**Accuracy Model**

1. **Preparation：**
   1. **Overview**

This model is used to evaluate the final optimization model obtained in Model 1. We used the professional gene sequence distance measurement software MEGA to calculate the distance, and believed that the distance DMEGA obtained was the real distance between the sequences. We believe that accuracy can be evaluated by the component of the difference between the predicted and ideal values. The wider the gap is, the less accurate the model is.

* 1. **Definition of mutation Rate**

We defined the mutation rate c as the percentage of the number of modifications in the total length of the original sequence.

1. **Description:**
   1. **Accuracy evaluation model**

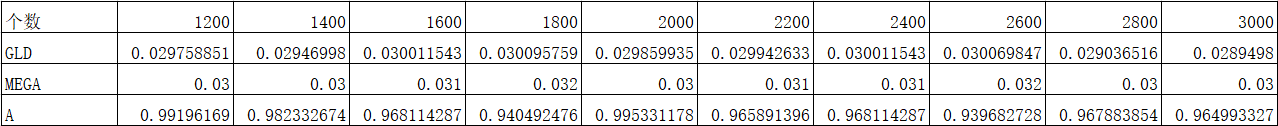
Similar to the procedure for finding the best alpha value, we take any sufficiently long sequence M, N (the length of M,N respectively is greater than 1001),. Use MEGA to calculate the distance between M and N as DMEGA(MN), the ideal value. GLDMN was calculated using Model 1. We define accuracy A as:

The result obtained here is in the interval [0,1]. The larger the result is, the higher the accuracy is.

* 1. **Evaluation process**
     1. **Certain variation rate, change length:**

A sequence M2000 with a length of 2000 bases was randomly generated, and the mutation rate was set as 5% (i.e., 20 modifications, 40 deletions, 40 additions) to obtain a new sequence N2000, with a computational accuracy. Experiment I times, and make:

Do the same thing for A1200,A1400……A3000, and draw the image.



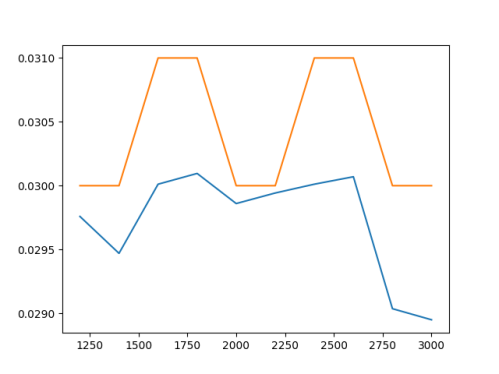
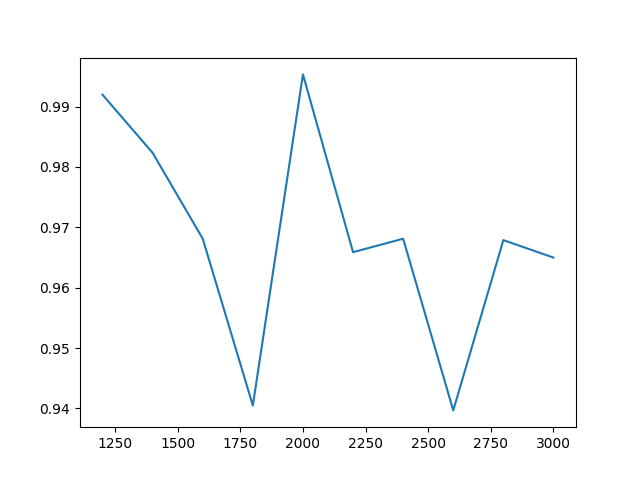


图1 准确性随个数变化图像 图2 预测值与理想值对比

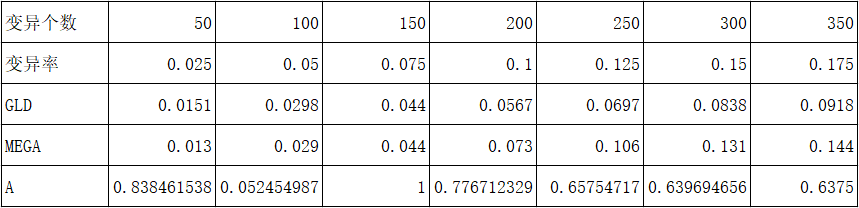
According to the above analysis, when the variation rate is constant and small, the sequence length is relatively short.

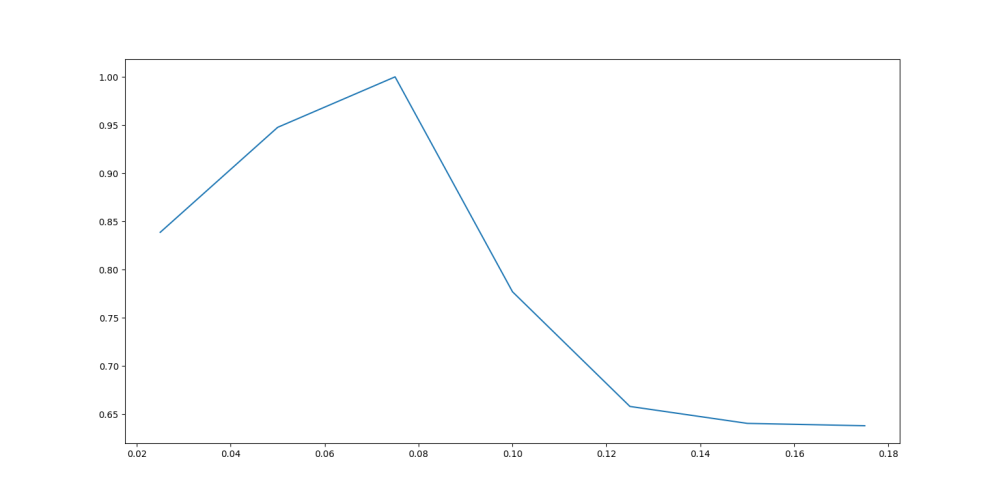
The accuracy fluctuates with length, but the accuracy is stable at about 95%, and the predicted value is relatively small compared with the actual value. Therefore, the accuracy of the model is higher when the variation rate is small.

* + 1. **The length is fixed, change the mutation rate**

The sequence M5% with A length of 2000 bases was randomly generated, and the variation rate was set as 5% (i.e., 40 modifications, 30 deletions and 30 additions) to obtain a new sequence N5%, with a computational accuracy. Experiment I times, and make:

By the same token, A2.5%, A7.5%...... A17.5%, and draw the image:





By analyzing the above chart, it can be concluded that when the sequence length remains unchanged and the variation rate changes:

The accuracy will change significantly, with the specific trend of increasing first and then decreasing, reaching the peak when the variation rate is about 7.5%. Then it will decline rapidly and level off when the variation rate is greater than 12.5%. Accuracy will not fall below 60%. The data in the icon is consistent with that in 2.2.1, indicating high reliability.

* + 1. **Concrete example – COVID-19:**

The model was used to analyze the distance between alpha and beta sequences (length of 30000+ bases) of two sequence of COVID-19, and it was found that the predicted distance value was about 0.01556 and the ideal value was 0.018, and the result was 86.4%. This result accords with the conclusion of the above two accuracy analyses.

* 1. **Conclusion**

From the above accuracy analysis, it can be concluded that the accuracy of this model is less affected by the sequence length and more affected by the variation rate. This indicates that the model is suitable for sequences of biologically close groups of organisms. The model has high accuracy in the determination of homologous sequence and homologous sequence distance.