Let's Run global and local alignments. To remind you, global alignment matches the entire length of the sequences while local alignment looks for subsequences of both sequences that match optimally to each other. There are a lot of free and user friendly bioinformatics tools to perform such alignments. For the following example we will use the free online emboss package tools.

Enter the emboss package using the following link:

http://emboss.bioinformatics.nl/

Choose the **Needle**- program for running **global** alignment or the **Water**- program for running **local** alignment.

Compare the human and horse glycophorin protein sequences which are erythrocyte membrane proteins.
Both sequences appear in Fasta Format where the sequence begins with a single-line description, followed by lines of sequence data. The description line must begin with a greater-than (">") symbol in the first column. Usually each line contains 60 letters, although the format itself accepts lines of any length.

```
>uniprot|P02724|GLPA_HUMAN Glycophorin A precursor (PAS-2)
(Sialoglycoprotein alpha) (MN sialoglycoprotein) (CD235a antigen).
MYGKIIFVLLLSAIVSISASSTTGVAMHTSTSSSVTKSYISSQTNDTHKRDTYAATPRAH
EVSEISVRTVYPPEEETGERVQLAHHFSEPEITLIIFGVMAGVIGTILLISYGIRRLIKK
SPSDVKPLPSPDTDVPLSSVEIENPETSDQ

>uniprot|P02726|GLP_HORSE Glycophorin HA.
QTIATGSPPIAGTSDLSTITSAATPTFTTEQDGREQGDGLQLAHDFSQPVITVIILGVMA
GIIGIILLLAYVSRRLRKRPPADVPPPASTVPSADAPPPVSEDDETSLTSVETDYPGDSQ
```

First, run a Needle with default values.
Paste the human sequence at the first box and the horse sequence at the second one (you can also enter either an accession number or upload a sequence file). Don't forget to give informative names for each sequence as they will be presented at the final output.

Both programs (Needle and Water) use the affine gap penalty that is constructed of two parts:
Gap opening penalty and gap extension penalty using the following formula: total gap cost is equal to gap opening penalty plus gap extension penalty multiplied by the number of gaps minus one.

For proteins, the default value for gap opening penalty is 10 and for gap extension penalty is 0.5.

Notice that both parameter values can be changed.

The substitution matrix default value is Bloum62 and can also be changed.
In this example we will use a PAM matrix to score the global alignment between the two closely related proteins sequences. Therefore, change the substitution matrix to EPAM100.

Now click on the Run Needle button. The alignment will appear on the next page.

Before analyzing the result let's run Needle again and change the substitution matrix to EPAM200. Finally, run the program for the third time and change the substitution matrix to EPAM50.

Let's go to the first output page where we used the PAM100 as the substitution matrix value. The input parameters used are presented at the middle of the page. You can see that we ran PAM100 with gap penalty equaling 10 and extension equaling 0.5. The alignment appears at the end of the page where the entire length of each sequence is aligned as we run global alignment. The human sequence appears at the upper line and the horse sequence appears at the lower one. Since the alignment is too long to be presented in a single line, the alignment is broken into segments of 50 letters each. The numbers on the left and right sides of each segment represent the location of the first and last amino acid in the original sequence. The numbers that appear at the right side in the last segment represent the location of the last amino acids at each sequence and therefore represent sequence length. So we can see that the human sequence length is 150 amino acids and the horse sequence length is 120 amino acids.

The gap in the alignment is represented by a hyphen,
identity is represented with a solid line,
high similarity with two dots
and low or no similarity with one dot.
A good alignment should include a large area with a lot of lines and two dots.

The alignment length appears at the middle part of the page and counts all gaps and letters in the upper or lower line of the alignment. The length here is 170 which is larger than the length of the human sequence as it includes also gaps that have been introduced in both sequences.

The score of the alignment is 213 but this number has no meaning by itself. Rather, calculated percentages will provide meaningful information.

The percent of Identity is the number of identical amino acids (represented by lines) divided by the alignment length, which in our case gives 28.8% identity.

The percent of similarity includes the number of amino acids with high similarity (represented by two dots) plus the number of identical amino acids divided by the alignment length, which gives 39.4% similarity.

The percent of gaps is the number of gaps divided by the alignment length, which gives 41.2%.

Let's see how changing the substitution matrix affects the alignment:

Go to the second output page where the substitution matrix value was changed to 200.

Note that the higher the numbers in the PAM matrix name, the greater the evolutionary distance (for example Pam 200 is used for more distant sequences than PAM100). Therefore, the percent of identity is approximately the same as in the previous output page but the similarity percentages is much higher now as amino acids that were not considered matching at the PAM100 have now a higher similarity score. For that reason, we also get fewer gaps as they are replaced by similarities.

Look at the third output page where the substitution matrix value is 50. The percent of identity still stays the same as in the first output page where the substitution matrix value is 100. However, the percent of similarity is lower and equals 35.8%. Whereas more gaps were introduced giving a high gap percentage value of 46.6% in comparison to the 33.3% that was received at the first alignment. Thus, we can see that the selection of the weight matrix does have an effect on the results.

Now, run the sequences in Water for local alignment.
First, run the sequences with the default parameter values and compare it to the global alignment.

For the local region, we can see that the local alignment gives better alignment, more specifically - there are fewer gaps because we don't force the alignment to include the entire sequence but just to locate the part in which the sequence most resembles one another. The percentages of similarity and identity are also higher in comparison to the global alignment. Please note that the numbers at the end of the alignment don't represent the sequences original lengths anymore, as the alignment did not include the entire sequences but just aligned parts of them.

How can we know whether to use global or local alignment?
How should we decide which weight matrix and gap penalty to use?
And do the results have biological meaning?
The truth is that there is no one good answer for these questions. It largely depends on the sequences and on the experience of the user. What I can suggest is to try several alternatives. If the results are somewhat different but basically consistent, it is a good sign to adopt the results and consider their biological significance. If, however, the results change drastically when a moderate change is made to the selected parameters, then you should be very cautious when considering if indeed the results are meaningful.