Aligning coding sequence with MACSE

MACSE provides various subprograms to allow complex alignment strategies. The most direct approach is to simply use the alignSequences subprogram. Here we present some examples where this suprogram is used to solve frequent problems (e.g. alignment including pseudogenes, or NGS contigs)

Basic examples of coding sequence alignments

example 1

The files and command lines for the following examples are available in the 01_basicAlignment directory of the demo archive.If you have a small set of coding sequences you can simply use MACSE with defaut options. The only required option is the name of the input alignment file.

macse.jar -prog alignSequences -seq tmem184a.fasta

example 2

In previous example the name of the output files are generated automatically based on the name of the input file. Alternatively you can provide the name of the two output files:

- the nucleotidic alignment (-out_NT)
- the amino acid alignment (-out_AA)

macse.jar -prog alignSequences -seq tmem184a.fasta -out_NT tmem184a_nuc.aln -out_AA tmem184a_amino.aln

example 3

If your sequences do not use the standart genetic code, you can indicate the default genetic code (def_gc) that MACSE should uses to align your sequences for (def_gc)For instance for MATK chloroplatic gene from Poacea you need to specify the genetic code 11.The genetic codes are those provided by the NCBI (http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi)

macse.jar -prog alignSequences -seq poacea_matk.fasta -def_gc 11

Note that MACSE alignment is correct even thougt some sequences do not start on the first position of a codon since MACSE is able to handle those problematic cases by introducing a frameshift in the first and last codon (such frameshifts are not strongly penalized).

Aligning dataset with a mix of genes and pseudogenes.

MACSE can take two different sequence files as input "-seq" and "-seq_lr" containing sequences refered respectively as "reliable" and "less reliable" sequences. You can then use different stop and frameshift costs for those two kinds of sequences using options "fs" and "stop" to assign cost to standart sequences and option "fs_lr" and "stop_lr" to assign cost to the so called less reliable sequences.

example 1

The files and command lines for the following examples are available in the 02_aligningPseudogenes directory of the demo archive.If you have a small set of sequences and have no particular idea of which ones are protein coding sequences and which ones are pseudogene sequences, you can simply use MACSE with default options (stop and frameshift have a penalty of 7.0 30.0.

macse.jar -prog alignSequences -seq AMELX.fasta

macse.jar -prog alignSequences -seq ENAM_all.fasta

example 2

In previous example the same penalties are used for all sequences. If you know which sequence are prodeins coding sequences and which are pseudogenes, it is preferable to have them in different files that so that you can specify to MACSE which file contains the coding sequences and which one contain the pseudogenes. You can hence let MACSE use different cost for the sequences from the two files assigning standart cost to one file (option "fs" and "stop") and different cost for others (option "fs_lr" and "stop_lr"). For pseudo-genes "-fs_lr 10.0 -stop_lr 10.0" is often a good initial parameter set.

macse.jar -prog alignSequences -seq ENAM_genes.fasta -seq_lr ENAM_pseudos.fasta -fs_lr 10.0 -stop_lr 10.0

Since frameshift and stop cost are lower for identified pseudo-gene sequences, MACSE is able to detect more of them, hence producing a better alignment (see Figure Below).

Aligning NGS data (reads or contigs).

MACSE can take two different sequence files as input "-seq" and "-seq_lr" containing sequences refered respectively as "reliable" and "less reliable" sequences.

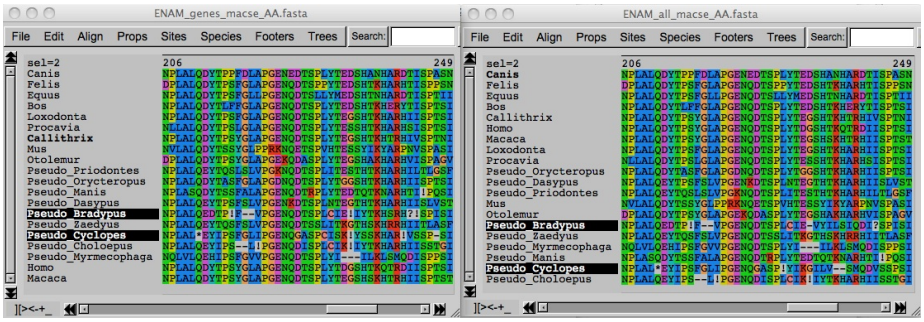

Figure 1: Using lower frameshift and stop penalties for identified pseudo-genes (left alignment) allows to produce better alignment than if all sequences are conidered equivalent (right alignment). This can be seen, for instance, in this alignment for Bradypus and Cyclopes sequences, which have an additional frameshift in the left alignment.

You can then use different stop and frameshift costs for those two kinds of sequences using options “fs” and “stop” to assign cost to standart sequences and option “fs_lr” and “stop_lr” to assign cost to the so called less reliable sequences.

example 1

The files and command lines for the following examples are available in the 03_aligningNGSContigs directory of the demo archive.When aligning a set of sequence containing NGS data it is preferable, as for pseudogenes, to split the sequences in two files and to use costs for frameshift and stop codon taking into account this heterogeneity. For NGS raw sequences (reads or contigs) “-fs_lr 10.0 -stop_lr 10.0” is often a good initial parameter set. You can of course adapt those cost depending on the sequencing technology since some induce more errors in homopolymers lengths (seen as frameshift by MACSE) while others induce more nucleotide reading errors (which can cause a standart codon into a STOP codon)..The ideal case is to have a set or reliable sequences (i.e. from public databases or manually annotated) that you can provide to MACSE to guide the alignment process. In this case we identify that the reads are similar to ensembl sequences ENSG00000119777 (TMEM214) and use orthologuous sequences to guide read alignments and hence detecting errors within them:

macse.jar -prog alignSequences -seq TMEM214.fasta -seq_lr TMEM214.fasta -fs_lr 10.0 -stop_lr 10.0”

Note that the more reliable sequence you have the better it is. If you collect numerous reliable sequences or have a clean alignment corresponding to your current contigs, you can use the enrichAlignment to rapidly add your contig to your initial alignmnent (see corresponding sections for detail exemples)

Transforming coding sequence alignments with MACSE

MACSE provides subprograms, which allows to split or merge alignments

Extracting a fraction of coding sequence alignments

example 1

The files and command lines for the following examples are available in the 01_splitAlignment directory of the demo archive.If you have an alignment and you can restrict it to a subset of sequences the names of which are listed in a separated file.For instance considering tmem184a_NT.aln, which contains an alignment for 20 mammals you can restrict it to the 5 primate sequences specified in primate_seqId.list.

macse.jar -prog splitAlignment -align tmem184a_NT.fasta -subset primate_seqId.list

Aligning coding sequence with MACSE

MACSE provides various subprograms to allow complex alignment strategies. The most direct approach is to simply use the alignSequences subprogram. Here we present some examples where this suprogram is used to solve frequent problems (e.g. alignment including pseudogenes, or NGS contigs)

Basic examples of coding sequence alignments

example 1

The files and command lines for the following examples are available in the 01_basicAlignment directory of the demo archive.If you have a small set of coding sequences you can simply use MACSE with defaut options. The only required option is the name of the input alignment file.

macse.jar -prog alignSequences -seq tmem184a.fasta

example 2

In previous example the name of the output files are generated automatically based on the name of the input file. Alternatively you can provide the name of the two output files:

- the nucleotidic alignment (-out_NT)
- the amino acid alignment (-out_AA)

macse.jar -prog alignSequences -seq tmem184a.fasta -out_NT tmem184a_nuc.aln -out_AA tmem184a_amino.aln

example 3

If your sequences do not use the standart genetic code, you can indicate the default genetic code (def_gc) that MACSE should uses to align your sequences for (def_gc)For instance for MATK chloroplatic gene from Poacea you need to specify the genetic code 11.The genetic codes are those provided by the NCBI (http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi)

macse.jar -prog alignSequences -seq poacea_matk.fasta -def_gc 11

Note that MACSE alignment is correct even thougt some sequences do not start on the first position of a codon since MACSE is able to handle those problematic cases by introducing a frameshift in the first and last codon (such frameshifts are not strongly penalized).

Aligning dataset with a mix of genes and pseudogenes.

MACSE can take two different sequence files as input "-seq" and "-seq_lr" containing sequences refered respectively as "reliable" and "less reliable" sequences. You can then use different stop and frameshift costs for those two kinds of sequences using options "fs" and "stop" to assign cost to standart sequences and option "fs_lr" and "stop_lr" to assign cost to the so called less reliable sequences.

example 1

The files and command lines for the following examples are available in the 02_aligningPseudogenes directory of the demo archive.If you have a small set of sequences and have no particular idea of which ones are protein coding sequences

and which ones are pseudogene sequences, you can simply use MACSE with default options (stop and frameshift have a penalty of 7.0 30.0.

macse.jar -prog alignSequences -seq AMELX.fasta

macse.jar -prog alignSequences -seq ENAM_all.fasta

example 2

In previous example the same penalties are used for all sequences. If you know which sequence are prodeins coding sequences and which are pseudogenes, it is preferable to have them in different files that so that you can specify to MACSE which file contains the coding sequences and which one contain the pseudogenes. You can hence let MACSE use different cost for the sequences from the two files assigning standart cost to one file (option "fs" and "stop") and different cost for others (option "fs_lr" and "stop_lr"). For pseudo-genes "-fs_lr 10.0 -stop_lr 10.0" is often a good initial parameter set.

macse.jar -prog alignSequences -seq ENAM_genes.fasta -seq_lr ENAM_pseudos.fasta -fs_lr 10.0 -stop_lr 10.0

Since frameshift and stop cost are lower for identified pseudo-gene sequences, MACSE is able to detect more of them, hence producing a better alignment (see Figure Below).

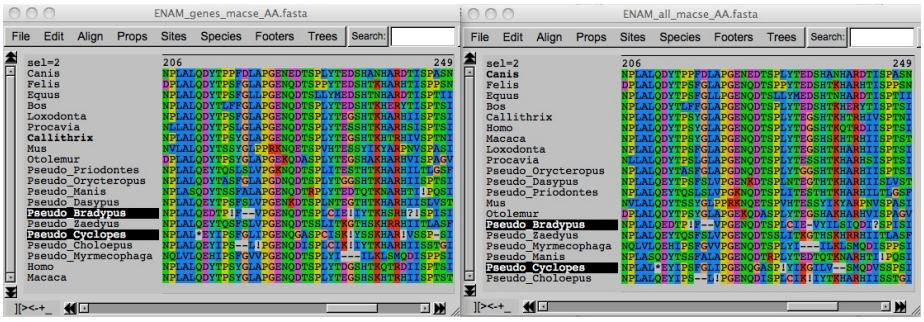

Figure 2: Using lower frameshift and stop penalties for identified pseudo-genes (left alignment) allows to produce better alignment than if all sequences are conidered equivalent (right alignment). This can be seen, for instance, in this alignment for Bradypus and Cyclopes sequences, which have an additional frameshift in the left alignment.

Aligning NGS data (reads or contigs).

MACSE can take two different sequence files as input "-seq" and "-seq_lr" containing sequences refered respectively as "reliable" and "less reliable" sequences.

You can then use different stop and frameshift costs for those two kinds of sequences using options “fs” and “stop” to assign cost to standart sequences and option “fs_lr” and “stop_lr” to assign cost to the so called less reliable sequences.

example 1

The files and command lines for the following examples are available in the 03_aligningNGSContigs directory of the demo archive.When aligning a set of sequence containing NGS data it is preferable, as for pseudogenes, to split the sequences in two files and to use costs for frameshift and stop codon taking into account this heterogeneity. For NGS raw sequences (reads or contigs) “-fs_lr 10.0 -stop_lr 10.0” is often a good initial parameter set. You can of course adapt those cost depending on the sequencing technology since some induce more errors in homopolymers lengths (seen as frameshift by MACSE) while others induce more nucleotide reading errors (which can cause a standart codon into a STOP codon)..The ideal case is to have a set or reliable sequences (i.e. from public databases or manually annotated) that you can provide to MACSE to guide the alignment process. In this case we identify that the reads are similar to ensembl sequences ENSG00000119777 (TMEM214) and use orthologuous sequences to guide read alignments and hence detecting errors within them:

macse.jar -prog alignSequences -seq TMEM214.fasta -seq_lr TMEM214.fasta -fs_lr 10.0 -stop_lr 10.0”

Note that the more reliable sequence you have the better it is. If you collect numerous reliable sequences or have a clean alignment corresponding to your current contigs, you can use the enrichAlignment to rapidly add your contig to your initial alignmnent (see corresponding sections for detail exemples)

Transforming coding sequence alignments with MACSE

MACSE provides subprograms, which allows to split or merge alignments

Extracting a fraction of coding sequence alignments

example 1

The files and command lines for the following examples are available in the 01_splitAlignment directory of the demo archive.If you have an alignment and you can restrict it to a subset of sequences the names of which are listed in

a separated file.For instance considering tmem184a_NT.aln, which contains an alignment for 20 mammals you can restrict it to the 5 primate sequences specified in primate_seqId.list.

macse.jar -prog splitAlignment -align tmem184a_NT.fasta -subset primate_seqId.list