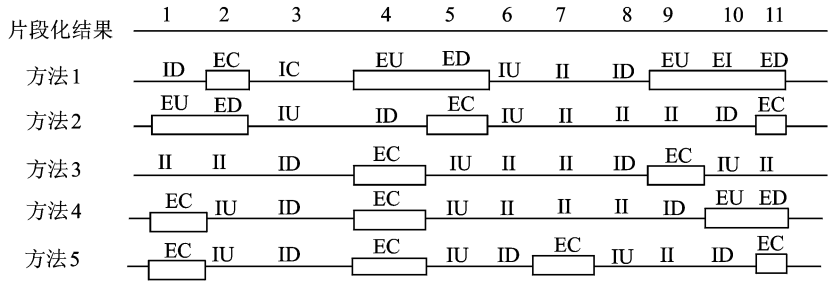


图 2　基因证据区间片段化处理和类型标志

1.4　基因片段所属类型的后验概率计算

1.4.1　相关定义

为了说明基因片段类型的后验概率计算过程，对一些符号做如下定义：

定义 1　预测方法集：$M=\{m_1,\cdots,m_i,\cdots,m_n\}$。用于表示参与提供基因证据的所有预测方法，其中 m_i 表示第 i 个预测方法，n 为预测方法总数。

定义 2　基因片段类型集：$T=\{t_1,\cdots,t_i,\cdots,t_l\}$。用于表示定义的所有基因片段类型（具体类型定义见 1.3 节），其中，t_i 表示第 i 种基因片段类型，l 为定义的基因片段类型总数。

定义 3　全基因组基因片段总集：$s=\{s_1,\cdots,s_i,\cdots,s_k\}$。用于表示所有基因证据区间上的基因片段汇总，其中，s_i 表示全基因组基因片段总集上的第 i 个片段，k 为全基因组基因片段总集中的基因片段总数。

定义 4　各种预测方法对全基因组基因片段总集中的基因片段的类型识别结果矩阵

$$\boldsymbol{E}_{n\times k}=\begin{bmatrix}E_{1,1} & \cdots & E_{1,k}\\ \vdots & E_{i,j} & \vdots\\ E_{n,1} & \cdots & E_{n,k}\end{bmatrix}$$

式中：$E_{ij}\in T$ 表示预测方法 m_i 对全基因组上的第 j 个基因片段的类型识别结果。

定义 5　$\alpha_{m,t}$ 表示预测方法 m 对基因片段类型为 t 的基因片段识别的假阳性，$\beta_{m,t}$ 为相应的假阴性，其中，$m\in M$，$t\in T$。

定义 6　ξ_t 表示 t 类型基因片段在全基因组基因片段总集中所占的比例，其中 $t\in T$。

定义 7　各种预测方法对全基因组基因片段总集中的 k 个基因片段进行“是否为 t 基因片段类型”的识别结果矩阵

$$\boldsymbol{A}(t)_{n\times k}=\begin{bmatrix}A(t)_{1,1} & \cdots & A(t)_{1,k}\\ \vdots & A(t)_{i,j} & \vdots\\ A(t)_{n,1} & \cdots & A(t)_{n,k}\end{bmatrix}$$

式中：$A(t)_{i,j}=f(E_{i,j},t)=\begin{cases}1 & t=E_{ij}\\ 0 & t\neq E_{i,j}\end{cases}$，$t\in T$。即，$m_i$ 预测方法对第 j 个基因片段进行类型识别，识别结果为 t 时，$A(t)_{i,j}=1$；识别结果不为 t 时，$A(t)_{i,j}=0$。

1.4.2　基因片段各种类型后验概率的计算

（1）各种预测方法对各种基因片段类型识别的性能和各种类型基因片段占有比例的估计

依据文献[12]提出的“诊断性能估计模型”，构造如下公式

$$L\propto\prod_{j=1}^{k}\left[\xi_t\prod_{i=1}^{n}\beta_{m_i,t}^{(1-A(t)_{i,j})}(1-\beta_{m_i,t})^{A(t)_{i,j}}+(1-\xi_t)\prod_{i=1}^{n}\alpha_{m_i,t}^{A(t)_{i,j}}(1-\alpha_{m_i,t})^{(1-A(t)_{i,j})}\right]\quad(1)$$

在式(1)上用最大似然估计法估计出各种预测方法 m_i 对 t 类型的基因片段识别的假阳性 $\alpha_{m_i t}$ 和假阴性 $\beta_{m_i t}$，以及 t 类型基因片段在全基因组上的占有率 ξ_t。

经过 l 次计算获得 n 种预测方法对 l 种基因片段类型的识别假阳性矩阵 $\alpha_{n\times l}$ 和假阴性矩阵 $\beta_{n\times l}$，以及各种类型片段在全基因组证据区间上占有率向量 θ，计算结果如下

$$\alpha_{n\times l}=\begin{bmatrix}\alpha_{m_1,t_1} & \cdots & \alpha_{m_1,t_l}\\ \vdots & \alpha_{m_i,t_j} & \vdots\\ \alpha_{m_n,t_{t1}} & \cdots & \alpha_{m_n,t_{tl}}\end{bmatrix}$$

α_{m_i,t_j} 表示预测方法 m_i 对 t_j 类型片段识别的假阳性；

$$\beta_{n\times l}=\begin{bmatrix}\beta_{m_1,t_1} & \cdots & \beta_{m_1,t_l}\\ \vdots & \beta_{m_i,t_j} & \\ \beta_{m_n,t_{t1}} & \cdots & \beta_{m_n,t_{tl}}\end{bmatrix}$$

β_{m_i,t_j} 表示预测方法 m_i 对 t_j 类型片段识别的假阴性；$\theta=\{\xi_{t_1},\cdots,\xi_{t_i},\cdots,\xi_{t_l}\}$，为 l 种类型片段在全基因组证据区间上的占有率。

(2)基因片段的各种类型后验概率的计算

$$p_{t,j}=\frac{\xi_t\prod_{i=1}^{n}\beta_{m_i,t}^{(1-A(t)_{i,j})}(1-\beta_{m_i,t}^{A(t)_{i,j}})}{\xi_t\prod_{i=1}^{n}\beta_{m_i,t}^{(1-A(t)_{i,j})}(1-\beta_{m_i,t}^{A(t)_{i,j}})+(1-\xi_t)\prod_{i=1}^{n}\alpha_{m_i,t}^{A(t)_{i,j}}(1-\alpha_{m_i,t})^{(1-A(t)_{i,j})}} \tag{2}$$

所有基因片段为各种类型的后验概率矩阵

$$\boldsymbol{P}_{l\times k}=\begin{bmatrix}p_{t_1,1} & \cdots & p_{t_1,k}\\ \vdots & p_{t_i,j} & \vdots\\ p_{t_l,1} & \cdots & p_{t_l,k}\end{bmatrix}$$

式中：$p_{t_i,j}$ 表示全基因组上第 j 个片段为 t_i 类型片段的后验概率。

1.5 基因证据区间上的一致基因整合

1.5.1 相邻基因片段类型转移约束模型

基因证据区间中相邻片段的类型受合理的基因结构约束，表 1 反应了相邻基因片段间的类型转移约束关系。

表 1 相邻基因片段类型转移约束关系表

	1(EU)	2(ED)	3(EC)	4(EI)	5(IU)	6(ID)	7(IC)	8(II)
1(EU)	0	1	0	1	0	0	0	0
2(ED)	0	0	0	0	1	0	1	0
3(EC)	0	0	0	0	1	0	1	0
4(EI)	0	1	0	1	0	0	0	0
5(IU)	0	0	0	0	0	1	0	1
6(ID)	1	0	1	0	0	0	0	0
7(IC)	1	0	1	0	0	0	0	0
8(II)	0	0	0	0	0	1	0	1

表 1 中，如果第 i 行第 j 列为 1，则表示相邻两个基因片段允许从第 i 类型向第 j 类型转移，反之不允许。比如，第 2 行第 5 列为 1，表示相邻两个基因片段的类型允许为(ED,IU)，即允许“下游外显子片段”后面紧跟着“上游内含子片段”；第 2 行第 1 列为 0，表示相邻两个基因片段类型不允许为(ED,EU)，即不允许“下游外显子片段”后面紧跟着“上游外显子片段”。

1.5.2 一致基因结构求解

以 1.2 节聚类得到的单个基因证据区间为一致基因结构求解范围。设单基因证据区间中片段总数为 q，基因片段类型种类数为 l，各个基因片段对应各种类型的后验概率矩阵为 $\boldsymbol{SP}_{l\times q}$，相邻基因片段类型转移约束矩阵为 $\boldsymbol{C}_{l\times l}$，如表 1 所示。求解基因证据区间上的一致基因结构，可以归纳为：$\boldsymbol{C}_{l\times l}$ 矩阵约束下，在 $\boldsymbol{SP}_{l\times l}$ 矩阵中搜寻一条步长为 q 的最大后验概率路径。本文采用类似 Viterbi 解码的动态规划法实现最优路径的求解，其时间复杂度为 $O(l^2q)$。求解过程用 l 条始终以 $t_1,\cdots,t_l$ 结尾的局部最优路径记录中间求解状态，对 l 条局部最优路径进行 q 次更新，每更新 1 次对应着 1 个新基因片段的类型加入到局部最优路径中。为了便于说明算法，作如下定义：LOP_t_i 表示以 t_i 类型结尾的局部最优路径；PRO_LOP_t_i 表示以 t_i 类型结尾局部最优路径对应的后验概率。

求解算法如下：

(1)对应于第 1 个基因片段，用 $t_1,\cdots,t_l$ 片段类型初始化 l 条局部最优路径，对应的后验概率为第 1 个基因片段对应的各种片段类型的后验概率。

(2)对应后续 $q-1$ 个基因片段，进行逐次迭代计算更新 l 条局部最优路径及其后验概率。其中，对第 $x(2\leqslant x\leqslant q)$ 个基因片段对应的局部最优路径更新计算时，每个局部最优路径 LOP_t_i 和对应的后验概率 PRO_LOP_t_i 的更新过程如下：

for 第 $x-1$ 个基因片段对应的每个局部最优路径 LOP_t_j do

if(允许类型 t_j 向 t_i 转换 and PRO_LOP_t_j * 第 x 个基因片段为 t_i 的后验概率 > PRO_LOP_t_i) then

PRO_LOP_t_i = PRO_LOP_t_j * 第 x 个基因片段为 t_i 的后验概率；

LOP_t_i = LOP_t_j & t_i；

End if

End for

(3)在 l 条路径中，后验概率最大的局部最优路径作为全局最优路径返回。

1.6 AIMEGPR 算法关键步骤

本文以整合各种预测方法结果获得一致基因集为目的，利用“诊断性能评估模型”计算各种预测方法的性能，然后利用各种预测方法的性能参数计算出基因证据区间上各个基因片段为各种类型的后验概率，最后用类似 Viterbi 解码算法求解一致基因。

算法关键步骤如下：

(1)对所有预测结果进行聚类计算获得基因证据区间。

(2)结合所有预测方法结果，对基因证据区间进行片段化处理和片段类型标志。

(3)根据各种预测方法对基因片段的类型标志，构建似然估计函数，估计各种预测方法对各种类型的基因片段识别的假阳性矩阵$\alpha_{n\times l}$和假阴性矩阵$\beta_{n\times l}$以及各种类型基因片段在全基因组范围内所有基因证据区间上的占有率θ。

(4)利用步骤(3)计算获得的$\alpha_{n\times l}$，$\beta_{n\times l}$和θ，计算所有基因片段类型的后验概率矩阵$\boldsymbol{P}_{l\times k}$。

(5)从$\boldsymbol{P}_{l\times k}$中获得各个基因证据区间的基因片段类型后验概率矩阵$\boldsymbol{SP}_{l\times q}$。

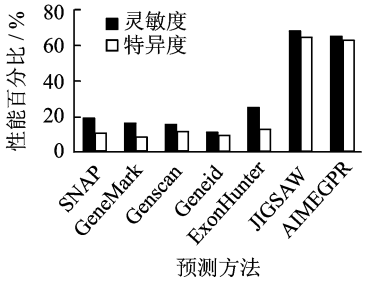
(6)在基因片段类型转移矩阵$\boldsymbol{C}_{l\times l}$的约束下，在各个基因证据区间对应的$\boldsymbol{SP}_{l\times q}$矩阵中求解基因证据区间对应的一致基因结构。

2　实验和结果分析

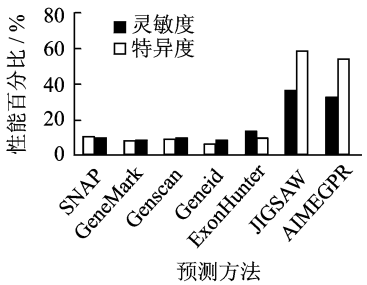
实验以 Encode 提供的人类第 1，2，3 三条染色体为实验基因组数据(http://genome.ucsc.edu/ENCODE/encode.hg18.html)，以 wgEncodeGencodeManualV3 注释的 9 436 个基因为参考基因。为了评估 AIMEGPR 的性能，选择 SNAP[13]，Genscan[14]，GeneMark[15]，Geneid[16] 等作为独立预测算法比较对象，选择基于学习的 JIGSAW 和基于投票法的 ExonHunter 作为多基因证据整合方法比较对象。在 9 436 个基因中，抽取 1 500 个基因用于 JIGSAW 的训练，剩余的 7 936 个基因用于各种预测方法的性能比较。SNAP，Genscan，GeneMark，Geneid 四个预测软件的预测结果作为 JIGSAW，ExonHunter 和 AIMEGPR 的基因整合源。

灵敏度(Sensitivity，SN)和特异度(Specificity，SP)是目前常用的评价指标，其中灵敏度＝真阳性/(真阳性＋假阴性)，特异度＝真阳性/(真阳性＋假阳性)。前者反应算法对真样本的识别能力，后者反应算法识别结果的可靠性。当预测的外显子和 ENCODE 中的外显子边界完全相同时，认为该外显子被准确预测；当预测的转录本和 ENCODE 中的转录本的所有外显子都一样时，认为该转录本被准确预测；对具有多个转录本的基因而言，只要有一条转录本被预测到，则认为基因被预测到。

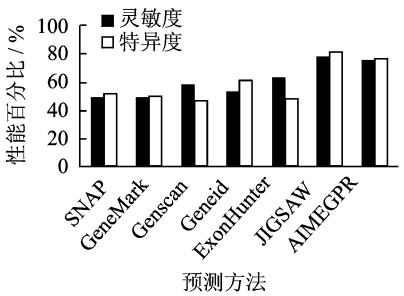
实验结果如图 3 所示，AIMEGPR 在基因、转录本以及外显子水平上的预测性能明显优于 4 种独立预测法以及基于投票法的 ExonHunter。ExonHunter 比 4 种独立预测算法性能优越的主要原因在于其基因结构建立在 4 种预测算法对各个基因元件类型的投票统计结果上。AIMEGPR 比 ExonHunter 算法性能优越的主要原因在于 AIMEGPR 充分考虑了 4 种独立预测算法在整个基因集上的统计结果，大大减少了在单个基因结构上投票偶然性。AIMEGPR 和规则学习法 JIGSAW 相比，虽然总体性能相当，但 AIMEGPR 算法不需要可靠基因集进行训练，具有很好的可操作性，这对于基因组新测序的物种而言，具有十分显著的优越性。



(a) 基因水平灵敏度和特异度比较



(b) 转录本水平灵敏度和特异度比较



(c) 外显子水平灵敏度和特异度比较

图 3　7 种方法在基因、转录本及外显子水平上的预测性能比较

3　结束语

不同的独立预测算法基于不同的原理，采用不

同的模型，其预测结果产生差异在所难免。本文设计的多基因预测结果整合算法 AIMEGPR 把多个预测结果综合起来考虑，计算基因证据区间上各个基因片段类型的后验概率，利用动态规划算法整合出一致基因结构。通过比较发现，该算法确实可以显著提高整合结果的可靠性。此外，AIMEGPR 不需要对软件进行复杂的训练，也不需要花费大量的人工成本去定制各种整合规则，在进行多基因预测结果时操作简便、便于使用。AIMEGPR 算法和大多数预测软件相似，只给出一个一致基因结构。当基因存在选择性剪接时，AIMEGPR 不能提供多个转录本结构，后续的研究工作将围绕这方面展开，以期能够进一步完善算法功能。

参考文献：

[1] 孙红卫，翁洋，朱允民. 基因预测准确性的度量标准分析[J]. 四川大学学报：自然科学版，2006，43(3)：649-654.
Sun Hongwei, Weng Yang, Zhu Yunmin. An analysis of the measures for gene prediction accuracy[J]. Journal of Sichuan University: Natural Science Edition, 2006, 43(3): 649-654.

[2] 张景祥，赵晓兵. 基于统计 CPG 岛的基因预测方法[J]. 生物数学学报，2012，27(2)：342-348.
Zhang Jingxiang, Zhao Xiaobing. The gene prediction method based on the statistics of the CPG islands [J]. Journal of Biomathematics, 2012, 27(2): 342-348.

[3] 周艳红，杨雷，王卉，等. 基于多级优化的真核生物基因结构预测[J]. 科学通报，2004，49(2)：140-145.
Zhou Yanhong, Yang Lei, Wang Hui, et al. Based on multi-level optimization of eukaryotic gene structure prediction[J]. Chinese Science Bulletin, 2004, 49(2): 140-145.

[4] 马玉韬，车进，关欣，等. 加窗窄通带滤波器蛋白质编码区预测算法[J]. 数据采集与处理，2013，28(2)：129-135.
Ma Yutao, Che Jin, Guan Xin, et al. Prediction algorithm for protein coding regions based on windowed narrow pass-band filter[J]. Journal of Data Acquisition and Processing, 2013, 28(2): 129-135.

[5] Korf I, Flicek P, Duan D, et al. Integrating genomic homology into gene structure prediction[J]. Bioinformatics, 2001, 17: 140-148.

[6] 张恩民，海荣，俞东征. 基因预测方法研究进展[J]. 中国媒介生物学及控制，2009，20(3)：271-273.
Zhang Enmin, Hai Rong, Yu Dongzheng. Research progress of gene prediction methods[J]. China J Vector Bio and Control, 2009, 20(3): 271-273.

[7] 童庆，郑浩然，王煦法. 基于统计组合与特征分类的基因预测算法[J]. 中国科学技术大学学报，2006，36(2)：1184-1189.
Tong Qing, Zheng Haoran, Wang Xufa. Gene-prediction algorithm based on the statistical combination and the feature classification[J]. Journal of University of Science and Technology of China, 2006, 36(2): 1184-1189.

[8] Brejova B, Brown D G, Li M, et al. ExonHunter: A comprehensive approach to gene finding[J]. Bioinformatics, 2005, 21: 57-65.

[9] Allen J E, Salzberg S L. JIGSAW: Integration of multiple sources of evidence for gene prediction[J]. Bioinformatics, 2005, 21: 3596-3603.

[10] 李校. 组合多重证据促进真核生物基因结构预测[D]. 成都：四川大学，2007.
Li Xiao. Improving eukaryotic gene structure prediction by combining multiple sources of evidence[D]. Chengdu: Sichuan University, 2007.

[11] 本杰明·卢因. 基因Ⅷ[M]. 第 1 版. 余龙，江松敏，赵寿元，译. 北京：科学出版社，2009：346-347.
Benjamin L. Genes Ⅷ[M]. First Edition. Yu Long, Jiang Songmin, Zhao Shouyuan, Trans. Beijing: Science Press, 2009: 346-347.

[12] Torrance-Rynard V L, Walter S D. Effects of dependent errors in the assessment of diagnostic test performance[J]. Stat Med, 1997, 16: 2157-2175.

[13] Korf I. Gene finding in novel genomes[J]. BMC Bioinformatics, 2004, 5: 59.

[14] Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA[J]. J Mol Biol, 1997, 268: 78-94.

[15] Lukashin A V, Borodovsky M. GeneMark.hmm: New solutions for gene finding[J]. Nucleic Acids Res, 1998, 26: 1107-1115.

[16] Parra G, Blanco E, Guigo R. GeneID in drosophila [J]. Genome Res, 2005, 10: 511-515.

作者简介：刘金定(1978-)，男，讲师，研究方向：生物信息与数据挖掘；朱毅华(1973-)，男，博士，副教授，研究方向：生物信息与数据挖掘；黄水清(1964-)，男，博士，教授，研究方向：计算机信息检索、数据挖掘，E-mail：sqhuang@njau.edu.cn。