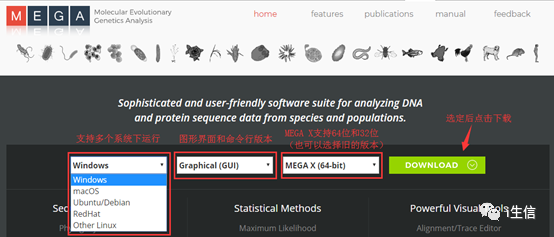
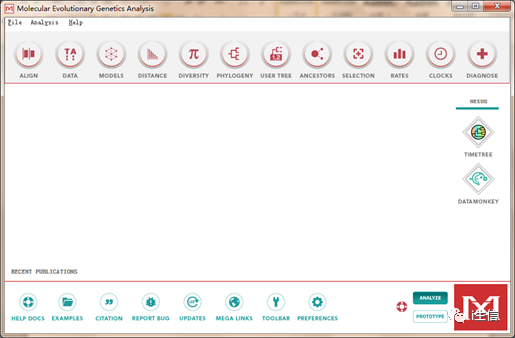
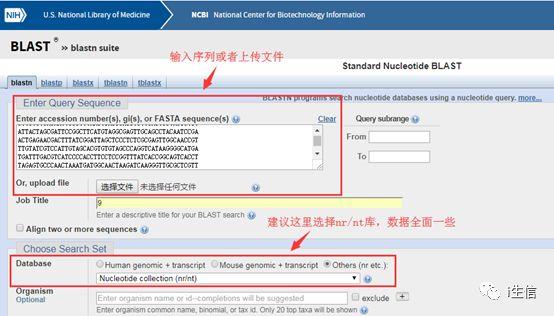
对于经常构建进化树的朋友来说，[MEGA](http://www.sci666.com.cn/tag/mega)应该是个老朋友了。MEGA从1993年的第一个版本问世一直锤炼到去年刚刚发布的MEGA-X，已经经历了26年，在这期间，MEGA共更新八个版本，先后在Molecular Biology and Evolution、Bioinformatics、Computer Applications in the Biosciences等期刊上发表共十篇论文，总引用量已经超过11万。对于如此熟悉的一个老朋友，让我们今天一起来了解一下它的新版本MEGA-X，开发它更多的使用方法。 MEGA-X的官网网址是<https://www.megasoftware.net/>，它支持在Windows、MacOS 以及Linux 系统下运行，有图形界面和命令行两个版本可供选择，支持64 位和32 位，与之前的版本比较，MEGA-X **最大的特点是大数据运算能力增强，并且支持多种计算平台**。



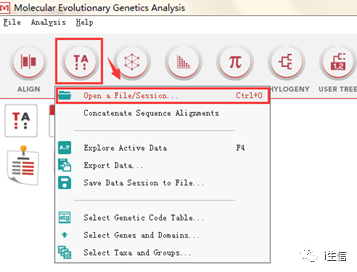
今天主要介绍的是在MEGA-X图形界面下构建系统发育树并且对发育树进行美化。下载安装好MEGA-X后，首先打开软件。



此处我们以一株细菌的16S rRNA序列为目标序列，首先在NCBI中进行Blast比对，下载将要一起比对建树的菌株序列。在NCBI中输入序列或者上传文件，选择数据库时可以选择「Nucleotide collection(nr/nt)」或者「16S ribosomal RNA sequences」数据库，一般来说nr/nt库信息比较全面。



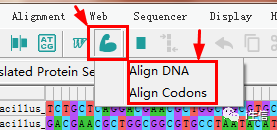
我们选择了10个不同种的16S rRNA序列进行下载。另外，此处还可以比对下载2-3条大肠杆菌（*Escherichia coli*）和沙门氏杆菌（*Salmonella*）的16S rRNA序列作为外类群（在Organism选项中进行物种限定），后面推断进化时间的时候可以用到。将所有下载的序列整理在一个文件中，为了方便后面的建树可以将菌株名称后面多余的信息在这里替换删除掉（只是名称上的信息，不要改动碱基序列），然后将文件的扩展名改为.fasta。在MEGA-X首页选择DATA，点击Open a File/Session，选择刚才的文件。



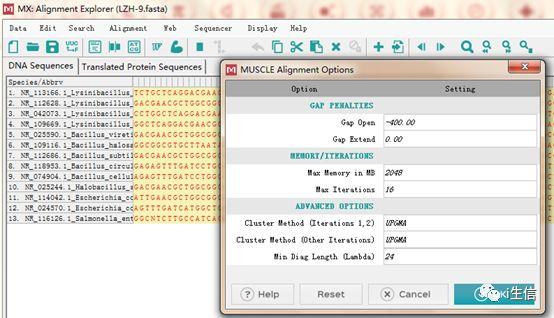
打开文件时询问「Analyze or Align File?」，此处点击Align。序列中可能会出现混合碱基符号，混合碱基符号指两种或多种碱基(核苷)混合物的表示符号，或未完全确定可能属于某两种或多种碱基(核苷)的符号：R表示A+G；Y表示C+T；M表示A+C；K表示G+T；S表示C+G；W表示A+T；H表示A+C+T；B表示C+G+T；V表示A+C+G；D表示A+G+T；N表示A+C+G+T。



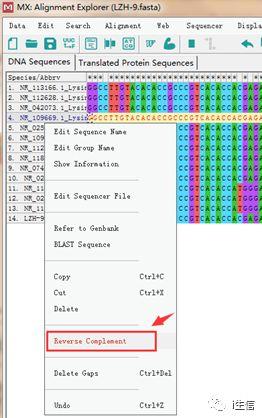
接下来选择序列比对的方法：Muscle或者ClustalW。ClustalW的基本原理是首先做序列的两两比对，根据该两两比对计算两两距离矩阵，是一种经典的比对方法，使用范围也比较广泛。Muscle的功能仅限于多序列比对，它的最大优势是速度，比ClustalW的速度快几个数量级，而且序列数越多速度的差别越大。方法可以通过点击图中上方Alignment或者下方的图标「W」和「Muscle」来选择。如果你的序列是DNA编码序列，就一定要选择Align Codons，因为序列通过密码子比对比DNA序列的比对会更加真实，避免间隙对比对结果产生的影响。MEGA可以比较方便快速地将密码子排列比对，后续作为输入文件在软件PAML或DATAMONKEY中进行进化压力的分析时就会比较方便。



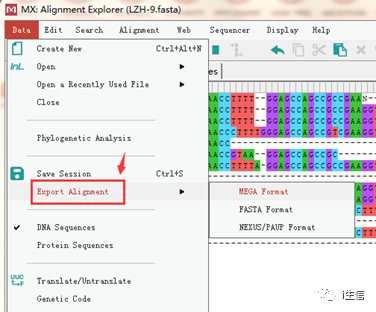
在这里我们选择Muscle进行序列排列，点击Align DNA，会出现一些参数选项，根据自己需要进行修改，在这里直接点击OK选择默认参数即可。



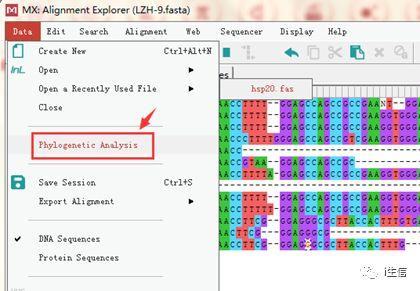
比对之后要去检查一下比对的情况，有的差异很大的或许是因为序列方向反了，这个时候要把它反转回来，右击这条序列，点击Reverse Complement，反转后一定要再次点击Muscle比对，检查是否大部分都对齐了。



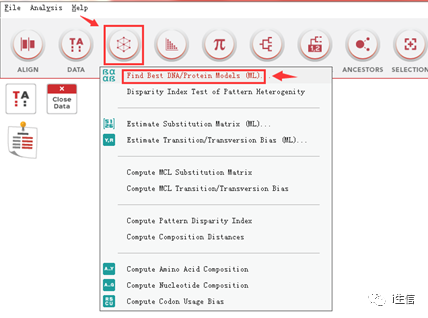
这里我们可以将最后对比后的文件导出，可以导出保存为MEGA格式。



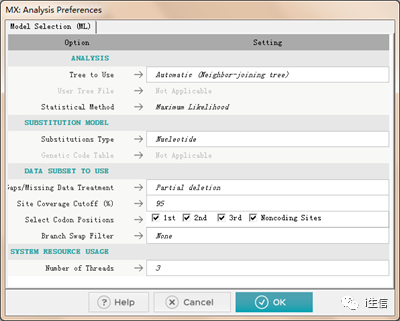
然后点击Data中的Phylogenetic Analysis直接进行系统发育分析。



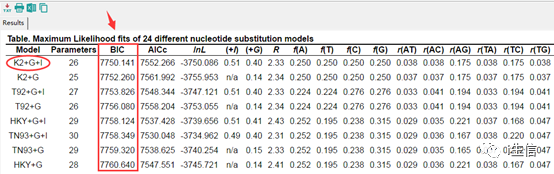
分析后返回主页面，接下来我们要选择一个最优的模型，提高建树的精确度。如果想要快速建树可以省去这一步，直接选择默认的模型。点击MODELS中的Find Best DNA/Protein Models(ML) 软件就会根据你的数据帮你计算寻找最适合的模型。



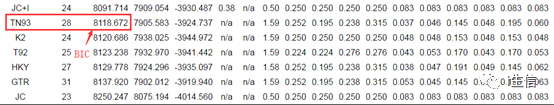
分析时选择默认参数，开始进行分析计算。



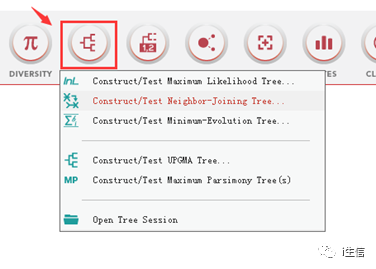
不久就会显示分析计算结果。具有最低BIC分数（BayesianInformation Criterion）的模型被认为是最好地描述替代模式。对于每个模型，还给出了AICc值（Akaike Information Criterion, corrected，值越低拟合程度越好），以及用来计算上述两个分值的最大似然值（lnL）和参数数量（包括分支长度）。在这里就可以看到，BIC分数最低的模型是K2+G+I，K2+G+I在这里就是最好的模型。



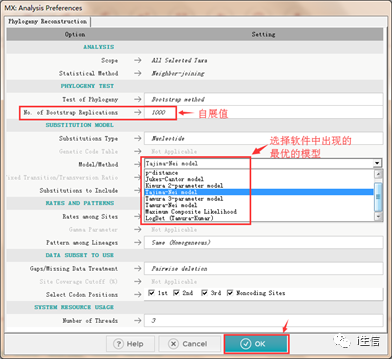
但因为实际在后面的模型选择中，软件有时没有提供组合的模型来选择，所以我们继续看下面的BIC分数，可以找到单个模型中得分最小的，就是我们在这里要选择的最优模型。看到这里的BIC值最低的单个模型是TN93（Tamura-Nei）。



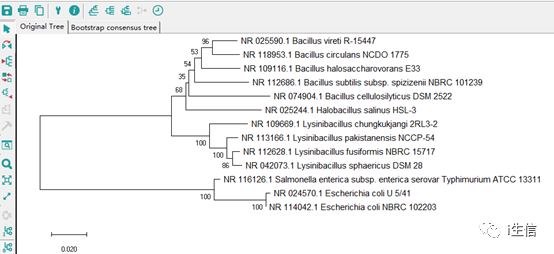
模型选好后，就可以点击PHYLOGENY进行方法的选择了。构建系统发育树有三种主要的建树方法，分别是距离法、最大节约法(maximumparsimony, MP)和最大似然法(maximum likelihood,ML)。最大似然法考察数据组中序列的多重比对结果，优化出拥有一定拓扑结构和树枝长度的进化树，这个进化树能够以最大的概率导致考察的多重比对结果；距离树考察数据组中所有序列的两两比对结果，通过序列两两之间的差异决定进化树的拓扑结构和树枝长度，基于距离的方法有UPGMA、ME（Minimum Evolution，最小进化法）和NJ（Neighbor-Joining，邻接法）等；最大节约法考察数据组中序列的多重比对结果，优化出的进化树能够利用最少的离散步骤去解释多重比对中的碱基差异。在这些方法中，如果模型合适，ML的效果较好。对于近缘序列，有人喜欢MP，因为用到的假设最少，远缘序列上一般用NJ或者ML，这两个方法都是需要选择模型的。对于相似性很低的序列，NJ往往出现Long-branch attraction（LBA，长枝吸引现象），这种现象有时候会严重干扰进化树的构建。其实当序列的相似性比较高时，各种方法都会得到不错的结果，模型之间的差别也不是很大。所以平时我们一般推荐用两种不同的方法进行建树，如果得到的进化树类似，则结果较为可靠。这里我们先选择Neighbor-Joining法建树。



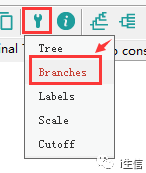
点击方法后，选择自展值，即重复建树以进行检验的次数，一般选择1000以上才比较可靠。模型的话选择之前计算好的最优的模型，点击OK。



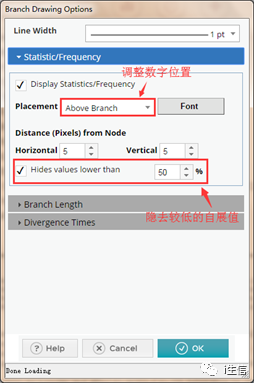
稍等片刻，就初步构建好了这个树，接下来我们对这棵树进行一些调整和美化。首先我们看到每棵树的前面都标有自展值，这主要是对进化树进行评估的一个百分比值。因为进化树的构建是一个统计学问题，我们所构建出来的进化树只是对真实的进化关系的评估或者模拟。如果我们采用了一个适当的方法，那么所构建的进化树就会接近真实的「进化树」。这里的数值表示我们将该树重复构建1000（之前设置的数值）次，得到相同结果的次数占重复次数的百分比值。一般Bootstrap的值>70%，则认为构建的进化树非常可靠，50%-70%认为基本可靠，小于50%认为不可靠（不同的人对于这个阈值有不同的划分）。如果Bootstrap值太低，则有可能进化树的拓扑结构有错误。当Bootstrap的值小于50%时，我们一般选择隐去。



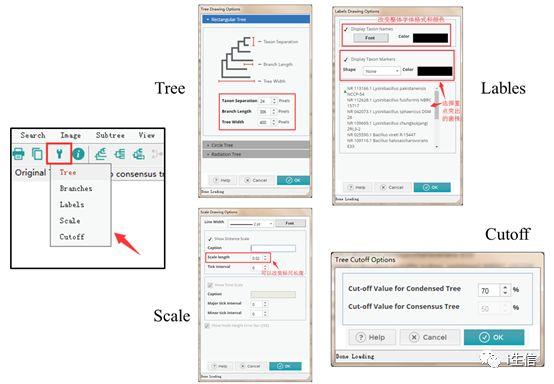
点击左上的工具图标，隐去较低的自展值可以点击此处的Branches。



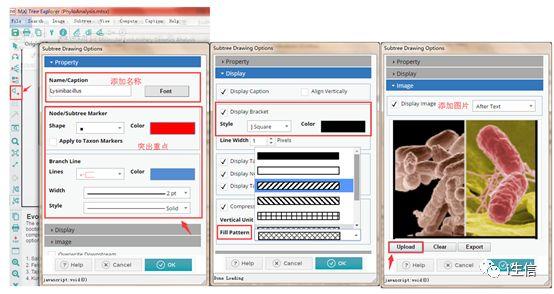
设置隐去50以下的数值。



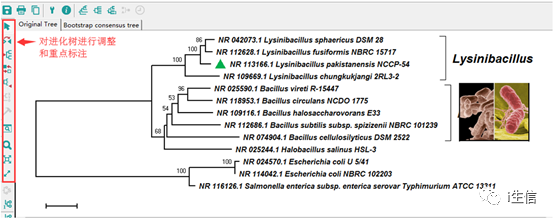
这个工具栏里的其他选项还可以对字体和线条粗线长短等进行修改，并且将想要突出的菌株和分枝信息进行重点标注。



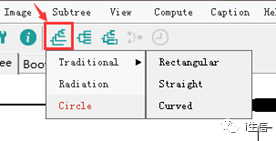
左边的工具栏里还可以对进化树进行修改，比如改变树根，调整分枝的上下位置，以及重点标记突出某个分枝信息。



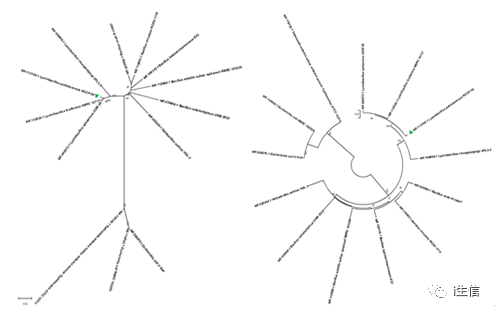
最后美化结果。



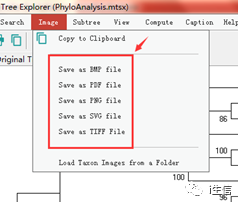
在这里还可以点击树形图标将进化树修改成为圆圈型和松针型。



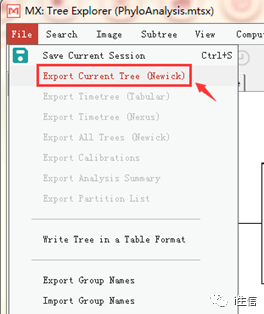
结果图展示如下。



为了数据更加可靠，大家可以再选择另外一种方法进行建树，和上一种方法构建出的进化树进行比对。然后最后导出储存图片，这里提供EMP、SVG （矢量图，可进一步通过绘图软件如AI编辑）、PDF、PNG和TIFF （位图）几种格式来导出。



记得将进化树的文本文件也保存一下。点击File中的Export Current Tree(Newick)保存Newick格式。



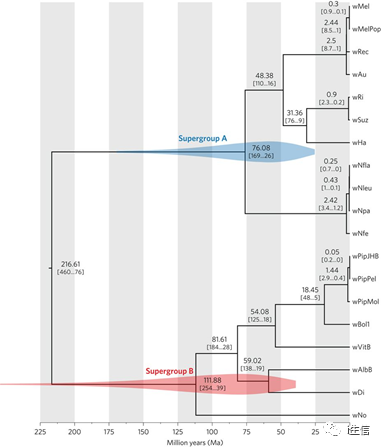
 保存后的树文件格式如下，导出的树文件便于之后在iTOL、Evolview、Figtree等工具中进行更进一步的美化，比如添加分类颜色、标记以及热图和条形图等。



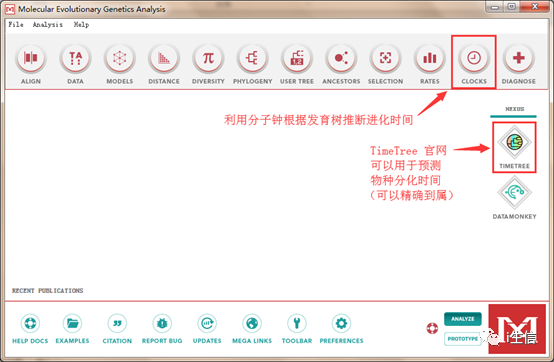
MEGA的使用非常方便，作为一个历史悠久并且极具创新精神的软件，它一直在提高自己的计算速度，丰富自己的计算平台。今天主要介绍了用图形界面MEGA-X建树的操作步骤，下次我们来介绍下根据MEGA-X构建的进化树来推断物种进化时间以及链接到DATAMOKEY等软件去计算进化压力等，期待一下吧。

参考文献：Kumar S, Stecher G, LiM, et al. MEGA-X: Molecular Evolutionary Genetics Analysis across computingplatforms[J]. Molecular Biology & Evolution, 2018.

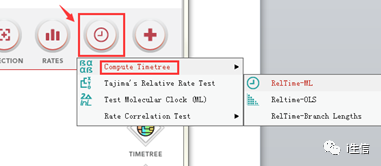
除了构建进化树之外，[MEGA](http://www.sci666.com.cn/tag/mega)-X还有很多其他功能值得探索，今天要介绍的是用MEGA-X中的「CLOCKS」来推断物种分化时间。 在文献中，我们有时会看到下面这种进化树，上面标注了每个分支的进化时间，可以帮助作者探究环境对生物的影响以及宿主与寄生微生物的分化一致性等。那这些进化时间都是怎么推断的呢？在MEGA-X中我们可以通过「分子钟」技术来追溯生物进化的历程，这种技术通过直接对比不同物种间基因的碱基排列顺序来推断进化时间，揭示生物进化历史。分子钟有时也被称为「进化钟」或「基因钟」，由于地球上的所有生命都会从上一代体内继承DNA 遗传基因，且这种遗传基因会随着时间的变化而逐渐改变（即按一定的速率发生碱基突变），因此运用分子钟技术便可以将生物进化的起始时间大致地预测出来。



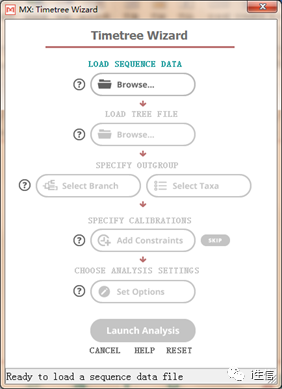
（图片引自文献Gerth M *et al*. . **Nature Microbiology**, 2016） 接下来我们来试一下用MEGA-X的RelTime方法来推断一下进化时间，首先在推断前需要准备的文件有以下几个：可用于推断分化时间的系统发育树（注意这里的发育树上需要有外类群以及校准点）的序列文件以及Newick格式的树文件。 在分子钟推断进化时间的过程中，最重要的找到至少一个校准点（即已确定分化时间的发育树节点），只有校准点准确，其他推断信息才会有意义并且可靠。校准信息主要依赖于化石记录，但也有一些其他来源，比如年龄地质事件，因此一定要对文献进行详细评估得出校准信息。听起来比较复杂，但是有一个很神奇的网站可以帮助我们——TimeTree，这是一个通过生命树（the tree of life）不同分支的特异性分子钟时间推算来预测不同物种分化时间的数据库，可以通过MEGA-X直接访问（右上角）。除此之外，还有一些非常有用的网络资源可以提供化石信息，例如化石校准数据库（https://fossilcalibrations.org/）和古生物学数据库（https://paleobiodb.org/）。无论使用哪种方式寻找校准点，在进行分子钟分析之前都要确保校准信息是准确的，获得了系统发育和校准边界，就可以用MEGA-X中的RelTime来推断进化时间了。 打开MEGA-X软件，点击右上的钟表图标「CLOCKS」。



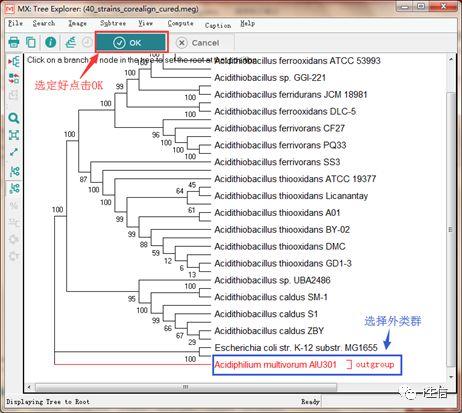
点击CLOCKS后点击Compute Timetree。可以选择下面三个选项：RelTime-ML，RelTime-OLS和RelTime-Branch Length。RelTime-ML是最传统常用的方法，大多数情况选这个就可以。



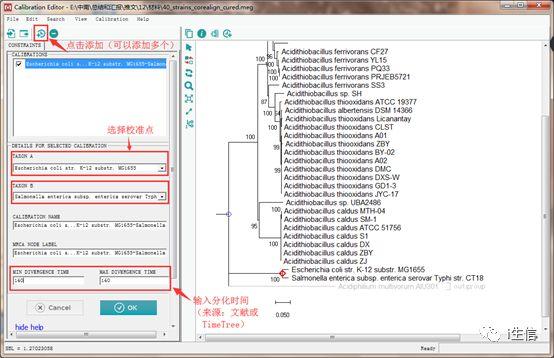
打开后，展示以下界面，需要导入的文件有序列信息和树文件，然后选择外类群（outgroup）和校准点。在这里我们导入了.meg格式的数据文件和.nwk格式的树文件。



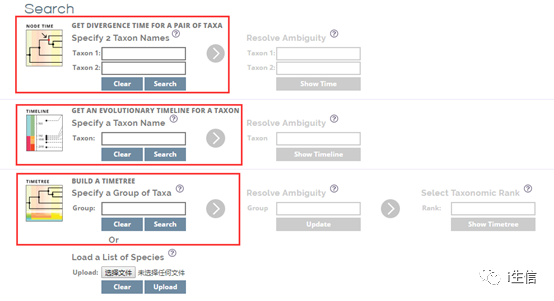
导入文件后，选择外类群，所谓外类群是指为了探知内类群（指定被研究的对象）的演化关系而借助比较的外部类群，它与内类群在演化关系上是最为接近的类群，具有最相近的祖先，且在进化程度上低于内类群。



下一步添加校准点。点击Add Constraints后，再点击左上角的「添加时钟」的图标，就可以进行选择，然后输入校准点分化时间的最大值和最小值，在这里分化时间主要可以依据文献中的数据或者从TimeTree中进行推断（TimeTree 仅仅能推断到属）。此处输入第一个校准点，大肠杆菌和沙门氏菌，它们的分化时间是140MYA（百万年前）（U. Bergthorsson*et al*..**Molecular Biology & Evolution**, 1998）。



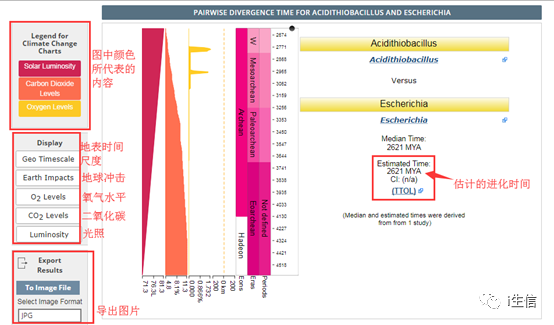
在这里我们再添加一个校准点，这个校准点我们用TimeTree来搜索进化时间（只能精确到属水平），可以点击右侧TimeTree图标进入官网，也可以通过网址http://www.timetree.org/直接访问，搜索选项中有以下三个选择：NODE TIME（输入两个物种，推断它们之间分化的时间节点）、TIMELINE（输入一个物种，推断它在自然界中出现的时间）、TIMETREE（输入大于属水平的一类物种如芽孢杆菌科，推断其下级分类学单位的物种之间分化的时间节点）。



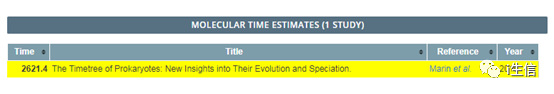
在这里我们选择NODE TIME进行搜索。输入两个物种的名称，在这里我们输入*Escherichia*（埃希氏菌属）和*Acidithiobacillus*（嗜酸硫杆菌属）进行搜索*。*



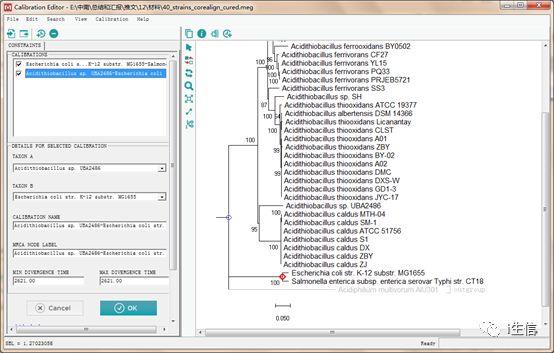
结果显示如下，左边图中不同的颜色表示不同的地球条件，比如大气中氧气、二氧化碳、光照条件的变化等。中间的框中可以选择图片对哪些信息进行展示，哪些进行隐藏，最后调整好之后可以选择导出图片，格式可以选择JPG、PDF、SVG、PNG几种。右侧显示了TimeTree推断的进化时间，也就是我们输入校准点信息时要用到的时间信息，在这里显示为2621MYA。



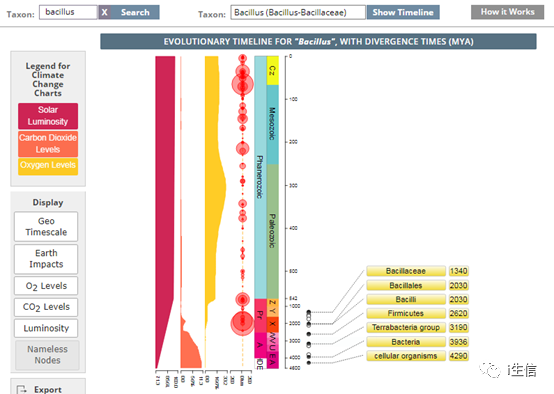
在左图中右上方有一个黑色的实心圆圈，点击这个圆圈，在图片的下方就会显示引用的文献。



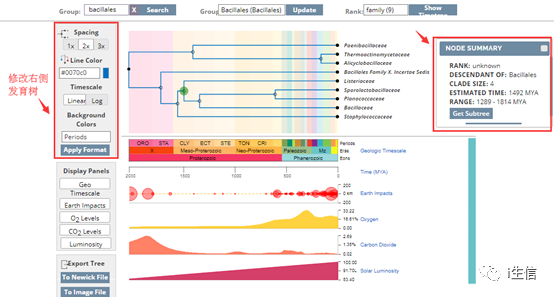
然后我们根据TimeTree的推断的校准时间来添加第二个校准点。



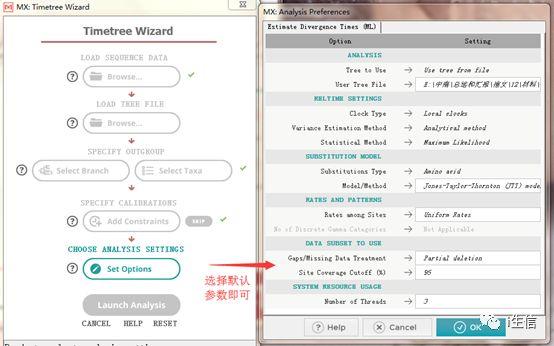
在进行下一步前，我们再演示一下TimeTree中其他两种方法TIMELINE、TIMETREE的操作。TIMELINE中要求输入一个物种分类名称，可以帮你推断一下该物种的分化时间，比如在此处我们输入*Bacillus*，搜索后显示结果如下图，在这里分化时间只能定位到科水平。



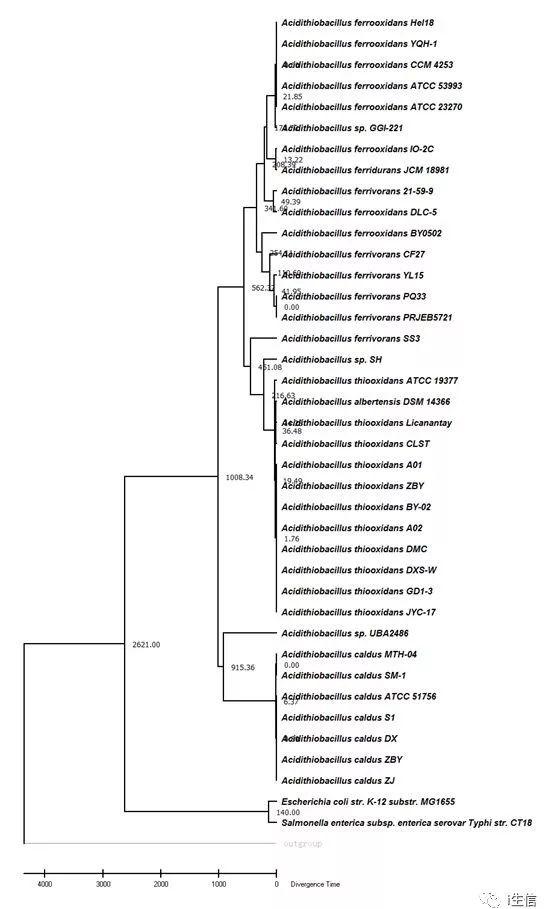
TIMETREE根据输入的物种名称，构建该分类学单位以下物种的系统发育树，点击每个圆圈都会有更详细的信息展开，实心的圆直接展示NCBI中分类的节点，空心圆是无名的未标记的分类节点。左侧的工具栏中可以对进化树和图像进行修改。并且可以导出图片和树文件。



TimeTree的简单使用介绍到这里，更加详细的内容可以去参考TimeTree的文献。可以看到，TimeTree在线软件只能预测到数据库中包含的物种的分化时间，而一些不在数据库里的物种的分化时间我们可以借助MEGA-X来完成推断。回到MEGA-X中，点击OK，进入下一步，点击Set Options，选择默认参数即可。



稍等片刻，就会出现结果。每个分支上都有通过分子钟推断的进化时间。让我们再给它伸展下枝条，对枝条的长短、粗细以及字体的格式进行简单的修改美化，最后这棵利用分子钟推断了进化时间的系统发育树就建好了。进一步的美化可以将SVG格式图片导入到AI中进行操作。



通过以上对MEGA-X以及一些其他网站工具的介绍，相信大家对物种进化时间的推断有了简单了解，还有更多工具和方法值得我们探索，MEGA-X使用的介绍就到这里，以后有机会再介绍下链接到DATAMOKEY 等网站计算进化压力以及运行命令行版MEGA-X来构建进化树。

参考文献：1. Sudhir Kumar, Glen Stecher, Michael Li, Christina Knyaz, Koichiro Tamura, MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms, *Molecular Biology and Evolution*, Volume 35, Issue 6, June 2018, Pages 1547–1549, https://doi.org/10.1093/molbev/msy0962. Sudhir Kumar, Glen Stecher, Michael Suleski, S. Blair Hedges, TimeTree: A Resource for Timelines, Timetrees, and Divergence Times, *Molecular Biology and Evolution*, Volume 34, Issue 7, July 2017, Pages 1812–1819, https://doi.org/10.1093/molbev/msx116  
3. Beatriz Mello,Estimating TimeTrees with MEGA and the TimeTree Resource, *Molecular Biology and Evolution*, Volume 35, Issue 9, September 2018, Pages 2334–2342, https://doi.org/10.1093/molbev/msy133  
4. U Bergthorsson, H Ochman, Distribution ofchromosome length variation in natural isolates of Escherichia coli.Molecular Biology and Evolution, Volume15, Issue 1, 1 January 1998, Pages 6–16, https://doi.org/10.1093/oxfordjournals.molbev.a025847