CONTIGuator

Introduction

CONTIGuator is a Python script for Linux environments whose purpose is to speed-up the bacterial genome finshing process, taking advantage of the high number of near genomes that can be used to align and resolve the relative position of the contigs obtained with the latest sequencing technologies and therefore to design a set of PCR primers in order to fill the gaps and take a step further in the finishing process. It also can be used to obtain a first insight of the genome structure using the well-known artemis comparison tool (ACT).

CONTIGuator uses the megaBlast algorithm to create a so-called “contig profile”, where each contig and the regions of the reference genomes are divided into regions of high similarity; this results in an higher number of PCR primers, generated by a run of ABACAS using primer3 and Mummer.

The outputs of the program can be visualized with the Artemis comparison tool (ACT), where the user can obtain a clear insight into the structural genomic features of the draft genome (a result of the contig profiling step) and visualize the position and lenght of the putative PCR products. Moreover, if the BioPython version used is above 1.58, one publication-quality pdf map will be produced for each putative replicon.

For the impatient

- **Default run WITHOUT primer picking:**
    - python CONTIGuator.py -r references.fna -c contigs.fna
- **Default run WITH primer picking:**
    - python CONTIGuator.py -r references.fna -c contigs.fna -P -A
- **Default run with many outputs:**
    - Add the -M option
- **Default run showing the maps automatically:**
    - Add the -l option
- **Default run with no N to fill the gaps:**
    - Add the -N option

Cite the program

If you are willing to use CONTIGuator, please consider adding the following citation to your manuscript: *CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes* Marco Galardini, Emanuele G Biondi, Marco Bazzicalupo and Alessio Mengoni Source Code for Biology and Medicine 2011, 6:11 doi:10.1186/1751-0473-6-11

Requirements

CONTIGuator was developed and tested in a Linux environment and has a few software requirements, listed below with their websites:

- Python (python.org)
- BioPython (biopython.org)
- Blast+ (www.ncbi.nlm.nih.gov/BLAST/)

- Perl (perl.org) (optional)
- ABACAS (abacas.sourceforge.net) (optional)
- MUMmer (mummer.sourceforge.net) (optional)
- Primer3 (primer3.sourceforge.net) (optional)

To view the results the Artemis comparison tool (ACT) is needed: it can be both downloaded or launched as a Java web-applet (sanger.ac.uk).

To take advantage of the tblastn search of reference proteins from unmapped regions in the excluded contigs, one or more ptt files should be present in the same directory as the reference genome FASTA file; there should be one ptt file for each reference replicon. If the reference genome FASTA file was downloaded from NCBI, the ptt file name is already in the right format, otherwise the ptt file name should be in the format SEQUENCEID.ptt.

Command line options

Inputs

- c fasta file containing all the contigs
- r fasta file containing the reference genome; if it contains more than one sequence (i.e. more than one replicon) it can both be in a single file or in more files (just use -r multiple times)
- f prefix for the output directories

Blast parameters

- e Blast e-value treshold [Default: 1e-20]
- -b Use the blastn algorithm instead of megablast (for distant genomes)
- -t Threads to be used by Blast (useful for large genomes)

“Parse Blast” mode

- p parse ready-made Blast output (no Blast runs will be performed) [Default: no]
- x Blast XML output to be parsed [Default: blast.xml]

Contig profiling parameters

- L minimal lenght of a contig to be accepted in the analysis [Default: 1000 bp]
- C minimal coverage of a contig to be accepted in the analysis [Default: 20%]
- B minimal lenght of a significant Blast hit [Default: 1100 bp]

Primer picking

- P do primer picking [Default: no]
- A use default parameters for primer picking [Default: no]

Output options

- f prefix for the output directories
- M prepare even more outputs
- n How many Ns should be used to fill the gaps
- N Do not use N to fill the gaps

ACT options

- a ACT binary location
- l Open the maps automatically

Logging

- V verbose
- D development
- G debug mode

Outputs

The outputs of CONTIGuator are various files and they are divided by folders

“Map_” folders According to the number of input reference replicons, there will be the same number of directories whose name starts by “Map_”, followed by the ID of the reference replicon. Inside each directory there will be a series of files that can be used as input for ACT

- Reference.embl: pseudo-contig ACT file
- PseudoContig.fsa: pseudo-contig fasta file
- PseudoContig.crunch: ACT comparison file
- PseudoContig.embl: pseudo-contig ACT file
- MappedContigs.txt: names (and lenghts) of the contigs mapped to the particular reference molecule
- A shell script to open the ACT map

If Biopython version is above 1.58:

- A pdf file containing the "manual" version of the map viewable with ACT (publication quality)

If options -M was selected:

- AlignDetails.tab: tab-delimited file containing details about the mapped hits
- AlignedContigsHits.fsa: mapped hits fasta file (on contigs)
- UnAlignedContigsDetails.tab: details on unmapped regions (on contigs)
- UnAlignedReferenceDetails.tab: details on unmapped regions (on reference)
- AlignedReferenceHits.fsa: mapped hits fasta file (on reference)
- UnAlignedContigsHits.fsa: unmapped regions fasta file (on contigs)
- UnAlignedReferenceHits.fsa: unmapped regions fasta file (on reference)

If the primer picking option was selected (-P) the folder will contain other files

- PCRPrimers.tsv: table containing details about the PCR primers generated

“UnMappedContigs” folder This folder contains those contigs that CONTIGuator was unable to map in fasta format, divided in categories

- Excluded.fsa: all the excluded contigs
- Multi.fsa: contigs mapped to more than one replicon
- Short.fsa: contigs below the lenght treshold (-L)
- NoCoverage.fsa: contigs below the coverage treshold (-C)
- CoverageBorderLine.fsa: contigs near the coverage treshold (-C)
- Discarded.fsa: contigs discarded due to duplicated hits
- UnMappedContigsHits.tab: contains the list of the excluded contigs with the number of tblastn hits
- UnMappedReferenceRegions.tab: contains the reference genome unmapped regions with at least one tblastn hit

- UnMappedContigs.txt: names (and lenghts) of the contigs not mapped to any particular reference molecule

Other files In addition to these files, a log file (called “CONTIGuator.log”) is present in the source directory: the amount of log output can be modulated using the options -V and -D.

Visualization

The outputs of CONTIGuator can be loaded into the Artemis comparison tool (ACT).

One way to open the maps is to use the shell script that CONTIGuator may have generated in each “Map_” directory: just launch each script (i.e. by double-clicking); if you are a lazy user (like me) just add the -l option and CONTIGuator should show each map automatically. Otherwise you can open the maps manually by starting ACT.

The reference molecule is on top, while the pseudo-contig is on bottom. The contigs coloured in light-red or in red are those that are supposed to overlap each other. On the reference track, the red blocks represent the regions of the reference with a significant hit with a region of a contig, while the green blocks represent those regions having a tblastn hit.

Another way to visualize the maps is to open the pdf map that is present in each "Map_" folder: these maps can be opened and edited in vector graphics programs like Adobe Illustrator or inkscape for publication purposes. The only requirement is that the user has a version of biopython equal or greater than 1.59.