分子人类学中的统计方法

吴晋渊

2021 年 11 月 4 日

1 系统发育

较高等的动物有减数分裂。一个个体携带的一对同源染色体中，有一条来自父亲而另一条来自母亲。父亲和母亲在产生精子和卵子的时候是通过减数分裂，即来自他们每个人的父母的两条染色体又只有一条进入了配子中。这样表面上看，似乎在遗传信息的代际传递中，实际上只有两条染色体是重要的：一个个体的两条同源染色体是从N代祖父母的该种染色体中取了两条，或者说对该个体的该种染色体有贡献的每一代祖父母中都只有两个个体。不过，应当注意，减数分裂中会出现同源区段重组，从而，实际上，每一代祖父母中，对一个个体的某种染色体有贡献的个体数目要多于2。虽然如此，由于同源区段重组发生几率是不大的，对某一个体的某种染色体的贡献超过某一阈值的祖父母的个数随着代际数目的增加只是线性增加。这件事在图1中展示得很清楚。另一方面，祖先的数目随着代际数目上升是指数增长的。因此，一个家族中，大部分的祖先对某个后代都几乎没有遗传信息的贡献。一个直接的结果就是，使用系统发育方法得到的族谱和（往往是记录父系血统的）族谱会非常不同，因为显然没有什么规定了主要的遗传物质贡献都来自每一代人的父系。

虽然每个个体继承的祖先的遗传信息都只是很少一部分，但是如果种群不断扩张，每一代的后代个体都足够多，每个祖先的遗传物质都应该被一部分现存后代继承了（当然，可能有各种突变）。实际上种群当然不可能不断扩张，总会有几支血脉灭绝的，于是一个种群的现存后代的遗传信息就只来自其祖先的一小部分。

以上我们都只是在讨论个体，不过其实我们可以更加细粒度一点，由于一条染色体上不同的部分可以来自不同的祖先，实际上可以讨论一个**区段**的祖先等等。

如果不考虑基因突变，那么通过已知同源的区段的DNA序列比对，我们可以粗略地确定现存个体的祖先：一些个体肯定有一个共同的祖先，另一些个体肯定具有另一个祖先，等等。基因突变让辨认族谱的工作稍微麻烦了一些，但是它带来的好处远远大于坏处：它能够让我们确定族谱的具体结构甚至分支时间。这是因为基因突变的速率是可以确定的，那么我们就能够通过类似于版本学的方法辨认出哪两个序列的关系更加紧密——总的来说，两个序列越相似它们的关系就越紧密，并且通过一些统计模型我们可以定量（虽然仍然不是非常精确）地获知它们可以追溯到多早的一个祖先。

我们会发现，现存个体的DNA序列只是反映了一小部分祖先的DNA序列这件事实际上大大简化了我们重构族谱的工作，因为反正信息就这么多，族谱是能够很清楚地展示出来的。一个**区段**可以有明确的**进化树**或**系统发育树**：由于同源区段重组是以同源区段为基本单位的，我们可以暂时认为同源区段的变化都来自基因突变。这样，设我们有一些区段$\{S_i\}$，通过彼此比较，会发现它们大体上会来自几个不同的祖先（注意，一个现存个体的某个区段一定可以追溯到一个单一的祖先!），在第2节中用不同的颜色标记。然后我们根据一个通常是马尔可夫链的基因变异模型，会发现，比如说，第2节中的S_1和S_2很可能还有一个时间比较晚近的共同祖先，这个共同祖先和S_3再有某个（标记成红色的）共同祖先。

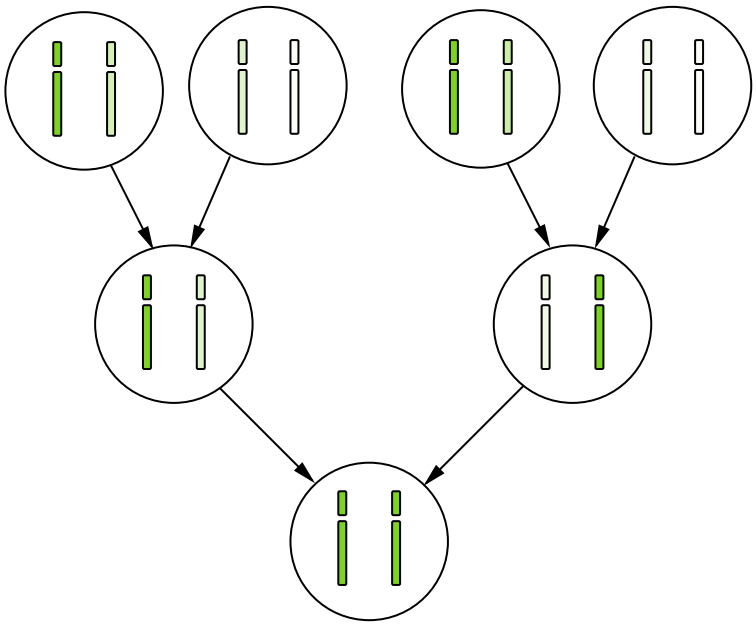

图 1: 染色体的传代：我们用颜色深浅来代表一个个体的祖先的染色体对某个后代个体的贡献，可以看到随着我们往上追溯族谱，大部分祖先的染色体的颜色都是很淡的

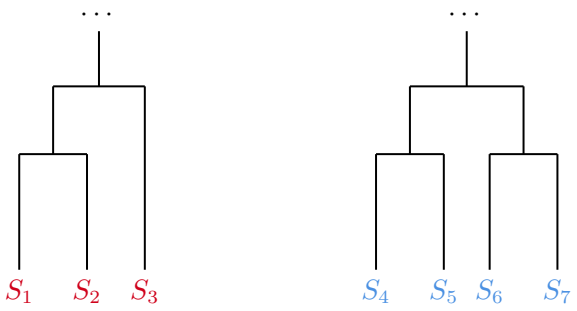

图 2: 进化树或者说系统发育树的例子：红色和蓝色表示两个不同的可能祖先的后代

注意我们这里说的是区段而不是基因：Y染色体自己可以看成一个大的区段，而我们可以画出非常漂亮的Y染色体进化树，从而将全人类定位到一个**Y染色体亚当**上面。

另一方面，关于个体，原则上树并不是足够好的描述方式，既然不同区段完全可以来自两个完全不相关的祖先，并且在常染色体上，一个个体携带同一个等位基因的两个变体，所以即使我们只测试常染色体的单一区段，原则上也不能够绘制出符合历史事实的进化树：我们必须使用**系统发育网**来代替系统发育树。不过这并不是说树在这类情况中毫无用处，因为无论如何，我们可以绘制**聚类树**而不是**进化树**。并且，由于出现在树中的样本通常不会太多，系统发育网的精确结构常常是揭示不出来的，此时聚类树实际上已经是对系统发育过程的足够好的描述了：我们可以假定历史上实际的演化过程大体上是树形的，在此假设之下建立树，然后评估这个树是否足够“稳定”，如果历史上的系统发育过程是非常网状的，那么使用树将不能够很好地拟合它，会出现多次计算得到的树完全不一样、一些结点始终不收敛等等暗示。

常染色体上有大量基因这件事虽然让绘制族谱变得困难，但是同样是有益的，因为这个事实让我们能够更清楚地看到历史上的种群瓶颈——如果一个群体中大量基因的最近共同祖先都能够追溯到同一个时间段，那么这个群体多半就是这个时间段的一小群个体产生的。我们可以据此分析人群的迁移。