LOCP User Manual

LOCP is designed to LOCate putative Pilus operons in Gram-positive prokaryotes. This is done in two steps: First, pilus-related sequences are identified by hmmsearch with a carefully selected set of profile Hidden Markov Models (HMM). Each protein sequence is then labeled either with the HMM match or with a null model if no HMM matched. Sequences labeled with HMMs are referred as hits and those labeled with null-model as gaps. Second, chromosome regions with statistically significant clustering of hits are located using hypergeometrical distribution and Monte Carlo simulations. These regions are referred as Pilus Like Dense Regions (PLDRs).

Input data:

You can copy-paste the analyzed sequences into the main text area. Alternatively, you can select a FASTA file directly from your local drive by clicking the "brows" button. Input sequences are assumed to be laid out in the same order as they appear in the genome. This condition can be violated only when analysing a set of contigs with a "limit to contigs" option. Input sequences must be in FASTA format and must include only IUPAC characters: ABCDEFGHIKLMNOPQRSTUVWYZX for protein sequences and ACGTRYKMSWBDHVN for DNA sequences. The submitted fasta files should be saved using ANSI, DOS or UTF-8 encoding.
When using "html" output format (see below), the maximum size of the input file is limited to 5MB. Larger files should be analysed using "text" output format, which can analyse files up to 40MB.

Input parameters:

Score threshold

PLDR score is calculated by summing scores of all HMM hits in a given PLDR. This threshold can take values from 0 to infinity, with no filtering at 0 and increased filtering as the threshold gets larger.

P-value threshold

P-value estimates the probability of obtaining at least the same number HMM hits as in a given PLDR by chance (by random sampling of sequences from the analyzed genome). This threshold can take value from 0 to 1, with maximum filtering at 0 and no filtering at 1.

P_adj-value threshold

P_adj-value is the P-value corrected for multiple hypothesis testing. LOCP corrects its P-values using Monte Carlo simulation that is implemented in bootstrap function. Like the P-value threshold, this threshold can take values from 0 to 1, with maximum filtering at 0 and no filtering at 1.

Output format

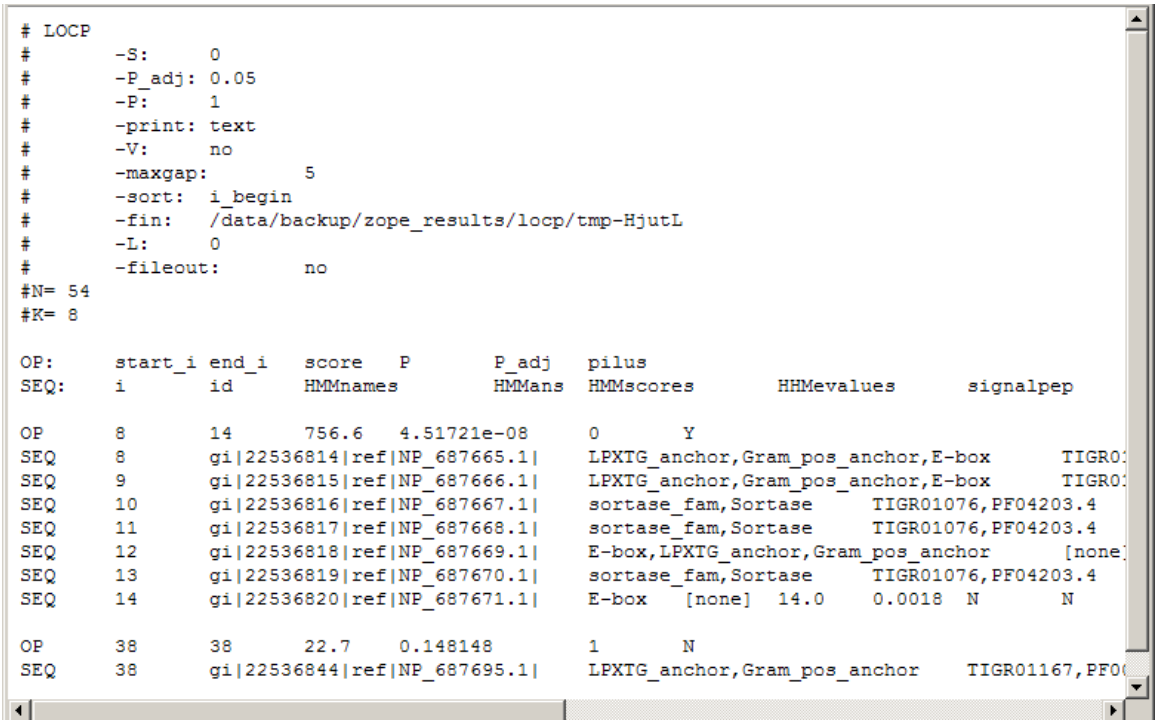
Output can be printed as a tab delimited text or a html page (default).

text:
Tab delimited text containing tab-delimited records describing PLDRs and sequences belonging to these PLDRs. Records describing PLRDs begin with an "OP" flag and those describing sequences with "SEQ" flag. Each PLDR record is followed by one or several records of sequences that belong to that PLDR. Each PLDR record is preceded by a newline. Text format displays all PLDRs, including PLDRs that did not pass threshold filtering.

OP record fields:
1. OP flag
2. start_i index of the first sequence
3. end_i index of the last sequence
4. score PLDR score
5. P PLDR P-value
6. P_adj PLDR P_adj-value
7. pilus Y if a given PLDR is a putative pilus operon and N otherwise

SEQ record fields:
1. SEQ flag
2. i sequence index
3. id sequence id field
4. HMMnames names of the matching HMM models[1]
5. HMMans annotation numbers of the matching HMM models[1]
6. HHMscores scores of the matching HMM models[1]
7. HMMevalues E-values of the matching HMM models[1]
8. signalpep Y if sequence is predicted to have N-terminal signal peptide, otherwise N[2]
9. C-anchor Y if sequence is predicted to end on the C-terminus with a transmembrane region followed by a cytoplasmic domain[2]

1. entries are comma delimited lists sorted by HMM bitscore.
2. signal peptide- and membrane topology are predicted with Phobius (Käll, *et al*., 2004).



```
# LOCP
#       -S:     0
#       -P_adj: 0.05
#       -P:     1
#       -print: text
#       -V:     no
#       -maxgap:        5
#       -sort:  i_begin
#       -fin:   /data/backup/zope_results/locp/tmp-HjutL
#       -L:     0
#       -fileout:       no
#N= 54
#K= 8

OP:     start_i end_i   score   P       P_adj   pilus
SEQ:    i       id      HMMnames        HMMans  HMMscores       HHMevalues      signalpep

OP      8       14      756.6   4.51721e-08     0       Y
SEQ     8       gi|22536814|ref|NP_687665.1|    LPXTG_anchor,Gram_pos_anchor,E-box      TIGR0
SEQ     9       gi|22536815|ref|NP_687666.1|    LPXTG_anchor,Gram_pos_anchor,E-box      TIGR0
SEQ     10      gi|22536816|ref|NP_687667.1|    sortase_fam,Sortase     TIGR01076,PF04203.4
SEQ     11      gi|22536817|ref|NP_687668.1|    sortase_fam,Sortase     TIGR01076,PF04203.4
SEQ     12      gi|22536818|ref|NP_687669.1|    E-box,LPXTG_anchor,Gram_pos_anchor      [none]
SEQ     13      gi|22536819|ref|NP_687670.1|    sortase_fam,Sortase     TIGR01076,PF04203.4
SEQ     14      gi|22536820|ref|NP_687671.1|    E-box   [none]  14.0    0.0018  N       N

OP      38      38      22.7    0.148148        1       N
SEQ     38      gi|22536844|ref|NP_687695.1|    LPXTG_anchor,Gram_pos_anchor    TIGR01167,PF0
```

Screenshot: text output.

HTML:
HTML document displaying the same information as text format and additional information on the sequence-HMM alignments. By default, only a summary of each PLDR is displayed. Detailed sequence records can be brought to view by clicking sequence id links. In html format operon records are merged with sequence records by adding score, P and P_adj fields to each sequence record. Consecutive operons are separated by an empty row. HTML format displays only putative pilus operons, that is PLDRs that have passed threshold filtering. Note that detailed sequence records are implemented using JavaScript and will be disabled if JavaScript is disabled in your browsers settings.

LOCP parameters

i	id	topHMMname	signalpep	C-anchor	score	P	P_adj
8	gi|22536814|ref|NP_6	LPXTG_anchor	Y	Y	756.6	4.51721e-08	0/1000
9	gi|22536815|ref|NP_6	LPXTG_anchor	Y	Y	756.6	4.51721e-08	0/1000
10	gi|22536816|ref|NP_6	sortase_fam	N	Y	756.6	4.51721e-08	0/1000

Sequence

id: gi|22536816|ref|NP_687667.1|
annotation: sortase family protein [Streptococcus agalactiae 2603V/R]

HMM hits

name	assession number	score	e-value
Alignment			
sortase_fam	TIGR01076	244.6	1.3e-72
Sortase	PF04203.4	208.0	1.3e-61

Phobius

FT DOMAIN 1 19 CYTOPLASMIC
FT TRANSMEM 20 40
FT DOMAIN 41 260 NON CYTOPLASMIC
FT TRANSMEM 261 281
FT DOMAIN 282 305 CYTOPLASMIC

11	gi|22536817|ref|NP_6	sortase_fam	Y	Y	756.6	4.51721e-08	0/1000
12	gi|22536818|ref|NP_6	E-box	Y	Y	756.6	4.51721e-08	0/1000
13	gi|22536819|ref|NP_6	sortase_fam	N	N	756.6	4.51721e-08	0/1000
14	gi|22536820|ref|NP_6	E-box	N	N	756.6	4.51721e-08	0/1000
i	id	topHMMname	signalpep	C-anchor	score	P	P_adj

Screenshot: HTML output

Output sorting

Before printing, located PLDRs can be sorted by: **sequence index** (= the running number of the first sequence in a given PLDR), **Score**, **P-value** or **P_adj-value**.

Alphabet

This option determines whether input data are interpreted as protein or dna sequences. If “dna” is selected, input LOCP will convert input dna sequences to all possible open reading frames prior to the main analysis.

Minimum ORF length

Only open reading frames longer than or equal to this threshold value will be analyzed.

Limit to contigs option

This option restricts putative operons to a single contig. This restriction can be useful when analysing metagenomes, or other data wherein consecutive contigs do not

follow genomic order. LOCP interprets any set of consecutive sequences that have identical first id field as belonging to the same contig.
For example these sequences:
>gi|106748457|gb|AAQK01000001.1|_orf_frame:-3_pos:109
..
>gi|106748456|gb|AAQK01000002.1|_orf_frame:1_pos:57
..
>gi|106748456|gb|AAQK01000002.1|_orf_frame:1_pos:251
..
>gi|106748456|gb|AAQK01000002.1|_orf_frame:-2_pos:186
..
>gi|106748454|gb|AAQK01000004.1|_orf_frame:1_pos:438

would be arranged into three contigs: seq1, seq2-seq4 and seq5.

Default values

Default parameter values can be set by clicking “Default” or “Default Meta” buttons. “Default” is intended for protein sequences that are assumed to be in their genomic order, for example all the proteins encoded by the chromosome of the genome. “Default Meta” is intended for metagenomic DNA sequences and is not depended upon the order of the sequences.

System requirements

LOCP should work on any browser that supports JavaScript and dynamic HTML.
LOCP was tested on the following platforms:
Windows XP Firefox 2.0
Windows XP Internet Explorer 6.0
Mac OS X 10.5.5 Firefox 3.0
Mac OS X 10.5.5 Safari 3.2

References:

Käll, *et al*. (2004) A combined transmembrane topology and signal peptide prediction method. J.Mol.Biol., 338, 1027-1036.