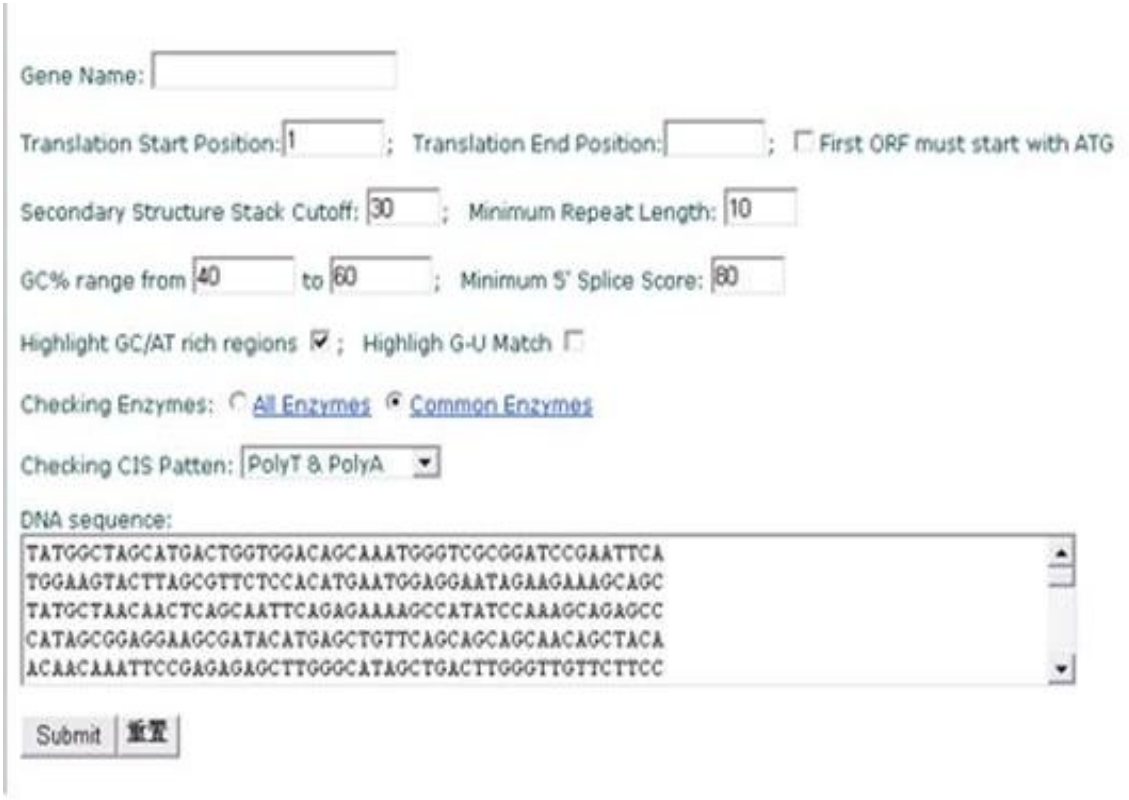
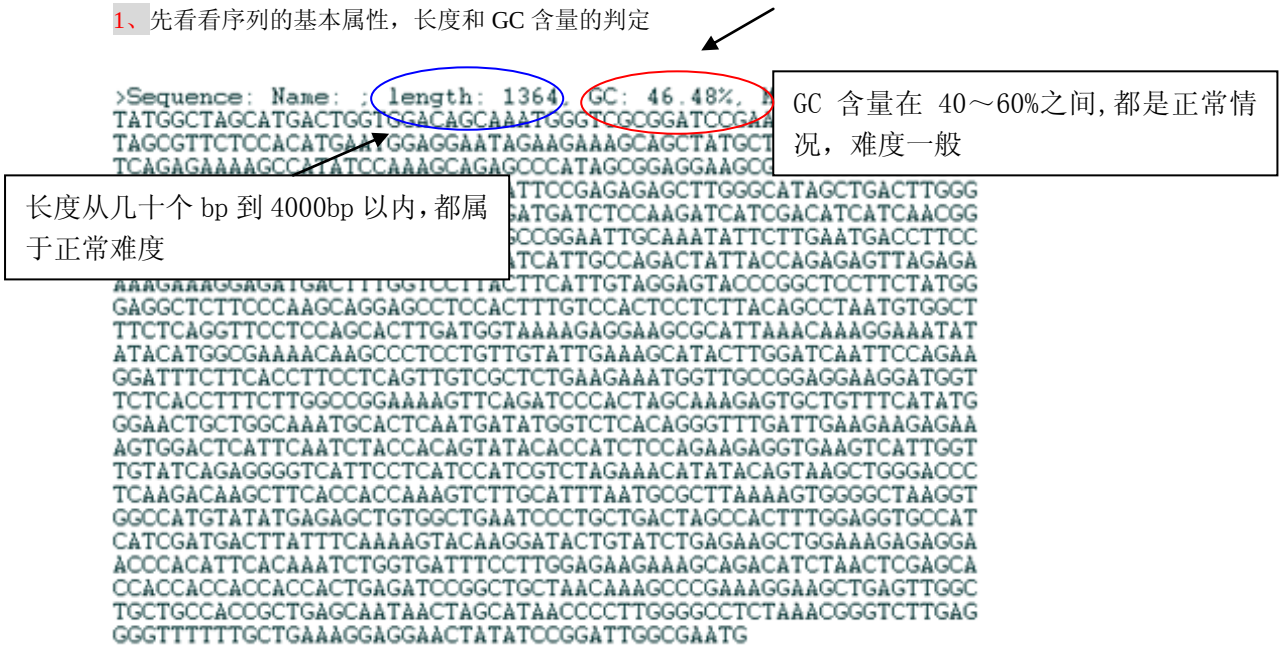
基因合成难度界定

一、我们收到订单之后，会利用在线的一个分析软件分析基因的结构，预测合成周期。如下图所示，我从 NCBI 数据库下载了一个基因的序列，举一个例子说明。



二、得到的分析结果如下，分别解释如下

1、先看看序列的基本属性，长度和 GC 含量的判定

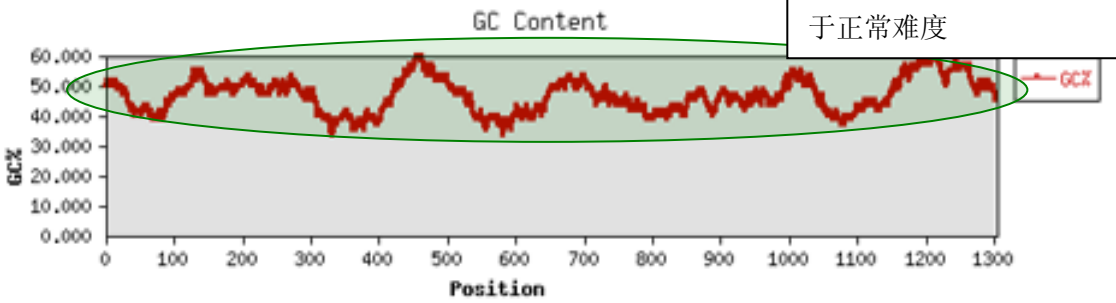


2、根据基因序列，可以推算出蛋白质序列，如果发现密码子数目不多，蛋白翻译错误，将会有信息提示。

```
>Protein: Translation Start Position: 1, Translation End Position: 1362
YG*HDVVTANGSRIRIHGST*RSPHEVRNRRKQLC*QLSNSEKSHIQSRAHSGGSDT*AV
QQQQQLQQQIPRELGHS*LGLFFRTQHSVDDLQDHRHHQRRMPPSVP*IAGIANILE*PS
RE*FQHHFHIIARLLPES*RKERR*LVSLLHCRSTRLLLVEALPKQEPPLCPLLLQPNVA
FSGSSST*V*KRKRIKQRKYIHGENKPSCCIESILGSIPDGFIHLPQLSLKNGCRRKDG
SHLSVPEKFRSH*QRVLFHMGTAGKCTQ*YGLTGFD*RRESGLIQSTTVTISRRGEVIG
CIRGVIPHPSSRNIYSKLGPSRQASPPKSCI*CA*KWG*GGHVYESCG*IPAD*PLVRCH
HR*LISKVQGYCI*EAGKRGTHIHKSGDFLGEESRHLTRAPPPPPLRSGC*QSPKGS*VG
CCHR*AITSITPWGL*TGLEGFFAERRNYIRIGE
Direct Protein Repeats(>6):
```

3、从整体的序列来看，分析 GC 含量在整个序列中的分布情况

GC 含量是锯齿状波动的，这是很正常的分布状态，即使小部分区域高达 80%或低于 30%，只要分布范围较窄，都属于正常难度



Download GC Content Profile

4、分析具体的 A、T、G、C 各个碱基的含量，理想的状态是各占 25%，局部浮动到 10～40%都是可以接受的。

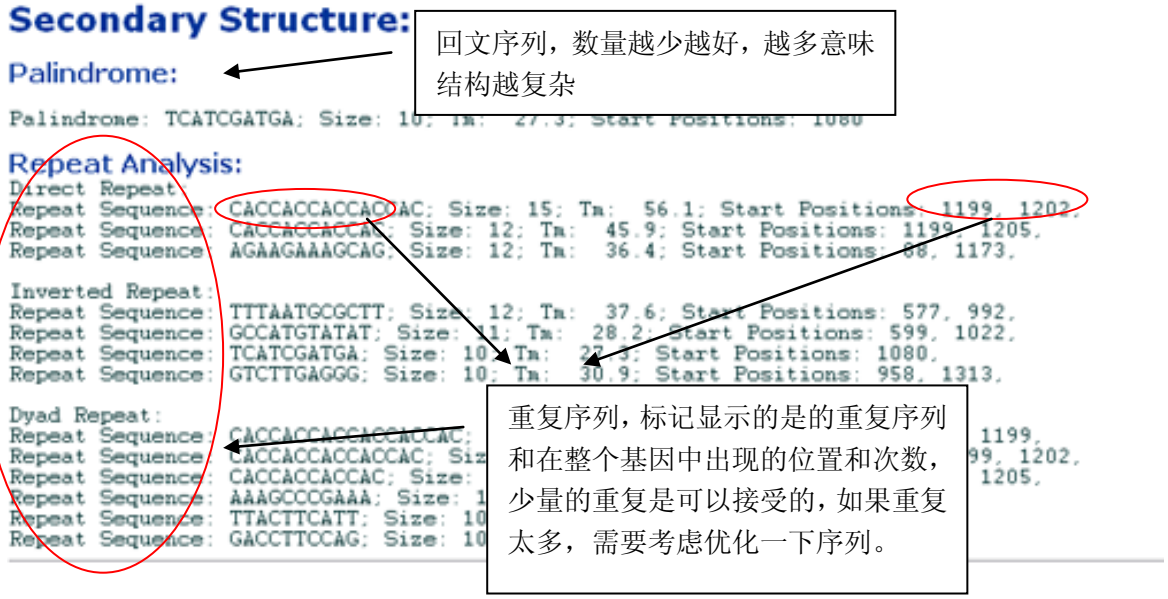
Base Composition

	A	C	G	T	X
Base Count	402	314	320	328	0
Base Frequency	29.47%	23.02%	23.46%	24.05%	0.00%

5、根据基因序列，预测蛋白结构，

回文结构、发卡结构等，对合成和测序都会带来难度，此图的预测结构中，二级结构较少，难度正常。

6、根据基因序列，直接分析基因结构


Secondary Structure:
Palindrome:
回文序列，数量越少越好，越多意味结构越复杂
Palindrome: TCATCGATGA; Size: 10; Tm: 27.3; Start Positions: 1080
Repeat Analysis:
Direct Repeat:
Repeat Sequence: CACCACCACCACCAC; Size: 15; Tm: 56.1; Start Positions: 1199, 1202,
Repeat Sequence: CACCACCACCAC; Size: 12; Tm: 45.9; Start Positions: 1199, 1205,
Repeat Sequence: AGAAGAAAGCAG; Size: 12; Tm: 36.4; Start Positions: 68, 1173,
Inverted Repeat:
Repeat Sequence: TTTAATGCGCTT; Size: 12; Tm: 37.6; Start Positions: 577, 992,
Repeat Sequence: GCCATGTATAT; Size: 11; Tm: 28.2; Start Positions: 599, 1022,
Repeat Sequence: TCATCGATGA; Size: 10; Tm: 27.3; Start Positions: 1080,
Repeat Sequence: GTCTTGAGGG; Size: 10; Tm: 30.9; Start Positions: 958, 1313,
Dyad Repeat:
Repeat Sequence: CACCACCACCACCAC;
Repeat Sequence: CACCACCACCAC; Siz
Repeat Sequence: CACCACCACCAC; Size:
Repeat Sequence: AAAGCCCGAAA; Size: 1
Repeat Sequence: TTACTTCATT; Size: 10
Repeat Sequence: GACCTTCCAG; Size: 10
重复序列，标记显示的是的重复序列和在整个基因中出现的位置和次数，少量的重复是可以接受的，如果重复太多，需要考虑优化一下序列。