

ER-flow Application Description Template

Application Name: Sequence assembly of small genomes
Application domain: bioinformatics, life-sciences, dna sequencing
Brief description of application Reconstruction of genomes by building contig sequences from raw DNA sequence fragments data: input data format: sff, fasta, or fastq input data value range output data format: fasta, report output data value range sample data: (link) tbd application (link) Newbler documentation (link) http://454.com/products/analysis-software/index.asp publication (link)
Execution environment DCI: (computing, data, VO, etc) middleware: pbs, glite workflow system: guse
Execution characteristics data size (per unit, typical number of units): input 1GB temporary output 10MB processing time (per unit): tbd memory usage: tbd disk usage: tbd
Target users Community, projects: (link) number of users: Biomedical and bioinformatics users Of general interest to researchers that sequence small genomes like bacteria and viruses. user type: end-user
Usage scenario for workflow in the ER-FLOW (how workflow will be reused, metaworkflow, how expected to contribute to project indicators, etc.). a) Researcher has performed shotgun sequencing of bacterial or viral genome and wants to reconstruct the original DNA/RNA from the sequence fragments b) Researcher has performed a metagenomics study (e.g. for virus discovery) and wants to create longer sequence fragments. With the larger sequence fragments he can do a more accurate search with Blast against known viruses in the public database.
Contact information (author) name: Barbera van Schaik e-mail b.d.vanschaik@amc.uva.nl , support@ebioscience.amc.nl Bioinformatics Laboratory, KEBB, Academic Medical Center, Amsterdam, NL