RNAplex tutorial

University of Leipzig

April 2010

Chapter 1

Introduction

RNAplex is a general RNA-RNA interaction prediction program designed for speed and precision. The following tutorial should provide the user of RNAplex with a good starting point on how to use RNAplex and what are the constraints to deal with. It should be noted that the perlscript used to realign target sequences and to find sRNA targets will be commented in another tutorial.

This tutorial assumes that the latest version of the ViennaRNA package is installed as well as RNAplex version 0.2. Test sequences, alignments and accessibility profiles are found in the directory ./sequence, ./alignments ./profiles, respectively

Chapter 2

RNAplex tutorial

2.1 Installing RNAplex

Download RNAplex from `www.bioinf.uni-leipzig.de/~htafer/RNAplex` or from `www.tbi.uni-leipzig.de/~htafer/RNAplex`. Unpack the tar file using `zcat RNAplex-0.2.tar.gz | tar -xvf -` Change into the `RNAplex` directory you just created. Type:

Compile

```
./configure --prefix=dirname
make
make install
```

to build the library and the programs in the `./Progs` directory and set the toplevel installation directory to `dirname` . If the default directory are preferred (`/usr/bin/` and `/usr/lib`) do not use the prefix option.

In case you set the `--prefix` option, do not forget to modify your environment variables. . For the C-shell the commands would be

Variable setting

```
setenv PATH ${PATH}:/wherever/RNAsnoop/Progs
setenv MANPATH ${MANPATH}:/wherever/RNAsnoop/man
```

You should now be able to execute the program and read the man page. Take a look at the `RNAplex` man page by typing, `man RNAplex`.

2.1.1 RNAplex command-line options

The extensive number of command-line options accepted by `RNAplex` allow the user to fine-tune the RNA-RNA target search. These options are

RNAplex options

```
-q : File containing query sequences. Must be used with the
   -t
option.
-t : File containing target sequence. Must be used with the
   -q
option.
-l : Maximum interaction length
-c : Fix extension penalty
-z : Size of the windows over which the best candidate for
   the
backtracking step is chosen
-T : Temperature at which the hybridization occurs
-e : Energy threshold for an interaction to be returned
-f : if -f is set to 1, no backtracking is done
-P : Reads alternative energy file
-V : Scaling factor to increase or decrease the effect of
   the
opening energies
-C : Constraint folding
-A : Use alignments as input
-a : Use RNAplfold accessibility profiles.
-I : Allows to draw an structure annotated alignment of the
   interaction
```

Figure 2.1: Description of RNAplex's command line options.

shown in figure 2.1.

2.2 Single sequence prediction

In single sequence mode, RNAplex can take input sequences in two different ways. Either from a file containing target and query sequences one below the other:

Input file structure (inputfile.fa)

```
>target1
CAGUCAGCUAGCUAGCAGCUAGCUAGCAUGCUAGCUAGCUAGCUAGCAGCUGA
>query1
CGAUCAGUCAGUCAGUCGAUCGAUGCUAGCAUGCAGCUA
>target2
CAGCUAGCAUGCAUGCAUGCAUGGUGUGUGUUGUGGCAGACACACGUAGCUAGCA
```

```
>query2
CAGCCACACACCACCACACAGUGUGUGCACACCACACAACGUCAGUCAG
```

In this case the command-line would look like

Simple RNAplex call

```
RNAplex < inputfile.fa
or
cat inputfile.fa  | RNAplex
```

If one wants two check all sequences in one file against all sequences in another files, then RNAplex can be called in the following way:

RNAplex with two files

```
RNAplex -t target.fa -q query.fa
```

We should note that for large numbers of query and target sequences, RNAplex performs faster with the $-q$ and $-t$ options.

If one is interested into all interactions below a given threshold then RNAplex should be called with the $-e$ option:

Suboptimals RNAplex call

```
RNAplex -e -10 < inputfile.fa
```

A threshold on the maximum size of the duplex can be set with the $-l$ option. An alternative way of constraining the length of a duplex is given by setting the $-c$ option which penalizes the addition of a nucleotide to the duplex. A good value for $-c$ is 30 (0.3kcal/(mol·nt))

Length limited target search

```
RNAplex -e -10 -l 25 -c 30 < inputfile.fa
```

The extent of overlap on the target between the various suboptimals is controlled by the parameter $-z$. If $-z$ is set to $-l$, then no overlap occurs between the suboptimal. On the other hand, for $-z1$ all possible interactions with an interaction energy below the energy threshold will be returned. Setting $-z$ to $-l$ is a good compromise between sensitivity and runtime.

If accessibility profiles are available, one can use them to improve the predictions returned by RNAplex. If none are available they can be generated by RNAplfold from the ViennaRNA-package http://www.tbi.univie.ac.at/RNA/. Here one should note that the $-u$ option of RNAplfold should be at least as long as the $-l$ option from RNAplex. We should further note that RNAplex expects the accessibility files to be in the $-a$ directory. Further RNAplex expects the accessibility-file names to follow the name pattern

from RNAplfold,i.e if RNAplfold is used on the inputfile.fa, then 4 files are produced:

1. target1_openen
2. query1_openen
3. target2_openen
4. query2_openen

The standard command-line call for RNAplfold is :

RNAplfold usage

```
RNAplfold -W 240 -L 160 -u 30 -O < inputfile.fa
```

Please have a look at the man page of RNAplfold for more information. If the accessibility profiles are stored in the folder $./profiles$, then RNAplex should be run with the following options:

Accessibility usage

```
RNAplex -e -10 -l 25 -a ./profiles < inputfile.fa
```

We should note that the effect of the accessibility can be modulated by setting the $-V$ option, which allows to multiply the opening energy by a given factor, e.g. $-V0.5$ would divide the opening energiesby two.

RNAplex can also be run with constraints $-C$ on the query sequences. The only constaint currently available is to force the binding of one of the nucleotide with the | symbol. To this aim the inpute files should be modified in order to contain the constraints:

Constrained usage

```
RNAplex -C -e -10 -l 25 -a ./profiles < inputfile.fa
RNAplex -C -q query.fa -t target.fa -e -10 -l 25 -a ./
   profiles
```

inputfile.fa

```
>target1
CAGUCAGCUAGCUAGCAGCUAGCUAGCAUGCUAGCUAGCUAGCUAGCAGCUGA
>query1
CGAUCAGUCAGUCAGUCGAUCGAUGCUAGCAUGCAGCUA
.|||||.
>target2
CAGCUAGCAUGCAUGCAUGCAUGGUGUGUGUUGUGGCAGACACACGUAGCUAGCA
>query2
CAGCCACACACCACCACACAGUGUGUGCACACCACACAACGUCAGUCAG
.|||||.
```

query.fa

```
>query1
CGAUCAGUCAGUCAGUCGAUCGAUGCUAGCAUGCAGCUA
.|||||.
>query2
CAGCCACACACCACCACACAGUGUGUGCACACCACACAACGUCAGUCAG
.|||||.
```

RNAplex can be run at different temperature $-T$ and with alternative parameter files $-P$.

Alternative temperature and energy parameters

```
RNAplex -T 63 -P DNA -C -e -10 -l 25 -a ./profiles <
   inputfile.fa
```

Please note that for the sake of consistency one should compute the accessibility profile under the same conditions, i.e. $-T$ and $-P$ options.

If only energy information are wanted, then the $-f$ option can be set to 1. In this case the backtracking step is switched off and no information about the structure is returned. In this mode the $-l$ option is uneffective as no information on the duplex length is available.

With option $-A$ RNAplex can use comparative information in order to improve specificity. Instead of single sequences, RNAplex takes as input a query and a target multiple sequences alignment files in $clustalw$ format. The user should ensure that both alignments contain the same number of sequences and that the sequences are ordered in the same way. Further no gap columns are allowed. Similar to the single sequence version, the $-a$, $-e$, $-l$, $-T$, $-P$,$-z$,$-f$ and $-V$ options can also be used. A typical call of RNAplex with alignments would be:

Comparative target search

```
RNAplex -A target.aln query.aln -e -10 -l 25 -a ./profiles
```

RNAplex can be used to generate structure annotated multiple sequence alignments of targets and queries. To this aim RNAplex needs a file containing results from a previous RNAplex run with comparative information as well as the two alignment files used to compute the interactions:

Input target alignment

```
ColiAPE_APECO1_62/253-270   -ATGATAACGAGGCGCAAA
eColi_536_ECP_0962/181-198  -ATGATAACGAGGCGCAAA
eColi_K12_b0957/181-198     -ATGATAACGAGGCGCAAA
eColi_OH_EDL_Z1307/181-198  -ATGATAACGAGGCGCAAA
ent638_Ent638_1469/180-197  GATGATAACGAGGCGCAA-
erwipecto_ECA1751/182-199   GATGATAATGAGGCGTAA-
salmPARATY_SPA1780/181-198  -ATGATAACGAGGCGCAAA
salmTyph_TY2_t1850/181-198  -ATGATAACGAGGCGCAAA
salmTyphimu_STM1070/181-198 -ATGATAACGAGGCGCAAA
salm_CHOL_SC1022/181-198    -ATGATAACGAGGCGCAAA
shigFlex_2A_SF0957/181-198  -ATGATAACGAGGCGCAAA
sodaGlos_SG1030/180-197     GATGATAACGAGGCGCAA-
yPs_YpsIP31758_2542/194-211 GATGATAATGAGGCGTAA-
```

Input query alignment

```
ColiAPE_micA/1-75       CGCGCAUUUGUUAUCAU
eColi_536_micA/1-75     CGCGCAUUUGUUAUCAU
eColi_K12_micA/1-75     CGCGCAUUUGUUAUCAU
eColi_OH_EDL_micA/1-75  CGCGCAUUUGUUAUCAU
ent638_micA/1-75        CGCGCAUUUGUUAUCAU
erwipecto_micA/1-72     CGCGCAUUUAUUAUCAU
salmPARATY_micA/1-76    CGCGCAUUUGUUAUCAU
salmTyph_TY2_micA/1-76  CGCGCAUUUGUUAUCAU
salmTyphimu_micA/1-76   CGCGCAUUUGUUAUCAU
salm_CHOL_micA/1-76     CGCGCAUUUGUUAUCAU
shigFlex_2A_micA/1-75   CGCGCAUUUGUUAUCAU
sodaGlos_micA/1-75      UGCGCAUUUGUUAUCAU
yPs_micA/1-74           CGCGCAUUUGUUAUCAU
```

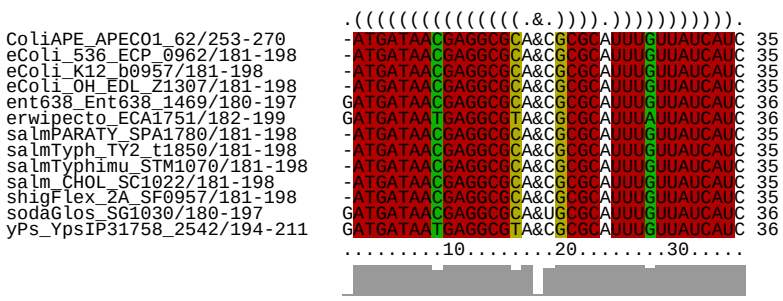
Structure annotated alignment

```
RNAplex -A target.aln query.aln > result
RNAplex -A target.aln query.aln -I result
```

The name of the returned postscript file is composed by the name of the first sequence of the target alignment, followed by the name of the first sequence of the query alignment, followed by the coordinates of the duplex on the target and query alignments:

Name pattern of the output files

```
ColiAPE_APECO1_62_253-270_ColiAPE_micA_1-75_1_17_1_18.ps
```



The target and query alignments are located left and right of the & column, respectively. The conservation profile is located at the bottom of the alignment, while the consensus structure is found on top of the alignment. The red, ocher and green colors stand for interactions with one, two and three type of base pairs. For more information on the stuctural annotation of alignments please look at the `RNAalifold` man page.