CgiHunter

CATCGTGGACGAT

User Manual

This documentation introduces the basic functions of the CgiHunterLight software tool for CpG island annotation of DNA sequences.

Table of content

1. Installation & Setup

CgiHunterLight has been developed in the programming language python version 2.4. To use the software python 2.4 has to be installed on your computer. A python interpreter can be obtained at http://www.python.org/download/ for Unix and Windows systems.

To install the CgiHunterLight software, download the archive and uncompress it.

Optionally, on 32-bit systems the python package 'psyco' can be installed for a performance improvement of the computation. It is available from http://psyco.sourceforge.net/download.html .

2. First Steps – The Test Run

The CgiHunterLight archive includes the sequences of the human chromosome 21 and 22 (assembly hg19). To generate CpG island annotations for them you can open a command line shell and change to the CgiHunterLight directory. Depending on your operating system you can start the annotation process by typing *python CgiHunterLight.py –x TestRun.xml* or *CgiHunterLight.py –x TestRun.xml* . This tells the python compiler to start the program *CgiHunterLight.py* and use the configurations in the *TestRun.xml* file to guide the annotation process. Three subfolders will be generated one for each of the chromosomes that are used to store intermediate files and the *Results* folder, which is used to store the final output.

If the program terminated successfully the *Results* folder will contain four files. *CGI_Shadowmap_hg19_chr21.bed* and *CGI_Shadowmap_hg19_chr22.bed* contain the CpG island annotations and can be directly displayed in a genome browser.
CGI_Shadowmap_hg19_chr21.txt and *CGI_Shadowmap_hg19_chr21.txt* are files to confirm the validity of the annotation. They contain at a nucleotide resolution the information of how many genome regions overlap an individual position that fulfill the given CpG island definition.

3. Command line

Direct calls to CgiHunterLight.py are possible, but it is strongly encouraged to only use them to execute predefined XML files (e.g. *python CgiHunterLight.py –x <name of xml file>*). If direct configuration of CgiHunter via command line is demanded, use the –h or --help option to find the following overview of all command line options.

Option	**Type**	**Description**
-x or --xml	string	The name of a xml file specifying a CGIH_Task. Overrides all other parameters
-X or --tasklist	string	Relative or absolute path of a task list. Each line of this file should contain the path to a xml file that will be executed as if called by –x.
-F or --force	flag	The execution of an xml file will be forced, even if it is marked as finished
-a or --assembly	string	The genome assembly/species the DNA sequence belongs to
-c or --chrom	string	The chromosome the DNA sequence belongs to
-d	int	Debug level – the higher the number the more status messages will be shown
-f or --comb	string	Filenmae of the prefilterstep. This option can be used, if the prefilterstep has already been performed
-m or --mask	string	If set to True, lower cased letters in the DNA sequence are handeld as repeats and are always considered as non-G and non-C. Default is False.
-o or --offset	int	This offset is added to all coordinates in the outputfiles. It should be used if fasta files are applied which din at start at the befinning of the chromosome.
-p or --path	string	The path to the directory where all output will be placed
-s or --source	string	Absolute or relative path of the fasta file that contains the DNA sequence
--gc	float	Minimal content of cytosine and guanine a region have to posses to qualify as a CpG island\n Only values between 50 and 100 are valid. Example: --gc 50
-r or --ratio	float	Minimal ratio of observed CpGs over expected CpGs a region have to posses to qualify as a CpG island. Only values between 50 and 100 are valid. Example: -r 60
--cw or --combwidth	float	Combwidth for the combing/filter step. Must be greater than 1. Default 1.2.
-w or --min_win	int	Minimal length threshold of CpG island
-W or --max_win	int	Optional upper bound to CpG island length. Per default set to a million base-pairs
-g or --getxml	string	If configuration is performed by command line this option can be used to retrieve the xml file that contains all options.

4. XML files

Each CgiHunter analyses can be completely specified as XML document. These documents can be either generated by using the CgiHunterGUI, by modifying the template document included in the CgiHunter distribution, by using the *–g* option of the command line or by your own custom techniques.
Each XML file has a status option that is initially set to ‘unfished’. Upon completion of a CgiHunter call this mark will be set to ‘finished’. CgiHunter will automatically ignore tasks that are marked as ‘finished’ to ensure that already processed files are not reanalyzed. To reset the status you can either load the XML file in the CgiHunterGUI and save it again under the same name or directly modify the status in the file.
If the computation of a file is interrupted, the status will contain the respective error message.
XML files can be executed directly from the command line (e.g. *python CgiHunter.py –x <name of xml file>*).