FASTA format

From Wikipedia, the free encyclopedia

In bioinformatics and biochemistry, the **FASTA format** is a text-based format for representing either nucleotide sequences or amino acid (protein) 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 near universal 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, Haskell, and Perl.

FASTA format

Filename extensions	.fasta, .fna, .ffn, .faa, .frn, .fa
Internet media type	text/x-fasta
Uniform Type Identifier (UTI)	no
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	www.ncbi.nlm.nih.gov/BLAST/fasta.shtml

Contents [hide]

Original format & overview [edit]

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.[citation needed] 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). Hence, 80 characters became the norm.[citation needed]

The first line in a FASTA file started either with a ">" (greater-than) symbol or, less frequently, a ";"[citation needed] (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) was the actual sequence itself in standard one-letter character string. Anything other than a valid character would be ignored (including spaces, tabulators, asterisks, etc...). 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. Below are 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
MADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTID
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 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 [edit]

The description line (defline) 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. An example of a multiple sequence FASTA file follows:

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

NCBI identifiers [edit]

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)	lcl|*integer* lcl|*string*	lcl|123 lcl|hmm271
GenInfo backbone seqid	bbs|*integer*	bbs|123
GenInfo backbone moltype	bbm|*integer*	bbm|123
GenInfo import ID	gim|*integer*	gim|123
GenBank	gb|*accession*|*locus*	gb|M73307|AGMA13GT
EMBL	emb|*accession*|*locus*	emb|CAM43271.1|
PIR	pir|*accession*|*name*	pir||G36364
SWISS-PROT	sp|*accession*|*name*	sp|P01013|OVAX_CHICK
patent	pat|*country*|*patent*|*sequence-number*	pat|US|RE33188|1
pre-grant patent	pgp|*country*|*application-number*|*sequence-number*	pgp|EP|0238993|7
RefSeq	ref|*accession*|*name*	ref|NM_010450.1|
general database reference (a reference to a database that's not in this list)	gnl|*database*|*integer* gnl|*database*|*string*	gnl|taxon|9606 gnl|PID|e1632
GenInfo integrated database	gi|*integer*	gi|21434723
DDBJ	dbj|*accession*|*locus*	dbj|BAC85684.1|
PRF	prf|*accession*|*name*	prf||0806162C
PDB	pdb|*entry*|*chain*	pdb|1I4L|D
third-party GenBank	tpg|*accession*|*name*	tpg|BK003456|
third-party EMBL	tpe|*accession*|*name*	tpe|BN000123|
third-party DDBJ	tpd|*accession*|*name*	tpd|FAA00017|
TrEMBL	tr|*accession*|*name*	tr|Q90RT2|Q90RT2_9HIV1

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 [edit]

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][8]

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
(i)	i	**i**nosine (non-standard)
R	A or G (I)	pu**R**ine
Y	C, T or U	p**Y**rimidines
K	G, T or U	bases which are **K**etones
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**ucleic acid
-	gap of indeterminate length	

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

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

FASTA file [edit]

Filename extension [edit]

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, fa[9]	generic FASTA	Any generic fasta file. See below for other common FASTA file extensions
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 acid sequences. 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 [edit]

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[10] performs lossless compression of these files using context modelling and arithmetic encoding. For benchmarks of FASTA files compression algorithms, see Hosseini et al., 2016,[11] and Kryukov et al., 2020.[12]

Encryption [edit]

The encryption of FASTA files has been mostly addressed with a specific encryption tool: Cryfa.[13][14] Cryfa uses AES encryption and enables to compact data besides encryption. It can also address FASTQ files.

Extensions [edit]

FASTQ format is a form of FASTA format extended to indicate information related to sequencing. It is created by the Sanger Centre in Cambridge.[3]

A2M/A3M are a family of FASTA-derived formats used for sequence alignments. In A2M/A3M sequences, lowercase characters are taken to mean insertions, which are then indicated in the other sequences as the dot (".") character. The dots can be discarded for compactness without loss of information. As with typical FASTA used in alignments, the gap ("-") is taken to mean exactly one position.[15] A3M is similar to A2M, with the added rule that gaps aligned to insertions can too be discarded.[16]

Working with FASTA files [edit]

A plethora of user-friendly scripts are available from the community to perform FASTA file manipulations. Online toolboxes are also available such as FaBox[17] or the FASTX-Toolkit within Galaxy servers.[18] 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[19]) 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.[20]

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.[21]

See also [edit]

- The FASTQ format, used to represent DNA sequencer reads along with quality scores.
- The SAM format, used to represent genome sequencer reads, generally but not necessarily after they have been aligned to genome sequences.[22]
- The GVF format (Genome Variation Format), an extension based on the GFF3 format.

References [edit]

1. ^ Lipman DJ, Pearson WR (March 1985). "Rapid and sensitive protein similarity searches". *Science*. **227** (4693): 1435–41. Bibcode:1985Sci...227.1435L. doi:10.1126/science.2983426. PMID 2983426.
2. ^ Pearson WR, Lipman DJ (April 1988). "Improved tools for biological sequence comparison". *Proceedings of the National Academy of Sciences of the United States of America*. **85** (8): 2444–8. Bibcode:1988PNAS...85.2444P. doi:10.1073/pnas.85.8.2444. PMC 280013. PMID 3162770.
3. ^ a b 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". *Nucleic Acids Research*. **38** (6): 1767–71. doi:10.1093/nar/gkp1137. PMC 2847217. PMID 20015970.
4. ^ "What is FASTA Format?". *zhanglab.ccmb.med.umich.edu*. explains the FASTA format
5. ^ *NCBI C++ Toolkit Book*. National Center for Biotechnology Information. Retrieved 2018-12-19.
6. ^ Tao Tao (2011-08-24). "Single Letter Codes for Nucleotides". *[NCBI Learning Center]*. National Center for Biotechnology Information. Retrieved 2012-03-15.
7. ^ "IUPAC code table". NIAS DNA Bank. Archived from the original on 2011-08-11.
8. ^ "anysymbol". *MAFFT - a multiple sequence alignment program*.
9. ^ "Alignment Fileformats". 22 May 2019. Retrieved 22 May 2019.
10. ^ Pinho AJ, Pratas D (January 2014). "MFCompress: a compression tool for FASTA and multi-FASTA data". *Bioinformatics*. **30** (1): 117–8. doi:10.1093/bioinformatics/btt594. PMC 3866555. PMID 24132931.
11. ^ M. Hosseini, D. Pratas, and A. Pinho. 2016. A survey on data compression methods for biological sequences. *Information* **7**(4):(2016): 56
12. ^ Kryukov K, Ueda MT, Nakagawa S, Imanishi T (July 2020). "Sequence Compression Benchmark (SCB) database—A comprehensive evaluation of reference-free compressors for FASTA-formatted sequences". *GigaScience*. **9** (7): giaa072. doi:10.1093/gigascience/giaa072. PMC 7336184. PMID 32627830.
13. ^ Pratas D, Hosseini M, Pinho A (2017). "Cryfa: a tool to compact and encrypt FASTA files". *11th International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB)*. Advances in Intelligent Systems and Computing. Vol. 616. Springer. pp. 305–312. doi:10.1007/978-3-319-60816-7_37. ISBN 978-3-319-60815-0.
14. ^ Hosseini M, Pratas D, Pinho A (2018). *Cryfa: a secure encryption tool for genomic data. Bioinformatics*. Vol. 35. pp. 146–148. doi:10.1093/bioinformatics/bty645. PMC 6298042. PMID 30020420.
15. ^ "Description of A2M alignment format". *SAMtools*.
16. ^ "soedinglab/hh-suite: reformat.pl". *GitHub*. 20 November 2022.
17. ^ 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.
18. ^ 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". *Genome Biology*. **15** (2): 403. doi:10.1186/gb4161. PMC 4038738. PMID 25001293.
19. ^ 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". *BMC Research Notes*. **11** (1): 403. doi:10.1186/s13104-018-3268-y. PMC 5838971. PMID 29506565.
20. ^ Pagès, H; Aboyoun, P; Gentleman, R; DebRoy, S (2018). "*Biostrings: Efficient manipulation of biological strings*". *Bioconductor.org*. R package version 2.48.0. doi:10.18129/B9.bioc.Biostrings.
21. ^ 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". *Nucleic Acids Research*. **36** (Web Server issue): W465–9. doi:10.1093/nar/gkn180. PMC 2447785. PMID 18424797.
22. ^ https://samtools.github.io/hts-specs/SAMv1.pdf[bare URL PDF]

External links [edit]

- Bioconductor
- FASTX-Toolkit
- FigTree viewer
- Phylogeny.fr
- GTO

V · T · E **Bioinformatics** [hide]

Databases	• Sequence databases: GenBank, European Nucleotide Archive and DNA Data Bank of Japan • Secondary databases: UniProt, database of protein sequences grouping together Swiss-Prot, TrEMBL and Protein Information Resource • Other databases: Protein Data Bank, Ensembl and InterPro • Specialised genomic databases: BOLD, Saccharomyces Genome Database, FlyBase, VectorBase, WormBase, Rat Genome Database, PHI-base, Arabidopsis Information Resource and Zebrafish Information Network
Software	BLAST · Bowtie · Clustal · EMBOSS · HMMER · MUSCLE · SAMtools · SOAP suite · TopHat
Other	Server: ExPASy · Ontology: Gene Ontology · Rosalind (education platform)
Institutions	Broad Institute · China National GeneBank (CNGB) · Computational Biology Department (CBD) · Microsoft Research - University of Trento Centre for Computational and Systems Biology (COSBI) · Database Center for Life Science (DBCLS) · DNA Data Bank of Japan (DDBJ) · European Bioinformatics Institute (EMBL-EBI) · European Molecular Biology Laboratory (EMBL) · Flatiron Institute · J. Craig Venter Institute (JCVI) · Max Planck Institute of Molecular Cell Biology and Genetics (MPI-CBG) · US National Center for Biotechnology Information (NCBI) · Japanese Institute of Genetics · Netherlands Bioinformatics Centre (NBIC) · Philippine Genome Center (PGC) · Scripps Research · Swiss Institute of Bioinformatics (SIB) · Wellcome Sanger Institute · Whitehead Institute
Organizations	African Society for Bioinformatics and Computational Biology (ASBCB) · Australia Bioinformatics Resource (EMBL-AR) · European Molecular Biology network (EMBnet) · International Nucleotide Sequence Database Collaboration (INSDC) · International Society for Biocuration (ISB) · International Society for Computational Biology (ISCB) (Student Council (ISCB-SC)) · Institute of Genomics and Integrative Biology (CSIR-IGIB) · Japanese Society for Bioinformatics (JSBi)
Meetings	Basel Computational Biology Conference ([BC^2]) · European Conference on Computational Biology (ECCB) · Intelligent Systems for Molecular Biology (ISMB) · International Conference on Bioinformatics (InCoB) · International Conference on Computational Intelligence Methods for Bioinformatics and Biostatistics (CIBB) · ISCB Africa ASBCB Conference on Bioinformatics · Pacific Symposium on Biocomputing (PSB) · Research in Computational Molecular Biology (RECOMB)
File formats	CRAM format · **FASTA format** · FASTQ format · NeXML format · Nexus format · Pileup format · SAM format · Stockholm format · VCF format
Related topics	Computational biology · List of biobanks · List of biological databases · Molecular phylogenetics · Sequencing · Sequence database · Sequence alignment

Category · Commons

Categories: Bioinformatics | Biological sequence format