进化树的构建和分析

生物信息学导论

孙铭浩

（大致的）目录

1.进化的背景知识

2.系统发生树的结构

3.序列比对

4.系统发育树的构建

1.背景知识

1.1 生物进化与分子进化

Nothing in biology makes sense without the light of evolution.

生物进化的本质是分子进化。因为分子进化，导致筛选，基因控制的蛋白质合成改变，形状改变。改变到一定程度，基因组差异度↑，形态结构↑。

自然选择直接作用于个体，本质上却是作用于基因（尤其是编码蛋白质的基因）以及相关的分子。

因此，可以通过**比较同源蛋白质编码基因的碱基/同源蛋白质中氨基酸序列差异** 可以得到分子进化信息。（构建系统发生树的重要理论依据）

不变残基：同源蛋白质的很多位置上的氨基酸序列相同。说明有共同起源，在物种中变化不大。

可变残基：在其他位置的很多氨基酸残基在种属间差异大。可变残基多，或者说氨基酸序列差异大，说明蛋白质的亲缘关系远。当然，一般认为蛋白质一级结构（氨基酸序列）改变的速率在概率上是恒定的（clock,分子钟假说）。不同蛋白质的改变速率不同。但是统计学上的有意义的。

由于遗传物质突变的积累导致了蛋白质的突变。最终形成新物种。例如细胞色素c，进化速度（即形成稳定突变的速度）很慢。在物种中比较保守。人与primates的cytochrome c几乎没有差别，而人与马的有11个氨基酸的差别。可以通过衡量氨基酸差异的个数，判断系统发育的距离（体现在系统发生树的branch length）。还可以初步估计divergence time（进化分歧时间-通过速率+距离计算）。

1.2 进化的原因

(1) 达尔文 ：自然选择。过量繁殖-生存斗争-优胜劣汰-进化

(2) 综合进化论（the evolutionary synthesis）：综合考虑染色体遗传学、群体遗传学分类、地理、胚胎、生化等。代表人物：Dobzhansky, Huxley, Fisher等。

(3) **分子进化的中性学说**（neutral theory，）：木村资生。

观点：①分子水平上的大多数变化/变异，是由于等位基因发生中性突变。这些中性突变的基因的遗传漂变（genetic drift）。中性突变例如有同义突变（密码子简并性）、非功能性DNA突变等。自然选择不对这些变异起作用（不起明显作用）。例如蛋白质多态性、同工酶（isozyme）。

②中性突变不改变生活能力，通过随机婚配在群体中固定或者消失。

③分子进化的速率近乎恒定。对于每个蛋白质来说，每个位点的蛋白质替代的速率都是恒定的，称为分子进化钟（**molecular evolutionary clock**）。这个替代速率不受群体大小、选择、环境等影响。功能强的分子保守性强，突变率低。

1.3 物种分类和种系发生/系统发生（phylogeny）

生物进化的主要分类依据：古生物、考古、比较胚胎、比较解剖、进化、数值计算等。分类系统：界（kingdom）、门（phylum）、纲（class）、目（order）、科（*family*）、属（*genus*）、种（species）。

动物的命名：双名法（binominal nomenclature）。每一个动物有一个学名（science name），由两个拉丁文或者拉丁化的文字组成。前为属名，后为种名。有的加上亚种名，或者命名人的名字。

系统发生：生物可能有共同祖先。将现存的生物类群（作为末梢）与祖先按照亲缘关系连接成动物进化系统。形成进化树/系统发生树（phylogenetic tree）。

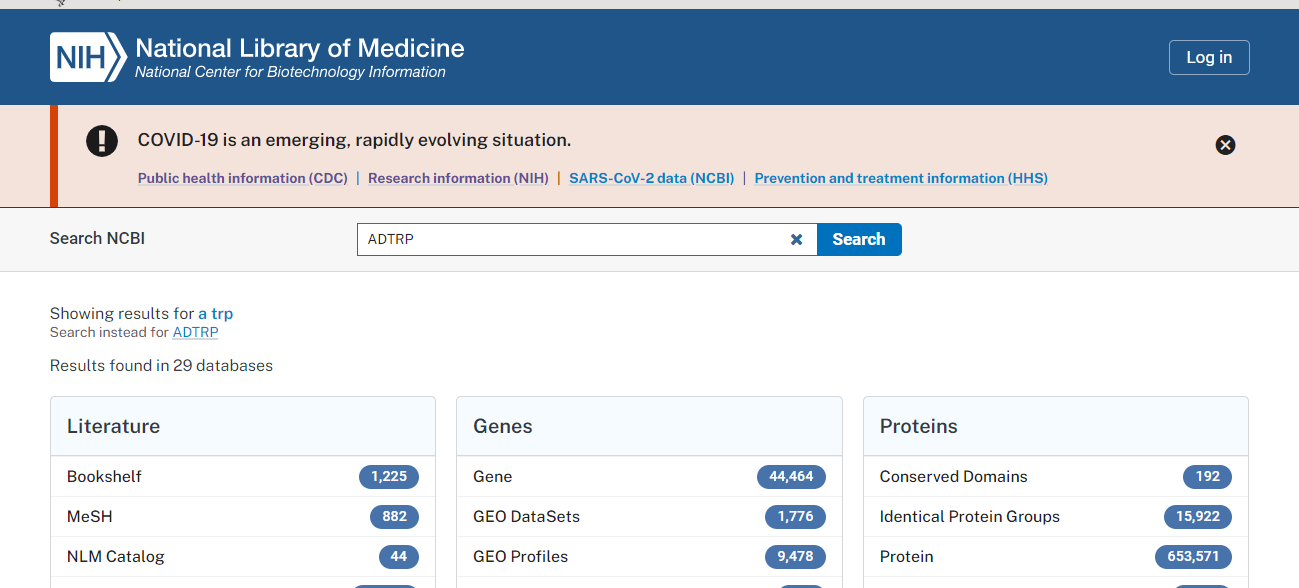
系统发生树：生物信息学的重要研究领域之一，对于得到的基因组，可以通过系统发育分析判断这个物种与其他物种的关系、共同祖先等。尤其是对于病毒（临床上常用）、细菌等造成大规模传染病、流行病的病原体的系统发育、变异方式等。

2.总论 - 构建系统发育树的基本步骤

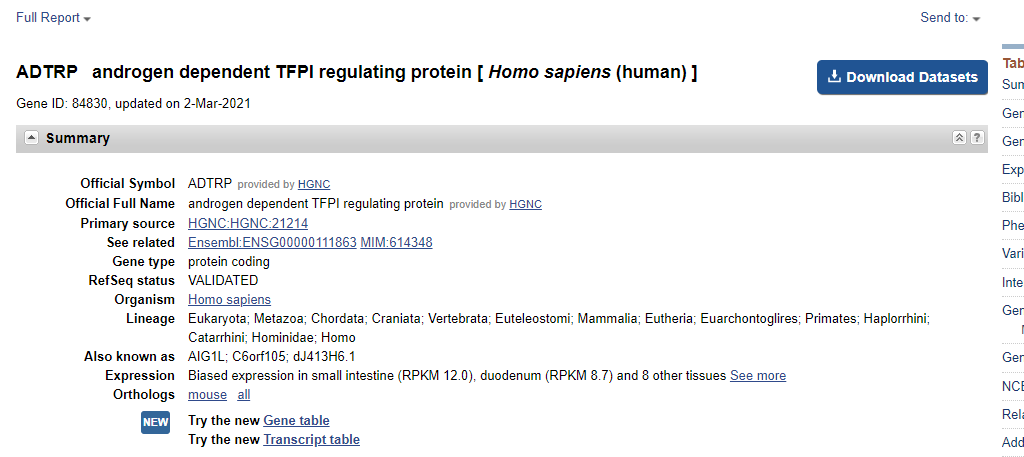
从头测序/在线数据库查询（例如NCBI，ENA等）→ 得到全基因组序列/某分子的序列 → BLAST搜索相近的序列，并根据相近程度筛选 → 得到数据表格，汇总得到全部需要进行比对的序列 → 序列比对 → 使用建树软件和特定算法得到系统发育树 → 进化树的解释和分析。例如研究发现xxx流感的基因组序列与曾经在xxx地曾经暴发的流感相像，说明可能是由于xxx地区的人带到xxx流感的，体现了病毒的亲缘关系等。

3.研究基因的选取和信息查阅

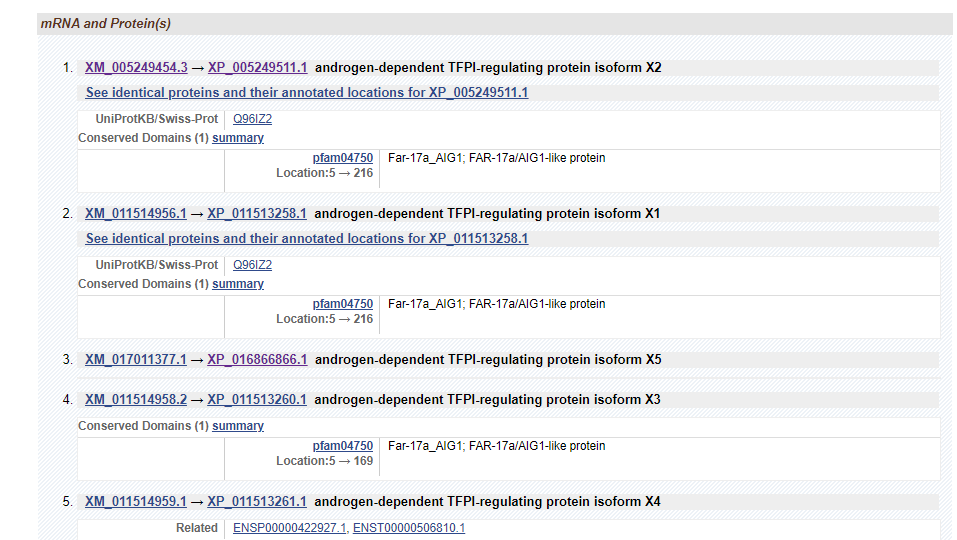
选定感兴趣的研究基因。本课程项目实践中选择ADTRP（Androgen-dependent TFPI-regulating protein，雄性激素依赖的TFPI调控蛋白，调节内皮细胞中抑制剂TFPI的表达和细胞相关的抗凝活性(体外)；TFPI：Tissue factor pathway inhibitor，组织因子途径抑制剂，一种抑制促凝早期反应的抗凝蛋白）。在NCBI和UniProt（https://www.uniprot.org/uniprot/Q96IZ2）上查阅资料了解其更多信息。



来到这样一个界面，存放了ADTRP基因相关的说明：

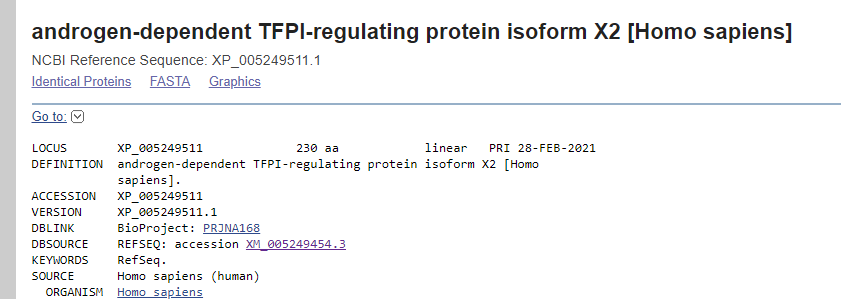


在下面找到其对应的mRNA和蛋白质



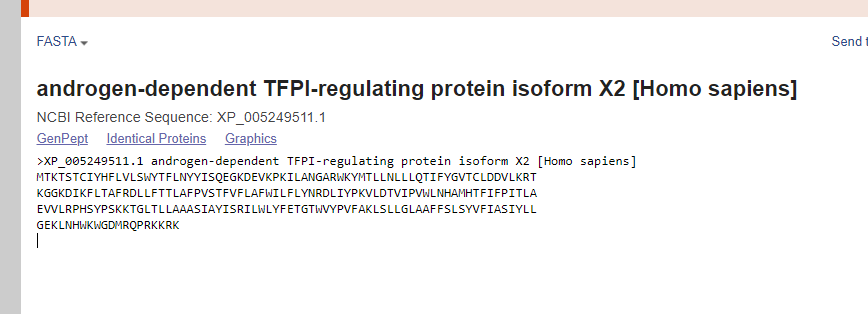
这里有很多条记录表示有不同的isoform或者不同的研究中测定的。

[XM\_005249454.3](https://www.ncbi.nlm.nih.gov/nuccore/XM_005249454.3) → [XP\_005249511.1](https://www.ncbi.nlm.nih.gov/protein/XP_005249511.1)  # M表示mRNA；P表示蛋白质。X/N表示这个数据（测序？）的质量。



4.序列比对、查询相近同源基因和数据整理。

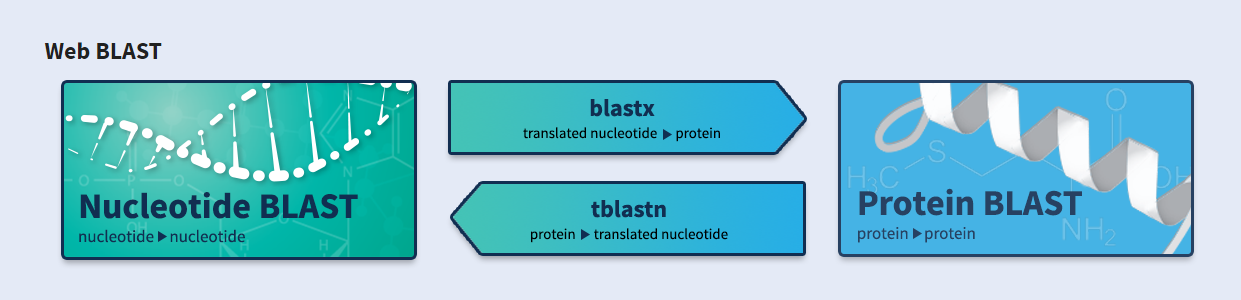
接上步。点击FASTA查看该基因的序列文件。

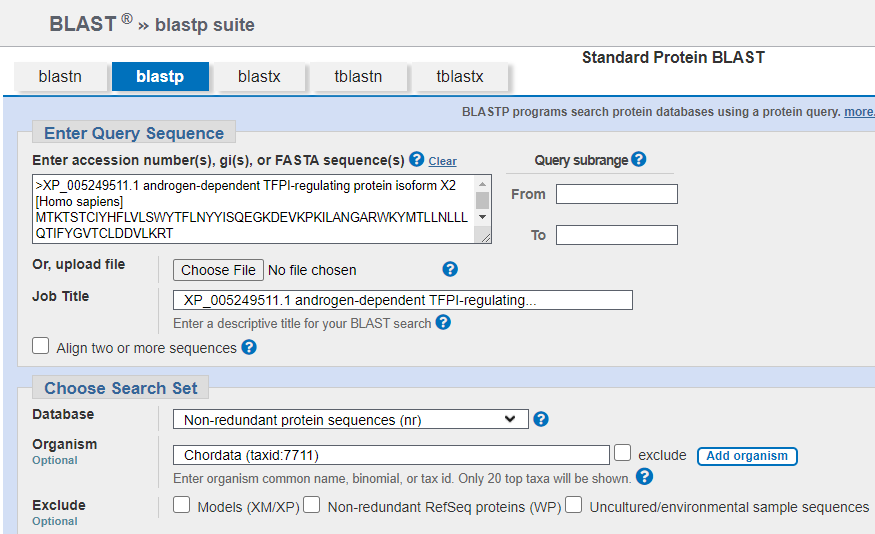


通过右上角send to 可以保存在某个位置。

根据这个ADTRP的fasta序列，以这个作为基准，使用BLAST寻找其他物种有相似的序列。

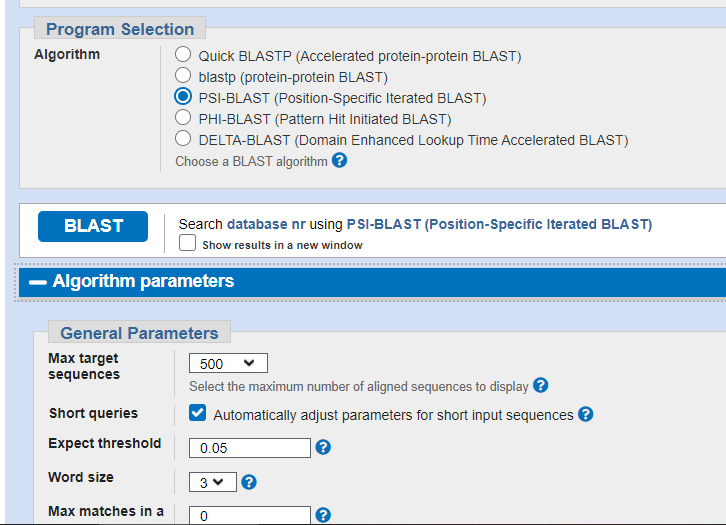
这里是进行蛋白质氨基酸序列的计算，所以选取BLASTp进行比对和查询。





将fasta文件粘贴到blastp中。选取的database，也就是查询的数据库，通常默认。

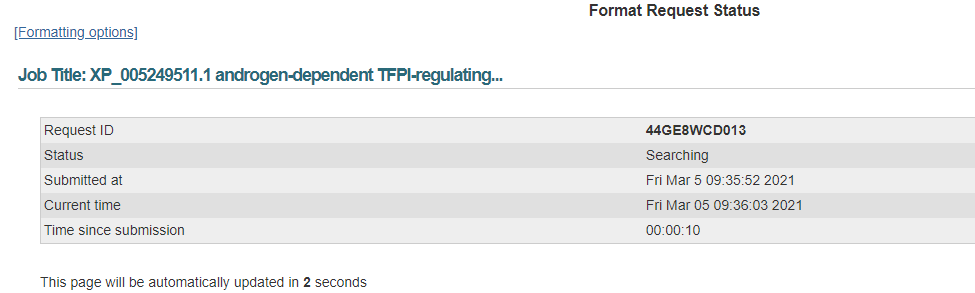
Organism：在指定的门类中查询。这里是在脊索动物门中查询输入后有相应的表单）



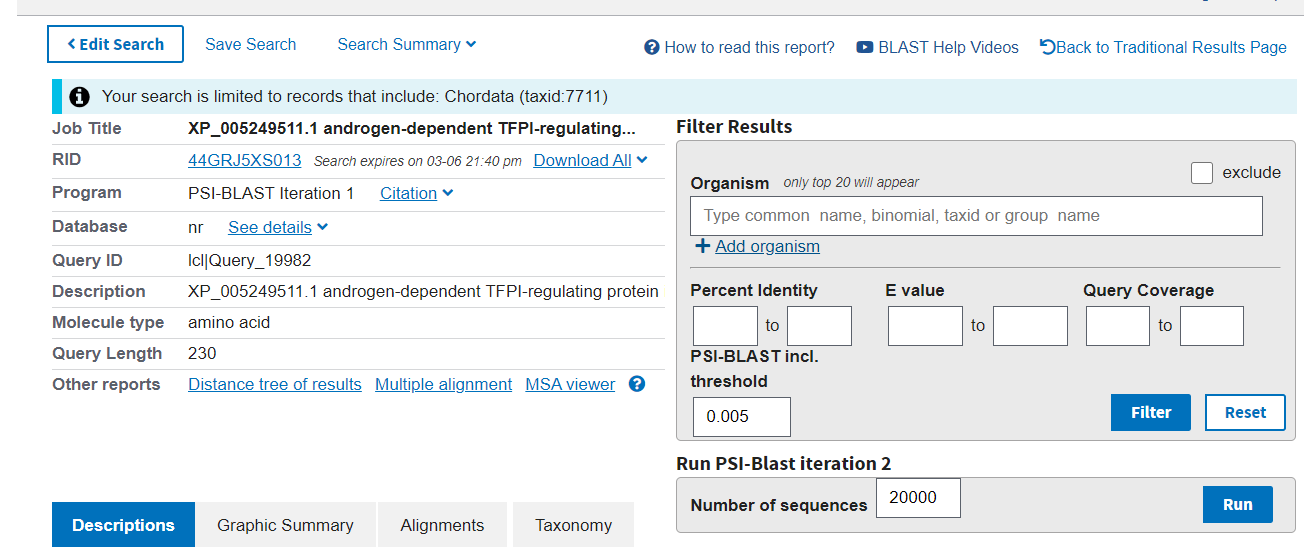
选取PSI-BLAST model，因为这个可以迭代（就是第一次查询后得到一堆结果，然后以这一堆结果作为query，再次查询，新出现的标注，我们需要不断迭代，直到不出现新的为止）。这里max target sequences选取20000，因为尽可能查询多个物种的这个序列。

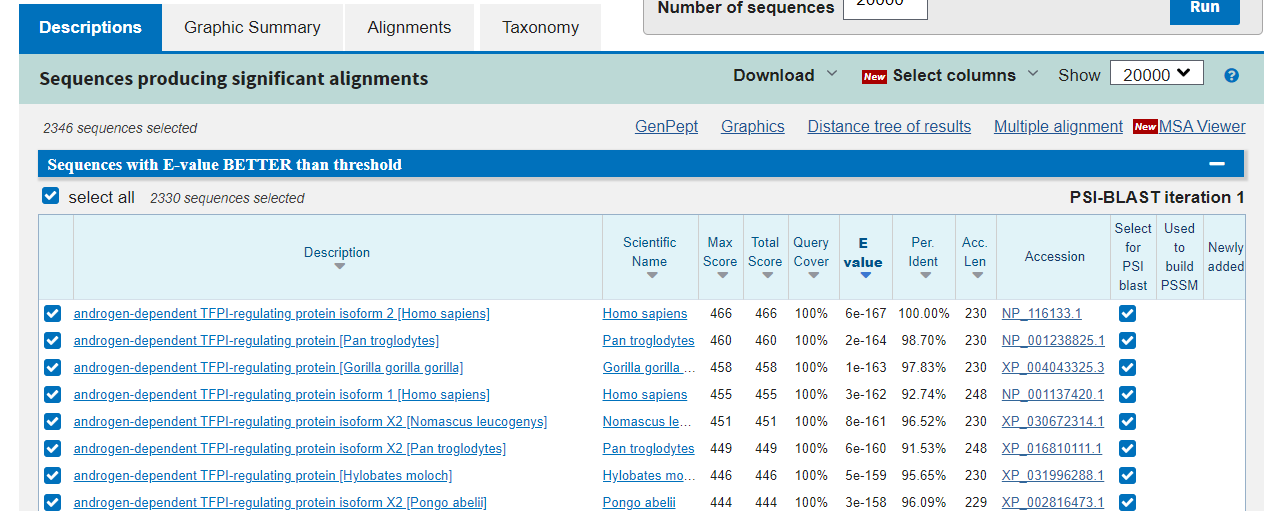
\*这里iteration的过程很容易崩盘。

根据原始序列查询中：

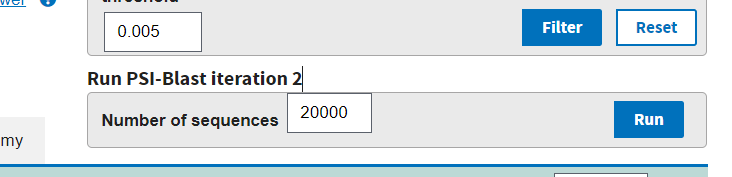


BLAST的结果：

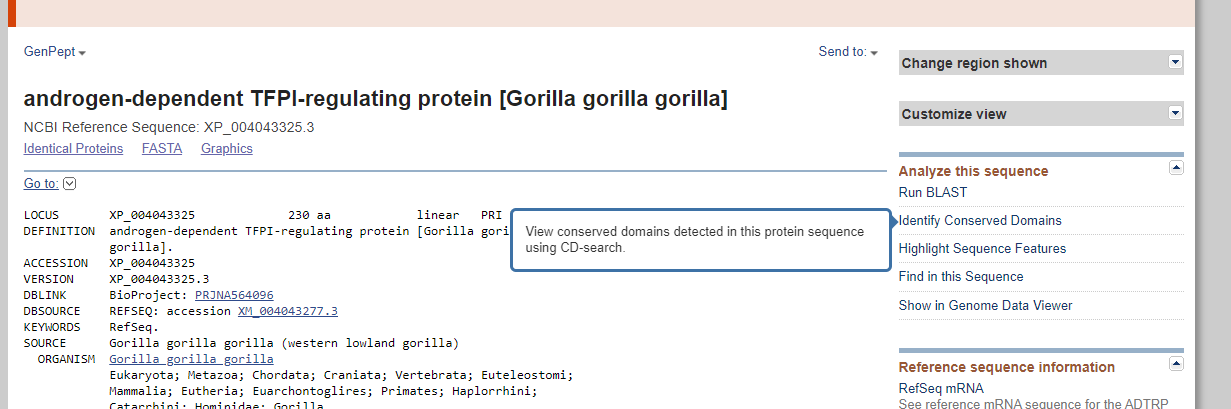


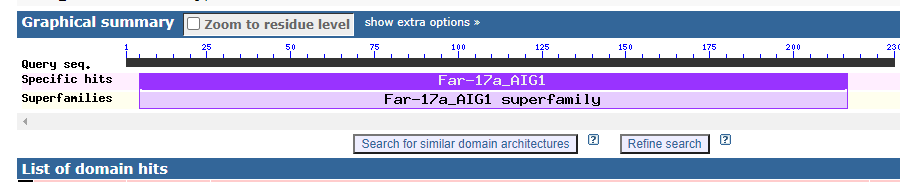


然后进行iterate第2、3、4、5......次。



然后需要点开每个蛋白，查看”identify conserved domains”，查看对应蛋白的结构域。

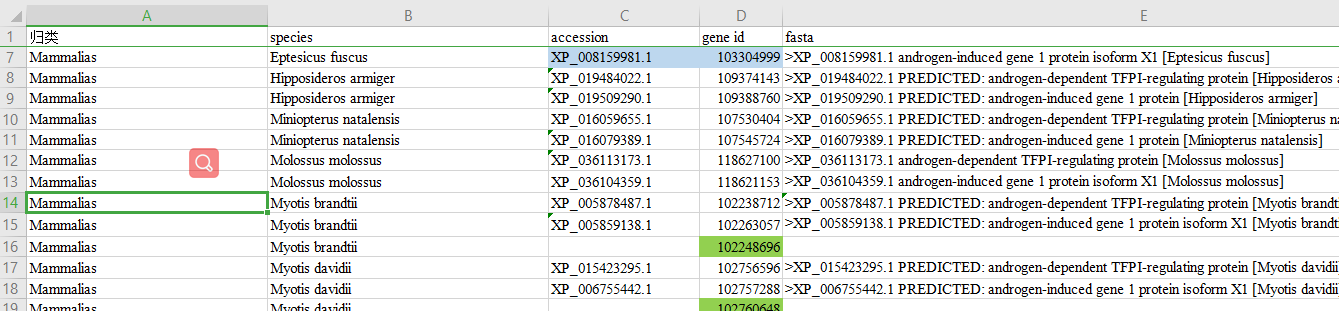


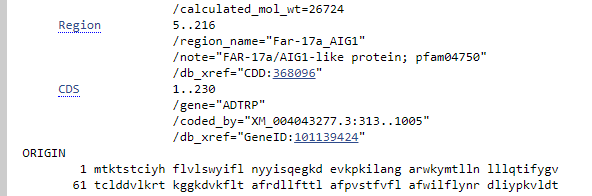


必须满足相应结构的蛋白，并且序列的E-value小于一定的值的数据才能保留，因为只有序列相似，并且使用蛋白质结构域确认过确实相似之后，才能认为这的确是相似的。

\*有的方法是使用Pfam工具检测候选的序列中是否存在xxxx结构域，从而进行确认。

抄写到excel表格中。顺便CDS项目中还有GeneID

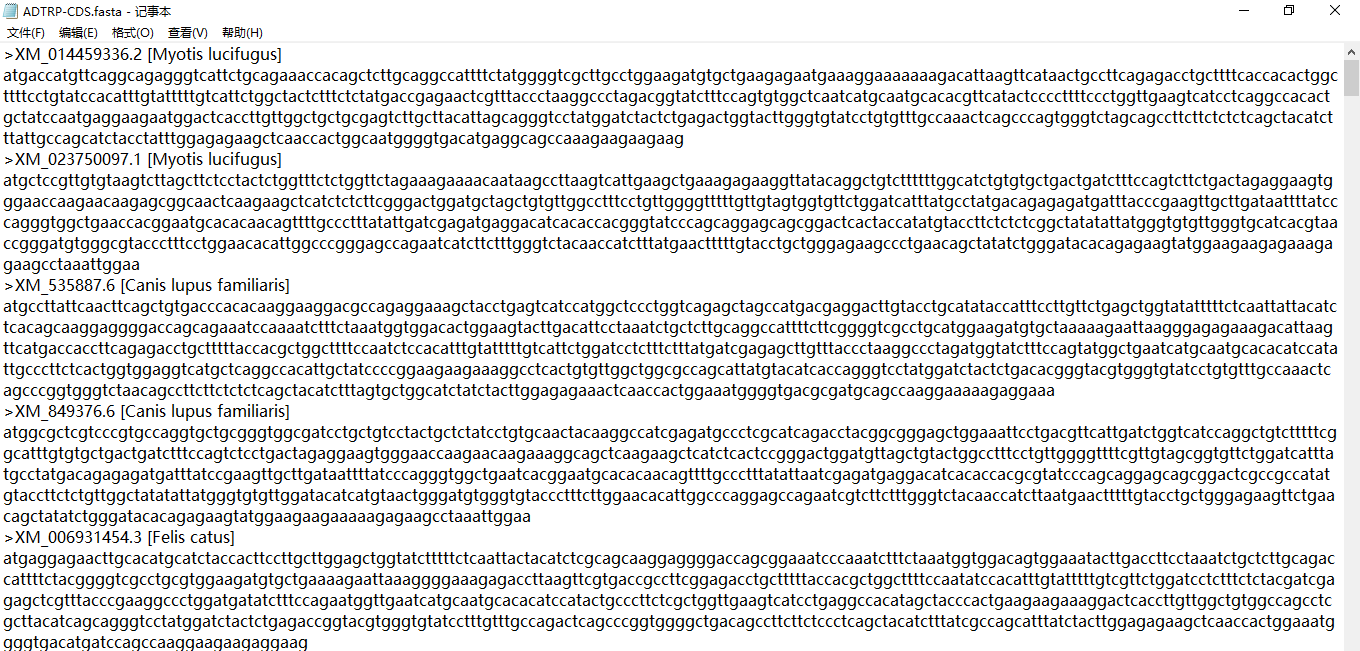




5.多序列比对(Sequence Alignment)

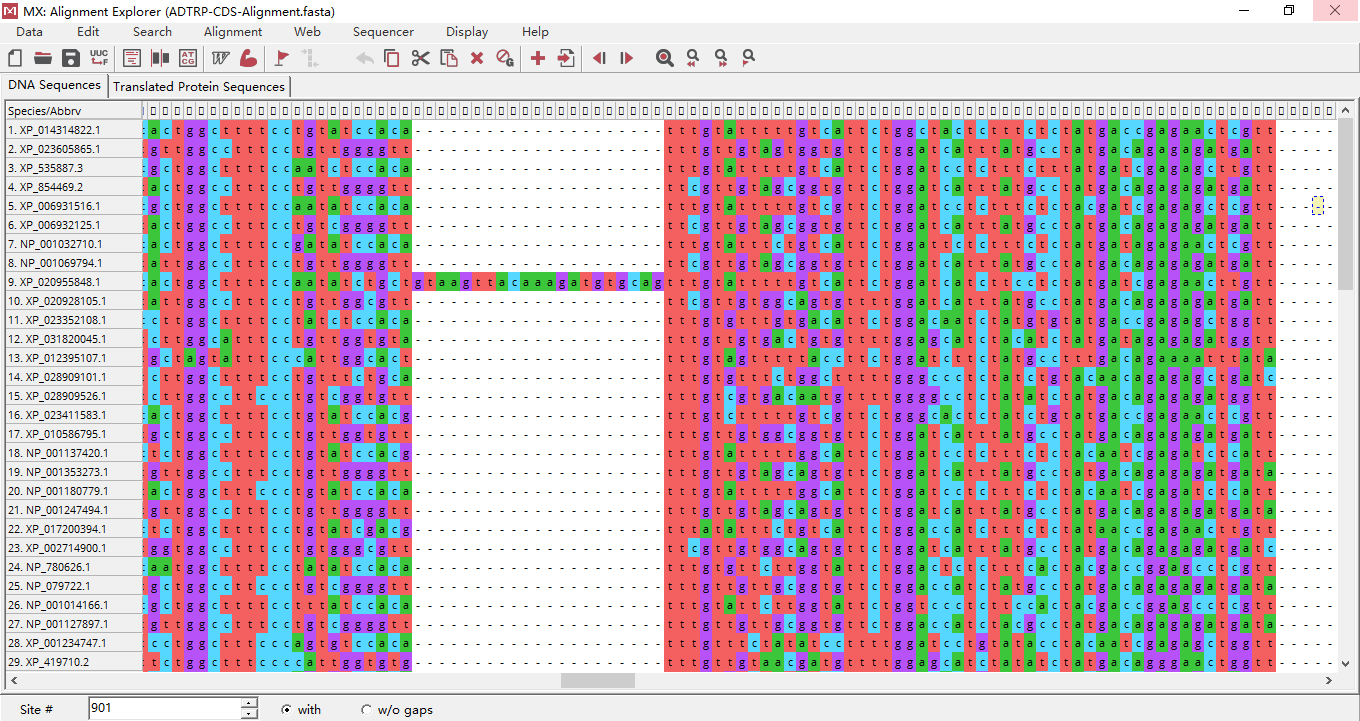
收集整理好的序列文件，将其转换成fasta格式文件，如图所示。

文件名：ADTRP-CDS.fasta



使用ClustalO进行多序列比对，比对结果文件：ADTRP-CDS-Alignment.fasta

比对结果在MEGA中的显示情况：



可见，-表示比对的gap，不同颜色表示不同的碱基。本次由于选择的物种跨越幅度较大，所以ADTRP分子序列差异性较大，故gaps较多。

6.导入到系统发生树软件建树

画树的算法：聚类分析(clustering analysis)，属于无监督的机器学习。没有事先给定的label，而是根据没有事先说明的特征，自动分成不同的类别，即将相似属性的对象聚到一起。

常用的建树方法：

基于距离的方法，例如UPGMA(unweighted pair group method using arithmetic average,非加权配对算术平均法或者说非加权组平均法) , NJ (Neighbor-joining，邻接法)

最大简约法：Maximum parsimony (MP)

最大似然法：Maximum likelihood (ML)

贝叶斯推断法：Bayesian Inference

\*从上到下计算速度越来也慢，但是精度越来越高。

本次使用iqtree的HKY和G4参数作为示例。

构建结果：见raxml-ng-HKY+G4-tree.support文件

\*注意：实际情况中建议根据数据情况选用不同的建树方法和模型。尤其是在分子进化分析中，建树完成后，可以将分子进化的系统发育树与物种的系统发育树作比较，选取比较接近物种发育树的结果。对于参数来说，没有最好，只有最适合。建议多修改参数和尝试，可以考虑使用组合优化、贝叶斯分类器等方法进行快速的选择和优化。

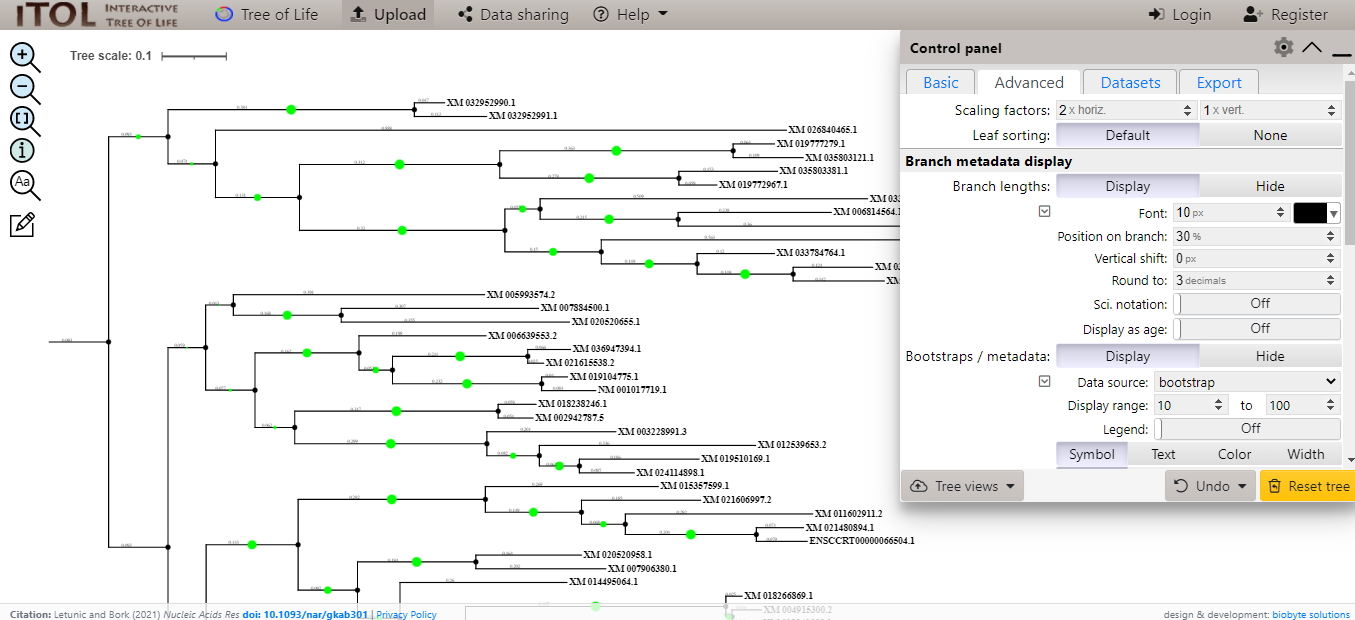
7.建树结果的可视化

得到建树文件，使用","分隔，每一个括号括起来表示发生树的一支。有时候括号后加上数字表示branch length或者bootstrap等，在建树结果可视化的时候可以根据实际的需要选择隐藏or显示。

常用的建树结果可视化工具：iTOL，TreeView，FigTree等。



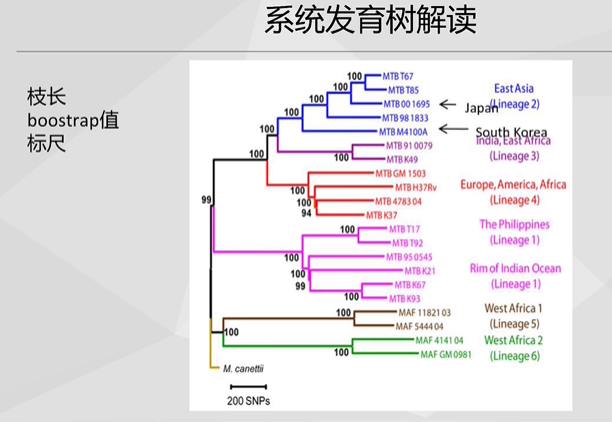
本次在iTOL中进行建树结果的可视化，文件见：



其中，在每一个branch上以数字形式标注了branch length，同时图的每一个branch长度都是按照branch length的比例作出的。绿色圆圈大小表示bootstrap值的大小，是结果可信度的衡量标准。还有其他选择，例如环状系统发生树、转换为进化时间、字体格式调整等。

8.建树结果的分析与解释

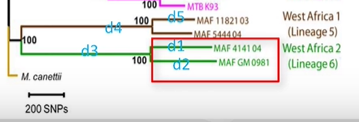
系统发生树显示了某一个基因/物种在生物进化中的情况，类似于人类的家谱。



系统发育树中的三个信息：branch length、bootstrap值（自展值）、标尺(scale)。

同一个支内的属于同一类。不同颜色的代表不同的类。同一类亲缘近，本质上就是序列相似度高。

①branch length代表相近程度。例如下面两个样本距离为d1+d2，第一个和第四个样本距离为d2+d3+d4+d5



2.bootstrap值：节点上的数字（本次进行了可视化）。用于检验进化树分支的可信度。简单来说就是将序列的位点重排，重排后以相同的方法画树。如果树的分支在重排后的树中也出现，就打1分。经过至少1000次重复后（即bootstrap值设置为1000），得到分值。换算成bootstrap值。

bootstrap值越小说明分支可信度低。通常认为>=75%的分支是可靠的。如果bootstrap值很小，可能说明是序列太相似或者差异过大，导致建树的计算不稳定。全基因组的bootstrap值通常高，稳定。但是单个基因的bootstrap值可能很低。

3.标尺/比例尺：



这里，一个单位距离为200个SNPs，即200个位置的氨基酸差异。有的时候是小数点，表示序列的差异程度。

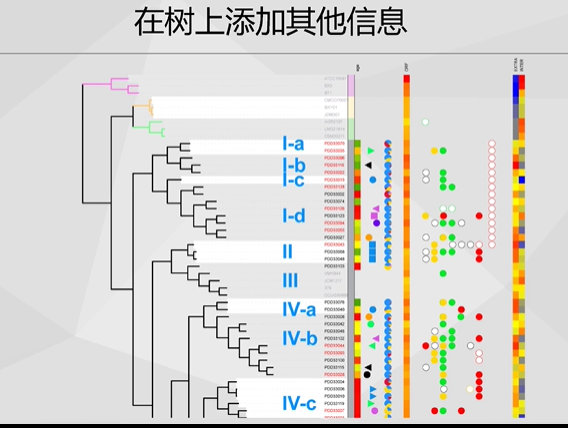


小数点数字表示遗传变异度为0.2。遗传变异度表示基因组中每个位点碱基的替换频率。遗传变异度=变异碱基树数/总碱基数（%）。衡量样本之间有多少差异。越长的差异越大。同时，根据分子进化的频率假设，可以大致认为这个代表了进化的时间。

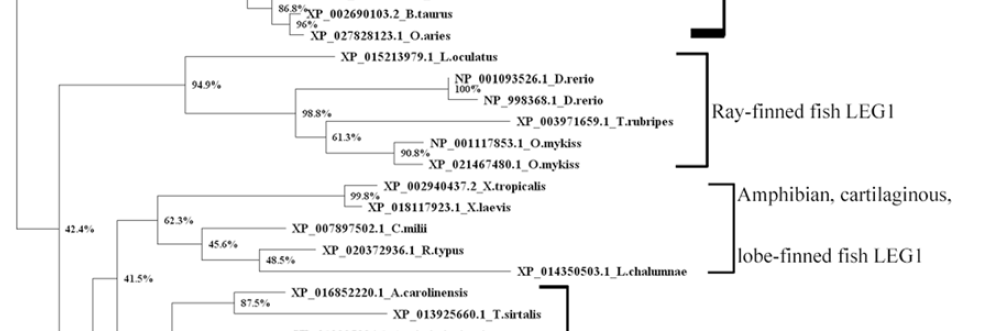
外节点：也叫叶节点，是末端的物种；

内节点：假定的祖先。

还有的树整合了其他的信息。



例如这里标注了对应支的物种的分类。



参考文献：

普通动物学

生物信息学教材

分子生物学

leg1a文献，以及其他一堆文献

bilibili基因学苑

How to read a phylogenetic tree

iTOL

MEGA

文献等