Seminar III: R/Bioconductor
GeneR

Amhed Missael Vargas Velazquez
avargas@lcg.unam.mx

September 25, 2009
What is GeneR?

GeneR is a package that allows direct use of nucleotide sequences within R software. Functions can be used to read and write sequences from main file formats (Embl, Genbank and Fasta) in order to perform a lot of manipulations and analyses.
What is GeneR?

GeneR is a package that allows direct use of nucleotide sequences within R software. Functions can be used to read and write sequences from main file formats (Embl, Genbank and Fasta) in order to perform a lot of manipulations and analyses.

▶ Authors
What is GeneR?

GeneR is a package that allows direct use of nucleotide sequences within R software. Functions can be used to read and write sequences from main file formats (Embl, Genbank and Fasta) in order to perform a lot of manipulations and analyses.

▶ Authors

▶ Maintainer
Y. d’Aubenton-Carafa
What is GeneR?

GeneR is a package that allows direct use of nucleotide sequences within R software. Functions can be used to read and write sequences from main file formats (Embl, Genbank and Fasta) in order to perform a lot of manipulations and analyses.

- **Authors**

- **Maintainer**
  Y. d’Aubenton-Carafa

- I think that Y. d’Aubenton-Carafa, entered the proyect at the end :)

GeneR is a very useful package which contains some functions for the manipulation of genetic data. It’s similar to Biostrings\textsuperscript{1}, However, GeneR contains more functions and it used for different things. In addition, it is related to GeneRfold\textsuperscript{2} package that allows the use of Vienna RNA library within R, meaning, tools for the prediction and comparison of RNA secondary structures.\textsuperscript{3} You can install the GeneR package in R using:

\begin{verbatim}
> source("http://bioconductor.org/biocLite.R")
> biocLite("GeneR")
\end{verbatim}

\textsuperscript{1}Biostrings was showed in the previous class by Isaac
\textsuperscript{2}A package created by Y. d’Aubenton-Carafa, A. Lucas; C. Thermes, the same creator as the GeneR package XD
\textsuperscript{3}It’s an excellent package to talk about, and it is also interesting and easy to use.
What is it used for?

- Reading and writing sequences
  Fast sequence retrieving even from very large sequence databanks, in Fasta, Embl or Genbank formats.
What is it used for?

- Reading and writing sequences
  Fast sequence retrieving even from very large sequence databanks, in Fasta, Embl or Genbank formats.

- Handling sequences
  The usual copy-paste of parts of sequences or other manipulations can be performed by functions using vectors of strands and positions.
What is it used for?

- Reading and writing sequences
  Fast sequence retrieving even from very large sequence databanks, in Fasta, Embl or Genbank formats.

- Handling sequences
  The usual copy-paste of parts of sequences or other manipulations can be performed by functions using vectors of strands and positions.

- Analyzing sequences
  To count oligo-nucleotides by mono, di or tri, to look for exact word positions or to shuffle sequences.
What is it used for?

- Reading and writing sequences
  Fast sequence retrieving even from very large sequence databanks, in Fasta, Embl or Genbank formats.

- Handling sequences
  The usual copy-paste of parts of sequences or other manipulations can be performed by functions using vectors of strands and positions.

- Analyzing sequences
  To count oligo-nucleotides by mono, di or tri, to look for exact word positions or to shuffle sequences.

- Manipulation of regions on a chromosome
  Tools to easily compute any subregions (intergenic regions, exons or more sophisticated regions), without an exhaustive texture on a whole chromosome.
What is it used for?

- Performing bioinformatic jobs
  Functions related to genetic and structural aspects of the sequences: ORF localization, translation, or RNA secondary structure determination\(^4\).

\(^4\)with extension of GeneR: GeneRfold package
I create a random sequence for the samples

\[ \text{> library(GeneR)} \]
\[ \text{> seq }\leftarrow \text{randomSeq(prob = c(0.2, 0.3, 0.2, 0.3), letters = c("T", "C", "A", "G"), n = 30)} \]

Insert a poly A into the end of the sequence

\[ \text{> seq }\leftarrow \text{insertSeq(seq, "AAAAAAAAAAAA", 30)} \]
\[ \text{> seq} \]

[1] "GAAACAGAGGCTCCTCTGGCTTCGTTTACAAAAAAAAAAAAAC"

---

So sorry my friends, but this is a brief description of the GeneR, so I'm not going to explain each function. ; p
Compute the reverse complementary

> strComp(seq)

[1] "GTTTTTTTTTTTGTAACGAAAGCCAGAGGAGCCTCTGTTTTTC"

Count di-nucleotides

> strCompoSeq(seq, wsize = 2)

     TT  TC  TA  TG  TX  CT  CC  CA  CG  CX  AT  AC
[1,] 0.1 0.05 0.05 0.05 0 0.05 0.05 0.1 0.05 0 0 0.05 0.

     GG  GX  XT  XC XA  XG  XX
[1,] 0.05 0 0 0 0 0 0

Translate the sequence string to a protein

> strTranslate(seq)

[1] "ETEAPLASFTKKK"

It can be in groups from 1 to 15
Most of the functions in the GeneR package use buffers.

- Why use buffers
Most of the functions in the GeneR package use buffers.

- Why use buffers
- To work on large sequences (i.e. a whole chromosome).
Most of the functions in the GeneR package use buffers.

- Why use buffers
- To work on large sequences (i.e. a whole chromosome).
- In addition, you can buffer fasta sequences from Ncbi
Buffering the complete genome of Nanoarchaeum equitans\textsuperscript{7} from Ncbi.

\begin{verbatim}
> seqNcbi("NC_005213", file = "toto.seq", submotif = TRUE + , type = "fasta")
[1] 1

> readFasta("toto.seq")
[1] 0
\end{verbatim}

Size of the genome.

\begin{verbatim}
> sizeSeq()
[1] 490885
\end{verbatim}

Looking for motifs\textsuperscript{8}.

\begin{verbatim}
> exactWord("ACTGA", seqno = 0, case.sensitive = TRUE)
\end{verbatim}
One of the most little genomes, i don't wanna break my computer

Also, there is a function named getOrfs, that is supposed used to know where find Open Reading Frames, however, is not working :(

Amhed Missael Vargas Velazquez avargas@lcg.unam.mx
DNA TO RNA

```r
> dnaToRna()
[1] 0
```

Or writing our new RNA file

```r
> writeFasta(seqno = 0, file = "Nan_rna.fa", name = + "MyRNA", comment = "RNA generated by DNA + of Nanoarchaeum equitans", append = TRUE)
[1] 1
```

You must remember, any function that uses the buffer, changes the content of the buffer.

We changed our DNA, so that if we use a `getSeq` you will see RNA

```r
> getSeq(seqno = 0, from = 1, to = 30)
[1] "UCUCCGACAGAUCUUCUUUGUUAACAAA"
```

You might prefer to change the number of the buffer for anything that you might do.
We already see in one of our class, how is constitute a bacterial genome...
So, why not use the functions to do a brief review the genome of the Rhizobium etli. We want to know:
We already see in one of our class, how is constitute a bacterial genome...
So, why not use the functions to do a brief review the genome of the Rhizobium etli. We want to know:
  - The size
We already see in one of our class, how is constitute a bacterial genome...
So, why not use the functions to do a brief review the genome of the Rhizobium etli. We want to know:

- The size
- The GC content
We already see in one of our class, how is constitute a bacterial genome...
So, why not use the functions to do a brief review the genome of the Rhizobium etli. We want to know:

- The size
- The GC content
- A GC Skew of the genome
Buffering the sequence

> seqNcbi("NC_007761", file = "Retli.seq", submotif = + TRUE, type = "fasta")

[1] 1

> readFasta("Retli.seq")

[1] 0

The size

> sizeSeq()

[1] 4381608

The GC content

> GCcontent()

    pgc  N
    G 0.6127221 0
For the GC skew, I create an object with the size for sectionate the genome

```r
> size <- sizeSeq()
```

And now we use the function `densityProfile`

```r
> dens <- densityProfile(ori = 398328 * (1:11), from = 1,
+ to = size, seqno = 0, fun = seqSkew, nbinL = 24, nbinR = 24, sizeBin = 16597)
```

At last, we plot :) 

```r
> plot(dens$skgc, main = "GC skew")
```

[1] 1
GC skew

Amhed Missael Vargas Velazquez
avargas@lcg.unam.mx

Seminar III: R/Bioconductor GeneR
GeneR has great tools:

- To find a region in the genome
- To manipulate sequences
- To do large jobs

As we see GeneR has the potential to be an excellent tool for conducting bioinformatics.
GeneR has great tools:

- To find a region in the genome
GeneR has great tools:

- To find a region in the genome
- To manipulate sequences
GeneR has great tools:
- To find a region in the genome
- To manipulate sequences
- To do large jobs

As we see Gene R has the potential to be an excellent tool for conducting bioinformatics.
I encourage you to explore the Help Options of this package and to use them, they’re user - friendly and fun XD .