分子进化算法介绍

元件注释
2011-09-07

BGI-RD
深圳

September 8, 2011

Outline

背景介绍

分子进化是在分子水平上对进化过程进行研究，始于 20 世纪 60 年代。分子水平的进化研究致力于两大问题：重建物种件的进化关系以及了解进化过程的动力于机制。前者属于系统学领域，后者是通过估计核苷酸置换速率以及采用序列数据来检测突变与选择模型进行研究 [Yang]。
本次组会介绍如何从比对数据出发，得到物种间完整的进化关系，以及对结果的评估。涉及的概念和方法：

- 二叉树
- 简约规则，Sankoff 算法
- 碱基替换矩阵，进化距离估计
- 最大似然规则,Felsenstein pruning 算法
- 树的检验和模型比较:Bootstrap LRT

树

物种（基因）间的谱系关系可以用一棵树表示，叶子节点代表可以观测的物种，内枝节点代表已灭绝的祖先，枝长通常是某种距离的测度。几乎所有进化关系都可以用二叉树表示（严格的星状树存在争议，另外，星状结构可以分解成内枝很短的几棵子树）。不同的建树方法可以产生有根树或无根树，其中分子钟的存在是一个关键的假设，通常这一假设仅在近缘物种间成立，对不能确定树根的情况，通常引入关系更远的物种作为外群来置根。

Newick 记法

$$(A : l1, (B : l2, C : l3))$$

用一对括号将临近物种归并到一个分支，冒号后的树脂是距离测度。

树的计数

- 无根树：
  $T_n = T_{n-1} * (2n-5) = 3 * 5 * 7 ... (2n-5) = (2n-5)!!/2$
- 有根树： $RT_n = (2n-3) * T_n$

数的增长大于指数速度。

- T3=1
- T5=15
- .....
- T10=2027025

树的搜索

- 添加法： $N = 3 + 5 + ... + (2N - 5)$
  - 不同阶段的分值不能比较
- 星状分解法:$N = n(n-1)/2 + (n-1)(n-2)/2 + ... + 3 = n(n^2-1)/6 - 7$
  - 每步树的规模相同，分支可以比较
  - 问题：星状树的似然值是怎么算的？这时候枝长还是进化距离的测度吗?
- 分枝交换

简约规则

简约规则选择同通过最少变化达到当前状态的树最为最优树。进化历程在这样的树上有最简洁的表现。
假定四个物种 1，2，3，4；
三棵无根树：

- T1:((1,2),(3,4)) 支持的位点构型：xxyy
- T2:((1,3),(2,4)) 支持的位点构型：xyxy
- T3:((1,4),(2,3)) 支持的位点构型：xyyx
- T*:((*,*),(*,*)) 非信息位点

Sankoff 算法

先计算子树的局部最优分值，逐层向上，直到根节点。
约定罚分规则如下：

碱基	T	C	A	G
T	0	1	1.5	1.5
C	1	0	1	1.5
A	1.5	1	0	1
G	1.5	1.5	1	0

简约法的缺点

- 计算过程枝使用树的结构信息，没有利用枝长
- 外枝显著大于内枝时倾向于将两个长枝聚合，‘长枝吸引’

这两个缺点反映的是同一个问题: 不能处理多重击中。下面的 ML 模型解决这个问题。

- 简约法需不需要计算两条序列间的距离？
- 有没有与距离相关的假定？
- 如果简约法估计有偏差，这偏差是向上还是向下的？

距离定义

‘距离’是为了配合进化树的枝长而提出的形象说法，实际上我们真正关心的是分化时间。如果假定分化速率不变，那么二者没有区别。在似然模型中，分化速率不变是最重要的假设。
进化分析中，通常是根据比对数据估计分化距离（时间）。
最简单的距离定义是位点差异比例，序列相似度越大，认为分化时间越短。

- 认为进化时间与位点差异是线性关系
- 分化初期这种定义勉强靠谱
- 位点经过多次变化的，称为‘多重击中’，这些位点破坏了线性关系
- 位点差异不是时间的线性函数，不能直接作为距离的度量
- 位点差异距离是对真实分化时间的高估/低估？

利用序列差异估计分化距离，难点是序列差异不是分化时间的线性函数，需要处理‘多重击中’
下面用马尔科夫链周全地解决多重击中的问题。

JC-69 矩阵

马尔科夫链第一要素：每一次变化只与当前状态有关。
这种关系是通过状态转移矩阵定义的，在当前语境，应该叫氨基酸替换速率矩阵。这是最简单的替换矩阵，只有 1 个参数 a，假设在分化过程中每个碱基以等速率向其它三种碱基转移。

碱基	A	C	G	T
A	-3a	a	a	a
C	a	-3a	a	a
G	a	a-3a	a	
T	a	a	a	-3a

一点点计算

以碱基 x 为例，说明 M 链怎样处理多重击中
在 0 时刻，x 变异，其比例设为 100 在 t 时刻，假如 x 比例为 $p(t)$, 则在下一个 $t+1$ 时刻，

$$p(t+1) = p(t) * (1-3a) + (1-p(t) * a$$

这是个离散过程，但是因为一个单位时间（比如定义一个繁殖世代，这取决于 a 的大小）相对于整个进化历史非常短，因此可以近似成一个连续的过程。

$$p^{'}(t) = -4a * p(t) + a$$

解此微分方程:$p(t) = \frac{1}{4} + c * e^{-4at}$, 利用边界条件 $p(0) = 1$ 确定 $c = 3/4$, 所以:

$$p(t) = \frac{1}{4} + \frac{3}{4} * e^{-4at}$$

尽管有以上结果，但是分化时间 t 仍然是不能直接得到的。
因为每个碱基有三个变化方向，所以定义 $d = 3at$, 差异位点比例 $p^* = x/n$ 作为对 $[1-p(t)]$ 的估计，有：

一个位点的计算

$$f(x_i|\tau_j, v_j, \theta) = \sum_y [\pi_{y_{root}} (\prod_{k=1}^{s} p_{y_{\sigma(k)}, x_{ik}}(v_k, \theta) (\prod_{k=s+1}^{2s-2} p_{y_{i\sigma(k)}}(v_k, \theta))]$$

谢谢