Anton Kratz, <anton.kratz@gmail.com>   24.12.2009

Manual for GGMS – General Genomic Mapping System

Align short and long sequences with SOAP and BLAT

Installation
BLAT - latest unix source code as of Nov 2009:
`http://users.soe.ucsc.edu/~kent/src/blatSrc34.zip`
and SOAP
`http://soap.genomics.org.cn/soap1/soap_1.11.tar.gz`
have to be installed first.

usage:
`perl ggms.pl input1.fa input2.fa`
input1 → a multifasta with the sequences which are to be aligned
input2 → a multifasta with the database (genome) to search against
The output will be in the file `final.fa`.

example usage and explanation:
`perl ggms.pl sequences.fa hg18.fa`
First, all long entries from `sequences.fa` are aligned with BLAT against each entry from `hg18.fa` separatedly. Then all short entries (short means smaller than 50 bp, can be changes in the soure code) from `sequences.fa` are aligned with SOAP against all entries from `hg18.fa`. The SOAP output is converted into PSL format. All alignments are concatenated, sorted and filtered for best hits only (using `pslSort` and `pslReps` from the BLAT installation).

config file
The user should edit the configuration file `.ggms_config` first, after having installed SOAP and BLAT. It contains the path to the SOAP binary in the first line, without the name of the binary itself and without trailing slash. So if the SOAP binary is `/home/kratz/GGMS/soap_1.11/soap`, then the entry in the config file must be `/home/kratz/GGMS/soap_1.11`. Same for BLAT in the second line. If SOAP or BLAT are not at these locations, then `ggms.pl` tries to find the files somewhere in the users home directory, after that the config file is automatically changes to contain the found paths.

example config file
My config file looks like this:
`/home/kratz/GGMS/soap_1.11 /home/kratz/bin/i686`

Parallel execution
It is possible to run several instances of GGMS in parallel, the temporary files are stored in sub-directories under `/tmp` which contain the process-ID and are therefore unique.