Data formats and scripting languages

Why formats?

- Data that has been generated by equipment or retrieved from a database is written to a file in a particular configuration (format).
- A *file format* is a formal set of rules about how data are to be encoded into a file.
- Otherwise, we cannot guarantee that the computer will retrieve the original data from a file.

Tabular data

- Tables are a fundamental mode of organizing data
- Rows represent observations
- Columns represent variables

An excerpt of a table summarizing infectious disease prevalence in Britain (1965):

Cases	Eng. & Wales	Grt. Lnd	Scot.	N.Ire.	Eire
Diptheria	0	0	1	0	
Dysentry	482	85	136	5	2
Encephatlis, acute	2	0		0	
Enteric fever, typhoid	3	1	4	1	
Measles	3268	153	37	40	73

British Medical Journal from 1965 (Graph and Table of Infectious Disease)

Comma-separated values (CSV)

- A common format for writing tabular data to a text file.
- Each row appears on a separate line.
- Values are separated by a reserved symbol (delimiter)
- If the value contains the delimiter, enclose the value in quotes.

```
Cases,Eng. & Wales,Grt. Lnd,Scot.,N.Ire.,Eire
Diptheria,0,0,1,0,
Dysentry,482,85,136,5,2
"Encephatlis, acute",2,0,,0,,
"Enteric fever, typhoid",3,1,4,1,
Measles,3268,153,37,40,73
```

Sequence data formats

- Sequence data are more complicated.
- The relative position of a nucleotide or amino acid is significant.
- Different sequence lengths breaks a standard assumption of tabular data.
- We may need to associate sequences with complex metadata.

Nucleotides

- IUPAC defines the following symbols to represent nucleotides and their mixtures:

Symbol	Bases	Symbol	Bases	Symbol	Bases
W	A,T	B	A,C,T	N	A,C,G,T
R	A,G	D	A,G,T	X	A,C,G,T
K	G,T	H	A,C,T		
Y	C,T	V	A,C,G		
S	C,G				
M	A,C				

Can you think of a mnemonic for 2-fold mixtures?

FASTA

- One of the most common file formats for sequence data.
- Every sequence record starts with a > symbol, followed by the sequence label (header).
- The sequence appears on subsequent lines until the next >.

```
>KP728283.1 Zaire ebolavirus isolate Ebola virus/H.sapiens-wt/CHE/2014/M
GGATCTTTTGTGTGCGAATAACTATGAGGAAGATTAATAATTTTCCTCTCATTGAAATTTATATCGGAAT
TTAAATTGAAATTGTTACTGTAATCATACCTGGTTTGTTTCAGAGCCATATCACCAAGATAGAGAACAAC
```

NEXUS

- The NEXUS format was designed to incorporate many different data types, including sequences.
- Data are organized into *blocks* enclosed by `BEGIN` and `END` tags.

```
#NEXUS
BEGIN DATA;
DIMENSIONS  NTAX=4 NCHAR=140;
FORMAT DATATYPE=DNA GAP=- MISSING=?;
MATRIX
AF084930.1 Proteus vulgaris  GGATCCGGGGAGGAAAGTCCGGGCTCC...
P.aeruginosa RNase P RNA     AGAGUCGAUUGGACAGUCGCUGUCGCG...
;
END;
```

FASTQ format

- FASTQ is essentially an extension of the FASTA format to include quality scores.

```
@SRR5261740.1 1 length=295
AAGCAGTGGTATCAACGCAGAGTACATGGGGACAGTGACCCTGATCTGGTAAAGCTC...
+SRR5261740.1 1 length=295
BBBBBFFFBFFFGGGGGGGGGGHHHHHHHHHGGGGGHHHHGHHHGHHHHHHHHHGHHHH...
```

- A *quality score* is the predicted probability that the base call is incorrect.
- To save space, this probability is transformed into a single character (more on this later!)

Genbank format

- A very complex format that contains a diverse amount of information:

```
LOCUS       NC_014372               18935 bp    cRNA    linear   VRL 13-A
DEFINITION  Tai Forest ebolavirus isolate Tai Forest
            virus/H.sapiens-tc/CIV/1994/Pauleoula-CI, complete genome.
ACCESSION   NC_014372
```

```
FEATURES             Location/Qualifiers
     source          1..18935
                     /organism="Tai Forest ebolavirus"
                     /mol_type="viral cRNA"
                     /isolate="Tai Forest virus/H.sapiens-tc/CIV/1994/Pa
                     /host="Homo sapiens"
                     /db_xref="taxon:186541"
                     /country="Cote d'Ivoire"
                     /collection_date="Nov-1994"
                     /note="Ivory Coast ebolavirus"
```

Converting between file formats

- One of the fundamental tasks in bioinformatics is the conversion of data from one format to another.
- Different programs write data to files in different formats, even when the data contain the same information.
- Converting formats is often a required step to feed the output of one program as input for another (building *pipelines*).

EMBOSS Seqret

- https://www.ebi.ac.uk/Tools/sfc/emboss_seqret/

How do we really do it?

- Web interface not adequate for building an analysis pipeline
- Several open-source programs for converting formats:
  - BioPython.SeqIO module
  - seqmagick, essentially a front-end for SeqIO
- but often we just have to do it ourselves -- this is why scripting languages are so popular.

How Perl saved the Human Genome Project

An article in *The Perl Journal* by Lincoln Stein

- Project start date 1990, involving many groups.
- Estimated 1 to 10 terabytes needed to complete project.
- Different groups came up with different data exchange formats.

> *Despite the fact that everyone was working on the same problems, no two groups took exactly the same approach.*

How Perl saved the Human Genome Project (2)

> *The long range solution to this problem is to come up with uniform data interchange standards that genome software must adhere to. [...] However, standards require time to agree on...*

- Perl enabled different groups to rapidly convert outputs to the other group's format.

How Perl saved the Human Genome Project (3)

> *Some groups attempted to build large monolithic systems on top of complex relational databases; they were thwarted time and again by the highly dynamic nature of biological research. By the time a system that could deal with the ins and outs of a complex laboratory protocol had been designed, implemented and debugged, the protocol had been superseded by new technology and the software engineers had to go back to the drawing board.*

What is Perl?

- Perl is a programming language.
- It is an *interpreted* language:

	Compiled	Interpreted
Running	Compile once	Interpret every time
Openness	Distribute binaries	Distribute source
Performance	Faster	Slower
Development	Difficult	Easy
Analogy	Buying a Tesla	Renting a bike

Perl is a scripting language

- "Scripting language" is difficult to define.
- Generally a script is a program that automates the execution of tasks.
- A scripting language is often developed using a compiled language - it operates at a higher (more abstract) level.
- (Like making macaroni and cheese with a Kraft Dinner mix instead of growing a wheat field, harvesting the grain, milling the grain into flour...)

Perl was the backbone of bioinformatics

- Perl (Practical Extraction and Reporting Language) was developed by Larry Wall (first release 1987)
- The *lingua franca* of bioinformatics for many years.
- Reputation for enabling developers to do work quickly ("there's more than one way to do it"), but can be difficult to read.

A bit of Perl code by Illumina to find genome positions with an N

```
my $a=();
my $start=0;
#====finding Ns in the specified genomic range=====
open(IN, "$ARGV[0]"); #genome fasta
my @lines = <IN>;
close IN;
my $pos = $start;
my %Ns = ();
for($i=1;$i<@lines;$i++) #The 1st line is header
{
  chomp $lines[$i];
  @a = split //,$lines[$i];
  for($j=0;$j<@a;$j++) {
    if($a[$j] eq 'N') {
      $Ns{$pos} = 1;
    }
    $pos++;
  }
}
```

Python

- Another scripting language developed by Guido van Rossum (first release 1990).
- "There should be one - and preferably only one - obvious way to do it."
- Notorious for whitespace requirements ("Readability counts.").
- Has overtaken Perl in popularity, even bioinformatics.

A conversion of the Illumina Perl code to Python

```
import sys

# read filename from command line
infile = sys.argv[1]

handle = open(infile, 'rU')
_ = handle.readline()  # discard header line

Ns = []  # span multiple lines
for line in handle:
    # we assume that the line consists entirely of nucleotide sequence
    seq = line.strip('\n')
    for base in seq:
        Ns.append(1 if base == 'N' else 0)
```

Suggested readings

- How Perl Saved the Human Genome Project