

FASTA format

In bioinformatics, **FASTA format** is a text-based format for representing either nucleotide sequences or peptide sequences, in which nucleotides 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.^[4]

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

FASTA format

Developed by	David J. Lipman William R. Pearson ^{[1][2]}
Initial release	1985
Type of format	Bioinformatics
Extended from	ASCII for FASTA
Extended to	FASTQ format ^[3]
Website	<div>www.ncbi.nlm.nih.gov/BLAST/fasta.shtml (https://www.ncbi.nlm.nih.gov/BLAST/fasta.shtml)</div>

Contents

Original format & overview

Description line

NCBI identifiers

Sequence representation

FASTA file

Filename extension

Compression

Encryption

Extended Format

Working with FASTA files

References

External links

Original format & overview

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).

In the original format, a sequence was represented as a series of lines, each of which was no longer than 120 characters and usually did not exceed 80 characters. This probably was to allow for preallocation of fixed line sizes in software: at the time most users relied on Digital Equipment Corporation (DEC) VT220 (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. Also, the width of a standard printed page is 70 to 80 characters (depending on the font).

The first line in a FASTA file started either with a ">" (greater-than) symbol or, less frequently, a ";" (semicolon) 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 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 character string. Anything other than a valid character 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
MDSKGSSQKGSRLLLLVSNLLLCQGVVSTPVCNPGNCGVSLRDLFDRAVMVSHYIHDLS
EMFNEFDKRYAQGGFITMALNSCHTSSLPTPEDKEQAQQTTHHEVLSLILGLLRSWNDPLYHL
VTEVRGMKGAPDAILSRAIEIEEENKRLLEGMEMIFGQVIPGAKETEPYPVWSGLPSLQTKDED
ARYSAFYNNLLHCLRRDSSKIDTYLKLNCRIIYNNNC*

>MCHU - Calmodulin - Human, rabbit, bovine, rat, and chicken
ADQLTEEQIAEFKEAFSLFDKDGDTITTKELGTVMRSLGQNPTAEALQDMINEVDADGNGTID
FPEFLTMMARKMKDSTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDDEMIREA
DIDGDGQVNYEEFVQMMTAK*

>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLITMATAFMGYVLPWQMSFWGATVITNLFSAIPYIGTNLV
EWIWWGGSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSKIPFHPYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPHLIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMLRPLSQLFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFPLIAGX
IENY

```

A multiple sequence FASTA format would be obtained by concatenating several single sequence FASTA files in a common file (also known as multi-FASTA format). 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 (i.e multi-FASTA) file if taken together.

Nowadays, modern bioinformatic programs that rely on the FASTA format expect the sequence headers to be preceded by ">", and the actual sequence, while generally represented as "interleaved", i.e. on multiple lines as in the above example, may also be "sequential" when the full stretch is found on a single line. Users may often need to perform conversion between "Sequential" and "Interleaved" FASTA format to run different bioinformatic programs.

Description line

The description line (define) or header/identifier line, which begins with '>', gives a name and/or a unique identifier for the sequence, and may also contain additional information. In a deprecated practice, the header line sometimes contained 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. Some databases and bioinformatics applications do not recognize these comments and follow the NCBI FASTA specification (<https://www.ncbi.nlm.nih.gov/blast/fasta.shtml>). An example of a multiple sequence FASTA file follows:

```

>SEQUENCE_1
MTEITAAMVKELRESTGAGMMDCKNALSETNGDFDKAVQLLREKGLGKAACKADRLAAEG
LVSVKVSDDFTIAAMRPSYLSYEDLDMTFVENEYKALVALEKEENEERRRLKDPNKPEHK
IPQFASRKQLSDAILKEAEKIKEELKAQKGPEKIWDNIIPGKMNSFIADNSQLDSKLT

```

```

MGQFYVMDKKTVQVIAEKEKEFGGKIKIVEFICFEVGEGLKKTEDFAAEVAAQL
>SEQUENCE_2
SATVSEINSETDFVAKNDQFIALTKDTTAHIQSNLSQSV EELHSSTINGVKFEEYLSQI
ATIGENLVRRFATLKAGANGVNGYIHTNGRVGVVIAAACDSA EVASKSRDLLRQICMH

```

NCBI identifiers

The NCBI defined a standard for the unique identifier used for the sequence (SeqID) in the header line. This allows a sequence that was obtained from a database to be labelled with a reference to its database record. The database identifier format is understood by the NCBI tools like `makeblastdb` and `table2asn`. The following list describes the NCBI FASTA defined format for sequence identifiers.^[5]

Type	Format(s)	Example(s)
local (i.e. no database reference)	<code>lc1 integer</code> <code>lc1 string</code>	<code>lc1 123</code> <code>lc1 hmm271</code>
GenInfo backbone seqid	<code>bbs integer</code>	<code>bbs 123</code>
GenInfo backbone moltype	<code>bbm integer</code>	<code>bbm 123</code>
GenInfo import ID	<code>gim integer</code>	<code>gim 123</code>
GenBank (https://www.ncbi.nlm.nih.gov/Genbank/index.html)	<code>gb accession locus</code>	<code>gb M73307 AGMA13GT</code>
EMBL (http://www.embl-heidelberg.de)	<code>emb accession locus</code>	<code>emb CAM43271.1 </code>
PIR (http://pir.georgetown.edu)	<code>pir accession name</code>	<code>pir G36364</code>
SWISS-PROT (http://www.ebi.ac.uk/swissprot)	<code>sp accession name</code>	<code>sp P01013 OVAX_CHICK</code>
patent	<code>pat country patent sequence-number</code>	<code>pat US RE33188 1</code>
pre-grant patent	<code>pgp country application-number sequence-number</code>	<code>pgp EP 0238993 7</code>
RefSeq (https://www.ncbi.nlm.nih.gov/projects/RefSeq)	<code>ref accession name</code>	<code>ref NM_010450.1 </code>
general database reference (a reference to a database that's not in this list)	<code>gn1 database integer</code> <code>gn1 database string</code>	<code>gn1 taxon 9606</code> <code>gn1 PID e1632</code>
GenInfo integrated database	<code>gi integer</code>	<code>gi 21434723</code>
DDBJ (http://www.ddbj.nig.ac.jp)	<code>dbj accession locus</code>	<code>dbj BAC85684.1 </code>
PRF (http://www.prf.or.jp)	<code>prf accession name</code>	<code>prf 0806162C</code>
PDB (http://www.rcsb.org/pdb)	<code>pdb entry chain</code>	<code>pdb 1I4L D</code>
third-party GenBank (https://www.ncbi.nlm.nih.gov/Genbank/index.html)	<code>tpg accession name</code>	<code>tpg BK003456 </code>
third-party EMBL (http://www.embl-heidelberg.de)	<code>tpe accession name</code>	<code>tpe BN000123 </code>
third-party DDBJ (http://www.ddbj.nig.ac.jp)	<code>tpd accession name</code>	<code>tpd FAA00017 </code>
TrEMBL	<code>tr accession name</code>	<code>tr Q90RT2 Q90RT2_9HIV1</code>

The vertical bars ("|") in the above list are not separators in the sense of the Backus–Naur form, but are part of the format. Multiple identifiers can be concatenated, also separated by vertical bars.

Sequence representation

Following the header line, the actual sequence is represented. 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:^{[6][7]}

Nucleic Acid Code	Meaning	Mnemonic
A	A	<u>Adenine</u>
C	C	<u>Cytosine</u>
G	G	<u>Guanine</u>
T	T	<u>Thymine</u>
U	U	<u>Uracil</u>
R	A or G	<u>puRine</u>
Y	C, T or U	<u>pYrimidines</u>
K	G, T or U	bases which are K etones
M	A or C	bases with a mino 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 ucleic acid
-	gap of indeterminate length	

The amino acid codes supported (22 amino acids and 3 special codes) are:

Amino Acid Code	Meaning
A	<u>Alanine</u>
B	<u>Aspartic acid</u> (D) or <u>Asparagine</u> (N)
C	<u>Cysteine</u>
D	<u>Aspartic acid</u>
E	<u>Glutamic acid</u>
F	<u>Phenylalanine</u>
G	<u>Glycine</u>
H	<u>Histidine</u>
I	<u>Isoleucine</u>
J	<u>Leucine</u> (L) or <u>Isoleucine</u> (I)
K	<u>Lysine</u>
L	<u>Leucine</u>
M	<u>Methionine/Start codon</u>
N	<u>Asparagine</u>
O	<u>Pyrrolysine</u>
P	<u>Proline</u>
Q	<u>Glutamine</u>
R	<u>Arginine</u>
S	<u>Serine</u>
T	<u>Threonine</u>
U	<u>Selenocysteine</u>
V	<u>Valine</u>
W	<u>Tryptophan</u>
Y	<u>Tyrosine</u>
Z	<u>Glutamic acid</u> (E) or <u>Glutamine</u> (Q)
X	any
*	translation stop
-	gap of indeterminate length

FASTA file

Filename extension

There is no standard filename 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 fas, fa, seq, fsa
fna	fasta nucleic acid	Used generically to specify nucleic acids.
ffn	FASTA nucleotide of gene 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

Compression

The compression of FASTA files requires a specific compressor to handle both channels of information: identifiers and sequence. For improved compression results, these are mainly divided in two streams where the compression is made assuming independence. For example, the algorithm MFCompress [8] performs lossless compression of these files using context modelling and arithmetic encoding. For a benchmark on FASTA files compression algorithms, see [9].

Encryption

The encryption of FASTA files has been mostly addressed with a specific encryption tool: Cryfa.[10] Cryfa uses AES encryption and enables to compact data besides encryption. It can also address FASTQ files.

Extended Format


FASTA format was extended by FASTQ format from the Sanger Centre in Cambridge.[3]

Working with FASTA files

A plethora of user-friendly scripts are available from the community to perform FASTA file manipulations. Online toolbox are also available such as FaBox[11] or the FASTX-Toolkit within Galaxy servers.[12] For instance, these can be used to segregate sequence headers/identifiers, rename them, shorten them, or extract sequences of interest from large FASTA files based on a list of wanted identifiers (among other available functions). A tree-based approach to sorting multi-FASTA files (TREE2FASTA[13]) also exists based on the coloring and/or annotation of sequence of interest in the FigTree viewer. Additionally, Bioconductor.org's Biostrings package can be used to read and manipulate FASTA files in R.[14]

Several online format converters exist to rapidly reformat multi-FASTA files to different formats (e.g. NEXUS, PHYLIP) for their use with different phylogenetic programs (e.g. such as the converter available on phylogeny.fr.[15]

References

1. Lipman DJ, Pearson WR (March 1985). "Rapid and sensitive protein similarity searches". *Science*. **227** (4693): 1435–41. doi:10.1126/science.2983426 (https://doi.org/10.1126%2Fscience.2983426). PMID 2983426 (https://www.ncbi.nlm.nih.gov/pubmed/2983426). 

2. Pearson WR, Lipman DJ (April 1988). "Improved tools for biological sequence comparison" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC280013). *Proceedings of the National Academy of Sciences of the United States of America*. **85** (8): 2444–8. doi:10.1073/pnas.85.8.2444 (https://doi.org/10.1073%2Fpnas.85.8.2444). PMC 280013 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC280013). PMID 3162770 (https://www.ncbi.nlm.nih.gov/pubmed/3162770).

3. Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM (April 2010). "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants" (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2847217>). *Nucleic Acids Research*. **38** (6): 1767–71. doi:10.1093/nar/gkp1137 (<https://doi.org/10.1093%2Fnar%2Fgkp1137>). PMC 2847217 (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2847217>). PMID 20015970 (<https://www.ncbi.nlm.nih.gov/pubmed/20015970>).
4. "What is FASTA Format?" (<http://zhanglab.ccmb.med.umich.edu/FASTA/>). *zhanglab.ccmb.med.umich.edu*. explains the FASTA format
5. *NCBI C++ Toolkit Book* (https://ncbi.github.io/cxx-toolkit/pages/ch_demo#ch_demo.id1_fetch.html_ref_fasta). National Center for Biotechnology Information. Retrieved 2018-12-19.
6. Tao Tao (2011-08-24). "Single Letter Codes for Nucleotides" (https://www.ncbi.nlm.nih.gov/staff/tao/tools/tool_lettercode.html). *[NCBI Learning Center]*. National Center for Biotechnology Information. Retrieved 2012-03-15.
7. "IUPAC code table" (<https://web.archive.org/web/20110811073845/http://www.dna.affrc.go.jp/misc/MPsrch/InfolUPAC.html>). NIAS DNA Bank. Archived from the original (<http://www.dna.affrc.go.jp/misc/MPsrch/InfolUPAC.html>) on 2011-08-11.
8. Pinho AJ, Pratas D (January 2014). "MFCompress: a compression tool for FASTA and multi-FASTA data" (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3866555>). *Bioinformatics*. **30** (1): 117–8. doi:10.1093/bioinformatics/btt594 (<https://doi.org/10.1093%2Fbioinformatics%2Fbtt594>). PMC 3866555 (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3866555>). PMID 24132931 (<https://www.ncbi.nlm.nih.gov/pubmed/24132931>).
9. M. Hosseini, D. Pratas, and A. Pinho. 2016. A survey on data compression methods for biological sequences. *Information* **7**(4):(2016): 56
10. Pratas D, Hosseini M, Pinho A (2017). *Cryfa: a tool to compact and encrypt FASTA files. 11'th International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB), Springer. Advances in Intelligent Systems and Computing*. **616**. pp. 305–312. doi:10.1007/978-3-319-60816-7_37 (https://doi.org/10.1007%2F978-3-319-60816-7_37). ISBN 978-3-319-60815-0.
11. Villesen P (April 2007). "FaBox: an online toolbox for fasta sequences". *Molecular Ecology Resources*. **7** (6): 965–968. doi:10.1111/j.1471-8286.2007.01821.x (<https://doi.org/10.1111%2Fj.1471-8286.2007.01821.x>).
12. Blankenberg D, Von Kuster G, Bouvier E, Baker D, Afgan E, Stoler N, Galaxy Team, Taylor J, Nekrutenko A (2014). "Dissemination of scientific software with Galaxy ToolShed" (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4038738>). *Genome Biology*. **15** (2): 403. doi:10.1186/gb4161 (<https://doi.org/10.1186%2Fgb4161>). PMC 4038738 (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4038738>). PMID 25001293 (<https://www.ncbi.nlm.nih.gov/pubmed/25001293>).
13. Sauvage T, Plouviez S, Schmidt WE, Fredericq S (March 2018). "TREE2FASTA: a flexible Perl script for batch extraction of FASTA sequences from exploratory phylogenetic trees" (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5838971>). *BMC Research Notes*. **11** (1): 403. doi:10.1186/s13104-018-3268-y (<https://doi.org/10.1186%2Fs13104-018-3268-y>). PMC 5838971 (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5838971>). PMID 29506565 (<https://www.ncbi.nlm.nih.gov/pubmed/29506565>).
14. Pagès, H; Aboyoun, P; Gentleman, R; DebRoy, S (2018). "*Biostrings: Efficient manipulation of biological strings*" (<https://bioconductor.org/packages/release/bioc/html/Biostrings.html>). *Bioconductor.org*. R package version 2.48.0.
15. Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard JF, Guindon S, Lefort V, Lescot M, Claverie JM, Gascuel O (July 2008). "Phylogeny.fr: robust phylogenetic analysis for the non-specialist" (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2447785>). *Nucleic Acids Research*. **36** (Web Server issue): W465–9. doi:10.1093/nar/gkn180 (<https://doi.org/10.1093%2Fnar%2Fgkn180>). PMC 2447785 (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2447785>). PMID 18424797 (<https://www.ncbi.nlm.nih.gov/pubmed/18424797>).

External links

- Bioconductor (<http://www.bioconductor.org>)
- FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/)
- FigTree viewer (<http://tree.bio.ed.ac.uk/software/figtree/>)
- Phylogeny.fr (http://phylogeny.lirmm.fr/phylo.cgi/data_converter.cgi)

Retrieved from "https://en.wikipedia.org/w/index.php?title=FASTA_format&oldid=875159165"

This page was last edited on 24 December 2018, at 05:51 (UTC).

Text is available under the [Creative Commons Attribution-ShareAlike License](#); additional terms may apply. By using this site, you agree to the [Terms of Use](#) and [Privacy Policy](#). Wikipedia® is a registered trademark of the [Wikimedia Foundation, Inc.](#), a non-profit organization.