ADINIS*multiplugin*G Version 1.00 specification

1. Runtime environment

The plugin has been tested on the following operating systems:

- Windows (XP, Vista, 7)
- Ubuntu 64 bit (v. 10.10)

The plugin has been tested under Geneious software v5.6.2 running on Java Runtime Environment v1.6

2. Multiple alignment column filter

__Input of the filter__**:** Multiple alignment, the first sequence has to be a reference sequence that might contain an annotation. **The amplicons covering the same area of the reference sequence need to have approximately the same size and position** (Mathematically each two reads that should be processed together need to satisfy the amplicon subset formula defined in the Preprocessing step paragraph for threshold set to 90%, the threshold can be changed).

__Preprocessing step__**:** The goal of this step is to separate the non-reference sequences into sets that cover approximately the same area and then split them into amplicon sets (if they overlap). This identifies the set of rows that are relevant for each position of the reference sequence. **The filter is then iteratively applied on the individual sets not using sequences from the other sets for calculations.** If two sequences overlap (at positions between leading and trailing gaps), they belong to the same set. If two sequences from two different sets overlap, they are merged so all of the sequences from these two sets now belong to the same set**. The sets are then split into additional subsets representing amplicon sets covering different exons (if they overlap).** The amplicon subsets are calculated from the mutual overlap of the reads, i.e. if one read overlaps more than the specified percentage (provided in options) of the other read and vice versa, they belong to the same amplicon subset. The longest reads are examined first. **Amplicon subset formula:** for reads r1,r2 :

IF [overlap(r1,r2) / length(r1) > threshold & overlap(r1,r2)) / length(r2) > threshold] THEN r1 and r2 belong to the same amplicon subset (cover the same exon),

where overlap(r1,r2) = max(0, min(end(r1), end(r2)) – max(start(r1), start(r2))).

If the overlap part in both reads does not satisfy the threshold, then they belong to different amplicon subsets (cover different exons). **The threshold is chosen in the options window** (default 90%). If more than one amplicon subset has been found and there are some short reads located inside the overlapping section, an error message will appear containing indexes of the problematic reads that have to be deleted before applying the filter.

__Filter options (displayed in the options window):__

- __Maximal % of gaps in columns, where gap is in the ref. seq.__
  Definition of the condition: *# of gaps in non-ref sequences / # of non ref sequences <= specified %.* **Note that the reads that have leading/ trailing gaps in the examined position are not considered in this formula.** If the given column has a gap in the ref. seq. and the number of non-ref. gaps in this column is higher than the specified

threshold then the column will be deleted. The filter can be disabled when the value is set to 100 %. If more amplicon subsets are in this position, the gap is removed if and only if the condition is not satisfied for all of the amplicon subsets.

- Minimal % of non-reference bases/gaps in columns that do not match the reference base

  Definition of the condition: *# of non-matching positions / # of non-ref sequences >= given %.* **Note that the reads that have leading/ trailing gaps in the examined position are not considered in this formula.** Each **c**olumn that does not satisfy this condition is replaced by an identity column that consists of the reference nucleotide (**bases are replaced only in the relevant reads for this amplicon subset, i.e. positions with leading/trailing gaps are not replaced**). Columns with a gap in the reference sequence are not processed by this filter. The filter can be disabled when the value is set to 0 %.

  - Option "Count gaps and N bases as non-matching bases" (checkbox, default true)
    If checked, all the gaps and N bases are treated as non-matching bases (this is default and is equal to the previously stated filter condition), otherwise they will be not taken into account when calculating the percentage of the non-matching nucleotides. If not checked the definition of the filter changes to :
    # of non-matching positions - (gaps + N bases)/ # of non-ref sequences - (positions that contain gaps and N bases) >= given %. **Note that the reads that have leading/ trailing gaps in the examined position are not considered in this formula.**

Output: The output of the filter is a new alignment that satisfies the conditions given in Section Filter options. If the reference sequence contains an annotation, the output reference sequence also contains an annotation that is justified according to the positions of the removed columns. The output alignment also contains information about the positions of the removed columns.

3. Statistics graph

The statistics graph displays the positions of the filtered columns. The top part of the graph displays positions of removed insertions, the bottom part of the graph displays the number of bases replaced by the reference base. By moving the mouse over the graph, the tooltip window shows the statistics (specified below) about the column(s) in the same position (the statistics are calculated from the row set for the given position). The tooltip behaves as a standard window (can be moved) and will disappear only if the user clicks on its "X" button.

Displaying removed columns

If one or more columns have been removed between two bases and the zooming is sufficient to see both of these bases then the graph paints an orange rectangle(line) in the position between the two bases. If the zooming is not sufficient and the reference sequence displays intervals instead of individual bases, then positions in intervals, where one or more columns were displayed, is shown by a red line in the statistics graph. If the column removal happened between the intervals it will be displayed by an orange line in the statistics graph.

Displaying the number of bases replaced by the reference base

The number of bases replaced by the reference base is displayed in the bottom part of the graph. The graph draws a rectangle above the reference base that has a color ranging (color gradient) from green (0% of replaced bases) to red (max % of replaced bases set in the filter options).

Displaying statistics

The tooltip window displays the statistics of the **relevant sequences** for the given position in the reference sequence. **The statistics are displayed separately for each amplicon subset (two amplicon subsets double the tooltip size) in this position, the order is determined by the start position of amplicon subsets, i.e. the start of the left-most read.** The relevant sequences are selected in the same way as in the preprocessing step of the Multiple alignment column filter.

The tooltip structure :

# of residues (the number of residues on the position, may be larger than 1 if unzoomed)
Reference (reference sequence on the position)
# of relevant sequences (total, forward, and reverse separately)
% of gaps (total, forward, and reverse separately)
# of removed columns (before, inside, after)
# of replaced substitutions (replaced, maximum allowed number chosen in filter)
Base % of non-gaps (total, forward, and reverse separately)

Here is an example of an actual tooltip:

```
# of residues  1
Reference   T
# of relevant sequences  305 (f:166 r:139)
% of gaps  30.5% (f: 20% r: 10.5%) 1; 0; 2 //
# of replaced substitutions  10 , max:35
 Base   % of non-gaps
  A   4.92% (f: 3.93% r: 0.98%)
  C  91.48% (f:41.31% r:50.16%)
  G   0.00% (f: 0.00% r: 0.00%)
  T   3.61% (f: 0.33% r: 3.28%)
  N   0.00% (f: 0.00% r: 0.00%)
```

4. Sorting and filtering the sequences (forward, reverse)

Input: A multiple sequence alignment, the first sequence can be a reference sequence (configured in options).

Filter options (displayed in options window):

a. 1st sequence reference? (checkbox, default true).
b. Type of operation (see next paragraph).
c. Forward sequences first ? (checkbox, default true, only for sorting operation).

This plugin part enables the following operations (output):

a. Creating a new alignment that consists of either forward or reverse sequences and the reference sequence in the first place if chosen in options (the annotation of the reference sequence will be copied if present).
b. Sorting the selected sequence starting with forward or reverse sequences (based on the input provided in options). The reference sequence will stay in the first place if selected in options. No new sequence is created (**the sequence has to be saved in order to keep the ordering afterwards**).

5. A "rank" list of the detected deviations to quantify forward and reverse coverage and percentage deviation

Input: A multiple sequence alignment, the first sequence has to be a reference sequence.

Output: A table displays statistics about significant columns - same as in the tooltip window, excluding the number of residues (irrelevant) and # of removed columns. 1st column contains position + amplicon subset number if more amplicon subsets are present in the column. The table has checkboxes to show/hide insertions and substitutions. Substitutions and insertions are ordered (descending) by the number of bases that differ from the reference base(in case of insertions the reference base is gap).