Poster H-49

New EST trimming procedure applied to SUCEST sequences

Authors:
Christian Baudet (*Institute of Computing - Unicamp*)
Zanoni Dias (*Institute of Computing - Unicamp*)

Short Abstract: In order to improve EST trimming, we proposed a new set of procedures to detect regions that do not belong to the sequenced organism. We evaluated the proposed method using SUCEST ESTs. Based on the results, we concluded that our method suits projects that want to produce more reliable clusters.

Long Abstract:

In order to improve the EST trimming procedure and, as result, the clustering, we built a new set of procedures to detect regions that do not belong to the sequenced organism, or have low quality.

In this new set, the identification of different types of artifacts is independently made. This means that the detection of one artifact has no effect on further detections.

The following steps are performed: ribosomal sequences discard, low quality identification, vector identification, adapter identification, poly-A/T tail identification, slipped sequence identification and,
finally, good sequence identification.

Ribosomal sequences detection is performed through BLAST of all ESTs against a database, which is populated with ribosomal sequences of organisms that are phylogenetically close to the sequenced
organism. Every sequence that shows a hit with e-value lower than or equal to 1e-10 is discarded.

Low Quality identification is performed in two steps. In the first step, the sequence is processed by a maximum subsequence algorithm, similar to the one implemented by phred. In that algorithm, we set 11
to the minimum quality threshold. After the maximum subsequence being found, we use a 10 bases sliding window to search regions that have an average probability error higher than or equal to 20%. For each found region, we cut the sequence into two parts at the position of minimum quality. Each part is processed by the maximum subsequence algorithm.

Vector trimming is performed by the software cross_match to align all sequences with the vector sequence. The alignment is made with the
parameters "-minmatch 12" and "-minscore 20". The resulting aligned regions are considered as vector regions.

Adapter detection uses the software swat. All sequences are aligned with the adapter

sequence using the parameters "-gap_init -5", "-gap_ext -5", "-ins_gap_ext -5", "-del_gap_ext -5", "-end_gap -5" and a score matrix that scores every match with 1 and every mismatch with -2. The alignments whose size is greater than or equal to the adapter size minus 4 are identified as adapter artifacts.

Swat is also used in poly-A/T tail detection. Using the same above parameters, the sequence are aligned with 500 bp sequences of "A"s or "T"s. Alignment regions showing scores of at least 10 are considered
as poly-A/T tail artifacts.

The slipped sequence detection is based on consecutive identical bases regions identification. If a region in the sequence comprises at least 5 consecutive identical bases, it is identified as echo region,
otherwise as normal region. Any subsequence composed by at least 8 echoes regions that represent 25% or more of its all regions was considered a slippage artifact.

After identifying all artifacts above, our trimming procedure ends with the identification of the sequence that will be used for clustering.

All artifacts found are masked and all non-masked regions are analyzed. Every region that has size lower than 100 bases are discarded. Only the non-masked region that has the highest quality sum is preserved. If there are two regions with the same sum, the method chooses the greatest one. Finally, if both regions have the same size, the method chooses the one that is closer to the 5' end.

To evaluate our trimming procedure, we worked with the ESTs of the Sugarcane EST Project - SUCEST. We trimmed the sequences and clustered them using the software cap3.

The computer used to cluster the sequences had limited memory, so we had to select a smaller set of the 291,689 SUCEST ESTs.

We sorted all sequences by name and selected those that are in odd positions. By doing this, we selected, approximately, half of the sequences of each library. Finally, 145,845 ESTs are selected to be
processed and clustered. The sequences of this set have an average size of 834.64 +/- 182.26 bases and average quality of 23.08 +/-15.67.

We processed this set of sequences using both our trimming method and the trimming method adopted by the SUCEST project, in order to make comparisons. SUCEST trimming procedure generated 118,991 good sequences with average size of 643.82 +/- 141.32 bases and average quality of 27.69 +/- 15.39. Our set of trimming methods generated 126,986 good sequences with average size of 473.33 +/- 121.66 and average quality of 33.25 +/- 13.15.

After trimming the sequences, we clustered both sets of processed sequences using cap3 with default parameters. The set of sequences processed by the SUCEST methods generated a clustering (TS) with
20,202 singletons and 16,394 contigs. The set processed with our methods generated a clustering (BD) with 22,479 singletons and 17,486

contigs.

We compared both clusterings by analyzing external consistency, internal consistency, redundancy and number of full-length clusters.

Proportionally to the number of clusters of each clustering, external consistency shown that the clustering BD has a number of clusters
overlaps lower than that shown by clustering TS. This could indicate that our clustering separates clusters that, in fact, must be assembled together, with a frequency lower than that one of clustering TS.

Internal consistency evaluated the occurrence of discrepant bases in the alignment of the sequences of a cluster. Our clustering shown better results.

Redundancy analysis revealed that redundancy of clustering BD is smaller than that shown by clustering TS. This result reinforce the conclusion of the external consistency analysis.

The number of full-length clusters of clustering BD is lower than the one shown by the clustering TS. This may be consequence of the average size decrease of the sequences processed by our methods.

The main difference between the two trimming sets is due to our method having a more stringent low quality trimming. This characteristic makes the clusters more reliable, but the number of full-length clusters decreases.

Based on this analysis, we conclude that our trimming procedure suits projects that have the objective of producing more reliable clusters. If the objective is to discover full-length clusters for gene annotation, we recommend a reduction of the low quality trimming
stringency.