

Bioinformatics: Life Sciences on Your Computer

Week 2 Practice Problems - Part 1

These are practice problems for Pairwise Sequence Analysis.

1. Compare the two alignments below in the following questions:
 - a. What is the % identity of the BLAST alignment?
 - b. What is the % similarity of the BLAST alignment?
 - c. What is the % identity of the LALIGN alignment?
 - d. What is the % similarity of the LALIGN alignment?
 - e. What notation is used by BLAST to show identical amino acids in the alignment?
 - f. What notation is used by LALIGN to show identical amino acids in the alignment?
 - g. What notation is used by BLAST to show similar amino acids?
 - h. What notation is used by LALIGN to show similar amino acids?
 - i. Why are the results different?

Alignment 1 - BLAST

lysine biosynthesis protein LysX [Natronolimnobius innermongolicus JCM 12255]
 Sequence ID: [ref|ZP_21501989.1](#) Length: 295 Number of Matches: 1

Score	Expect	Method	Identities	Positives	Gaps
219 bits (558)	2e-66	Compositional matrix adjust.	116/259 (45%)	159/259 (61%)	9/259 (3%)
Query 4	LAIYYDRIRPDEKMLFEALDELGVYDKVYAPQLRVTFD-----QAVPWRVALERCVSQ 57				
	+ I+Y RIR DEK+L L E K+ + TFD + + ++RC++				
Sbjct 6	VGILYSRIRKDEKLLLNLRLRERDHEVAKIDVR--KQTFDISEAPEEFRDLDIVVDRCLAT 63				
Query 58	SRGHAVTRALEGLGVRVINPSHVIELCGDKLATNARLARAGLPTPRTGVAFDGEAALQLI 117				
	SR I+ E G+ V+N ++C DK+ + L +AG+PTP T VAF E A++ I				
Sbjct 64	SRSLYATQFFEAYGIPVVNSHETADICADKVKNSLALEQAGVPTPATKVAFTKETAMEAI 123				
Query 118	EEMGYPVVLKPTVGSWGRMVSKINDRDAAEAIIHKEVLGGPQHSVFYVQELVKKPERDI 177				
	EE GYP VLKP VGSWGR+++KI+ DAAEAI+EHK LG +H VFYVQE V+KP RDI				
Sbjct 124	EEFGYPCVLKPVVGSWGRMLMAKIDSSDAAEAILEHKATLGHYEHKVYVQEFVEKPGRDI 183				
Query 178	RAFVVDGVCIGAIYRTSEHWITNTARGAKASRCEVTPEIADLAVRAAQAVEGQIVAILDV 237				
	R DG + A+ R+S+HWITN A+GA+ E E +L +A+ AV G ++ IDL+				
Sbjct 184	RVVATDGEFVAAMVRSSDHWITNAAGAETDVFEPDDEAKELVKKASDAVGGGLLGIDLM 243				
Query 238	EDPERGLLVIEINHTEFEK 256				
	E E G V E+NHT+EFK				
Sbjct 244	E-TEDGYTVHEVNHTVEFK 261				

Alignment 2 - LALIGN

>>>gi|494469826|ref|WP_007259304.1|, 295 aa vs TMP.q2 library

Waterman-Eggert score: 704; 202.2 bits; E(1) < 1.1e-56y
44.4% identity (74.9% similar) in 259 aa overlap (4-261:2-256)

```

      10      20      30      40      50      60      70      80
ref|W  VNVGILYSRIRKDEKLLLNELRERDHEVAKIDVRKQTFDISEAPEEFRDLDIVVDRCLATSRSLYATQFFEAYGIPVNVNS
      ..... : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : :
ref|YP AELAIYYDRIRPDEKMLFEALDELGVFPYDKVYAPQLRVTFDQAVP-WR---VALERCVSQSRGHAVTRALEGLGVRVINP
      10      20      30      40      50      60      70

      90      100     110     120     130     140     150     160
ref|W  HETADICADKVKNSLALEQAGVPTPATKVAFTKETAMEAIEEFGYPCVLKPVVGSWGRMLMAKIDSSDAEAILEHKATLG
      .. : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : :
ref|YP SHVIELCGDKLATNARLARAGLPTRTGVAFDGEAALQLIEEMGYPPVVLKPTVGSWGRMVSKINDRDAAEAIEHKEVLG
      80      90      100     110     120     130     140     150

      170     180     190     200     210     220     230     240
ref|W  HYEHKVFYVQEFVEKPGRDIRVVATDGEFVAAMVRSSDHWITNAAKGAETDVFEPDDEAKELVKKASDAVGGGLLGIDLM
      ..... : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : : :
ref|YP GPQHSVFYVQELVKKPERDIRAFVVDGVCIGAIYRTSEHWITNTARGAKASRCEVTPEIADLAVRAAQAVEGQIVVAIDL
      160     170     180     190     200     210     220     230

      250     260
ref|W  ET-EDGYTVHEVNHTVEFK
      : : : : : : : : : :
ref|YP EDPERGLLVIEINHTMEFK
      240     250
```

2. Go to the [FASTA alignment website](#) and compare the hemoglobin alpha chain (P69905) to the hemoglobin beta chain (P68871) with LALIGN (default settings).
 - a. What is the % identity of the longest alignment?
 - b. How many gaps in the alpha chain?
 - c. What is the score in bits?
3. Compare the alignment scores obtained with small and large gap penalties in the following example:
Go to the LALIGN program on the [University of Virginia FASTA server](#). This program aligns sequences by a local dynamic programming algorithm and includes end gap penalties. LALIGN produces several different alignments, with no two alignments including a match of the same two sequence positions.
 - a. What is the default scoring matrix?
 - b. Obtain two amino acid sequences in FASTA format: RECA from the bacterium Escherichia coli (gi|16130606) and RAD51 from yeast (gi|218469). These proteins have the same function, but have diverged enough that they have become difficult to align. Use LALIGN to align the above sequences with gap penalties of -12 and -2 (default). Two alignments are shown. The top alignment is usually the strongest possible alignment. What is the score in bits of the top alignment?
 - c. What is the length (overlap) of the top alignment?
 - d. What are the % identity, the % similarity of the top alignment.
 - e. How many gaps are in the yeast protein's alignment (218469)?
 - f. Repeat but with gap penalties of -8 and -2. How many gaps are now in the yeast protein's alignment?
 - g. What is the % identity and % similarity?

Discussion topic: Which looks more like a local alignment and which looks like a global alignment (look at gaps and identity)?

4. Use the same two sequences from question three. Check the box for "Align two or more sequences" (use blastp for protein) on the [NCBI server](#).
 - a. Click Search Summary. What are the default gap opening and gap extension penalties?
 - b. Which substitution matrix is used? Is this the same one LALIGN uses?
 - c. What is the percent identity of the alignment?
 - d. What is the percent similarity (positives)?
 - e. Explain the significance of the letters and + signs between the sequences.

Answers

- 1.a. 45%
b. 61% (BLAST uses "positives" for "similar")
c. 44.4%
d. 74.9%
e. A capital letter between the alignments
f. A double dot (or colon) - (:)
g. A plus sign (+)
h. A period or single dot (.)
i. They are different algorithms that use different methods, different penalties and different substitution matrices.
2. a. 43.4%
b. Two (one gap of length 1 and another of length 5)
c. 106.6
3. a. Blosum 50
b. 62.0
c. 230
d. 28.7% identity, 57.4% similarity
e. 4
f. 10
g. 29.7% identity, 56.8% similarity
- 4.a. 11,1
b. BLOSUM 62
c. 28%
d. 46%
e. Letters are for identity and + signs for conservative substitutions. Add the letters to the + signs for similarity (positives).