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!

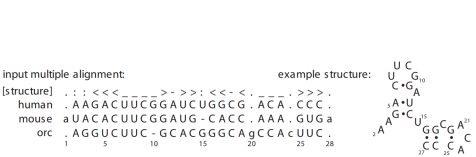
INFERNAL

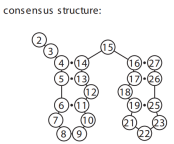
http://eddylab.org/infernal

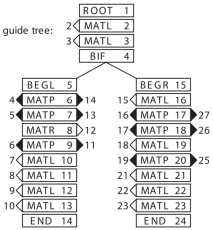
User's Guide Quote:

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.

CMbuild

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
```

CMbuild

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

```

CMbuild

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

$> cmbuild corona_5utr.cvm corona_5utr.out/results/result.stk
# or cmbuild --noss corona_5utr.cvm corona_5utr.aln
```

CMbuild

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

$> cmbuild corona_5utr.cvm corona_5utr.out/results/result.stk
# or cmbuild --noss corona_5utr.cvm corona_5utr.aln
```

Which one of the two approaches is more reasonable?

CMCALIBRATE

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

CMCALIBRATE

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

I prepared something beforehand...

Since `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

```
# One last step - now we search for sequences (or fragments),
# that fit to our covariance model.

# search in the fasta file for sequences matching
# the cmfile and get their positions as table

$> cmsearch --cpu 4 --tblout corona_5utr_cmsearch.csv corona_5utr.cvm \
        all_complete_acov.fa > cmsearch.log
```

RESULTS

```
┌──(co68mol@palinka)-(~/)-(Thu Oct 10 -- 04:18 PM)-(6 files, 31Mb)
└─(~/Conferences/brasil_ws/03_friday/02_infernal/playground)
└─ head -n 25 cov_5utr_cmsearch.csv
#target name         accession query name           accession mdl mdl from   mdl to seq from   seq to strand trunc pass   gc  bias  score   E-value inc description of target
#-------------------- --------- -------------------- --------- --- -------- -------- -------- -------- ------ ----- ---- ---- ------ --------- --- ---------------------
NC_009657.1          -         result               -         cm    1      295        1      293        +     no    1 0.42  0.0  267.3   1.9e-67 !   Scotophilus bat coronavirus 512, complete genome
MK211369.1           -         result               -         cm    1      295        3      295        +     no    1 0.40  0.3  263.9   1.5e-66 !   Coronavirus BtSk-AlphaCoV/GX2018A, complete genome
KP890336.1           -         result               -         cm    1      295        1      292        +     no    1 0.46  0.0  247.2   4.5e-62 !   Porcine epidemic diarrhea virus strain CH/HNYF/14, complete genome
MH891585.1           -         result               -         cm    1      295        1      292        +     no    1 0.46  0.0  246.7   6.2e-62 !   Porcine epidemic diarrhea virus isolate S14, complete genome
KU664503.1           -         result               -         cm    1      295        1      291        +     no    1 0.46  0.0  246.1     9e-62 !   Porcine epidemic diarrhea virus isolate ZJU/C1/2013, complete genome
KM887144.1           -         result               -         cm    1      295        1      291        +     no    1 0.46  0.0  246.1     9e-62 !   Porcine epidemic diarrhea virus isolate CHMZ013, complete genome
MH891590.1           -         result               -         cm    1      295        1      292        +     no    1 0.46  0.0  245.9     1e-61 !   Porcine epidemic diarrhea virus isolate S10, complete genome
MH891589.1           -         result               -         cm    1      295        1      292        +     no    1 0.46  0.0  245.9     1e-61 !   Porcine epidemic diarrhea virus isolate S6, complete genome
MH891587.1           -         result               -         cm    1      295        1      292        +     no    1 0.46  0.0  245.9     1e-61 !   Porcine epidemic diarrhea virus isolate S100, complete genome
MH891584.1           -         result               -         cm    1      295        1      292        +     no    1 0.46  0.0  245.9     1e-61 !   Porcine epidemic diarrhea virus isolate S32, complete genome
MH052682.1           -         result               -         cm    1      295        1      292        +     no    1 0.46  0.0  245.9     1e-61 !   Porcine epidemic diarrhea virus isolate KNU-1703, complete genome
MH243319.1           -         result               -         cm    1      295        1      292        +     no    1 0.46  0.0  245.4   1.3e-61 !   Porcine epidemic diarrhea virus isolate KNU-1807, complete genome
MH726365.1           -         result               -         cm    1      295        1      292        +     no    1 0.46  0.0  244.9   1.9e-61 !   Porcine epidemic diarrhea virus isolate GDS25, complete genome
KF577027.1           -         result               -         cm    1      295        1      291        +     no    1 0.46  0.0  244.7   2.1e-61 !   Porcine epidemic diarrhea virus strain PEDV/Belgorod/dom/2008, complete genome
KR003452.1           -         result               -         cm    1      295        1      292        +     no    1 0.46  0.0  244.1   3.1e-61 !   Porcine epidemic diarrhea virus isolate 15V018/BEL/2015, complete genome
KY929406.1           -         result               -         cm    1      295        1      292        +     no    1 0.46  0.0  244.1   3.1e-61 !   Porcine epidemic diarrhea virus strain PT-P96, complete genome
KY929405.1           -         result               -         cm    1      295        1      292        +     no    1 0.46  0.0  244.1   3.1e-61 !   Porcine epidemic diarrhea virus strain PT-P5, complete genome
LT807799.1           -         result               -         cm    1      295        1      291        +     no    1 0.46  0.0  244.0   3.3e-61 !   Porcine epidemic diarrhea virus isolate PEDV_GER_L00901-V215_1978 genome assembly, com
plete genome: monopartite
MK140812.1           -         result               -         cm    1      295        1      292        +     no    1 0.46  0.0  243.7   3.8e-61 !   Porcine epidemic diarrhea virus isolate CH/TP-3-1/2018, complete genome
KU297956.1           -         result               -         cm    1      295        1      292        +     no    1 0.46  0.0  243.7   3.8e-61 !   Porcine epidemic diarrhea virus strain SLO/JH-11/2015, complete genome
MK606368.1           -         result               -         cm    1      295        1      292        +     no    1 0.46  0.0  243.6   4.2e-61 !   Porcine epidemic diarrhea virus isolate CH-HB1-2018, complete genome
MK032691.1           -         result               -         cm    1      295        1      292        +     no    1 0.46  0.0  243.6   4.2e-61 !   Porcine epidemic diarrhea virus isolate KNU-1817, complete genome
MK392335.1           -         result               -         cm    1      295        1      292        +     no    1 0.46  0.0  243.6   4.2e-61 !   Porcine epidemic diarrhea virus isolate LW/L, complete genome
```

THANKS AND GOODBYE