Practical Session 11 Sequence alignment of Toll-like receptors

In this session, you will study the output of Sanger sequencing reactions and how to use these for further investigation. You will find out how an online stored mRNA sequence is composed of separately sequenced fragments or traces. You will learn how sequence alignment works and how to use a number of computer tools that can align two or more sequences. For all exercises, you will work with sequences of Toll-like receptors.

Wikipedia summarises for us that **Toll-like receptors** (TLRs) are a class of proteins that play a key role in the innate immune system. They are single, membrane-spanning, non-catalytic receptors that recognise structurally conserved molecules derived from microbes. Once these microbes have breached physical barriers such as the skin or intestinal tract mucosa, they are recognized by TLRs, which activate immune cell responses.

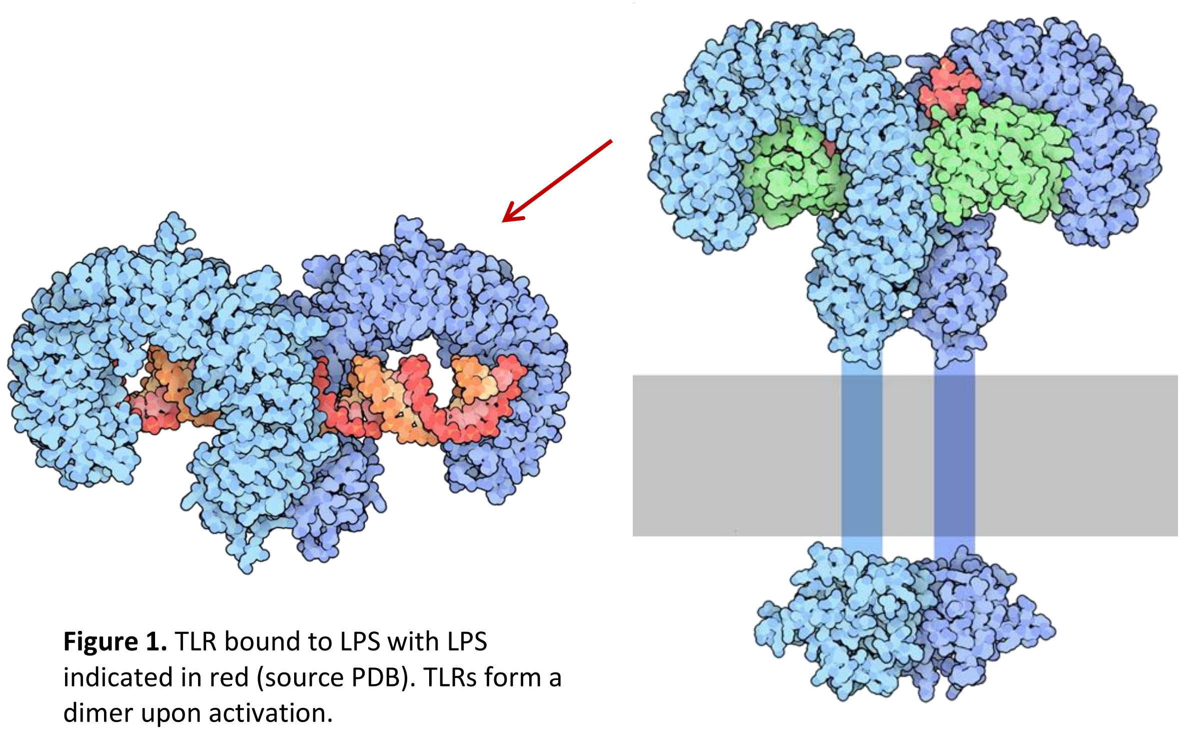
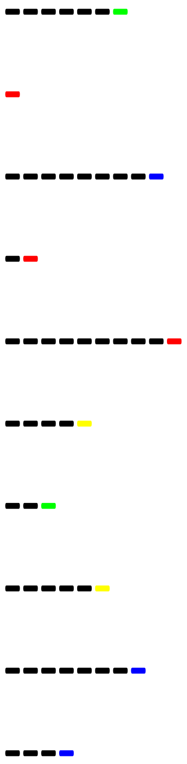


Figure 1. TLR bound to LPS with LPS indicated in red (source PDB). TLRs form a dimer upon activation.

Part 1 Sanger sequencing

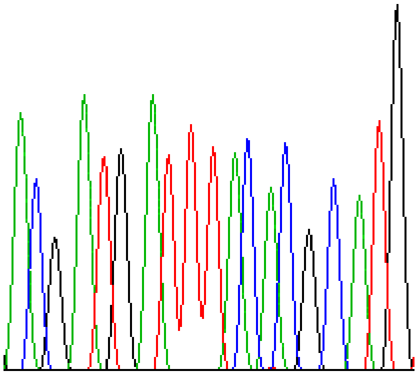
In the first lecture on Biological Databases, the procedure and output of Sanger sequencing has been discussed. Output is in the form of a chromatogram. Assignment 1 contains some introductory questions about reading the output of Sanger sequencing.

Assignment 1: The dashed lines represent the outcome of a Sanger sequencing reaction; the nucleotides had been labeled as follows:
ddATP: Green
ddGTP: Yellow
ddCTP: Blue
ddTTP: Red



a. What is the sequence of the input DNA of the reaction mentioned?
TTACGGACCT

This figure represents (part of) the raw output of a "dye-terminator sequencing" machine.



The colours represent the following dideoxynucleotides:

ddATP: Green
ddGTP: Black
ddCTP: Blue
ddTTP: Red

b. Write down the sequencing results of this output
ACGATGATTTACACGCATG

c. Sequencing is also used to determine the traces that are used to reconstruct genome sequences. Assuming an average trace length of 500 bp, how many would be needed to completely cover the human genome once? How many would you need more realistically?
3 billion / 500 = 6 million. This would be theoretically needed to cover once. More realistically you would need a – say – tenfold coverage, thus 60 million traces and probably even more

Assignment 2: Before the advent of Next Generation Sequencing, the sequences in Genbank also have been determined using Sanger sequencing. The Trace Archive contains the original traces from which the sequences have been composed. Open NCBI's Trace Archive (not the FTP version) by using the list of resources from the left-side menu on the NCBI page or using Google.

According to the NCBI Gene database, one of the traces that has been used to construct the reference sequence of TLR1, is the sequence with Genbank identifier BU623316. Look up this sequence in the Trace Archive by entering the following query in the search field: `ACCESSION="BU623316"`

a. The trace is given in FASTA format, what does this format look like?
One description line starting with > followed by the sequence in any formatting

Change the view to see the real chromatograms for the trace, by changing 'Show as FASTA' into 'Show as Trace'.

b. How long is this trace?
1,099 bases; you can find this by scrolling fully to the right in the trace

c. From which base to which base would you consider this trace to be most reliable?
From around base 100 till around base 475 (this is of course a matter of interpretation; you may have a slightly different answer). Note: the T's at the start may be real, but could also be an experimental artefact.

d. Make the quality scores visible by checking the checkbox below the sequence. Verify your answers to the previous question.
These bases indeed have high quality scores

e. What do you think of the overall quality of this trace?
It is not very good; many parts have not very clear signals (and as such have low quality scores)

f. Look TLR1 up in the nucleotide database. What is the RefSeq identifier of the TLR1 mRNA?
NM_003263

g. What is the length in basepairs of the RefSeq mRNA?
2,876 base pairs

h. The traces are much shorter than the mRNA, how can this happen?
It is not possible to sequence thousands of base pairs in one Sanger sequencing reaction

i. What would you need to be able to reconstruct the entire mRNA?
You would need to have multiple traces that build op the entire sequence. In order to be able to do this, they should also have overlapping parts (otherwise you do not know how they fit together)

Part 2 Pairwise sequence alignment

Assignment 3: Studying conserved TLR protein sequences using computer tools.

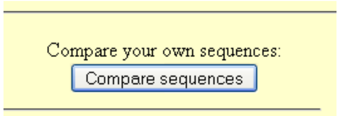
a. Look up the Refseq sequence of the human TLR1 (precursor) protein. What is the identifier of this protein?
NP_003254

b. Which database at the NCBI site can one best use to look up homologous sequences?
Homologene

c. Look up the homologue of TLR1 in mouse in this database. What is the Refseq identifier of the mouse TLR1 protein?
NP_001263374

The *needle* programme can build a Needleman-Wunsch global alignment (with gaps): http://www.ebi.ac.uk/emboss/align/

SSEARCH can build a Smith-Waterman local alignment: http://fasta.bioch.virginia.edu/
Tip: click this button to compare two sequences in *SSEARCH*:



d. Compare the human and mouse TLR1 protein sequences using both tools and complete the table:

Alignment length	795	782
Identity	582/795 (73.2%)	74.4%
Similarity	673/795 (84.7%)	90.3%
Gaps	9/795 (1.1%)	
Score	3062	3928

e. What could be causes of the small differences in outcome?

Needle performs a global alignment and SSEARCH a local one; furthermore, the default setting of several parameters are different:

Needle: Blosum62, opening gap: -10, extending gap: -0.5

SSEARCH: Blosum50, opening gap: -10, extending gap: -2

f. Change the scoring matrix in *needle* to Blosum50 and run again. What are the results now? Are they more or less similar to the *SSEARCH* results now and how could this be explained?

Alignment length	796
Identity	584/796 (73.4%)
Similarity	676/796 (84.9%)
Gaps	11/796 (1.4%)
Score	3928.5

More similar; now the scoring matrix is the same as the one that SSEARCH used (other differences remain, so the results are still not exactly the same)

Note: you will further investigate the effects of choosing several scoring matrices (BLOSUM and PAM matrices) in the next practical session on Blast.

g. Rerun both tools for the GAPDH protein, a strongly conserved enzyme that catalyses the sixth step in glycolysis and is known to be constitutively expressed in many cell types. The Refseq identifier for the human and mouse GAPDH proteins are NP_002037 and NP_032110, respectively. Use the standard settings.

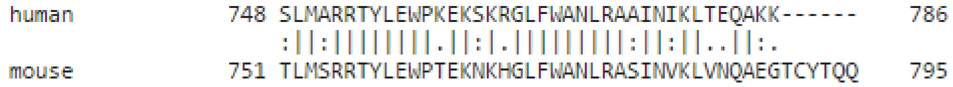
	needle	*SSEARCH*
Alignment length	335	332
Identity	313/335 (93.4%)	94.3%
Similarity	324/335 (96.7%)	98.5%
Gaps	2/335 (0.6%)	
Score	1623	2057

h. Is the TLR1 protein also very conserved?

TLR1 does not seem to be a strongly conserved gene. Already in the mouse the number of identities is only ~74% (short evolutionary distance) and the amount of positives is ~88%. Comparing to the values for GAPDH confirms this conclusion. Note: the fact that the score for TLR1 is higher than for GAPDH does not mean anything, scores cannot be compared between different sequence inputs (as they depend, for example, on the length of the sequence)

Assignment 4: Creating a dot plot of TLR protein sequences.

At the end of the sequences, the alignment produced by *needle* in question 3d looks as shown in the screenshot:



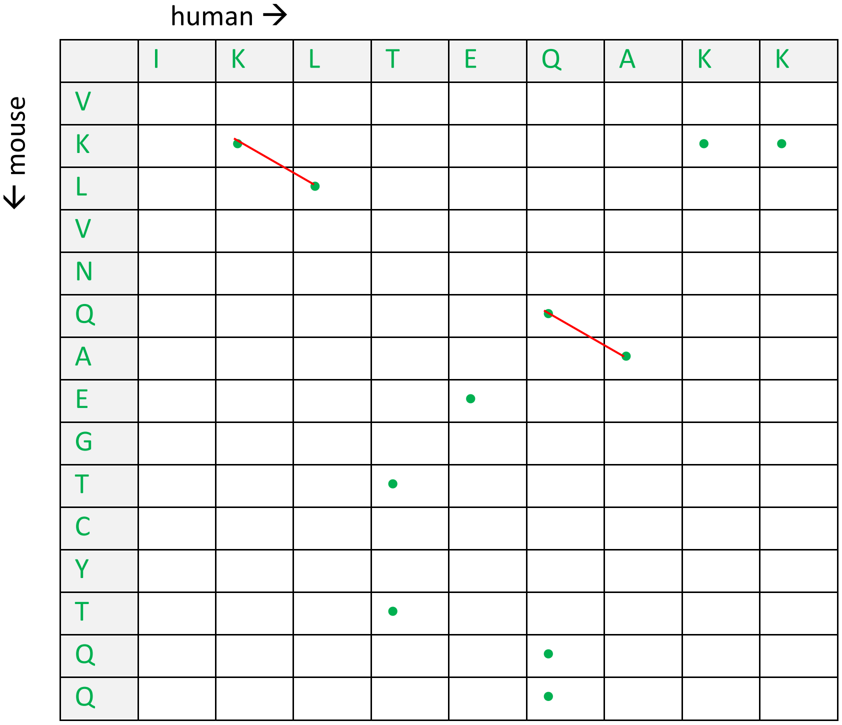
```
human            748 SLMARRTYLEWPKEKSKRGLFWANLRAAINIKLTEQAKK------    786
                     :||:||||||||.||:|.|||||||||:||:||..||:.
mouse            751 TLMSRRTYLEWPTEKNKHGLFWANLRASINVKLVNQAEGTCYTQQ    795
```

Create a dot plot of the last 15 positions of the alignment. Don't forget to connect sequential matches by lines.

Human: IKLTEQAKK------

Mouse (*Mus musculus*): VKLVNQAEGTCYTQQ

Dot plot:



	I	K	L	T	E	Q	A	K	K
V									
K		●						●	●
L			●						
V									
N									
Q						●			
A							●		
E					●				
G									
T				●					
C									
Y									
T				●					
Q						●			
Q						●			

Assignment 5: Manual scoring of the alignment of the previous question by using several scoring schemes.

Given the following two sequences:

Sequence 1: IKLTEQAKK

Sequence 2: VKLVNQAEGTCYTQQ

And the following two alignments:

(*consider a gap at the end of one of the sequences as NOT part of the alignment*)

```
IKLTEQAKK------                 IKLTEQAK-K-----
 ||  ||               and        ||  ||
VKLVNQAEGTCYTQQ                 VKLVNQAEGTCYTQQ
```

Score both alignments according to the following three methods:

i. % identity
ii. score per match: 2, per mismatch: -2; with gap penalties: opening: -1; elongation: -½
iii. score according to the PAM250 matrix (see lecture slides), with gap penalties: opening: -1½; elongation: -1

Which alignment is best according to each of these methods?

	i) % identity	ii) 2, -2, -1, -½	iii) PAM250, -1½, -½
MKWVTFISLL MKWVTLISFI	4/9 = 44.4%	8 - 10 = -2	4+5+6+0+1+4+2+0-2 = 20
MKWVT---FISLL MKWVTLISFI	4/10 = 40.0%	8 - 10 - 1 = -3	4+5+6+0+1+4+2+0-1½+0 = 20.5

According to the first and second methods, the first alignment is best; according to the third method the second alignment is best.

Part 3 Multiple sequence alignment

Assignment 6: Multiple alignment of TLR proteins.

In this question you will study multiple alignment of the sequences of TLR1, TLR2, TLR3, TLR5, and TLR6.

a. Have a look at the file 'pract_sequences.txt' that is provided at Student Portal. What does this file contain?

   The sequences of the given TLR proteins in FASTA format, pasted below each other

b. Build a multiple sequence alignment of all receptor sequences in the file using the online version of ClustalO (successor of ClustalW) at http://www.ebi.ac.uk/Tools/msa/clustalo/ (use **Internet Explorer** in the computer rooms). Tip: make sure the dropbox is set to 'Protein'. Are the family members highly conserved or not?

   At first sight, they do not seem highly conserved

c. Which proteins are most and least similar? What is the highest and what is the lowest percentage similarity? Tip: have a look at the 'Percentage Identity Matrix' from the 'Result Summary' tab.

   Highest: 68.88 (TLR1-TLR6); lowest 21.20 (TLR1-TLR3)

d. When looking at the Phylogram image from the 'Phylogenetic Tree' tab, what is the difference between the cladogram and real branch length phylogram as presented by this tool?

   The real branch length phylogram is a scaled (or additive) tree (showing the relative amount of differences between sequences), whereas the cladogram is ultrametric (making each branch equally long as a representation of equal time span)

On the 'Result Summary' tab there is a button that links to Jalview (using Internet Explorer in the computer rooms). Open this tool.

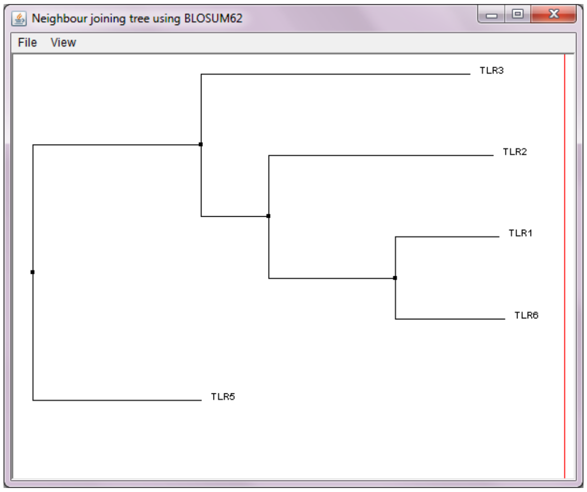
> Note that this button may not be there in your own browser (security settings), in case it is not present, visit www.jalview.org (in another browser window!) and click 'Launch Jalview Desktop' at the top right. Wait for the tool to run and close all windows it opens automatically. Then, on the 'Alignments' tab in ClustalO, click the 'Download Alignment File' button. In Jalview go to *File -> Input Alignment -> from file* and load the file (alternatively, you can copy-paste the link and use '*from URL*').

e. Using Jalview, colour the alignment by percentage identity, in order to see better where the differences are. What happens in regions where the sequences are very poorly conserved?

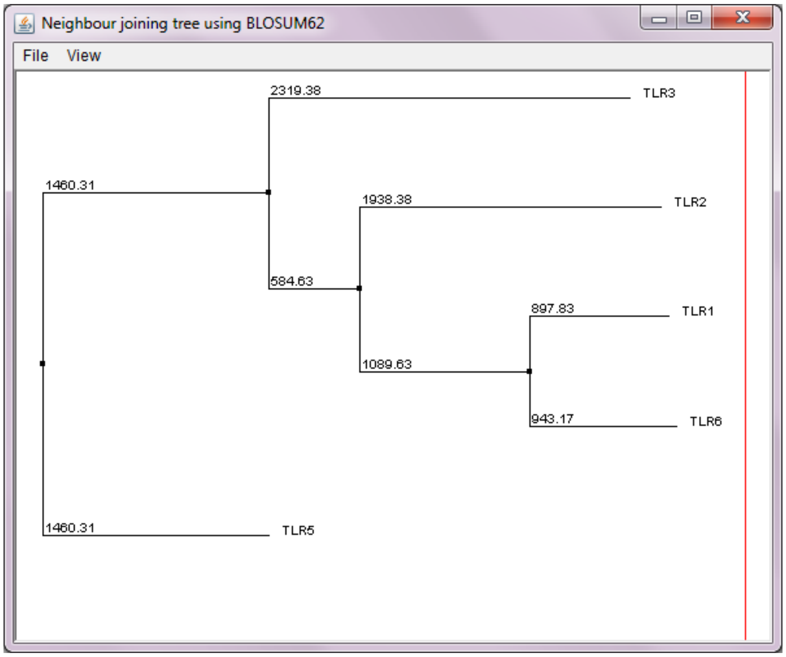
   There seem to have been insertions in TLR3 and TLR5 (or deletions in the others; that we cannot deduce from the alignment)

f. Open the Calculate Tree menu item, which distance method (percentage identity or BLOSUM62) would you use in this case and why?
BLOSUM62, since we are looking at protein sequences, not DNA

g. Sketch the tree using the Neighbour Joining method (you may make a screenshot).



h. From the View menu in the tree window, you can enable Show Distances. What is the distance to the common ancestor that joins the three receptors TLR1, TLR2, TLR6? And what is the longest pairwise distance between those three receptors? Tip: for the common ancestor, you may use the furthest member to calculate the distance.
The distance to the common ancestor is 1089.63+943.17=2032.8
Longest distance between the three receptors is 1938.38+2032.8=3971.18



Supporting literature:

Related to this session, you can consult some literature using the course books. 2 in Mount gives some information on sequencing and its output, and how we can store those using a computer. Chapter 3 in Pevsner / 4 in Zvelebil and Baum discuss (pairwise) sequence alignment. Chapters 4 and 5 in Pevsner give information on the BLAST tool. Chapter 10 in Pevsner discusses multiple sequence alignment. Some information on phylogenetic trees can be found in Chapter 11 in Pevsner / chapter 7 in Zvelebil and Baum

- Mount:
    - Chapter 2, Collecting and Storing Sequences in the Laboratory
        - Part on formats: limit to FASTA
- Pevsner:
    - Chapters 3, 10, 11
    - Chapter 4, Basic Local Alignment Search Tool (BLAST)
        - Excluding computation of E-value
    - Chapter 5, Advanced BLAST Searching
        - Limit to parts on PSI-BLAST and PHI-BLAST
- Zvelebil and Baum
    - Chapters 4, 7