**生物信息学上机实验**

**实验2 序列查询与比对**

**一、实验目的**

1． 了解序列比对的原理和用途；

2． 掌握利用NCBI信息中心的BLAST在线工具查找同源序列的方法；

3． 掌握利用PROTEIN和BLAST工具进行双序列比对的方法；

4. 掌握利用EMBL的Clustal Omega进行多序列比对的方法。

**二、实验内容：**

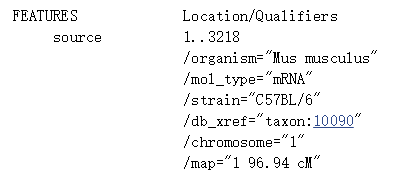
1. 研究人员在小鼠(Mus musculus)中发现了一个与有丝分裂相关的基因，通

过DNA测序，得到部分序列：

GATGAGCTGCTTATCCTACAACGAGAAGTCGGACATCTGGTCCTTGGGCTGCCTGCTGTATGAGCTGTGTGCACTAATGCCTCCCTTTACAGCTTTCAACCAAAAAGAGCTAGCTGGGAAAATCAGGGAAGGGAGGTTCAGGCGCATCCCCTACCGCTACTCTGATGGCTTGAATGACCTCATCACTCGGATGCTGAATTTAAAGGACTACCATCGACCTTCAGTGGAAGAAATTCTGGAGAGCCCTTTGATAGCAGACTTGGTTGCAGAAGAGCAAAGGAGAAATCTGGAGAGGAGAGGACGGCGCTCAGGCGAGCCTTCGAAGCTGCCGGACTCCAGCCCTGTGCTGAGCGAGCTCAAGTTGAAGGAAAGGCAACTGCAGGATCGAGAGCAAGCACTCAGAGCTCGGGAGGACATCCT

具体要求：

1. 通过BLAST工具检索https://blast.ncbi.nlm.nih.gov/Blast.cgi与此序列高度相似的序列，列出相应的查询结果。

2） 提取一条与其高相似，且为Refseq数据库中的标准序列（检索号为NM开头），说明：基因的标识符（Accession number）是什么？这个基因在小鼠中是哪个基因？这个基因在基因组上的定位是怎样的？

**参数设置：**

Database：/nr核酸数据库

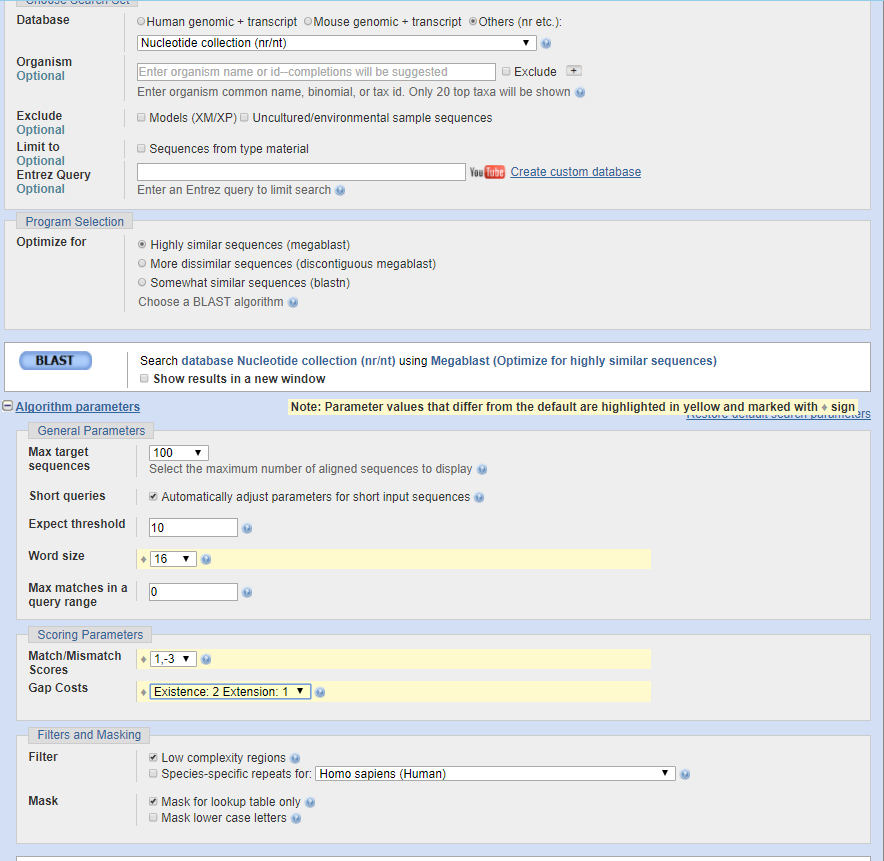
Program Selection：Highly similar sequences (megablast)

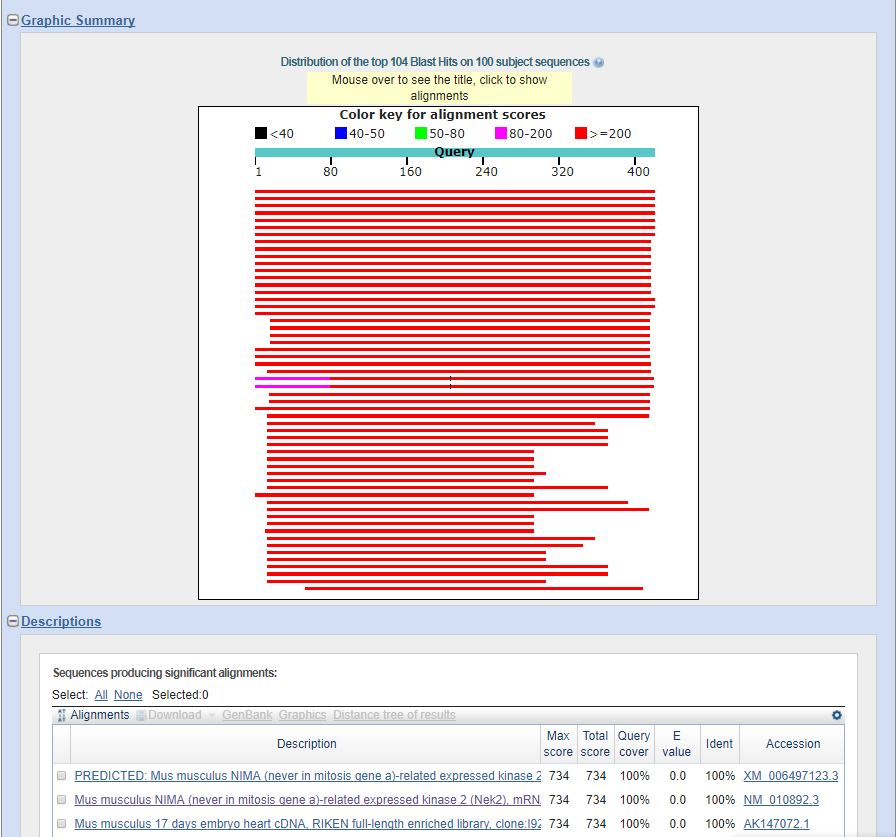
Algorithm parameters

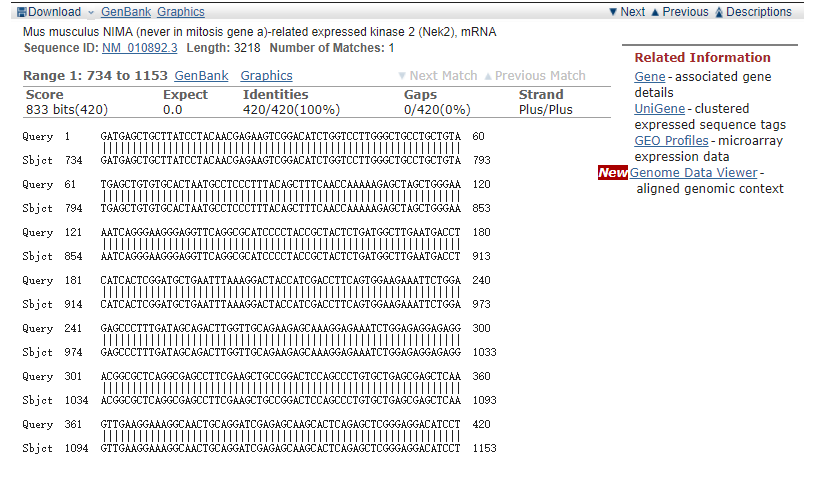
Word size:16

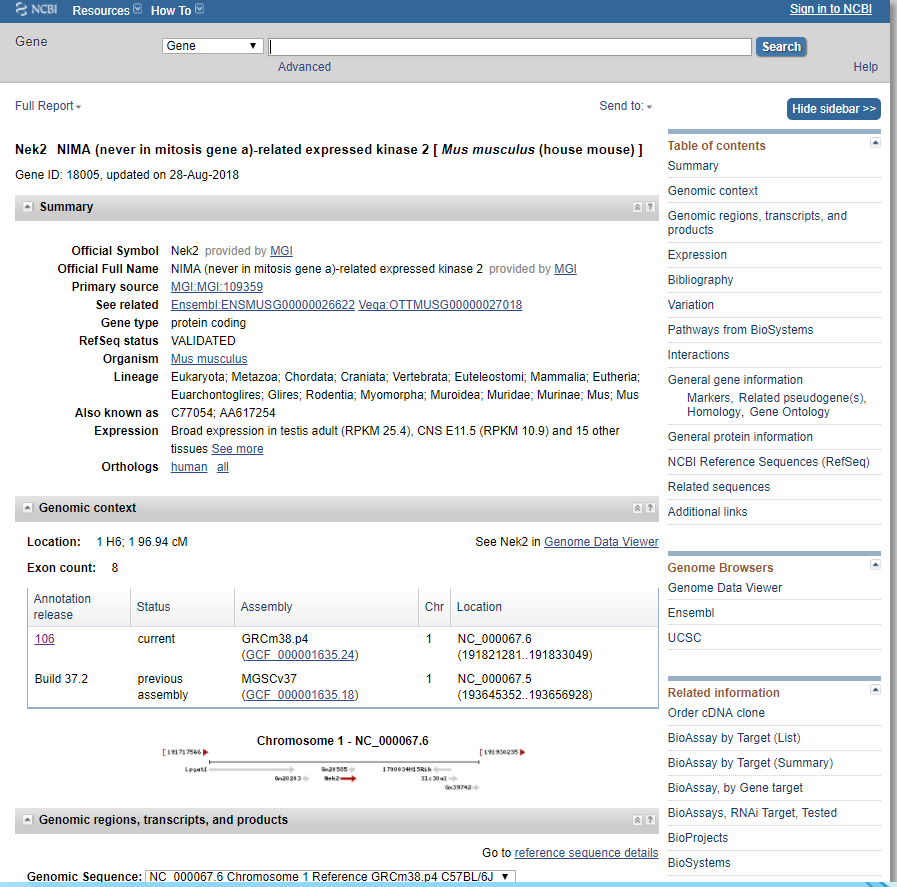
Match/Mismatch Scores:1,-3

Gap Costs: Existence:2， Extension:1

**答：1）参数设置：**

**查询结果：**

1. **对比结果：**

**序列信息：**

**基因标识符为：NM\_010892；**

**位于小鼠基因NIMA (never in mitosis gene a)-related expressed；**

**在基因组定义为：1 H6; 1 96.94 cM**

2. 用EMBL双序列比对工具<https://www.ebi.ac.uk/Tools/psa/> Needle [Protein](https://www.ebi.ac.uk/Tools/psa/emboss_needle)对下列两条序列作两次全局比对（Needle算法），第一次设置GAP OPEN=10，GAP EXTEND=0.5；第二次设置GAP OPEN=1，GAP EXTEND=5。其他参数默认。给出两次序列比对的结果（页面屏幕截图）。你觉得哪一次的序列比对的结果更佳，给出你的评判标准。并给出最佳方法得出的比对结论（相似性，同一性等）

>seq\_1

VKCRLNVLLWYQDAYGEVEINDGKLYDAYVSYSDCPEDRKFVNFILKPQLERRRGYKLFL

DDRDLLPRAEPSADLLVNLSRCRRLIVVLSDAFLSRAWCSHSFREGLCRLLELTRRPIFI

TFEGQRRDPAHPALRLLRQHRHLVTLLLWRPGSVTPSSDFWKEVQLALPRKVQYRPVEGD

PQTQLQDDKDPMLILRGRVPEGRAL

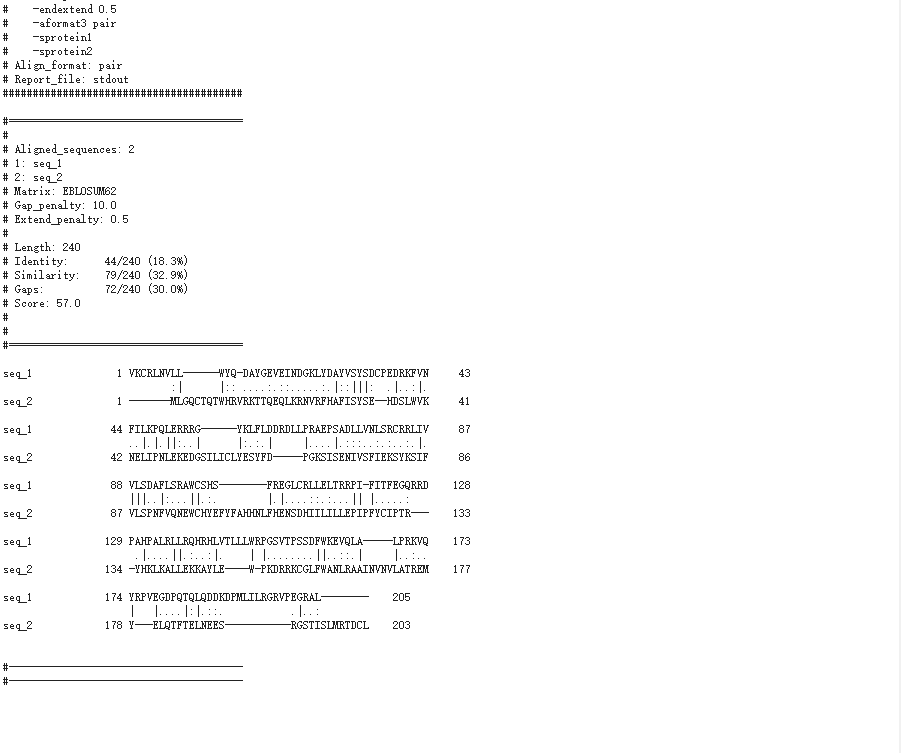
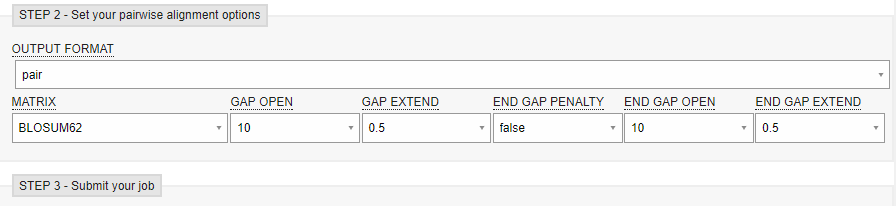
>seq\_2

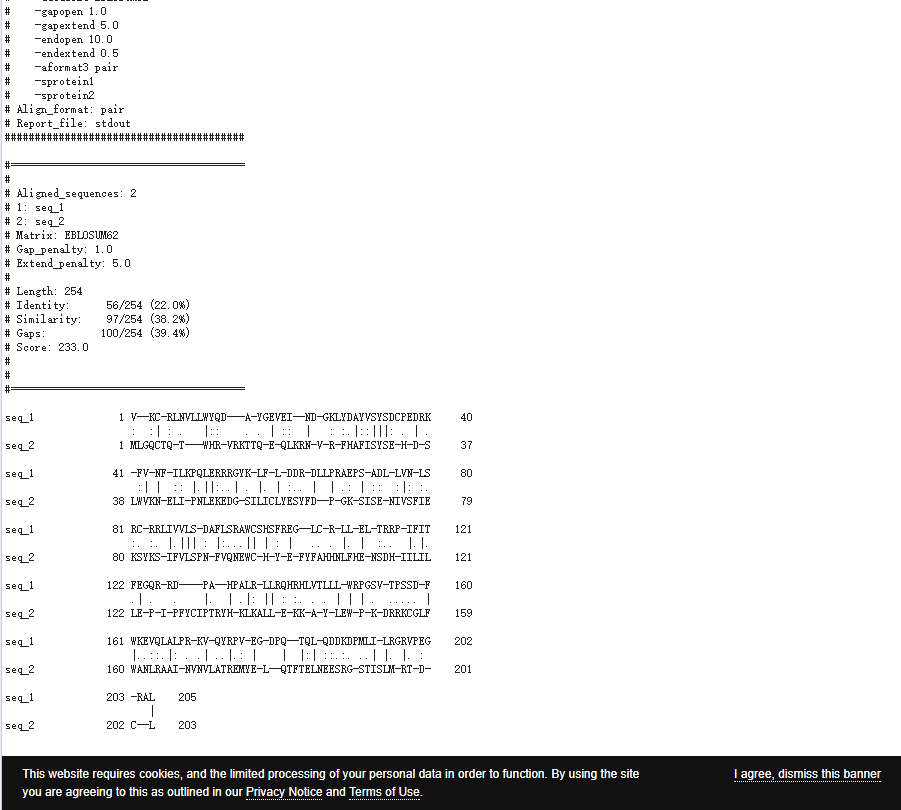
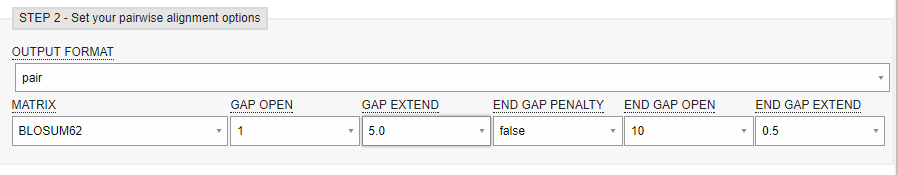
MLGQCTQTWHRVRKTTQEQLKRNVRFHAFISYSEHDSLWVKNELIPNLEKEDGSILICLY

ESYFDPGKSISENIVSFIEKSYKSIFVLSPNFVQNEWCHYEFYFAHHNLFHENSDHIILI

LLEPIPFYCIPTRYHKLKALLEKKAYLEWPKDRRKCGLFWANLRAAINVNVLATREMYEL

QTFTELNEESRGSTISLMRTDCL

 **答：1）第一次：**

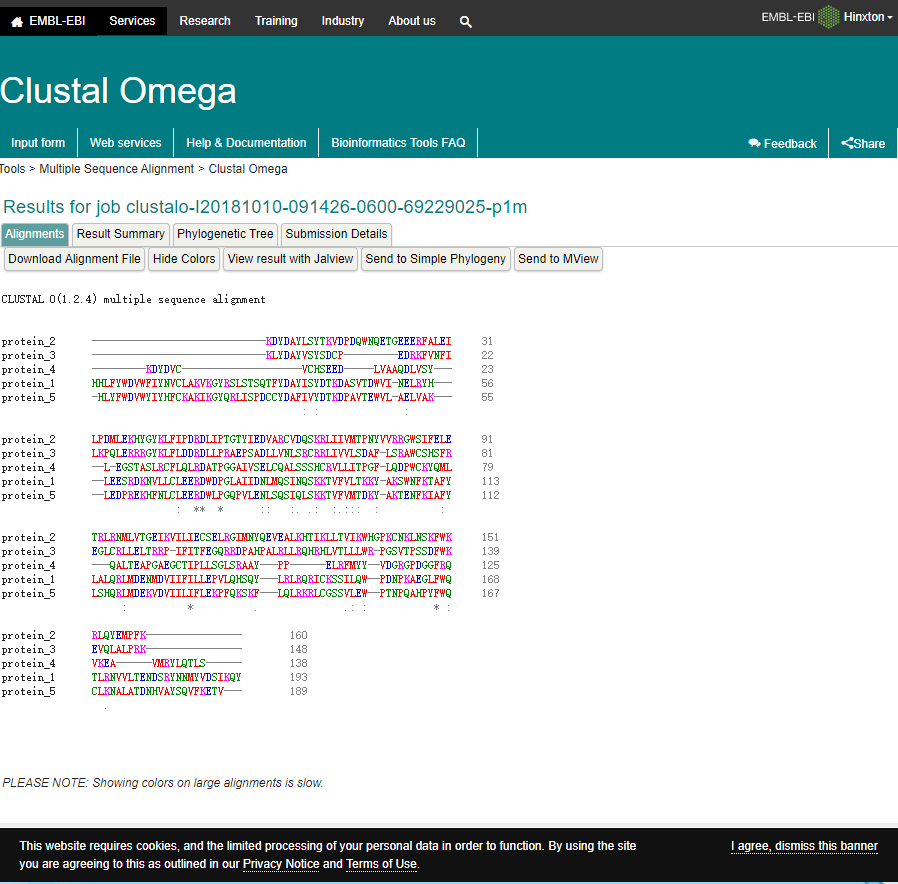
**第二次：**

**第二种结果更佳；虽然整体罚高于第一种的。但同一性22%与相似性60.2%远高于于第一种。**

3. 使用BLAST对上题中的双序列进行比对https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins&PROGRAM=blastp&BLAST\_PROGRAMS=blastp&PAGE\_TYPE=BlastSearch&BLAST\_SPEC=blast2seq&DATABASE=n/a&QUERY=&SUBJECTS=，保持参数一致，并比较和题2中的结果的差异（文字叙述）。

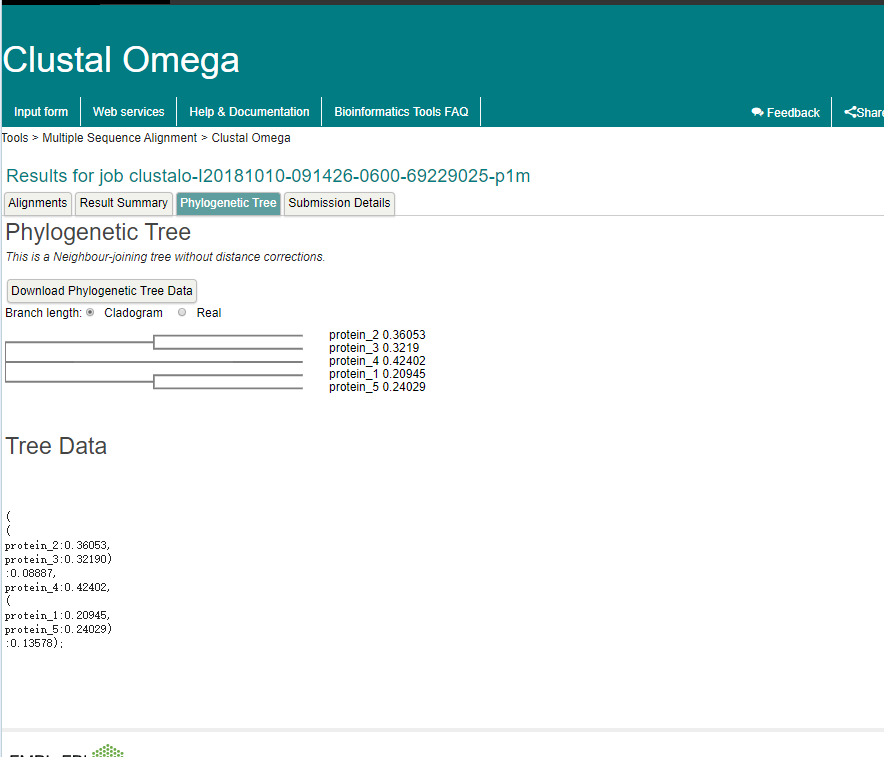
**答：BLAST同一性29%相似性为67%，均高于第二题。**

4. 为附件中exercise.fasta中的5条序列创建多序列比对。多序列比对工具自选https://www.ebi.ac.uk/Tools/msa/clustalo/。比对结果以Clustalx颜色方案标记，并按照序列两两一致度高低排序。将编辑后的多序列比对结果Show Colors，相似度矩阵Result Summary—Percent Identity Matrix及系统树保存为图片提交Phylogenetic Tree。文字叙述多序列比对结果（序列的相似性程度等。）

**答：颜色标记对比结果：**

**相似度矩阵：**



**系统发生树：**

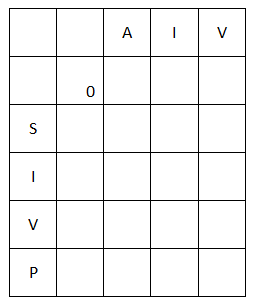
**结果显示为蛋白质序列1与序列5、序列2与序列3更为相似，相似率依次分别为55.03%、31.76%。**

5.用Needleman-Wunsch算法填充以下得分矩阵（分值和箭头），以获得序列1：AIV和序列2：SIVP的全局双序列比对，并给出它们的相似度与一致度。

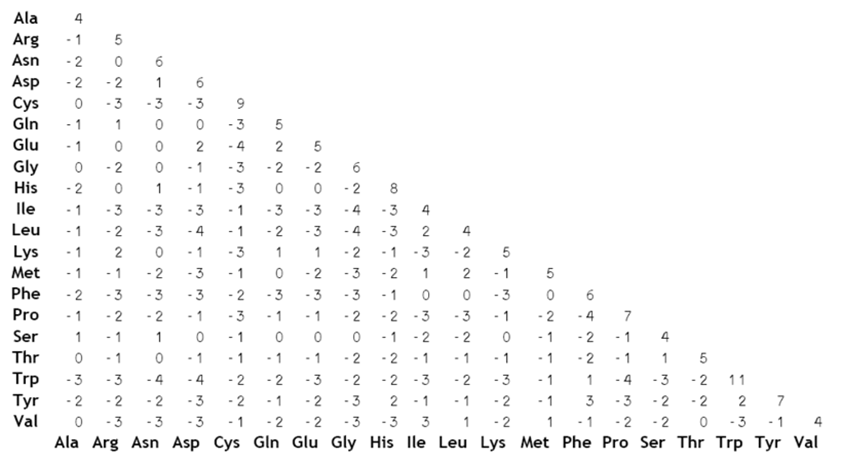
使用gap=-3，替换记分矩阵=BLOSUM62。

说明：得分矩阵可以用计算机文本编辑完成，也可以在纸上完成后手机拍照提交图片。注意用不同颜色或粗细区分最终的比对追溯箭头和其他箭头。

得分矩阵：



**BLOSUM62替换记分矩阵：**



答:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | A | I | V |
|  | 0 | -3 | -6 | -9 |
| S | -3 | 1 | -2 | -5 |
| I | -6 | -2 | 5 | 2 |
| V | -9 | -5 | 2 | 9 |
| P | -12 | -8 | -1 | 6 |

**对比结果：**

**1：A I V –**

**2：S I V P**

**同一性：50%：相似性：75%**