Bioinformatics

Contents

FASTA Format

Biological sequences are passed to software in a standardized format referred to as **FASTA**. FASTA is a plain text format that can be read in any text editor (TextEdit, Notepad, VIM, TextWrangler, etc.). Nucleic acids (DNA and RNA) and Proteins are represented by single letter nucleotides (A,T,C,G) or single letter amino acid (20 amino acids). FASTA sequences begin with a **>** character in the first line followed by some descriptive information about the sequence, like a sequence name. The next line consists of the sequence information. A FASTA file can contain multiple sequence entries all demarcated by a new line and a title line beginning with **>** .

EXAMPLE FASTA FILE

> Made-up nucleic acid sequence
ATATAGGGATTAGGATTAGAGGATAGAGGGGATTGCGCCG
> Another nucleic acid sequence in same file
GGGTCGGGCTAGCGGAATCGGATTCGGCATTCGGATATTCGGATTCGGAT

FASTA files are plain text but usually have an extension indicating it as a sequence file: .fasta, .fa, .fna or even .txt

A list of single-letter codes for nucleic acids follows below:

Nucleic Acid Code	Meaning	Mnemonic
A	A	**A**denine
C	C	**C**ytosine
G	G	**G**uanine
T	T	**T**hymine
U	U	**U**racil
R	A or G	pu**R**ine
Y	C, T or U	p**Y**rimidines
K	G, T or U	bases which are **K**etones

Nucleic Acid Code	Meaning	Mnemonic
M	A or C	bases with a**M**ino groups
S	C or G	**S**trong interaction
W	A, T or U	**W**eak interaction
B	not A (i.e. C, G, T or U)	**B** comes after A
D	not C (i.e. A, G, T or U)	**D** comes after C
H	not G (i.e., A, C, T or U)	**H** comes after G
V	neither T nor U (i.e. A, C or G)	**V** comes after U
N	A C G T U	**N**ucleotide
X	masked	
–	gap of indeterminate length	

Graphical Sequence Manipulation

The exercises described here regarding bioinformatics will utilize a free and open source software called Unipro UGENE.

- Okonechnikov K, Golosova O, Fursov M, the UGENE team. Unipro UGENE: a unified bioinformatics toolkit.Bioinformatics 2012 28: 1166-1167. doi:10.1093/bioinformatics/bts09
- Golosova O, Henderson R, Vaskin Y, Gabrielian A, Grekhov G, Nagarajan V, Oler AJ, Quiñones M, Hurt D, Fursov M, Huyen Y.**Unipro UGENE NGS pipelines and components for variant calling, RNA-seq and ChIP-seq data analyses**. PeerJ *2014* 2:e644. doi:10.7717/peerj.644