原核基因注释 -模型与软件

于秋林

结构与功能实验室
yuqiulin@genomics.cn

November 9, 2012

1. ORF 确定基因结构
2. Genemark
3. Glimmer
4. HMM 模型
   - 生物背景
   - HMM in augustus
   - 外显子建模
   - 特征序列 & 子模型
   - 利用外部证据推断基因结构

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。

Genemark

Genemark: 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，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？)

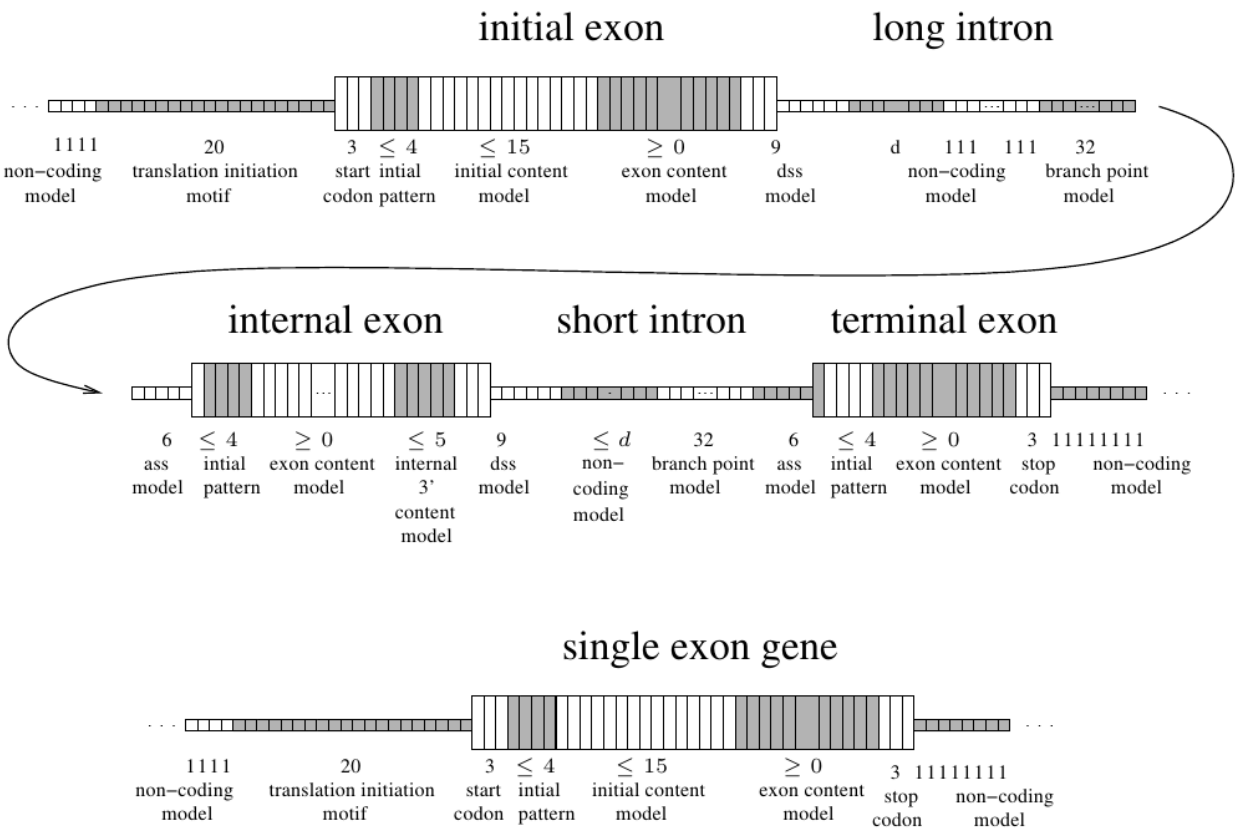
Glimmer

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 对训练集中不同强度模式充分利用，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。

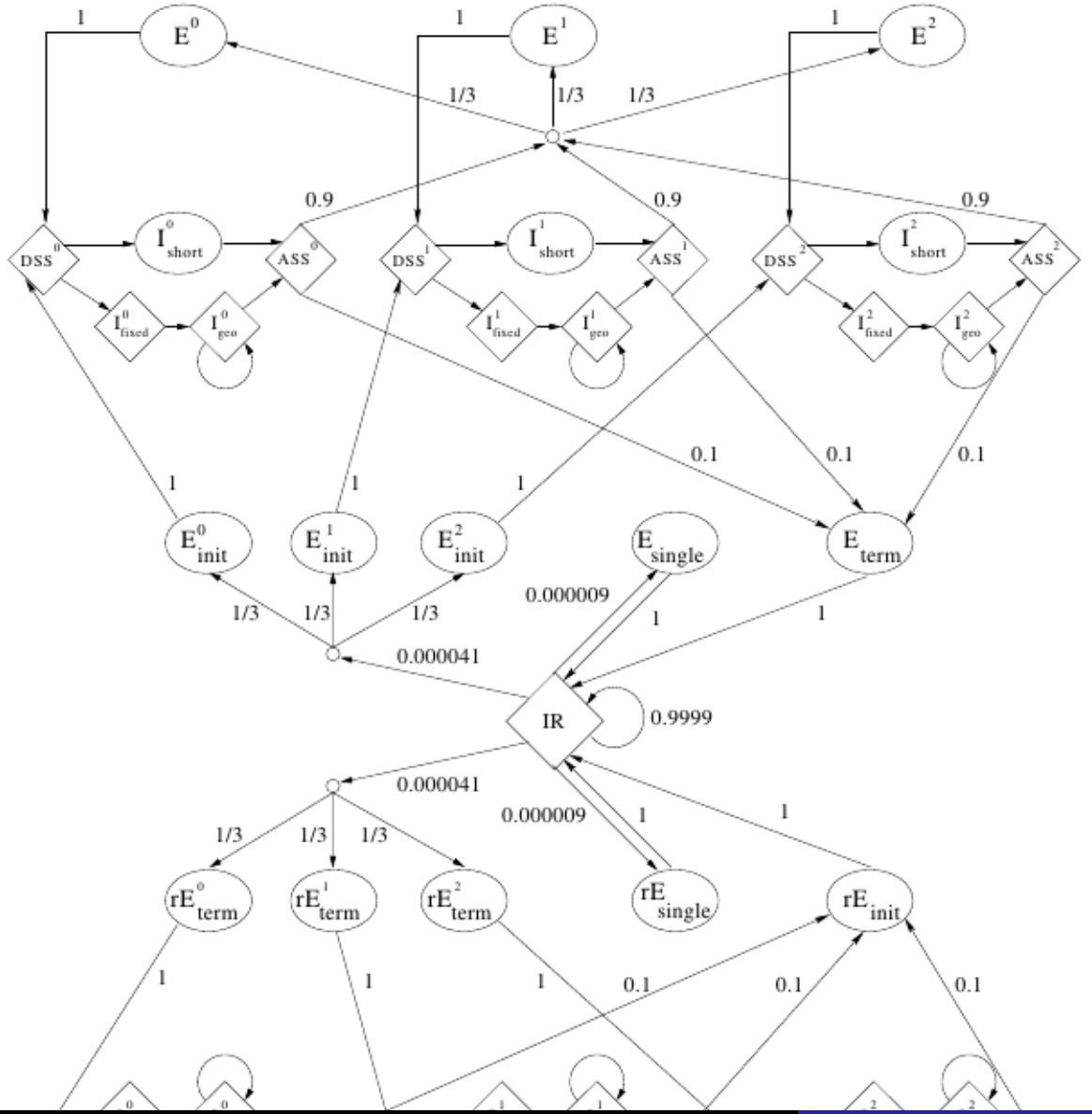
Glimmer

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。

基因结构可做状态划分


initial exon
long intron
1111 non−coding model
20 translation initiation motif
3 start codon
≤ 4 intial pattern
≤ 15 initial content model
≥ 0 exon content model
9 dss model
d
111 non−coding model
111
32 branch point model
internal exon
short intron
terminal exon
6 ass model
≤ 4 intial pattern
≥ 0 exon content model
≤ 5 internal 3' content model
9 dss model
≤ d non−coding model
32 branch point model
6 ass model
≤ 4 intial pattern
≥ 0 exon content model
3 stop codon
11111111 non−coding model
single exon gene
1111 non−coding model
20 translation initiation motif
3 start codon
≤ 4 intial pattern
≤ 15 initial content model
≥ 0 exon content model
3 stop codon
11111111 non−coding model

HMM in augustus



forward strand

reverse strand

外显子建模

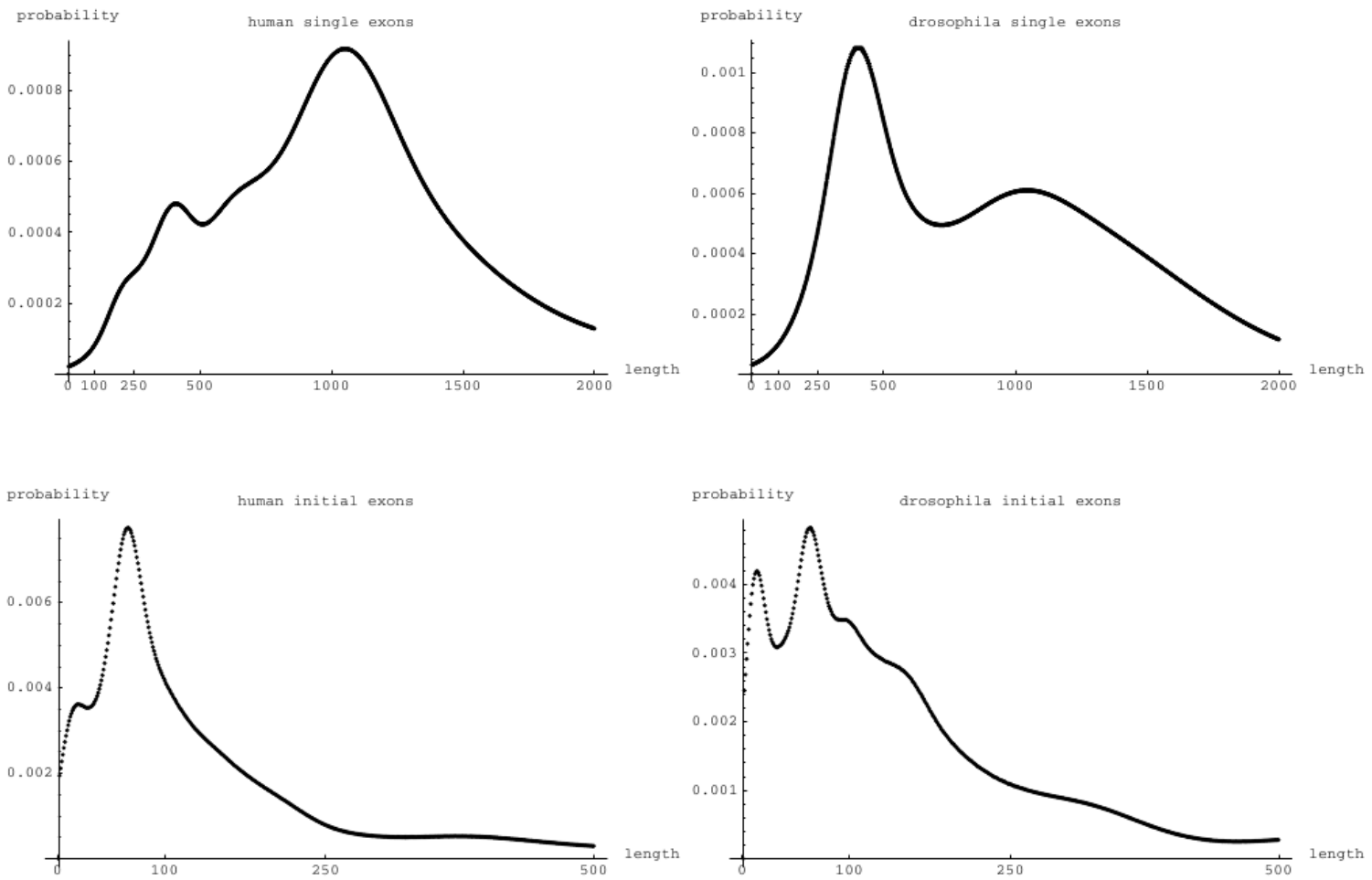


Figure : single and initial extron length distribution

外显子建模

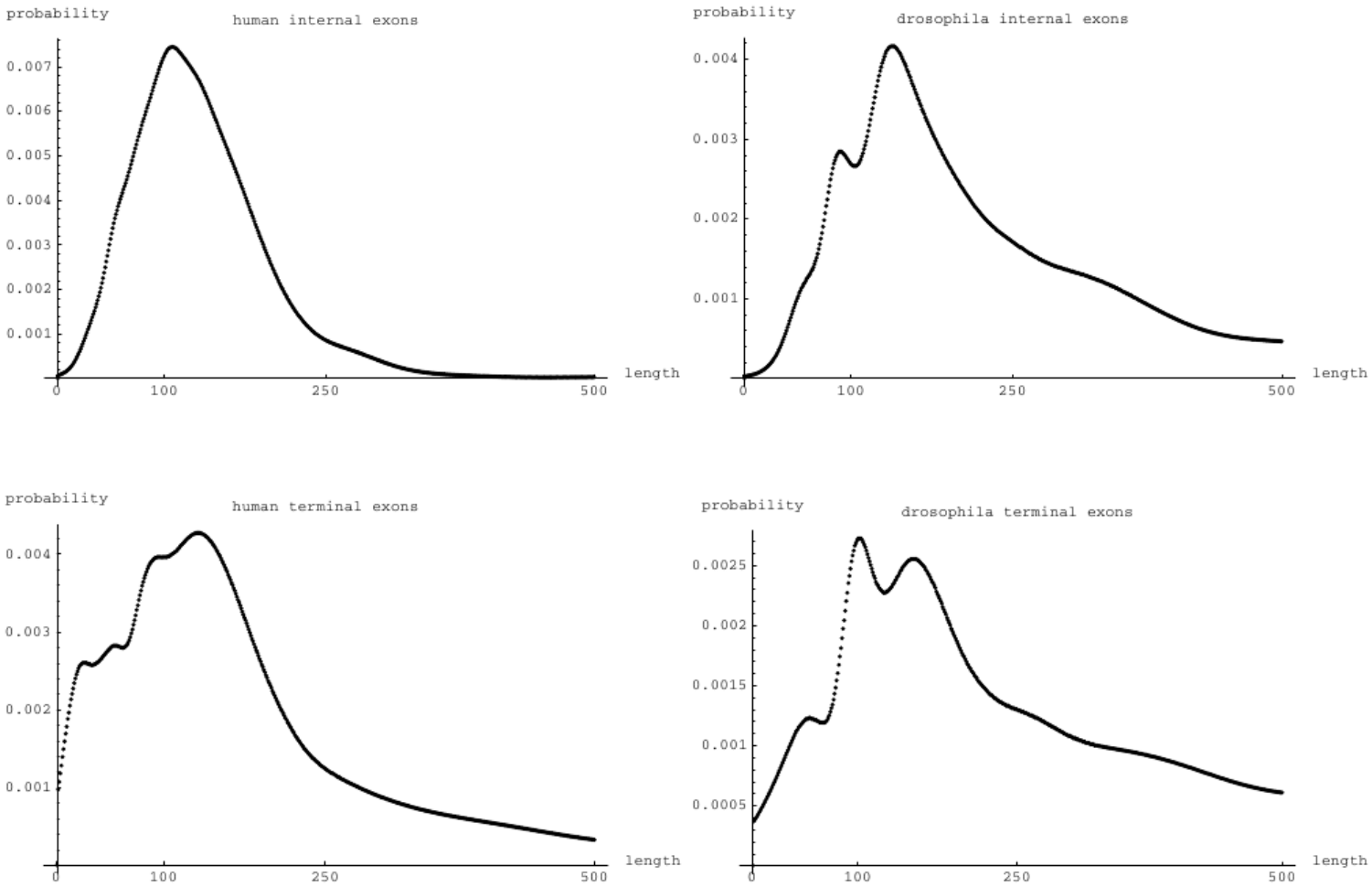


Figure : internal and terminal extron length distribution

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

强短信号

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 (内含子) 长度建模外，短模式对基因识别也非常重要。

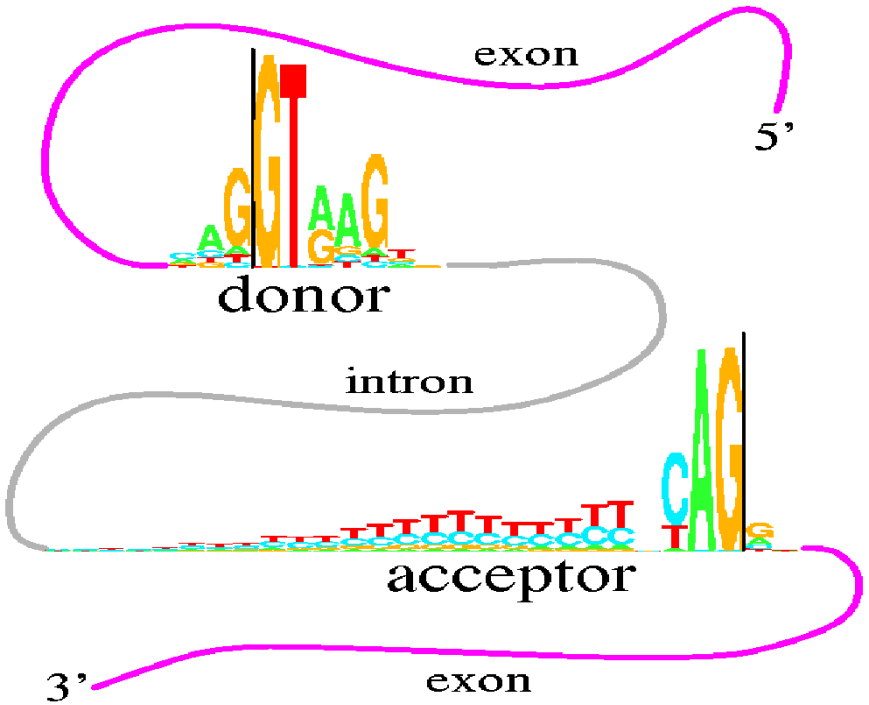
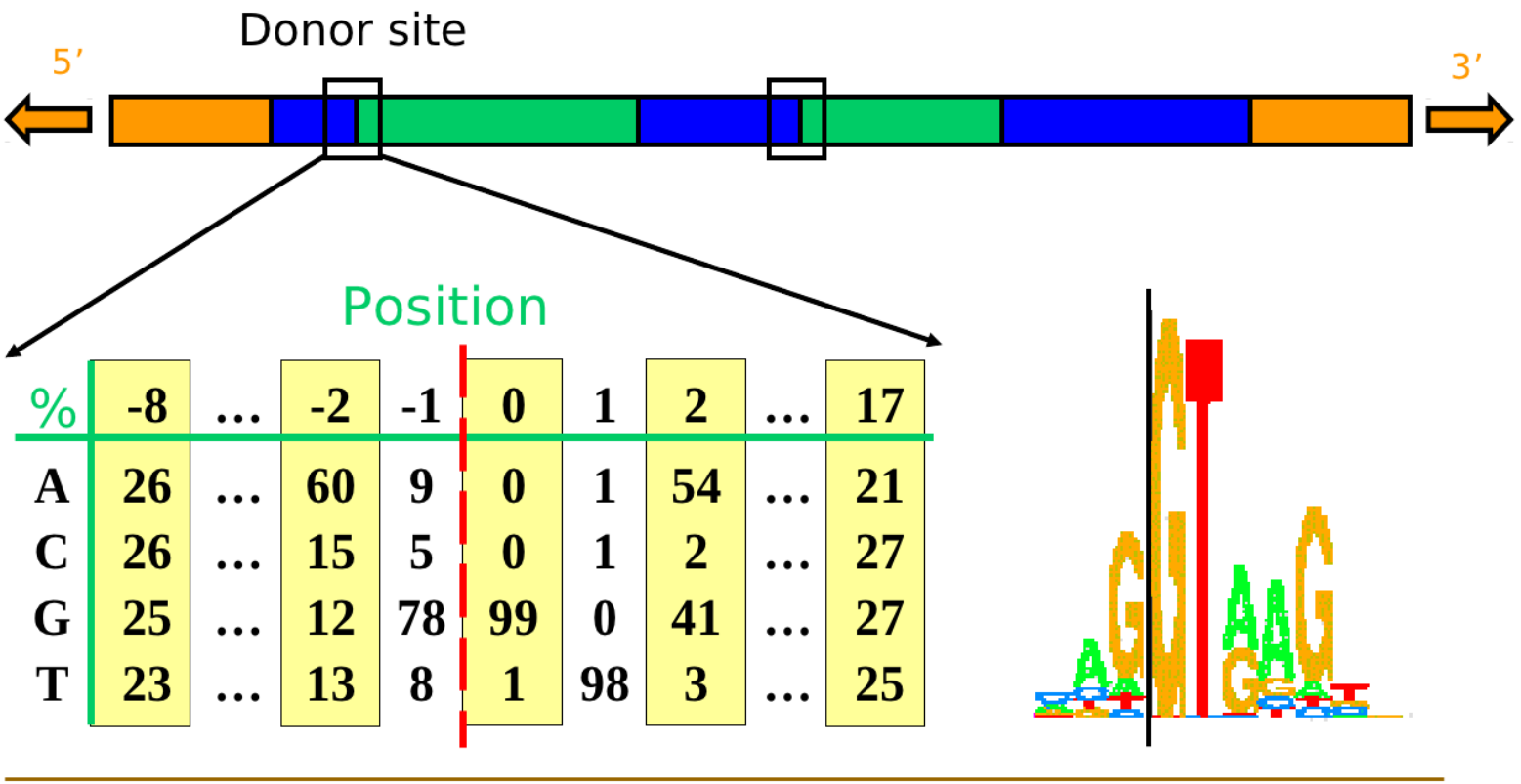


Figure : 外显子与内含子之间由 GT-AG 间隔，这是一个明显的短信号

偏倚数量化

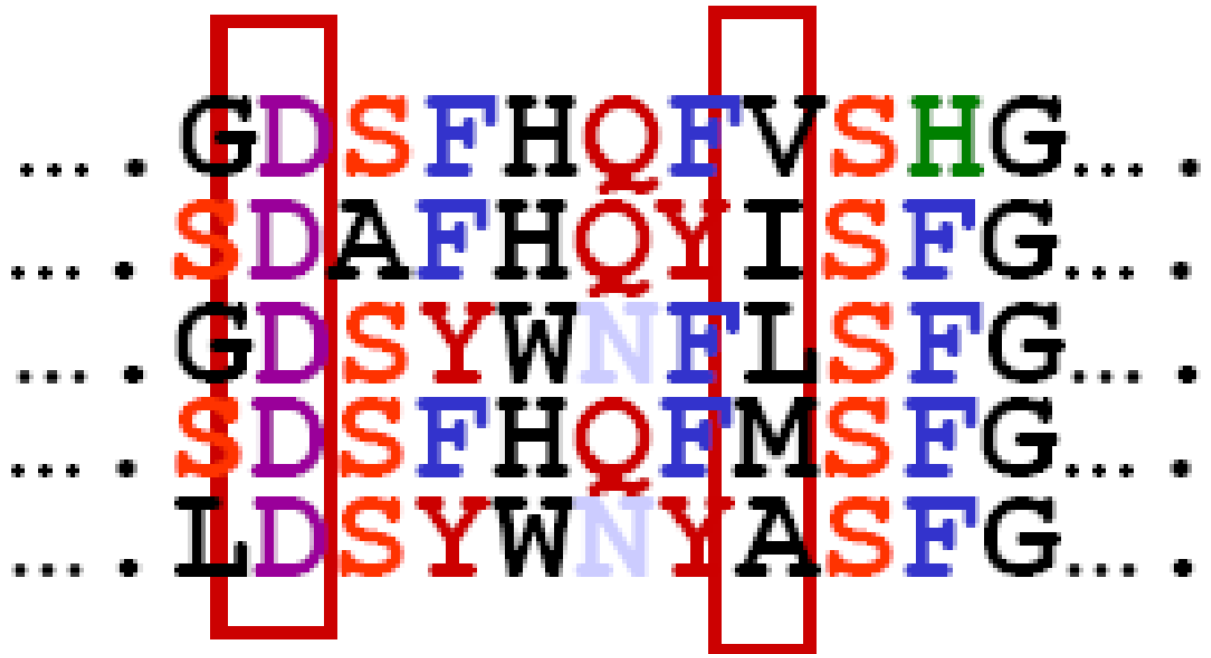


%	-8	…	-2	-1	0	1	2	…	17
A	26	…	60	9	0	1	54	…	21
C	26	…	15	5	0	1	2	…	27
G	25	…	12	78	99	0	41	…	27
T	23	…	13	8	1	98	3	…	25

From lectures by Serafim Batzoglou (Stanford)

Figure： 供体位点显示强烈偏倚

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

外部证据

最容易提升性能的部分，除了 augustus，genescan 等软件也在做这种努力。

- 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