FASTA format

In bioinformatics, **FASTA format** is a text-based format for representing either nucleotide sequences or peptide sequences, in which base pairs or amino acids are represented using single-letter codes. The format also allows for sequence names and comments to precede the sequences. The format originates from the FASTA software package, but has now become a standard in the field of bioinformatics.

The simplicity of FASTA format makes it easy to manipulate and parse sequences using text-processing tools and scripting languages like Python, Ruby, and Perl.

Format

A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line is distinguished from the sequence data by a greater-than (">") symbol in the first column. The word following the ">" symbol is the identifier of the sequence, and the rest of the line is the description (both are optional). There should be no space between the ">" and the first letter of the identifier. It is recommended that all lines of text be shorter than 80 characters. The sequence ends if another line starting with a ">" appears; this indicates the start of another sequence. A simple example of one sequence in FASTA format:

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

History

The original FASTA/Pearson format is described in the documentation for the FASTA suite of programs. It can be downloaded with any free distribution of FASTA (see fasta20.doc, fastaVN.doc or fastaVN.me --where VN is the Version Number).

A sequence in FASTA format is represented as a series of lines, which should be no longer than 120 characters and usually do not exceed 80 characters. This probably was because to allow for preallocation of fixed line sizes in software: at the time most users relied on DEC VT (or compatible) terminals which could display 80 or 132 characters per line. Most people preferred the bigger font in 80-character modes and so it became the recommended fashion to use 80 characters or less (often 70) in FASTA lines.

The first line in a FASTA file starts either with a ">" (greater-than) symbol or a ";" (semicolon) and was taken as a comment. Subsequent lines starting with a semicolon would be ignored by software. Since the only comment used was the first, it quickly became used to hold a summary description of the sequence, often starting with a unique library accession number, and with time it has become commonplace use to always use ">" for the first line and to not use ";" comments (which would otherwise be ignored).

Following the initial line (used for a unique description of the sequence) is the actual sequence itself in standard one-letter code. Anything other than a valid code would be ignored (including spaces, tabulators, asterisks, etc...). Originally it was also common to end the sequence with an "*" (asterisk) character (in analogy with use in PIR formatted sequences) and, for the same reason, to leave a blank line between the description and the sequence.

A few sample sequences:

```
;LCBO - Prolactin precursor - Bovine
; a sample sequence in FASTA format
```

```
MDSKGSSQKGSRLLLLLVVSNLLLCQGVVSTPVCPNGPGNCQVSLRDLFDRAVMVSHYIHDLSS
EMFNEFDKRYAQGKGFITMALNSCHTSSLPTPEDKEQAQQTHHEVLMSLILGLLRSWNDPLYHL
VTEVRGMKGAPDAILSRAIEIEEENKRLLEGMEMIFGQVIPGAKETEPYPVWSGLPSLQTKDED
ARYSAFYNLLHCLRRDSSKIDTYLKLLNCRIIYNNNC*
```

```
>MCHU - Calmodulin - Human, rabbit, bovine, rat, and chicken
ADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTID
FPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREA
DIDGDGQVNYEEFVQMMTAK*
```

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

A multiple sequence FASTA format would be obtained by concatenating several single sequence FASTA files. This does not imply a contradiction with the format as only the first line in a FASTA file may start with a ";" or ">", hence forcing all subsequent sequences to start with a ">" in order to be taken as different ones (and further forcing the exclusive reservation of ">" for the sequence definition line). Thus, the examples above may as well be taken as a multisequence file if taken together.

Format converters

FASTA files can be batch converted to or from MultiFASTA format using tools, some of which are available as freeware. Tools are also available for batch conversion from [chromatogram] formats (ABI/SCF) to FASTA.

Header line

The header line, which begins with '>', gives a name and/or a unique identifier for the sequence, and often lots of other information too. Many different sequence databases use standardized headers, which helps when automatically extracting information from the header. The header line may contain more than one header, separated by a ^A (Control-A) character.

In the original Pearson FASTA format, one or more comments, distinguished by a semi-colon at the beginning of the line, may occur after the header. Most databases and bioinformatics applications do not recognize these comments and follow the NCBI FASTA specification [1]. An example of a multiple sequence FASTA file follows:

```
>SEQUENCE_1
MTEITAAMVKELRESTGAGMMDCKNALSETNGDFDKAVQLLREKGLGKAAKKADRLAAEG
LVSVKVSDDFTIAAMRPSYLSYEDLDMTFVENEYKALVAELEKENEERRRLKDPNKPEHK
IPQFASRKQLSDAILKEAEEKIKEELKAQGKPEKIWDNIIPGKMNSFIADNSQLDSKLTL
MGQFYVMDDKKTVEQVIAEKEKEFGGKIKIVEFICFEVGEGLEKKTEDFAAEVAAQL
>SEQUENCE_2
SATVSEINSETDFVAKNDQFIALTKDTTAHIQSNSLQSVEELHSSTINGVKFEEYLKSQI
ATIGENLVVRRFATLKAGANGVVNGYIHTNGRVGVVIAAACDSAEVASKSRDLLRQICMH
```

Sequence representation

After the header line and comments, one or more lines may follow describing the sequence: each line of a sequence should have fewer than 80 characters. Sequences may be protein sequences or nucleic acid sequences, and they can contain gaps or alignment characters (see sequence alignment). Sequences are expected to be represented in the standard IUB/IUPAC amino acid and nucleic acid codes, with these exceptions: lower-case letters are accepted and are mapped into upper-case; a single hyphen or dash can be used to represent a gap character; and in amino acid sequences, U and * are acceptable letters (see below). Numerical digits are not allowed but are used in some databases to indicate the position in the sequence.

The nucleic acid codes supported are[2] :

Nucleic Acid Code	Meaning
A	**A**denosine
C	**C**ytosine
G	**G**uanine
T	**T**hymidine
U	**U**racil
R	G A (pu**R**ine)
Y	T U C (p**Y**rimidine)
K	G T U (**K**etone)
M	A C (a**M**ino group)
S	G C (**S**trong interaction)
W	A T U (**W**eak interaction)
B	G T U C (not A) (**B** comes after A)
D	G A T U (not C) (**D** comes after C)
H	A C T U (not G) (**H** comes after G)
V	G C A (not T, not U) (**V** comes after U)
N	A G C T U (a**N**y)
X	masked
-	gap of indeterminate length

The codes supported (24 amino acids and 3 special codes) are:

Amino Acid Code	Meaning
A	Alanine
B	Aspartic acid or Asparagine
C	Cysteine
D	Aspartic acid
E	Glutamic acid
F	Phenylalanine
G	Glycine
H	Histidine
I	Isoleucine

K	Lysine
L	Leucine
M	Methionine
N	Asparagine
O	Pyrrolysine
P	Proline
Q	Glutamine
R	Arginine
S	Serine
T	Threonine
U	Selenocysteine
V	Valine
W	Tryptophan
Y	Tyrosine
Z	Glutamic acid or Glutamine
X	any
*	translation stop
-	gap of indeterminate length

Sequence identifiers

The NCBI defined a standard for the unique identifier used for the sequence (SeqID) in the header line. The formatdb man page has this to say on the subject: "formatdb will automatically parse the SeqID and create indexes, but the database identifiers in the FASTA definition line must follow the conventions of the FASTA Defline Format."

However they do not give a definitive description of the FASTA defline format. An attempt to create such a format is given below (see also "The NCBI Handbook", Chapter 16, The BLAST Sequence Analysis Tool [3].).

```
GenBank                           gi|gi-number|gb|accession|locus
EMBL Data Library                 gi|gi-number|emb|accession|locus
DDBJ, DNA Database of Japan       gi|gi-number|dbj|accession|locus
NBRF PIR                          pir||entry
Protein Research Foundation       prf||name
SWISS-PROT                        sp|accession|name
Brookhaven Protein Data Bank (1)  pdb|entry|chain
Brookhaven Protein Data Bank (2)  entry:chain|PDBID|CHAIN|SEQUENCE
Patents                           pat|country|number
GenInfo Backbone Id               bbs|number
General database identifier       gnl|database|identifier
NCBI Reference Sequence           ref|accession|locus
Local Sequence identifier         lcl|identifier
```

The vertical bars in the above list are not separators in the sense of the Backus-Naur form, but are part of the format.

File extension

There is no standard file extension for a text file containing FASTA formatted sequences. The table below shows each extension and its respective meaning.

Extension	Meaning	Notes
fasta	generic fasta	Any generic fasta file. Other extensions can be fa, seq, fsa
fna	fasta nucleic acid	For coding regions of a specific genome, use ffn, but otherwise fna is useful for generically specifying nucleic acids.
ffn	FASTA nucleotide coding regions	Contains coding regions for a genome.
faa	fasta amino acid	Contains amino acids. A multiple protein fasta file can have the more specific extension mpfa.
frn	FASTA non-coding RNA	Contains non-coding RNA regions for a genome, in DNA alphabet e.g. tRNA, rRNA

References

[1] http://www.ncbi.nlm.nih.gov/blast/fasta.shtml
[2] "IUPAC code table" (http://www.dna.affrc.go.jp/misc/MPsrch/InfoIUPAC.html). NIAS DNA Bank. .
[3] http://www.ncbi.nlm.nih.gov/books/NBK21097/

External links

- What is FASTA Format? (http://zhanglab.ccmb.med.umich.edu/FASTA/) Explain the FASTA format.
- HUPO-PSI Standard FASTA Format (http://www.proteomecommons.org/data/fasta/hupo_standard.jsp) was describing another FASTA format as put forward by the Human Proteome Organisation's Proteomics Standards Initiative.
- Sequence ID (seqID) Fields in the FASTA Deflines of Sequences from NCBI (http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/formatdb_fastacmd.html#t1.1) describes the format of FASTA Deflines.
- FASTA File-Format Converter (http://www.bioinformaticsbox.com/tools/sequence_format_converter.php)

Article Sources and Contributors

FASTA format *Source*: http://en.wikipedia.org/w/index.php?oldid=435601724 *Contributors*: A. B., Akriasas, Alchemistmatt, Andrewrp, Applyalert1, Ascha, Augman85, BaChev, Beetstra, Blastwizard, Capricorn42, Cyrius, Dmb000006, Dongilbert, Ehamberg, Fedra, Grub, Gu margaret, Hydkat, J. Finkelstein, Jaredme, Lskatz, Lukaskoz, MHuyck, Maasha, Mandarax, Menat22, Mfursov, Miguel Andrade, Mikhail Dvorkin, Mirc007, Nihiltres, Nowak2000, Paulmkgordon, Ph.eyes, Ppgardne, Quuxplusone, Qwerty0, Reinyday, Rich Farmbrough, Rjwilmsi, SiobhanHansa, Spellsinger180, SteveChervitzTrutane, The Anome, TheObtuseAngleOfDoom, TheTweaker, ToddDeLuca, Torst, Versageek, Wzhao553, 90 anonymous edits

License