准确度模型

1. 准确度模型准备：
   1. 模型概述：

此模型用于评估模型1中所得出的最终优化模型。我们使用专业基因序列距离测算软件MEGA进行距离测算，并认为所得出的距离DMEGA是序列之间的真实距离。我们认为，准确性可以用预测值与理想值的差距分量进行评估。差距越大，准确度越低。

* 1. 变异率定义：

我们定义：变异率

1. 模型概述：
   1. 准确性评估模型：

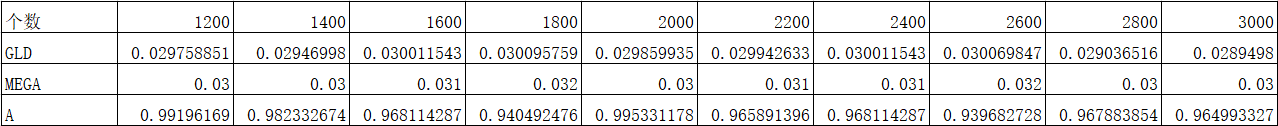
与求最佳alpha的值的过程相似，我们对任意足够长的序列M，N（M，N的长度大于1001），。使用MEGA算出M，N之间的距离DMEGA(MN)，作为理想值。使用模型1进行计算得出GLDMN。我们定义准确性A为：

这里所得到的结果在[0,1]区间内。结果越大，所得结果的准确度越高。

* 1. 评估过程：
     1. 变异率一定，改变长度：

随机生成长度为2000碱基的序列M2000，设定变异率为5%（即修改20次，删除40次，增加40个），得到新序列N2000，计算准确性A。实验i次，得到

同理求出A1200，A1400……A3000，并绘制图像



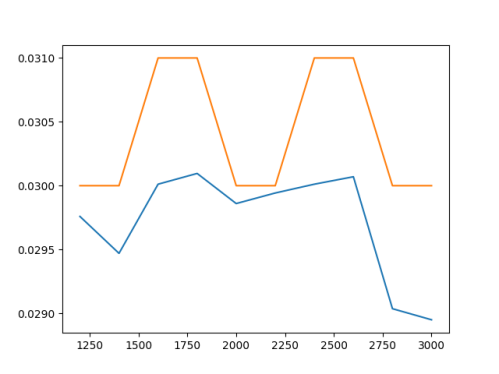
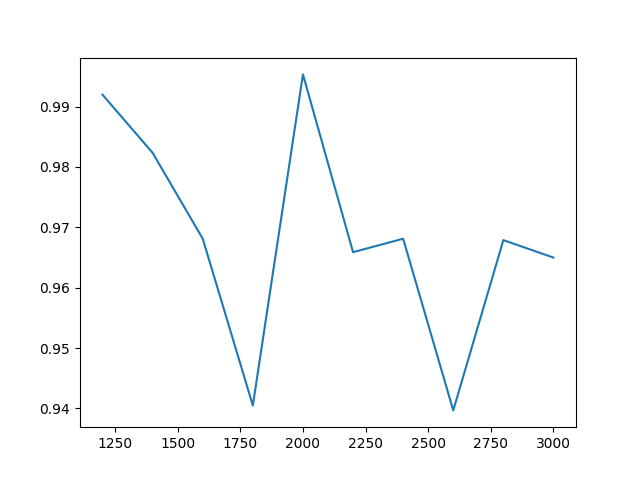


图1 准确性随个数变化图像 图2 预测值与理想值对比

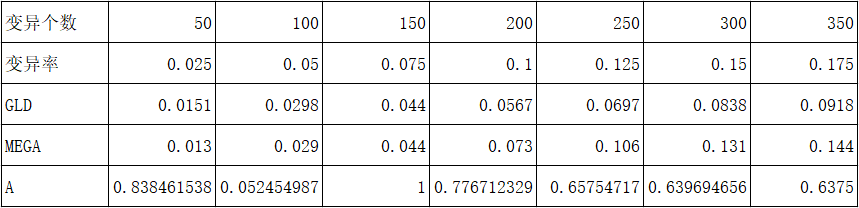
由以上分析可以得到，当变异率不变且较小，序列长度相对较短时：

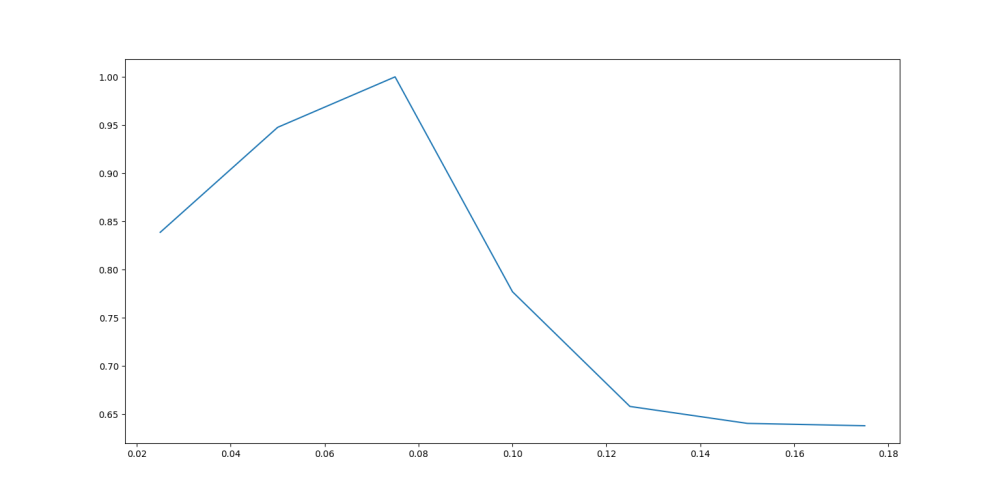
准确性随长度变化会有波动，但准确性稳定在95%左右，预测值较实际值相对较小。故在变异率较小时，模型准确性较高。

* + 1. 长度固定，变异率改变

随机生成长度为2000碱基的序列M2000，设定变异率为5%（即修改40次，删除30次，增加30个），得到新序列N2000，计算准确性A。实验i次，得到

同理求出A2.5%，A7.5%……A17.5%，并绘制图像





分析上图表可以得到，当序列长度不变，变异率变化时：

准确性会有显著变化，具体趋势为先增加后减小，在变异率为7.5%左右的时候达到顶峰。之后会迅速下降，在变异率大于12.5%后趋于平稳。准确率不会低于60%。图标中数据与2.2.1中相符，可信度较高。

* + 1. 具体举例

使用此模型对新型冠状病毒COVID-19两种毒株alpha与beta序列（长度为30000+碱基）进行距离分析，发现距离预测值约为0.01556，理想值0.018，结果为86.4%。这一结果符合上述两准确性分析所得结论。

* + 1. 总结：

由以上准确性分析可以得出：此模型的准确性受序列长度的影响较小，而受变异率影响较大。这说明，此模型适用于生物学距离较近的生物种群的序列。模型对同家族序列，同源序列距离的测定准确度高。