RNAcode: Getting started

January 5, 2011

1 Installing RNAcode on Unix like systems (Linux/OS X)

`RNAcode` is distributed as `tar.gz` compressed archive. Download the package to your machine and uncompress it, e.g.

```
# tar -xzf RNAcode-0.3.tar.gz
```

`RNAcode` uses the standard GNU installation system. So you can easily install it by running the following commands:

```
# ./configure
# make
# make install (requires root privileges)
```

This will install `RNAcode` in the `/usr/local/` tree. To install somewhere else run configure like that:

```
# ./configure --prefix=/home/stefan/programs
```

To test if `RNAcode` was successfully installed, run

```
# RNAcode -V
```

2 Obtaining multiple sequence alignments

To identify a coding region in your sequence of interest, you need one or more homologous sequences and create a multiple sequence alignment. The more sequences in your alignment the better the accuracy. However, if sequences are too similar ($> 90\%$) identity they only contribute little new information. Our benchmarks showed that alginments with 5–10 sequences and amean pairwise identity $< 90\%$ give good results.

There are different possible scenarios how to obtain alignments for your sequences.

2.1 Download pre-made alignments

These days, for many organisms the complete genomic sequence is known. Moreover, for many organisms also related species have been sequenced and pre-calculated multiple alignments are available for

download. Well known resources for major model organism are for example genome.ucsc.edu or ensemble.org, but also many independent smaller genome projects provide multiple alignments.

2.2 Create alignments of long genomic regions

If you want to analyze longer genomic regions ($> 1kb$) but cannot find pre-made alignments, we recommend using the MultiZ program suite to create alignments. It can be downloaded here http://www.bx.psu.edu/miller_lab/dist and comes with excellent documentation.

You will need homologous genomic sequences for your region in other species. For example, you can align the complete genomes of bacteria or align homologous loci in the megabase range of higher organisms with MultiZ.

2.3 Create alignments of individual short regions

If your sequence of interest is relatively short (a few hundred nucleotides) we recommend using a simple global alignment program like for example ClustalW. Use Blast to find homologous sequences in a sequence database (e.g. GenBank, www.ncbi.nlm.nih.gov/genbank/). Collect the significant hits that match to your region of interest and align the sequences afterwards with an alignment program.

3 Formatting the alignments

RNAcode can process alignments in two different formats: MAF and CLUSTAL W. You have to make sure that your alignment is in one of these formats before you can use RNAcode.

The MAF format was popularized by the USCS genome browser and is very useful to represent genome-wide alignments. The detailed specification can be found here: http://genome.ucsc.edu/FAQ/FAQformat.html. If you download alignments from an UCSC resource it is usually formatted as MAF. Also if you align your sequences using MultiZ the default output format is MAF.

For shorter alignments of individual regions, the CLUSTAL W format is useful. Apart from CLUSTAL W, many other alignment programs output their alignments in this format.

4 Pre-processing alignments

If your alignments contain blocks of long genomic regions it is usually no reasonable to score these long regions as a whole. The tar.gz package contains a script breakMAF.pl that allows you to easily pre-process your MAF files:

```
# scripts/breakMAF.pl examples/genomic.maf > genomic-preprocessed.maf
```

This command breaks blocks longer than 400 in shorter blocks of an average size of 200.

5 Running RNAcode

Analyze alignment with standard options and print detailed results page:

```
# RNAcode examples/coding.aln
```

Analyze alignment and show best non-overlapping hits below a p-value cutoff of 0.01 in gtf format:

```
# RNAcode --outfile out.gtf --gtf --best-only --cutoff 0.01 genomic-preprocessed.maf
```

Create color annotations for high scoring coding segments:

```
# RNAcode --eps coding.aln
```

Please refer to the more detailed README file that explains all options of `RNAcode` and how to interpret the results.

For details on the methodology refer to the paper:

RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data
Washietl S, Findeiß S, Müller S, Kalkhof S, von Bergen M, Hofacker IL, Stadler PF, Goldman N
RNA (2011), in revision