开发笔记 -基因注释概况: 模型与软件

于秋林

BGI-RD
深圳

March 22, 2012

- ORF 确定基因结构
- 偏倚数量化
- 内含子参数确定

概述

两类不同的基因：

- 原核生物基因：结构简单，冗余少，利用ORF即可界定多数基因。
- 真核生物基因：结构复杂，噪声大，需对不同子结构分别建模。

软件及模型：

- Genemark:Fixed-order Markov Model
- Glimmer:Interpolated Markov Model
- Augustus:Hidden Markov Model

基于马氏链的预测方法

不同物种的一些基因结构是相近的，特别是一些较保守的基因家族，在一些位点上有些短的特征信号非常保守，称为 motif。Genamark 和 Glimmer 采用的策略是先利用可信度高的样本构造训练集，将训练集打散成 kmer，统计 kmer 的相对频率（这个应该是与位置信息无关的？待确定，如果是与位置信息无关的，那么信息没有充分利用，难道是因为在后面的预测环节中位置信息是难以获取利用的，所以在这一步特征生成的过程中就没必要强调位置信息？？在真核基因预测中，motif 是与位置紧密相关的，有大把文献讲怎样找这些 motif，这也是因为真核基因结构复杂的必然需求），经过归一化的频率可处理成反映训练集特征的分值，对未知序列，选定一种 orf，顺次滑动累加分值，如果某段序列有持续高分，则暗示可能含有与训练集相近的结构，很可能是基因。

Genemark

最先取得成功的是原核基因预测，其中比较优秀的是 Genemark:
最初用于原核生物基因预测，首先用高分样本训练参数，然后采用 5 阶 Markov 模型对序列按照不同的读码框打分确定基因结构。后期使用 HMM 为真核基因结构建模，对应的版本是：GeneMark-E* 和 GeneMark.hmm-E.
开发者：Georgia Institute of Technology, Atlanta, Georgia, USA.

tradeoff:acuracy vs. feasibility vs. overfit

在一定范围内，马氏链的阶数越高越好（太高的话会 overfit？是个可以谈的话题，关系模型的弹性和准确，但现有的计算能力下这种担心只是乐观的一厢情愿），但通常不会高于 10，原因：

- 计算复杂度，这些串的概率都是用常量存储的，数量是随阶数指数增长的
- 太长的 motif 会导致支持数据不够 H.influenzae genome size 1.8mb,5-order,$averagefold = 1.8^6/4^(5+1) = 439$ (顺便统计一下每种 motif 的真实含量，搞清楚那个卡方阈值 400 到底怎么来的，肯定先从经验分布下手, 那个 95 置信区间是不是虚的？)

分辨率

由可靠基因构成的训练集和随机集合的碱基组成是很不同：

- 单碱基水平：GC 偏倚
- motif 水平：短的强信号
- 序列水平：熵分布……

精心构造的训练集在不同长度的 motif 分布都有特征（分布谱），分布谱反映了训练集合的特征。GeneMark 只能利用定长的 motif，对数据集的描述能力有限。

Glimmer

Glimmer 在定阶马尔科夫模型上做改进，提出可变阶的 Interpolated Markov Model。企图利用不同长度的 motif 更精细地描述数据集特征。
IMM 的特点是对训练集中不同强度的模式都可以充分利用，优先使用强的 long motif, 如果 long motif 没有足够的数据支持，IMM 对该 long motif 的次阶子串进行打分，并通过一种准确的加权策略利用次阶子串' 插值'出这个 long motif 分数(I=interpolated), 如果次级子串仍然没有足够的支持，这种'插值'还可以继续下去，直到子串短到可以被足够数据支持为止，最短即是单个字符。

Glimmer

Glimmer 的打分策略设计非常巧妙，细节见： IMM frame
1997 年的文章： Microbial gene identification using interpolated Markov models
1999 年的文章： Improved microbial gene identification with Glimmer
真核预测的版本：
http://www.cbcb.umd.edu/software/GlimmerHMM/ ，同样利用 HMM 对基因结构建模。

基因结构

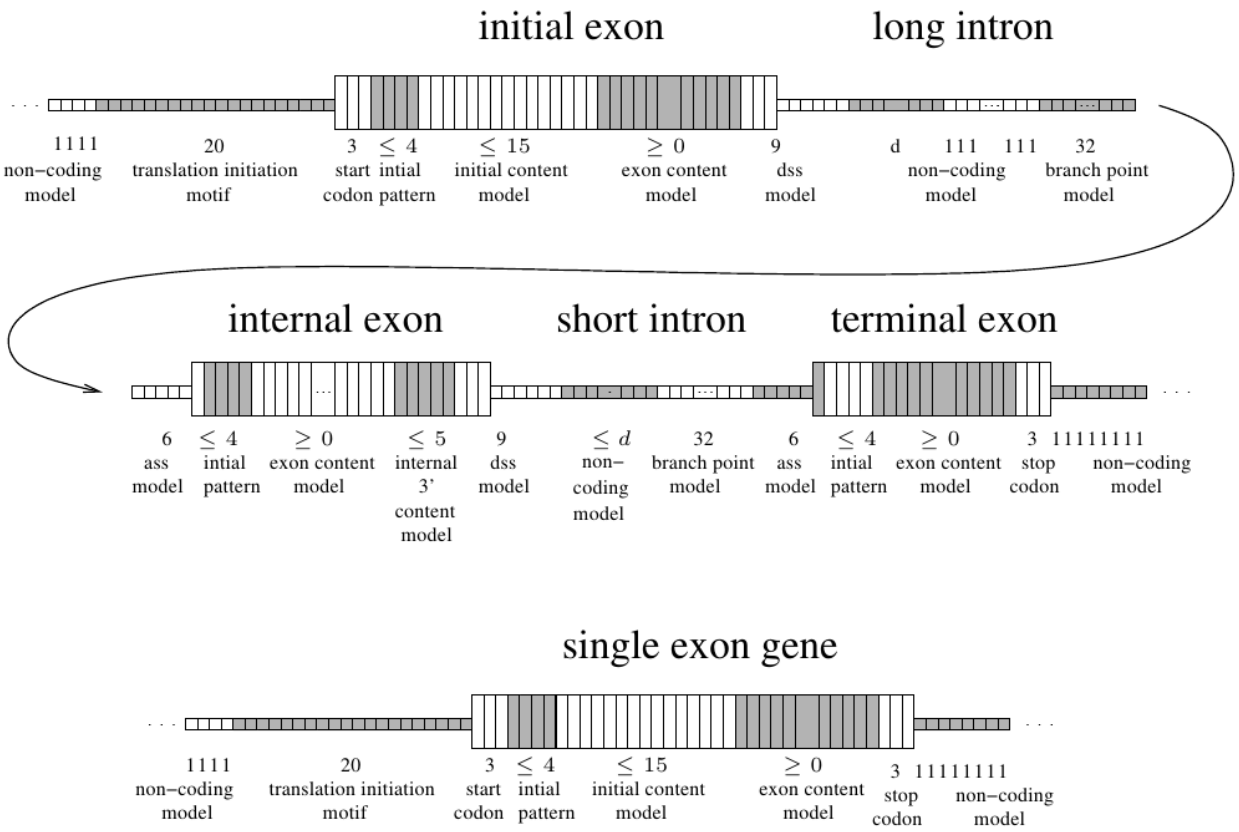


Figure: 真核基因结构

用复杂应对复杂

与原核生物不同，真核生物基因结构复杂，不能用简单的马氏链方法预测。复杂性体现在：

- 结构，强信号短特征 -子模型作为隐状态
- 大量噪声 -对 intron，间区细致建模
- 调控机制和表达模式 -利用外部证据，可以做的地方

用强模式构建隐状态

复杂不全是坏事：

如果结构中某部分有强特征，则可以针对该特征建立细致的子模型。在整个 augustus 框架中，每个子模型是 HMM 的隐状态。

为什么不用 HMM 预测原核生物？

1：原核基因相对均匀，没有强模式构建隐状态；2：状态间没有强依赖关系

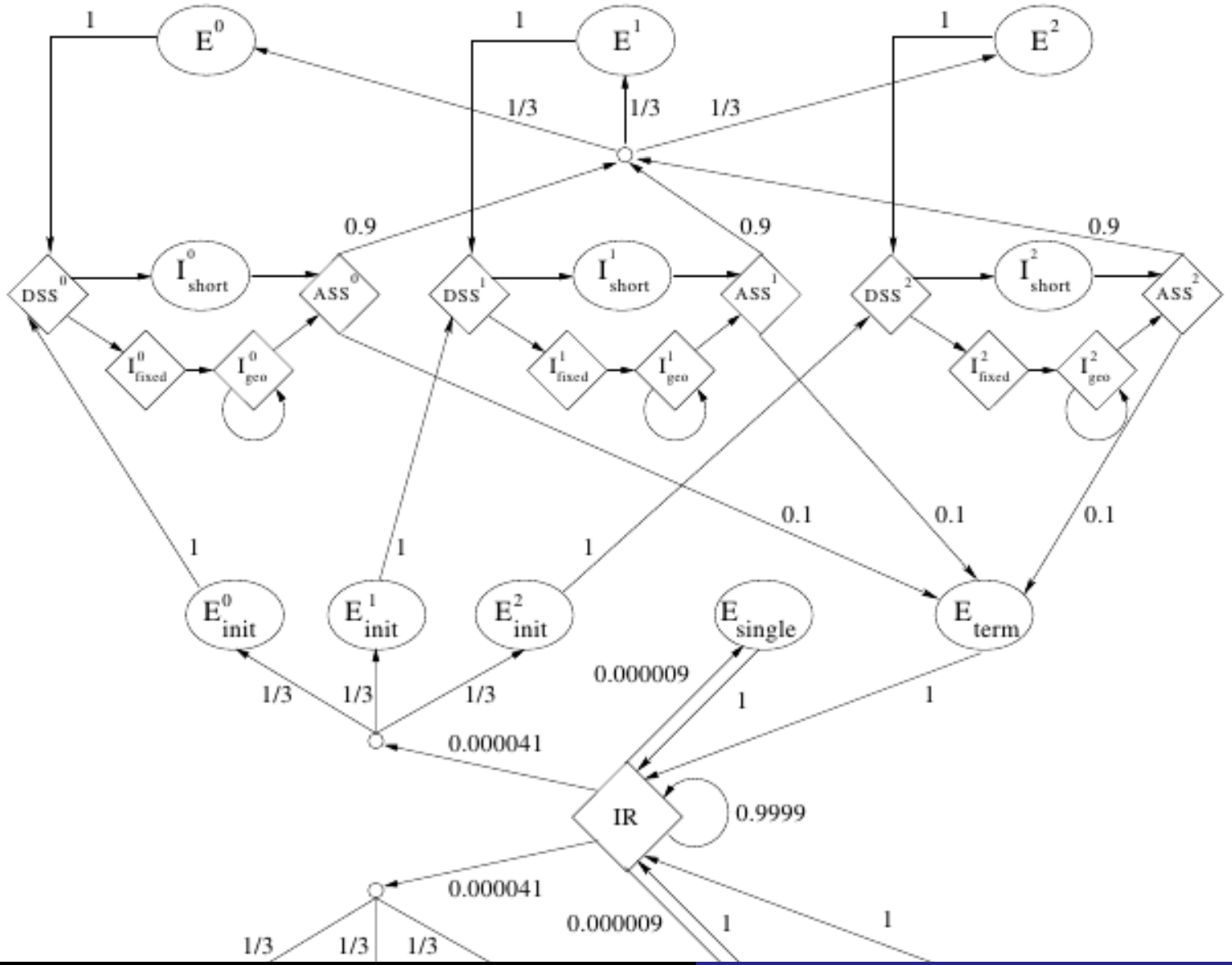
问题

一阶 HMM 是在用非依赖的方法处理依赖问题？马氏链或 HMM，所谓马氏性质未必界限分明

真正的简介

说三个算法，具体对 HMM 的算法链到之前的 HMM 教程。

HMM in augustus



fo
st

HMM in augustus

- 符号发射不定长, 目标是寻找好的 parse
- 多相位,2 方向

外显子建模

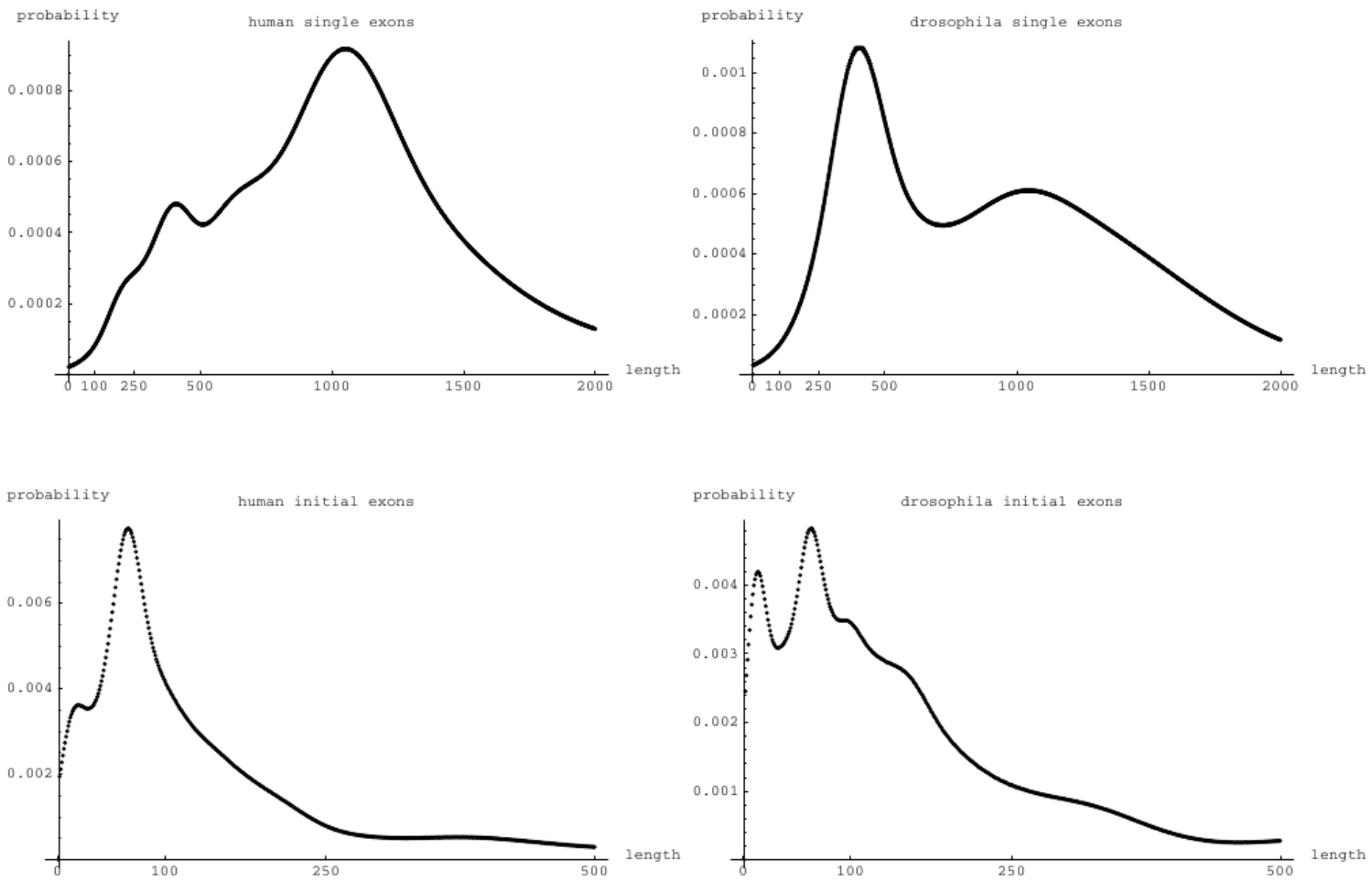


Figure: single and initial extron length distribution

外显子建模

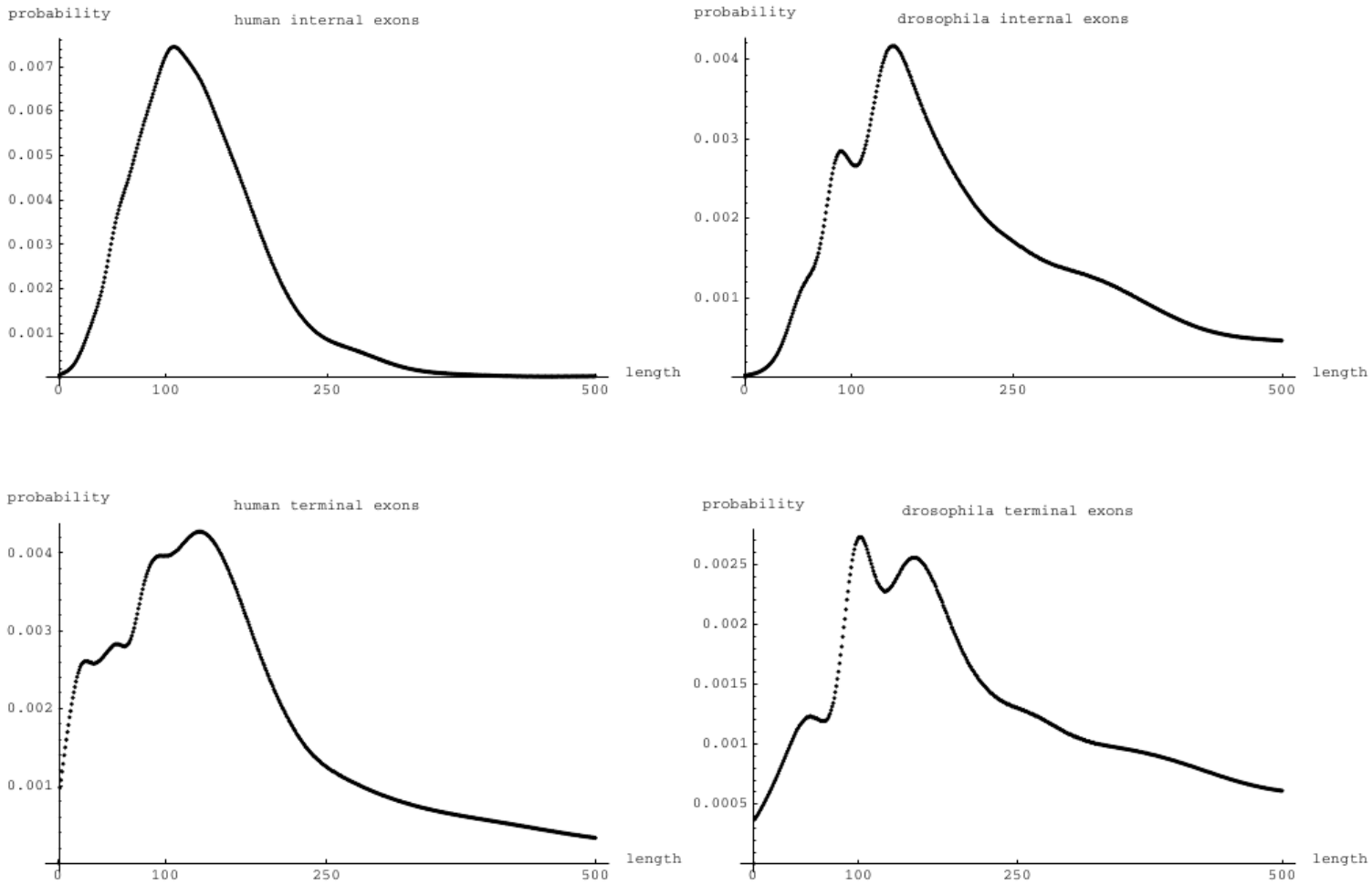


Figure: internal and terminal extron length distribution

外显子建模

- human: single, initial, internal, terminal: n = 462, n = 822, n = 4334, n = 822, respectively;
- Drosophila: single, initial, internal, terminal: n = 76, n = 324, n = 917, n = 324, respectively.
- 外显子分布窄，可以构造经验分布
- 密度估计利用高斯核函数

内含子建模

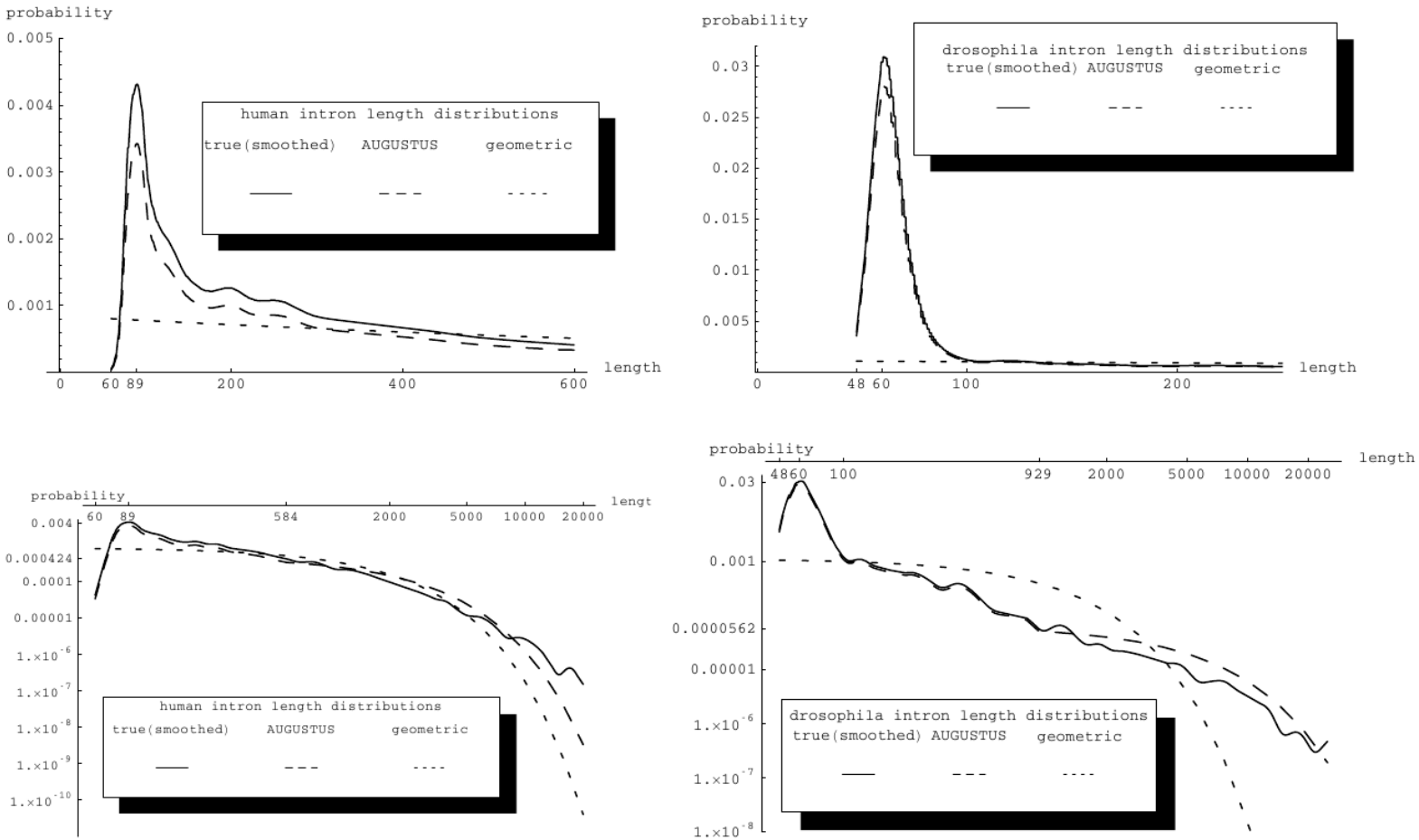


Figure: intron length distribution

内含子建模

显然：

- 内含子分布宽
- 内含子分布比外显子规律
- 峰后近似几何分布，对长内含子有低估，适当 shift

几何分布对内含子的低估问题：Reese et all,GENIE,a gene finder for Drosophila.

所以：

- 宽分布不适合采用经验分布描述（内存开销只是一方面）
- 采用经验分布 + 几何分布的混合模型，有 shift，但没有验证是否解决了对长内含子的低估

内含子建模

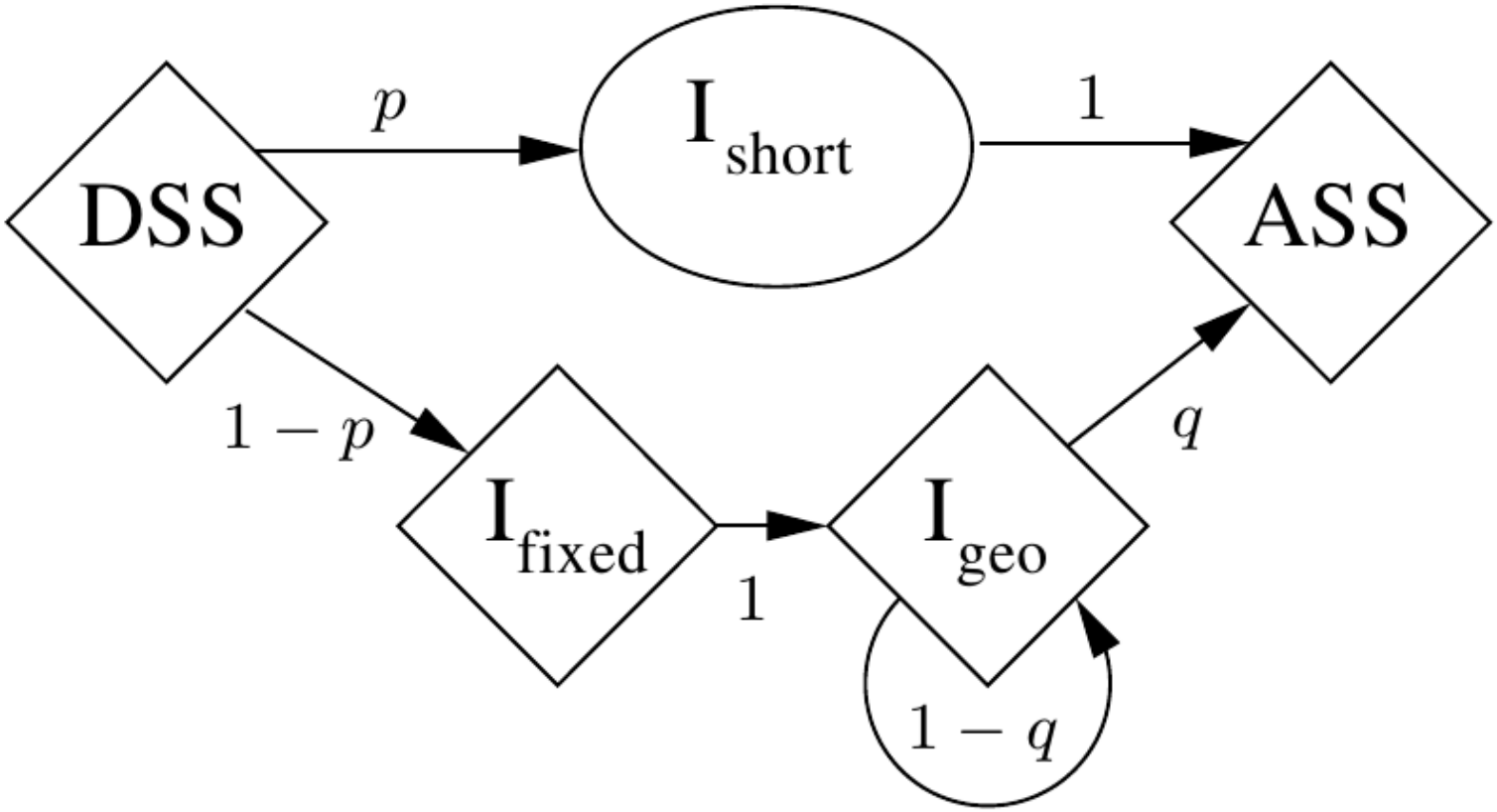


Figure: augustus 内含子模型

内含子参数 是由三个限制条件决定的。

强短信号

真核基因结构复杂证据之一是在基因上下游有丰富的特征模体(motif), 有效识别这些模体可以帮助检测潜在基因区域。
除了对外显子和内含子长度建模外，augustus 也对这些短模式建立子模型。

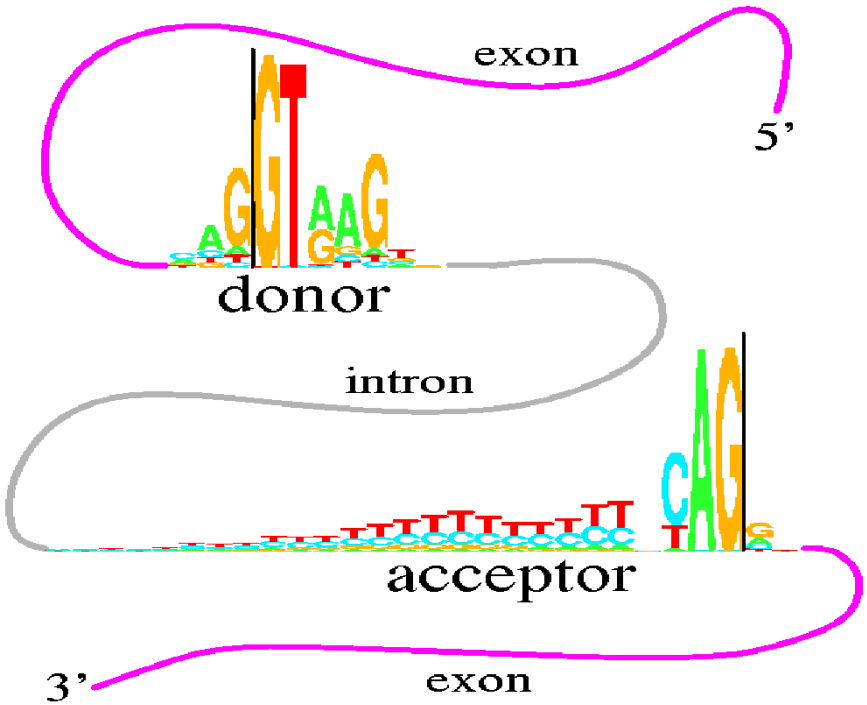


Figure: 外显子与内含子之间由 GT-AG 间隔，这是一个明显的短信号

两个滑动模型

对子模型建模 augustus 采用原核基因预测常用的两个小模型：
窗口加权字串模型 (WWAM), 插值马氏链 (IMM)。
这两个模型的训练与原核基因的短模式训练相似。

其他子模型

state	submodels
E_{single}	translation initiation motif, start codon, initial pattern, initial content model, exon content model, stop codon
E_{init}^{j}	translation initiation motif, start codon, initial pattern, initial content model, exon content model
E^{j}	initial pattern, exon content model, internal 3' content model
E_{term}^{j}	initial pattern, exon content model, stop codon
IR	non-coding model
DSS^{j}	dss model
ASS^{j}	branch point model, ass model
I_{short}^{j}	non-coding model
I_{fixed}^{j}	non-coding model
I_{geo}^{j}	non-coding model

转移矩阵

以上定义的所有模型均为‘发射模型’，是 HMM 中的微观部分，以（条件）概率分布（矩阵）形式存在，保证打分规则一致性。另外，即使所有模型已经完整定义，在预测阶段，HMM 沿未知序列滑动时，事先不知道 parse 是什么样，因此不能直接使用。子模型之间的转移由转移矩阵决定，仍然是概率形式，是 HMM 的宏观部分。

constraints_shadow_partial.txt
constraints_shadow_partial_utr.txt
states_shadow_2igenic.cfg
states_shadow.cfg
states_shadow_intronless.cfg
states_shadow_utr.cfg
states_singlestrand_2igenic.cfg
states_singlestrand.cfg

trans_shadow_atleastone.pbl
trans_shadow_complete.pbl
trans_shadow_complete_utr.pbl
trans_shadow_exactlyone.pbl
trans_shadow_intronless.pbl
trans_shadow_partial.pbl
trans_shadow_partial_utr.pbl
trans_singlestrand_atleastone.pbl
trans_singlestrand_complete.pbl
trans_singlestrand_exactlyone.pbl
trans_singlestrand_partial.pbl

模型的变体

geneModel=:

- partial(default)
- intronless
- complete
- atleastone
- exactlyone

- singlestrand=true
- hintsfile=hintsfile
- extrisnicCfgFile=cfgfile

GC content dependent training

- Burge：基因预测模型的多数参数与 GC 含量关系密切。
- Genescan 训练集分 GC%(<43%，43%-51%，51%-57%，>57%)4 个子集训练 4 组参数，预测时根据输入序列 GC 含量选择其中一组参数。

Genescan 没有充分利用训练集。

- Mario：Augustus 将训练集按 GC 含量分成 10 份，用所有训练数据单独训练每个子集
- 参与训练的子集权重由与被训练子集的平均 GC 含量 (α) 差异决定，权重决定训练层数
- 权重近似成 1-10 之间的整数：
  $w(\alpha, \beta) = cell(10 * exp(-200 * (\alpha - \beta)^2))$
- 在预测时，选择与输入序列 GC 含量接近的一组参数进行预测。
- 权重参数文件在： *augustus/config/species/centain − species/certain − species_weightmatrix.txt*
- Pitfall:augustus 每个子集的 size？

外部证据

这是最容易提升性能的部分，除了 augustus，genescan 等软件也在做这种努力。整合外部 hint 的参数文件
在:/augustus/config/extrinsic/**.cfg, 主要规定了激励、罚分规则。

- M manual anchor
- P protein database hit
- E est database hit
- C combined est/protein database hit
- D Dialign
- R retroposed genes
- T transMapped refSeqs

模块及变量介绍

数据集

整体分值计算

$$P(S|M) = \sum_{x=1or9}^{n} IMM_8(S_x)$$

motif 分值计算

$$IMM_k(S_x) = \lambda_k(S_{x-1}) * P_k(S_x) + [1 - \lambda_k S_{x-1}] * IMM_{k-1}(S_x)$$

λ 是表示数据可信度的权数， $P_k(S_k)$ 是一个估计值。

ORF

完整的基因结构包含起始密码子和终止密码子:

- A genome of length n is comprised of (n/3) codons
- Stop codons break genome into segments between consecutive stop codons
- The subsegments of these that start from the Start codon (ATG) are ORFs

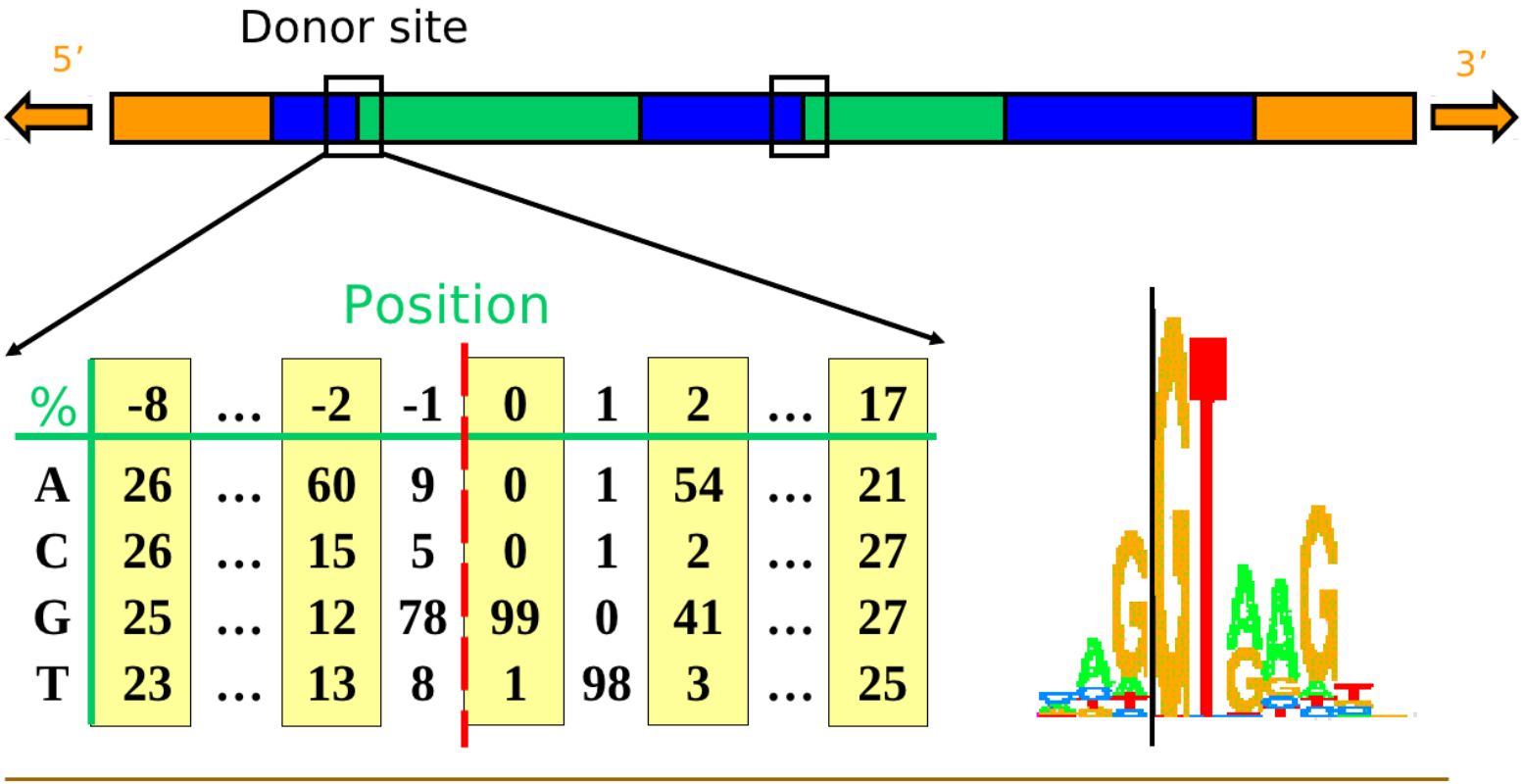
如果序列是随机的，终止密码子应该每 $21(21 = 64/3)$ 个密码子中出现一次，基因长度要大于此长度。设定合理的阈值确定长 ORF 即可将随机序列与基因分离。当确定一段 orf 后可以结合密码子使用偏倚，motif 位点特征等进一步分析确定是否是基因。

偏倚数量化

基因组在不同层面上存在组成偏倚。

- GC 含量 影响多数参数训练
- 密码子
- 信号位点碱基丰度 检测短模式

偏倚数量化



From lectures by Serafim Batzoglou (Stanford)

Figure: 供体位点显示强烈偏倚

偏倚数量化

在真核基因预测中，我们关心的子模型特征通常是碱基丰度偏倚。对偏倚数量化既是统计依据的需要也便于程序自动工作。
假定：

- 背景序列每个字出现概率是均匀分布 (1/20 或 1/4)
- 强信号处概率分布是不均匀的（但无须知道具体分布）

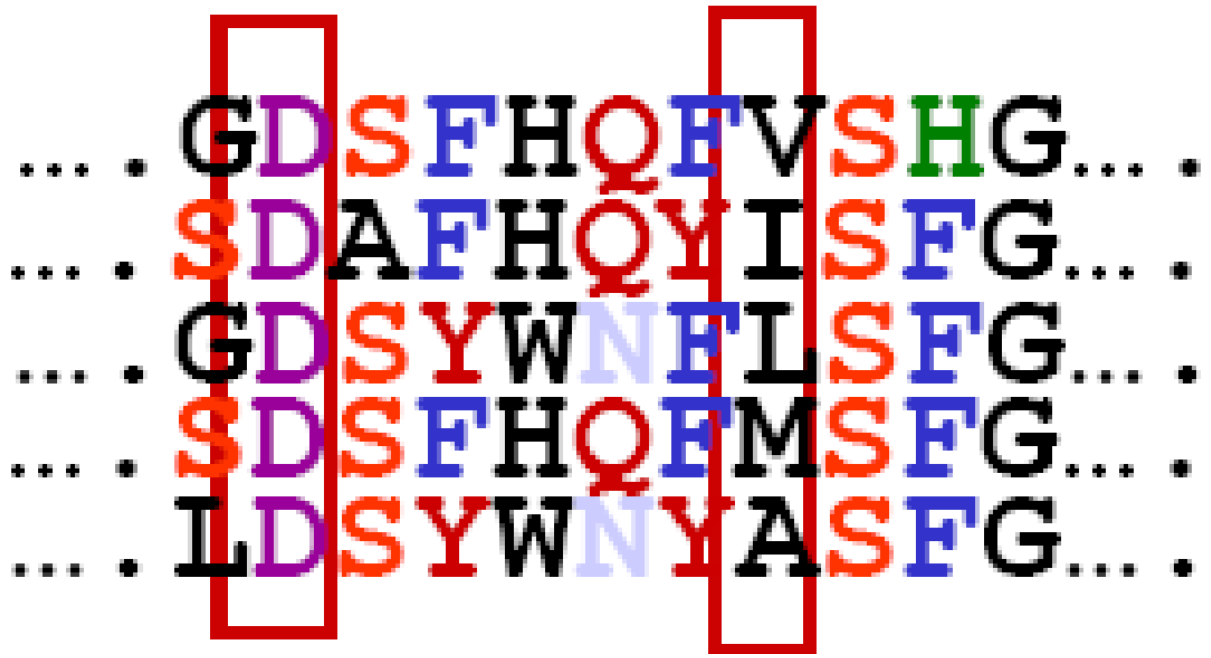
一个好的数量标准可以反映这种差异，它最好是这二者的函数。

出于同样的工程需要，这个量已被信号工程师定义过：

$$H = -\sum_{i=1}^{4or20} P_i * (log_2 Pi)$$

若以 2 为底取对数，信息强度的单位是 bit。

偏倚数量化



- 位点 1：
  $H_{bg} = -\sum_{i=1}^{20}(1/20) * \log_2(1/20) = 4.32bit$, $H_{site1} = 0bit$, 信号强度：4.32bit
- 位点 2：
  $H_{bg} = -\sum_{i=1}^{20}(1/20) * \log_2(1/20) = 4.32bit$, $H_{site2} = 4.32bit$ ，信号强度:0bit

内含子参数确定

内含子模型三个参数确定：

- $d + \frac{1}{q} = E[L]$
- $P(M = d) = P(M = d + 1)$
- $P(M = l) = (1 - p)(1 - q)^{l-d-1}q$