Design And *In Silico* Validation Of PCR-metabarcoding Primers

Daniel Marquina - Naturhistoriska Riksmuseet

First of all: system settings

```
> ssh username@milou.uppmax.uu.se
> interactive -n X -t X:00:00 -A g2016021
> cp -r /proj/g2016021/metabarcoding/ .
> cd glob/metabarcoding
> module load python/2.7.9
> python get-obitools.py
> ./obitools
> exit
> tar –zxvf ecoPCR.tar.gz
> cd ecoPCR/src/
> make
> export PATH=$PATH:/proj/g2016021/metabarcoding/ecoPCR/src
> cd ../../
> tar –zxvf ecoPrimers.tar.gz
> cd ecoPrimers/src/
> make
> export PATH=$PATH:/proj/g2016021/metabarcoding/ecoPrimers/src
```

PROPERTIES

→ Amplify a suitable marker:

1) Mutation rate: distinguish species.
2) Conserved regions: universal primers.
3) Appropriate length: variation without loss of information when degraded.
4) Reference libraries: taxonomic identification.

→ Amplify sequences of ALL the species belonging to the target taxon present in the sample.

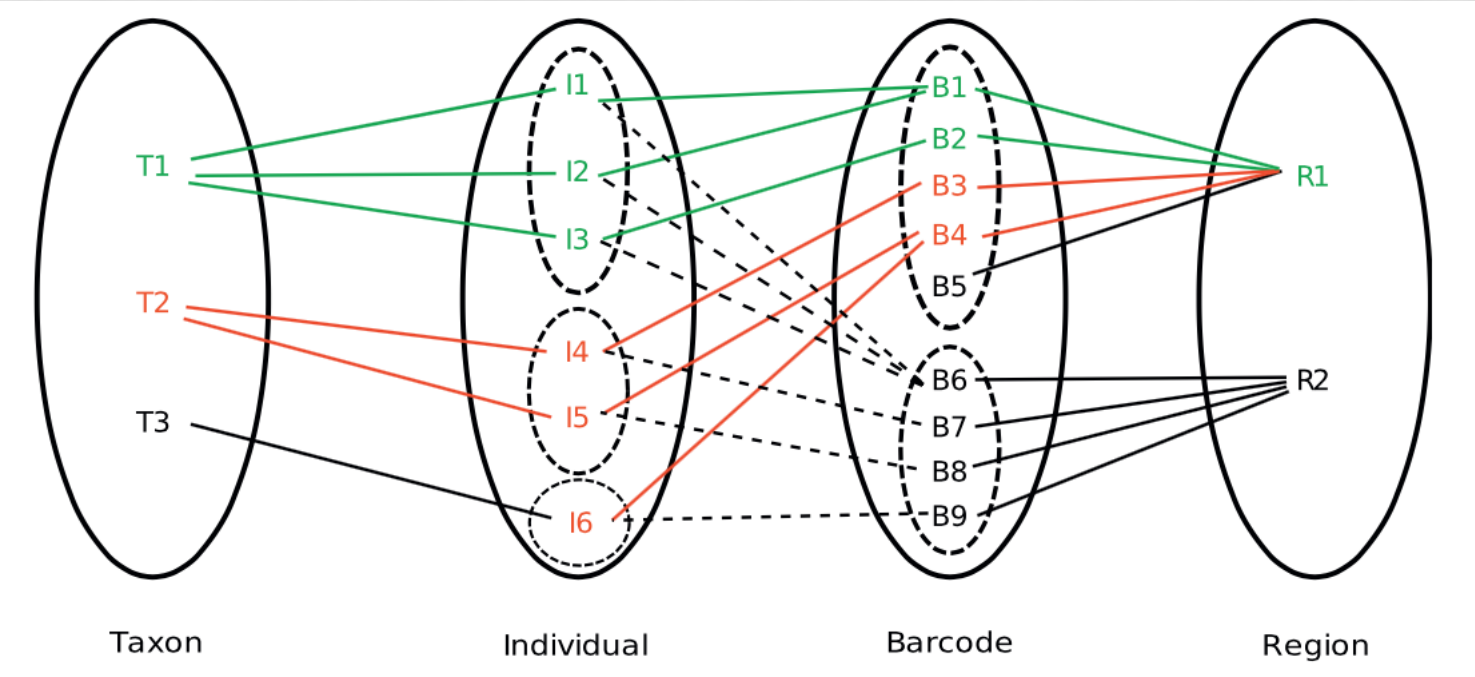
1) Ideally, amplify sequences of NONE of the species NOT belonging to the target taxon present in the sample. This can be a secondary (eDNA) or principal (dietDNA) requisite.
2) Amplify all sequences EQUITATIVELY = no amplification bias.

→ Region amplified of the 'suitable marker' should discriminate between closely related species.

Properties - Indexes

Bc: Taxonomic coverage Bc = no. sequences amplified / no. sequences present
PRIMER PROPERTY Bc = [0,1]

Bs: Resolution capacity Bs = no. taxa unambiguously identified/ no. sequences amplified
BARCODE PROPERTY Bs = [0,1]



Ficetola *et al.* 2010

ecoPrimers (Riaz *et al.* 2011)

→ Different software from OBITools, but in the same package.

→ Specifically developed for metabarcoding (of any taxonomic group).

→ Python based.

→ Algorithm: Strict Primer Algorithm (SPA).

→ Gives nice output.

```
#
# ecoPrimer version 0.3
# Rank level optimisation : species
# max error count by oligonucleotide : 3
#
# Restricted to taxon:
#     6960 : Hexapoda (superclass)
#
# strict primer quorum  : 0.70
# example quorum        : 0.90
# counterexample quorum : 0.10
#
# database : sixlegs
# Database is constituted of  1602 examples        corresponding to  1115 species
#                         and     0 counterexamples corresponding to     0 species
#
# amplifiat length between [50,500] bp
# DB sequences are considered as circular
# Pairs having specificity less than 0.60 will be ignored
#
     0  ATAGAAACCAACCTGGCT      TTACCTTAGGGATAACAG      53.6    1.7     47.7    27.0    8       7       GG      1514    0       0.945   1059    0       0.950   835     0.788   138     217     142.67          #16S

     1  ATAGAAACCAACCTGGCT      TACCTTAGGGATAACAGC      53.6    1.7     50.6    30.9    8       8       GG      1502    0       0.938   1048    0       0.940   824     0.786   137     216     141.67          #16S
     2  GATAGAAACCAACCTGGC      TACCTTAGGGATAACAGC      53.6    2.0     50.6    30.9    9       8       GG      1499    0       0.936   1046    0       0.938   822     0.786   138     217     142.67          #16S
     3  ATAGAAACCAACCTGGCT      GACCTCGATGTTGGATTA      53.6    11.4    51.8    38.1    8       8       GG      1498    0       0.935   1045    0       0.937   671     0.642   79      158     83.67           #16S
     4  GATAGAAACCAACCTGGC      TTACCTTAGGGATAACAG      53.6    2.0     47.7    27.0    9       7       GG      1511    0       0.943   1057    0       0.948   654     0.619   139     218     143.67          #16S
```

Strict Primer Algorithm

E

ATACGGCTACTAACT
ATACGGCTACTAACT
ATACGGCTAGTAACT
ATTCGGCTACTAAGT
ATTCGGCTACTAAGT
ATTCGGCTACTAAGT
ATTCGGCTACTAAGT

Words of length *L* present in at least *S* sequences of **E**
L: number (18-21)
S: percentage (defaut=70)

*Lp(**E**)*

ATACGGCTACTAACT
ATTCGGCTACTAAGT

Words of length *L* present in at least *S* sequences of **E**, and present in *T* sequences of **E** with no more than *m* mismatches.
T: percentage (defoult=90)
m: number (1-3)

*Lp'(**E**)*

ATACGGCTACTAACT
ATACGGCTAGTAACT
ATTCGGCTACTAAGT

Finds a space **D** within the interval of amplified sequence length [l_{min}-l_{max}] and creates Lp'(**D**). Pairs Lp'(**E**)-Lp'(**D**).

ATTCGGCTACTAAGT - ATTCGGCTACTAAGT

Bs
Bc

ecoPrimers

Ligths:

- Computes *Bc*/*Bs* from amplified sequences.
- Constrains no mismatches in 3'-end of the primer.
- Pairs primers within an interval of barcode length.
- Considers 'countersequences'.

Shadows:

- No degeneracy allowed. Mismatches are mismatches.
- Very taxonomy-constrained (EMBL, GB…) -> Whole genomes, no individual genes.

DegePrime (Hugerth *et al*. 2014)

→ Developed at the SciLifeLab.

→ Originally developed for 16S/metagenomics in procaryotes.

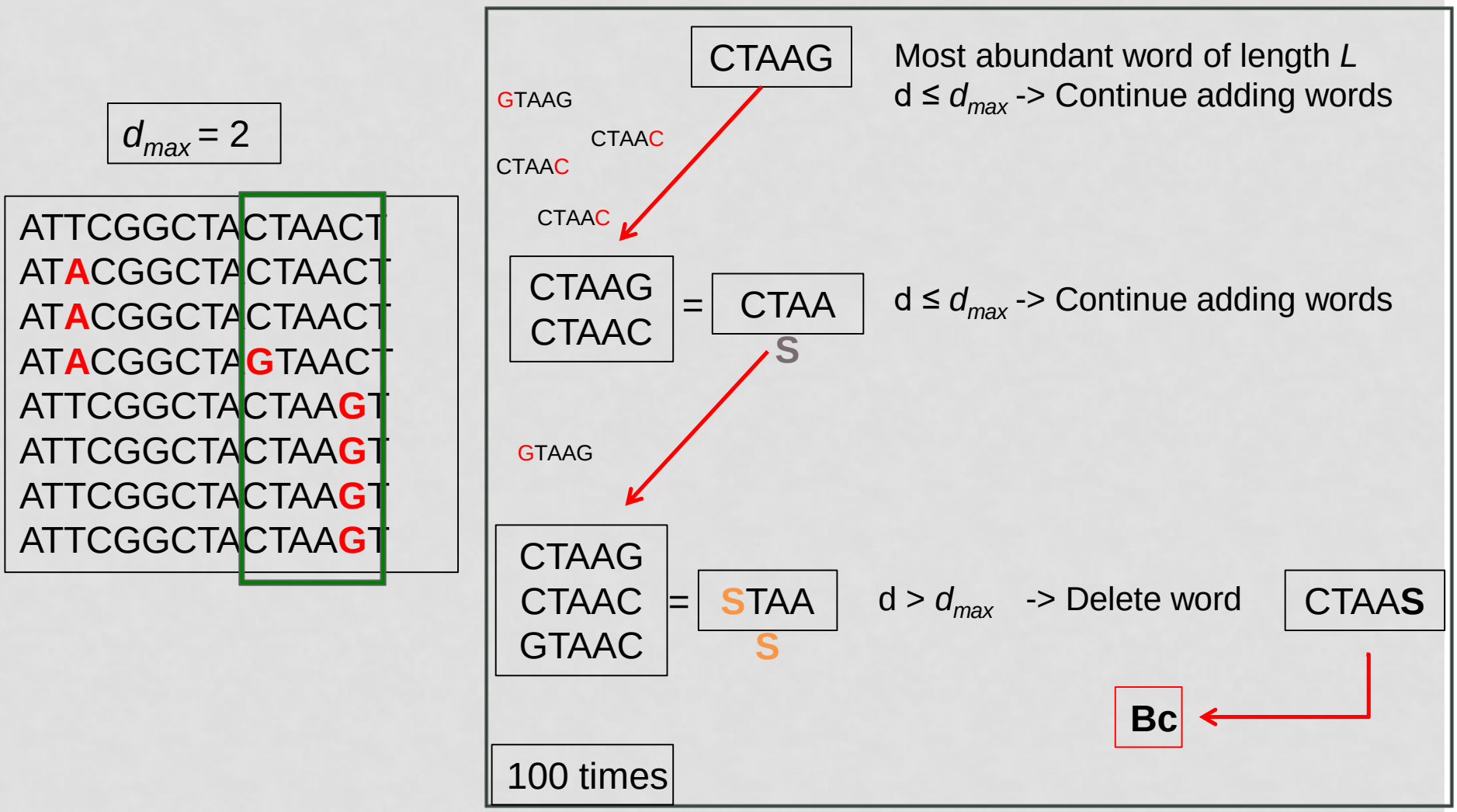
→ Perl based.

→ Algorithm: Weighed Randomized Combination.

→ Works over an alignment. It doesn't allow gaps, so the alignment should be modify in a program readable format (TrimAlignment.pl).

→ Output has to be edited to be useful.

Weighed Randomized Combination



DegePrime

Ligths:

- Allows degeneracy (not mismatches).
- Computes *Bc*.
- Gives measure of sequence diversity.

Shadows:

- No 3’-end constrain (but you can do it yourself).
- No pairing at length interval = No *Bs* index (but there are tools for doing it afterwards).
- Too much degeneracy when not needed (next slide).

ATTCGGCTACTAACT
ATACGGCTACTAACT
ATACGGCTAGTAACT
ATTCGGCTACTAAGT
ATTCGGCTACTAAGT

AT{T/A}CGGCTA{C/G}TAA{G/C}T

d = 1x1x2x1x1x1x1x1x1x2x1x1x1x2x1 = 8

ATTCGGCTACTAAGT
ATTCGGCTACTAACT

ATTCGGCTACTAA{C/G}T

d = 1x1x1x1x1x1x1x1x1x1x1x1x1x2x1 = 2

ATACGGCTAGTAACT
ATACGGCTACTAACT

ATTCGGCTA{C/G}TAACT

d = 1x1x1x1x1x1x1x1x1x2x1x1x1xx1 = 2

LET’S DO SOME SCIENCE NOW

ecoPrimers

What do we need?

- Set of mitochondrial genomes downloaded from GenBank (gb format) ✓
- Taxonomy repository download from NCBI ✓
- Format the taxonomy into OBITools format ✓ “ncbi20150906”
- Format the genomes into OBITools database ✓ “sixlegs”

```
> cd ~/metabarcoding/

> ecoPrimers -d sixlegs -e 3 -3 3 -l 50 -L 650 -r 6960 \

–c > Insectsprimers.ecoprimers
```

-d sixlegs: OBITools-format collection of genomes
-e 3: maximum number of mismatches in the second step
-3 3: number of nucleotides of the 3’ end constrained in a perfect match
-l 50 –L 650: minimum (*l*) and maximum (*L*) length of the potential barcode
-r 6960: NCBI taxid of the target group (when having other genomes too)
-c: sequences are circular (mtDNA)

```
> less Insectprimers.ecoprimers
```

```
#
# ecoPrimer version 0.3
# Rank level optimisation : species
# max error count by oligonucleotide : 3
#
# Restricted to taxon:
#     6960 : Hexapoda (superclass)
#
# strict primer quorum  : 0.70
# example quorum        : 0.90
# counterexample quorum : 0.10
#
# database : sixlegs
# Database is constituted of  1602 examples        corresponding to  1115 species
#                       and     0 counterexamples corresponding to     0 species
#
# amplifiat length between [50,500] bp
# DB sequences are considered as circular
# Pairs having specificity less than 0.60 will be ignored
#
     0  ATAGAAACCAACCTGGCT      TTACCTTAGGGATAACAG      53.6    1.7     47.7    27.0    8       7       GG      1514    0       0.945   1059    0       0.950   835     0.788   138     217     142.67

     1  ATAGAAACCAACCTGGCT      TACCTTAGGGATAACAGC      53.6    1.7     50.6    30.9    8       8       GG      1502    0       0.938   1048    0       0.940   824     0.786   137     216     141.67
     2  GATAGAAACCAACCTGGC      TACCTTAGGGATAACAGC      53.6    2.0     50.6    30.9    9       8       GG      1499    0       0.936   1046    0       0.938   822     0.786   138     217     142.67
     3  ATAGAAACCAACCTGGCT      GACCTCGATGTTGGATTA      53.6    11.4    51.8    38.1    8       8       GG      1498    0       0.935   1045    0       0.937   671     0.642   79      158     83.67
     4  GATAGAAACCAACCTGGC      TTACCTTAGGGATAACAG      53.6    2.0     47.7    27.0    9       7       GG      1511    0       0.943   1057    0       0.948   654     0.619   139     218     143.67
```

F primer

R primer

Sequences amplified

Taxa amplified

Bc index

Bs index

Taxa identified

Mean, max. and min. length of the barcodes produced.

DegePrime

What do we need?

- Set of mitochondrial genomes downloaded from GenBank (fasta format) ✓
- COI gen extracted from every genome ✓
- Alignment of the COI gen extracted ✓

```
> cd DegePrime

> perl TrimAlignment.pl –i COI_aligned.fasta -min 0.9 -o COI_trimmed

> perl DegePrime.pl -i COI_trimmed -d 12 -l 18 -o COIprimer

> less COIprimer
```

Pos	TotalSeq		UniqueMers	Entropy	PrimerDeg		PrimerMatching	PrimerSeq
20	1145	101	4.35512615524902		12	710	HAAYCATAARGATATTGG	
21	1145	96	4.23929291093935		12	706	AAYCATAARGATATTGGH	
22	1146	107	4.30799785685939		12	698	AYCATAARGATATTGGHA	
23	1148	130	4.49977750333136		12	678	YCATAARGATATTGGHAC	
24	1148	167	5.07112704195661		12	663	CATAARGATATTGGWACH	
25	1151	224	5.55058302792518		12	589	ATAARGATATTGGWACHT	
26	1151	223	5.54245740925846		12	590	TAARGATATTGGWACHTT	
27	1151	216	5.35036312521185		12	627	AARGATATTGGWACHTTA	
28	1151	215	5.34515272119099		12	628	ARGATATTGGWACHTTAT	
29	1151	214	5.35138851544978		12	626	RGATATTGGWACHTTATA	
30	1151	217	5.35568496151132		12	627	GATATTGGWACHTTATAY	
31	1151	217	5.3372916008392	12	630	ATATTGGWACHTTATAYT		
32	1151	213	5.31738112216666		12	633	TATTGGWACHTTATAYTT	
33	1151	223	5.52912247355897		12	599	ATTGGAACHTTATAYTTY	
34	1151	284	5.90074965532906		12	552	TTGGAACHTTATAYTTYA	
35	1151	283	5.90074965532905		12	552	TGGAACHTTATAYTTYAT	
36	1151	320	6.1616857699409	12	505	GGAACHTTATAYTTYATT		
37	1151	330	6.21183638827913		12	502	GAACHTTATAYTTYATTT	
38	1151	329	6.2042021670173	12	503	AACHTTATAYTTYATTTT		

-i *file*: input file
-o *file*: output file
-min: proportion of gaps “deleted”
-d: max degeneracy allowed
-l: length of the primer

```
> cat COIprimer > COIprimer.csv

> awk '{print $6," ",$1," ",$7," ",$4," ",$5}' COIprimer.csv > COIprimerfilter.csv

> echo "SeqMatched; Position; Sequence; Entropy; Degneracy" > COIprimer.csv

> sort -g -r -t" " -k1 COIprimerfilter.csv >> COIprimer.csv

> rm COIprimerfilter.csv COI_trimmed COIprimer

> less COIprimer.csv
```

SeqMatched;	Position;	Sequence;	Entropy;	Degneracy
959	686	ACAYTTATTYTGATTYTT	3.21773124654164	8
958	685	AACAYTTATTYTGATTYT	3.22689142492553	8
958	684	CAACAYTTATTYTGATTY	3.22689142492553	8
958	683	YCAACAYTTATTYTGATT	3.42525029766891	8
953	681	TAYCAACAYTTATTYTGA	3.47090117155666	8
950	682	AYCAACAYTTATTYTGAT	3.49294077700492	8
923	714	GAAGTHTAYATTYTAATT	3.35452347323419	12
911	911	WGCHACWATAATTATTGC	4.34891234548245	12
909	692	ATTYTGATTYTTTGGDCA	3.92213973138692	12
909	691	TATTYTGATTYTTTGGDC	3.92040362027581	12
908	910	CWGCHACWATAATTATTG	4.36594356035277	12
897	280	TAAATAAYATAAGHTTYT	3.50406186145596	12
896	281	AAATAAYATAAGHTTYTG	3.51302350097954	12
889	909	TCWGCHACWATAATTATT	4.5084895283526	12
879	689	YTTATTYTGATTYTTTGG	3.52505680452662	8
879	688	AYTTATTYTGATTYTTTG	3.52505680452662	8
879	687	CAYTTATTYTGATTYTTT	3.52505680452662	8
879	182	HATAATTTTYTTYATAGT	3.74491098498773	12

ecoPCR

What do we need?

- Set of mitochondrial genomes downloaded from GenBank (gb format) ✓
- Taxonomy repository download from NCBI ✓
- Format the taxonomy into OBITools format ✓ “ncbi20150906”
- Format the genomes into OBITools database ✓ “sixlegs”
- Primer pair ✓ HATAATTTTYTTYATAGT AARAATCARAATAARTGT
- OBITools package ✓

```
> cd ../

> ecoPCR -d sixlegs -e 0 -l 50 -L 500 HATAATTTTYTTYATAGT \
AARAATCARAATAARTGT > COIpcr.ecopcr
```

-d sixlegs: OBITools-format collection of genomes
-e 0: maximum number of mismatches primer-sequence
-l 50 –L 650: minimum (*l*) and maximum (*L*) length of the amplified barcode
HATAATTTTYTTYATAGT AARAATCARAATAARTGT: Forward and Reverse primers*

*it doesn’t matter the order of the primers

```
> less COIpcr.ecopcr
```

```
#@ecopcr-v2
#
# ecoPCR version 0.2
# direct  strand oligo1 : HATAATTTTYTTYATAGT              ; oligo2c :              ACAYTTATTYTGATTYTT
# reverse strand oligo2 : AARAATCARAATAARTGT              ; oligo1c :              ACTATRAARAAAATTATD
# max error count by oligonucleotide : 0
# optimal Tm for primers 1 :   nan
# optimal Tm for primers 2 :   nan
# database : sixlegs
# amplifiat length between [50,500] bp
# output in superkingdom mode
# DB sequences are considered as linear
#
FJ171325       |     16036 |   279481 | species             |   279481 | Polystoechotes punctatus       |   279480 | Polystoechotes                |   279479 | Polystoechotidae              |
     2759 | Eukaryota                     | D | TATAATTTTTTTTATAGT              |  0 |   nan | AAAAATCAAAATAAATGT              |  0 |   nan |   486 | TATACCTATTGTTATTGGAGGATTTGGTAATTGATTAGTTCC
TTTAATACTAGCAGCACCTGATATAGCTTTCCCACGAATAAATAATATAAGTTTTTGAATATTACCTCCTTCCCTTACTCTTTTATTAGCATCAAGTATAGTTGAAAGAGGGGCTGGTACAGGATGAACTGTCTATCCACCTCTTTCTGCAGGAATTGCTCATGCAGGAGCTTCTGTTGATTTAGCAATTTTTAG
TTTACATTTAGCCGGTGTATCATCGATTTTAGGTGCTGTAAATTTTATTACAACTGTAATTAATATACGTTTATCACATATAACTTTAGACCGAATACCTTTATTTGTATGATCTGTTGTTATTACAGCTTTATTATTACTTTTATCTTTACCTGTTCTTGCTGGAGCTATTACAATACTTCTTACTGATCGTAA
TTTAAATACATCATTTTTTGACCCTGCTGGAGGAGGTGATCCTATTTTATATCA |  Polystoechotes punctatus mitochondrion, complete genome
FJ171324       |     15877 |   559169 | species             |   559169 | Ascaloptynx appendiculatus     |   559167 | Ascaloptynx                   |   146494 | Ascalaphidae                  |
     2759 | Eukaryota                     | D | TATAATTTTTTTTATAGT              |  0 |   nan | AAAAATCAAAATAAGTGT              |  0 |   nan |   486 | AATACCTATTGTAATTGGTGGATTTGGAAATTGATTAGTTCC
ACTTATACTAGCCGCACCAGACATAGCTTTTCCACGAATAAATAATATAAGTTTTGATTATTACCTCCTTCATTAACACTTCTTCTGGCTTCATCTGTTGTCGAAAGAGGTGCTGGGACAGGTTGAACAGTTTACCCCCCTCTATCTGCTGGAATTGCTCATGCAGGTGCTTCTGTTGACTTAGCCATTTTCAG
TTTACATTTAGCTGGGGTATCCTCAATTTTAGGAGCTGTTAATTTTATTACGACAGTAATTAATATACGACTTTCTTATATAACACTTGATCGAATAGCTTTATTTGTTTGATCAGTTGTTATTACAGCAATTTTACTATTACTATTATTACCAGTTTTAGCAGGTGCAATTACTATATTATTAACTGATCGAAA
TCTAAATACATCATTCTTTGACCCAGCAGGAGGTGGAGACCCAATTTTATATCA |  Ascaloptynx appendiculatus mitochondrion, complete genome
```

F primer

R primer

Length of the barcode

Sequence of the barcode

```
> obitools
> ecotaxstat -d sixlegs -r 6960 COIpcr.ecopcr
```

```
COIpcr.ecopcr 100.0 % |##################################################\ ] remain : 00:00:00
rank                          ecopcr          db      percent
class                              4           4       100.00
family                           190         298        63.76
genus                            461         763        60.42
infraclass                         3           3       100.00
infraorder                        19          21        90.48
kingdom                            1           1       100.00
order                             22          29        75.86
parvorder                          2           2       100.00
phylum                             1           1       100.00
species                          673        1115        60.36
species group                     14          18        77.78
species subgroup                   8          13        61.54
```

```
> ecotaxspecificity -d sixlegs -e 14 COIpcr.ecopcr
```

-e 14: number of base errors to be considered the same species for determination = 0.03 of the barcode's length -485- (species identification threshold=97%)

```
Alignment  : 0959 x 0982 -> 4604   99.5 % |##################################################\ ] remain : 00:00:00
rank                       taxon_ok      taxon_total     percent
order                            20               22       90.91
infraclass                        3                3      100.00
superfamily                      81               83       97.59
parvorder                         2                2      100.00
species group                    14               14      100.00
superkingdom                      1                1      100.00
kingdom                           1                1      100.00
phylum                            1                1      100.00
infraorder                       19               19      100.00
subfamily                       210              213       98.59
class                             4                4      100.00
species                         593              673       88.11
superorder                        1                2       50.00
suborder                         20               21       95.24
```

```
> exit
```