# Theoretical and practical metagenomic approaches to viral discovery

Practical Session: Infernal and covariance models

Kevin Lamkiewicz, Manja Marz

25.10.2019

European Virus Bioinformatics Center



#### INFERNAL

http://eddylab.org/infernal

#### User's Guide Quote:

#### How to avoid reading this manual

If you're like most people, you don't enjoy reading documentation. You're probably thinking: 113 pages of documentation, you must be joking!



## **INFERNAL**

http://eddylab.org/infernal

#### User's Guide Quote:

#### How to avoid reading this manual

If you're like most people, you don't enjoy reading documentation. You're probably thinking: 113 pages of documentation, you must be joking!



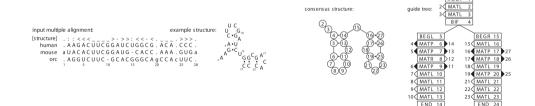
#### INFERNAL

http://eddylab.org/infernal

#### User's Guide Quote:

#### How to avoid reading this manual

If you're like most people, you don't enjoy reading documentation. You're probably thinking: 113 pages of documentation, you must be joking!





#### Use case

Given a metagenomic sample, let's assume we already assembled the reads to contigs. We suspect that viruses related to the Alphacoronaviruses are present in our sample, however, the sequence similarity is not sufficiently high for conventional methods like blast or hidden markov-models.



#### Use case

Given a metagenomic sample, let's assume we already assembled the reads to contigs. We suspect that viruses related to the Alphacoronaviruses are present in our sample, however, the sequence similarity is not sufficiently high for conventional methods like blast or hidden markov-models.

#### Restriction

Since scanning with Infernal takes quite a while, we won't use a real metagenomic sample today. Instead, we will download some genomes from the NCBI database - the procedure is the same for "real metagenomes", but faster.



In order to build a covariance model, we need an alignment. CMbuild can handle stockholm and CLUSTAL format.

```
1 # go to: https://www.ncbi.nlm.nih.gov/nuccore
2 # txid693996[Organism:exp] AND complete genome[Title]
3 # download them: all_complete_acov.fasta
```



In order to build a covariance model, we need an alignment.

CMbuild can handle stockholm and CLUSTAL format.

```
# go to: https://www.ncbi.nlm.nih.gov/nuccore
# txid693996[Organism:exp] AND complete genome[Title]
# download them: all_complete_acov.fasta

# build cm from aln alignment
$ > mlocarna --stockholm corona_5utr.fasta

# or clustalw corona_5utr.fasta
```



In order to build a covariance model, we need an alignment. CMbuild can handle stockholm and CLUSTAL format.

```
# go to: https://www.ncbi.nlm.nih.gov/nuccore
# txid693996[Organism:exp] AND complete genome[Title]
# download them: all_complete_acov.fasta
# build cm from aln alignment
# s> mlocarna --stockholm corona_5utr.fasta
# or clustalw corona_5utr.fasta
# or clustalw corona_5utr.cvm corona_5utr.out/results/result.stk
# or cmbuild corona_5utr.cvm corona_5utr.aln
```



In order to build a covariance model, we need an alignment. CMbuild can handle stockholm and CLUSTAL format.

```
# go to: https://www.ncbi.nlm.nih.gov/nuccore
# txid693996[Organism:exp] AND complete genome[Title]
# download them: all_complete_acov.fasta

# build cm from aln alignment
# s> mlocarna --stockholm corona_5utr.fasta
# or clustalw corona_5utr.fasta

# or clustalw corona_5utr.cvm corona_5utr.out/results/result.stk
# or cmbuild corona_5utr.cvm corona_5utr.aln
```

Which one of the two approaches is more reasonable?



## **CMCALIBRATE**

```
1  # now we calibrate our covariance model
2  # thus, the emission and transition probabilities are trained
3  # this takes quite a while
4  $> cmcalibrate --cpu 4 corona_5utr.cvm
```



## **CM**CALIBRATE

```
1  # now we calibrate our covariance model
2  # thus, the emission and transition probabilities are trained
3  # this takes quite a while
4  $> cmcalibrate --cpu 4 corona_5utr.cvm
```

#### I prepared something beforehand...

Since  ${\tt cmcalibrate}$  takes quite a while, even for such a small dataset, I already calibrated the covariance model beforehand. You find the CM here:

TODO



## **CMSEARCH**

```
1  # One last step - now we search for sequences (or fragments),
2  # that fit to our covariance model.
3
4  # search in the fasta file for sequences matching
5  # the cmfile and get their positions as table
6
7  $> cmsearch --cpu 4 --tblout corona_5utr_cmsearch.csv corona_5utr.cvm \
8  all_complete_acov.fa > cmsearch.log
```



## **RESULTS**

(codina) [boal toka) (-O, flux, oct 10 44.18, 1994) (-f. 11.0e, 11.0e) -(-C, Conference optimal), and (0.7, 14.04) (-0.7, 14.04) (-0.7) (-0.7, 14.04) (														
→ head -n 25 cov_Sutr_cusearch.csv														
#target name		on query name												nc description of target
*														
NC_009657.1														Scotophilus bat coronavirus 512, complete genome
MK211369.1										1 0.40				Coronavirus BtSk-AlphaCoV/GX2018A, complete genome
KP890336.1										1 0.46		247.2		Porcine epidemic diarrhea virus strain CH/HNYF/14, complete genome
MH891585.1														Porcine epidemic diarrhea virus isolate S14, complete genome
KU664503.1										1 0.46				Porcine epidemic diarrhea virus isolate ZJU/G1/2013, complete genome
KM887144.1														Porcine epidemic diarrhea virus isolate CHM2013, complete genome
MH891590.1										1 0.46				Porcine epidemic diarrhea virus isolate S10, complete genome
MH891589.1														Porcine epidemic diarrhea virus isolate S6, complete genome
MH891587.1										1 0.46	0.0			Porcine epidemic diarrhea virus isolate S100, complete genome
MH891584.1														Porcine epidemic diarrhea virus isolate S12, complete genome
MH052682.1														Porcine epidemic diarrhea virus isolate KNU-1703, complete genome
MH243319.1													1.3e-61 !	Porcine epidemic diarrhea virus isolate KNU-1807, complete genome
MH726365.1										1 0.46			1.9e-61 !	Porcine epidemic diarrhea virus isolate GDS25, complete genome
MF577027.1													2.1e-61 !	Porcine epidemic diarrhea virus strain PEDV/Belgorod/dom/2008, complete genome
KR003452.1														Porcine epidemic diarrhea virus isolate 15V010/BEL/2015, complete genome
KY929406.1														Porcine epidemic diarrhea virus strain PT-P96, complete genome
KY929405.1														Porcine epidemic diarrhea virus strain PT-P5, complete genome
LT897799.1														Porcine epidemic diarrhea virus isolate PEDV_GER_L00901-V215_1978 genome assembly, com
	: monopartite													
MK140812.1										1 0.46	0.0			Porcine epidemic diarrhea virus isolate CH/TP-3-1/2018, complete genome
KU297956.1		result				95 1								Porcine epidemic diarrhea virus strain SLO/JH-11/2015, complete genome
MK606368.1		result				95 1								Porcine epidemic diarrhea virus isolate CH-HB1-2018, complete genome
MK032691.1		result				95 1								Porcine epidemic diarrhea virus isolate KNU-1817, complete genome
MK392335.1		result				95 1								Porcine epidemic diarrhea virus isolate LW/L, complete genome
m332333. I														For the epidemic diarriles wirds isotate two, complete genome



## THANKS AND GOODBYE

GCAUGCAUGCUAGCUAGCAUGCAUGCAUGCAUGCAGUCCAGUCCGAGAUACCCUAACUCUAGGGUAUCUCGGACCUCAAAAGAGGGU



## THANKS AND GOODBYE

GCAUGCAUGCUAGCUAGCAUGCAUGCAUGCAUGCAGUCCGAGAUACCCUAACUCUAGGGUAUCUCGGACCUCAAAAGAGGGU

