Assembly Sequences with CAP3

CAP3 is a contig assembly program. It allows to assembly long DNA reads (up to 1000 bp). Binaries can be downloaded from http://seq.cs.iastate.edu/cap3.html Huang, X. and Madan, A. (1999) CAP3: A DNA Sequence Assembly Program, Genome Research, 9: 868-877.

Parameters in GUI

Parameter	Description	Default value
Output file	Write assembly results to this output file in ACE format..	result.ace
Quality cutoff for clipping	Base quality cutoff for clipping (-c).	12
Clipping range	Set a number which unit is base. It will get the refGenes in n bases from peak center. (--distance).	100
Quality cutoff for differeneces	Base quality cutoff for differences (-b).	20
Maximum difference score	Max qscore sum at differences (-d). If an overlap contains lots of differences at bases of high quality, then the overlap is removed. The difference score is calculated as follows. If the overlap contains a difference at bases of quality values q1 and q2, then the score at the difference is max(0, min(q1, q2) - b), where b is Quality cutoff for differences. The difference score of an overlap is the sum of scores at each difference.	200
Match score factor	Match score factor (-m) is one of the parameters that affects similarity score of an overlap. See Overlap similarity score cutoff description for details.	2
Mismatch score factor	Mismatch score factor (-n) is one of the parameters that affects similarity score of an overlap. See Overlap similarity score cutoff description for details.	-5
Gap penalty factor	Gap penalty factor (-g) is one of the parameters that affects similarity score of an overlap. See Overlap similarity score cutoff description for details.	6
Overlap similarity score cutoff	If the similarity score of an overlap is less than the overlap similarity score cutoff (-s), then the overlap is removed. The similarity score of an overlapping alignment is defined using base quality values as follows. A match at bases of quality values q1 and q2 is given a score of m * min(q1,q2), where m is Match score factor. A mismatch at bases of quality values q1 and q2 is given a score of n * min(q1,q2), where n is Mismatch score factor. A base of quality value q1 in a gap is given a score of -g * min(q1,q2), where q2 is the quality value of the base in the other sequence right before the gap and g is Gap penalty factor. The score of a gap is the sum of scores of each base in the gap minus a gap open penalty. The similarity score of an overlapping alignment is the sum of scores of each match, each mismatch, and each gap.	900
Overlap length cutoff	An overlap is taken into account only if the length of the overlap in bp is no less than the specified value (parameter -o of CAP3).	40
Overlap percent identity cutoff	An overlap is taken into account only if the percent identity of the overlap is no less than the specified value (parameter -p of CAP3).	90
Max number of word matches	This parameter allows one to trade off the efficiency of the program for its accuracy (parameter -t of CAP3). For a read f, CAP3 computes overlaps between read f and other reads by considering short word matches between read f and other reads. A word match is examined to see if it can be extended into a long overlap. If read f has overlaps with many other reads, then read f has many short word matches with many other reads. This parameter gives an upper limit, for any word, on the number of word matches between read f and other reads that are considered by CAP3. Using a large value for this parameter allows CAP3 to consider more word matches between read f and other reads, which can find more overlaps for read f, but slows down the program. Using a small value for this parameter has the opposite effect.	300
Band expansion size	CAP3 determines a minimum band of diagonals for an overlapping alignment between two sequence reads. The band is expanded by a number of bases specified by this value (parameter -a of CAP3).	20
Max gap length in an overlap	The maximum length of gaps allowed in any overlap (-f). I.e. overlaps with longer gaps are rejected. Note that a small value for this parameter may cause the program to remove true overlaps and to produce incorrect results. The parameter may be used to split reads from alternative splicing forms into separate contigs.	20

Assembly reverse reads	Specifies whether to consider reads in reverse orientation for assembly (originally, parameter -r of CAP3).	True
CAP3 tool path	The path to the CAP3 external tool in UGENE.	default
Temporary directory	The directory for temporary files.	default

Parameters in Workflow File

Type: cap3

Parameter	Parameter in the GUI	Type
out-file	**Output file**	*string*
clipping-cutoff	**Quality cutoff for clipping**	*numeric*
clipping-range	**Clipping range**	*numeric*
diff-cutoff	**Quality cutoff for differeneces**	*numeric*
diff-max-qscore	**Maximum difference score**	*numeric*
match-score-factor	**Match score factor**	*numeric*
mismatch-score-factor	**Mismatch score factor**	*numeric*
gap-penalty-factor	**Gap penalty factor**	*numeric*
overlap-sim-score-cutoff	**Overlap similarity score cutoff**	*numeric*
overlap-length-cutoff	**Overlap length cutoff**	*numeric*
overlap-perc-id-cutoff	**Overlap percent identity cutoff**	*numeric*
max-num-word-matches	**Max number of word matches**	*numeric*
band-exp-size	**Band expansion size**	*numeric*
max-gap-in-overlap	**Max gap length in an overlap**	*numeric*
assembly-reverse	**Assembly reverse reads**	*boolean*
path	**CAP3 tool path**	*string*
tmp-dir	**Temporary directory**	*string*

Input/Output Ports

The element has 1 *input port*:

Name in GUI: Input sequences

Name in Workflow File: in-data

Slots:

Slot In GUI	Slot in Workflow File	Type
Dataset name	**dataset**	*string*
Input URL(s)	**in.url**	*string*