Parsing Data

Talk to a Teacher
http://spoken-tutorial.org
National Mission on Education through ICT
http://sakshat.ac.in

Snehalatha Kaliappan
IIT Bombay

15 August 2015

Learning Objectives

Learning Objectives

- **Download FASTA and GenBank files from NCBI database website**

Learning Objectives

- **Download FASTA and GenBank files from NCBI database website**
- **Parse data files using functions in Sequence Input/Output (SeqIO) module**

Pre-requisites

Pre-requisites

- **Familiar with Undergraduate Biochemistry or Bioinformatics**

Pre-requisites

- **Familiar with Undergraduate Biochemistry or Bioinformatics**
- **Basic Python programming**

Pre-requisites

- Familiar with Undergraduate Biochemistry or Bioinformatics
- Basic Python programming
- Refer to Python Spoken Tutorials at http://spoken-tutorial.org

System Requirements

System Requirements

- **Ubuntu OS version 14.10**

System Requirements

- **Ubuntu OS version 14.10**
- **Python version 2.7.8**

System Requirements

- **Ubuntu OS version 14.10**
- **Python version 2.7.8**
- **IPython interpreter version 2.3.0**

System Requirements

- **Ubuntu OS version 14.10**
- **Python version 2.7.8**
- **IPython interpreter version 2.3.0**
- **Biopython version 1.64**

System Requirements

- **Ubuntu OS version 14.10**
- **Python version 2.7.8**
- **IPython interpreter version 2.3.0**
- **Biopython version 1.64**
- **Mozilla Firefox browser 35.0**

Data Files

Data Files

- **Data File Formats:**
  FASTA, GenBank, Swiss-Prot, EMBL
- **http://www.ncbi.nlm.nih.gov/nucleotide**

Parsing

Parsing

- **Extracting data from data files is called parsing**

Parsing

- **Extracting data from data files is called parsing**
- **Most file formats can be parsed using functions available in SeqIO module**

Parsing

- Extracting data from data files is called parsing
- Most file formats can be parsed using functions available in SeqIO module
- Functions of SeqIO: parse, read, write and convert

Summary

- Download **FASTA** and **GenBank** files from **NCBI** database website.
- Use **parse** and **read** functions of **SeqIO** module:
- To extract data such as record id's, description and sequences from **FASTA** and **GenBank** files

Assignment

Assignment

- **Download FASTA files for nucleotide sequence of your choice from NCBI database.**
- **Convert the sequences in the file to their reverse complements**

About the Spoken Tutorial Project

- **Watch the video available at http://spoken-tutorial.org /What_is_a_Spoken_Tutorial**
- **It summarises the Spoken Tutorial project**

About the Spoken Tutorial Project

- Watch the video available at http://spoken-tutorial.org/What_is_a_Spoken_Tutorial
- It summarises the Spoken Tutorial project
- If you do not have good bandwidth, you can download and watch it

Spoken Tutorial Workshops

The Spoken Tutorial Project Team

- **Conducts workshops using spoken tutorials**
- **Gives certificates to those who pass an online test**
- **For more details, please write to contact@spoken-tutorial.org**

Acknowledgements

- Spoken Tutorial Project is a part of the Talk to a Teacher project
- It is supported by the National Mission on Education through ICT, MHRD, Government of India
- More information on this Mission is available at http://spoken-tutorial.org/NMEICT-Intro