R바이오 프로세싱 - 1

발표자 : 이승우

오늘 발표할것은

~~반 삼질 반 스트레스 반 카페인~~

~~- 주피터 노트북과 R 커널을 셋팅해볼 것이며 (되어있다면 패스합니다)~~

- R Bioconductor, SeqinR 설치 해볼것이며
- Bio Database(NCBI, EMBL, DDBJ)에 대하여 알아보고
- 실제로 한번 입력해보도록 하겠습니다
- 그리고 문제 - Exercises 까지 발표해보는

시간을 가져보도록 하겠습니다

IRKerenl, R Bioconductor, SeqinR 을 설치해봅시다

IRKernel 설치

- Jupyter 에는 다른 언어들과 연동할수 있도록 커널을 만들어 지원합니다, 여기에서는 IRKernel 을 사용해보도록 하겠습니다
- R GUI를 연 후에 (우분투는 터미널에서 R 입력하세요)

  install.packages(c('pbdZMQ', 'repr', 'devtools'))

  devtools::install_github('IRkernel/IRkernel')

  IRkernel::installspec() // 설치 완료!

데체 뭐가 문제일까?

[처음시도**]** R을 새로운 버전으로 설치 시도 → 근데 안됨

[다시 생각해본다**]** 그럼 내가 뭘 잘못했지... 아까 보니까 Personal Package Library 뭐시기가 안된다고 하던데...
다시 설정해봐? → 안됨 // mkdir R/lib 일단 해보세요

[아아아 모르겠다**]** 스택오버플로우에 검색! → /etc/apt 뭐시기에 deb 에다가 cran 명령줄 추가하시고 뭘 apt-get 으로 설치하세요! → 된다!

[오오오오옷**?!]** 설치되었네? 내 환경(VirtualBox, Ubuntu, Port-Forwareded)에서 jupyter notebook --ip=0.0.0.0 을 입력해서 지원하는 언어를 보자! → 보이는구나!

Q. 근데 왜 **--ip=0.0.0.0** 을 붙이셨어요**?**

A. 버추얼박스**(VM)** 환경이라 네트워크 실제 물리적인 컴퓨터에서 돌 수 있도록 하기 위하여 설정해놓았습니다

입력할때는 http://localhost:포트번호 입력하셔야 합니다

그럼 여기서 정리

~~처음부터 찬찬히 프로세스~~

[주의!] 일단 저 위의 방법대로 설치해보시고 안되면 이 방법으로 시도해보시길 추천해드립니다

이 링크와 이 링크의 방법을 사용하였습니다

- 일단 R 의 버전을 확인합시다 **R --version** 을 입력하면 출력됩니다

최신 버전은 16년 7월 25일 자로 **3.3.1** 입니다

- 후에 **sudo apt-get update** 와 **sudo apt-get upgrade** 를 사용하여 업그레이드 합니다
- 그런 후에 R의 최신 버전이 존재하는지 확인합니다 **sudo apt-cache showpkg r-base** 를 입력합니다
- **sudo apt-get install r-base** 를 입력하여 설치를 다시 시도해봅니다
- **/etc/apt/sources/list** 를 vim 같은 에디터로 여신 후에 다음 구문을 추가합니다

  deb http://cran.rstudio.com/bin/linux/ubuntu trusty/
- 그 후에 **sudo apt-get install libssl-dev** 를 입력하여 설치합니다

다른것들도 설치해봅시다
Bioconductor & SeqinR

R GUI 환경에서 입력합니다

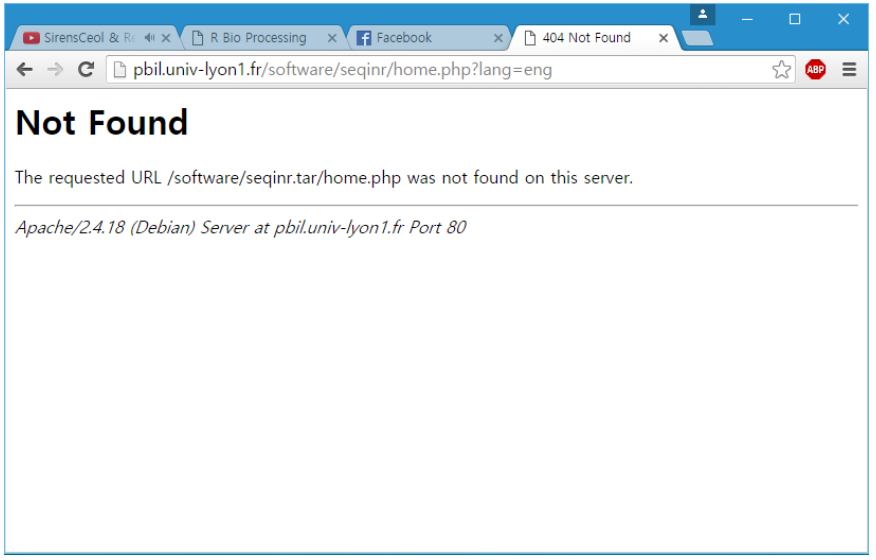
- 바이오컨덕터 설치와 사용하는 방법
  - 설치하는 방법과 사용하는 방법은 이렇습니다 (코드가 같습니다)

    source("http://bioconductor.org/biocLite.R")

    biocLite()
  - 바이오컨덕터 패키지를 설치하려면

    biocLite(c("GenomicFeatures", "AnnotationDbi")) // GenomicFeatures, AnnotationDbj 를 설치하는 명령어입니다
- SeqinR 을 설치...

해보려고 했는데...


SirensCeol &
R Bio Processing
Facebook
404 Not Found
pbil.univ-lyon1.fr/software/seqinr/home.php?lang=eng
Not Found
The requested URL /software/seqinr.tar/home.php was not found on this server.
Apache/2.4.18 (Debian) Server at pbil.univ-lyon1.fr Port 80

~~링크가 사라졌습니다~~

R GUI 환경에서 입력합니다

- 바이오컨덕터 설치와 사용하는 방법
  - 설치하는 방법과 사용하는 방법은 이렇습니다 (코드가 같습니다)

    source("http://bioconductor.org/biocLite.R")

    biocLite()
  - 바이오컨덕터에서 패키지를 설치하려면 (밑 코드는 예시입니다!)

    biocLite(c("GenomicFeatures", "AnnotationDbi")) // GenomicFeatures, AnnotationDbj 를 설치하는 명령어입니다
- SeqinR 을 설치합니다 → 공식 홈페이지 링크

  install.packages("seqinr") 을 입력하여 설치합니다

  library(seqinr) 을 입력하여 호출 한 후에 무수히 많은 함수를 사용할 수 있습니다

3. FASTA 포맷

gi|186681228|ref|YP_001864424.1| phycoerythrobilin:ferredoxin oxidoreductase

MNSERSDVTLYQPFLDYAIAYMRSRLDLEPYPIPTGFESNSAVVGKGKNQEEVVTTSYAFQTAKLRQIRA

AHVQGGNSLQVLNFVIFPHLNYDLPFFGADLVTLPGGHLIALDMQPLFRDDSAYQAKYTEPILPIFHAHQ

QHLSWGGDFPEEAQPFFSPAFLWTRPQETAVVETQVFAAFKDYLKAYLDFVEQAEAVTDSQNLVAIKQAQ

LRYLRYRAEKDPARGMFKRFYGAEWTEEYIHGFLFDLERKLTVVK

규칙들은 IUB/IUPAC 에 의하여 결정됩니다

몇가지 규칙에 대하여 이야기해보면

- 소문자는 허용되나 대문자로 맵핑됩니다
- 하이픈이나 대쉬는 gap of indeterminate length 로 사용할수 있습니다
- 아미노산 시퀀싱에서, U 와 * 은 허용되는 문자입니다
- any numerical digits in the query sequence should either be removed or replaced by appropriate letter codes

NCBI Sequence Database 에 대하여

NCBI 는 미국이 운영하고 있는 가장 큰 DNA와 단백질 시퀀싱 데이터가 있습니다, 유럽과 일본에도 있습니다

- 여기에서 미국은 NCBI 데이터베이스를 운영하고 있습니다 → www.ncbi.nlm.nih.gov
- 유럽은 EMBL 데이터베이스를 운영하고 있습니다 → www.ebi.ac.uk/embl
- 일본은 DDBJ 데이터베이스(혹은 은행) 을 운영하고 있습니다 → www.ddbj.nig.ac.jp

이들은 매 밤시간에 동기화되며, 거의 같은 데이터들이 있으나 그것들이 불리는 방식은 틀릴 수 있습니다

~~한국형 저장소는 없는겁니까 필요한곳에 완하고 왜 맨날 실패하는 한국형 알파 음음 당신 누구야~~

주피터 노트북 예제와
Exercises

오늘 발표한것은

1. length() 함수를 사용하여 백터나 리스트의 길이를 구해봤고
2. table() 함수를 사용하여 벡터나 리스트에서의 갯수를 출력할 수 있고 (ACGT 의 갯수들)
3. GC() 함수를 이용하여 DNA Sequence 에서의 GC Percentage 를 구해봤으며
4. count() 함수를 사용하여 각 유전자 갯수를 구해봤습니다

감사합니다