**Help - About Nucleotide And Protein Sequence Formats**

Sequence formats are simply the way in which the amino acid or DNA sequence is recorded in a computer file. Different programs expect different formats, so if you are to submit a job successfully, it is important to understand what the various formats look like.

In order to successfully submit a job it is important to understand what the various sequence formats used for describing biological sequences are and what their basic structure is. The job submission forms are fairly flexible but cannot cope with too much inconsistency.  
You can submit sequence to the search and analysis programs in any of the formats mentioned in the options your [chosen tool](http://www.ebi.ac.uk/help/programs.html).

If you are submitting sequences to ClustalW2 or pratt you may the normal format, as described below, just making sure that the sequences follow each other and are separated from each other with the format´s separator. In the case of EMBL format this would be `//´.

In order to aid the user with the process of converting sequences to appropriate formats please use the following link: [READSEQ](http://www.ebi.ac.uk/readseq/index.html).

**Examples of Sequence Formats:**

* [ALN/ClustalW2 format](http://www.ebi.ac.uk/help/formats.html#aln)[download](http://www.ebi.ac.uk/help/formats/aln.txt)
* [AMPS Block file format](http://www.ebi.ac.uk/help/formats.html#amps)[download](http://www.ebi.ac.uk/help/formats/amps.txt)
* [ClustalW2](http://www.ebi.ac.uk/help/formats.html#ClustalW2) [download](http://www.ebi.ac.uk/help/formats/ClustalW.txt)
* [Codata](http://www.ebi.ac.uk/help/formats.html#codata) [download](http://www.ebi.ac.uk/help/formats/Codata.txt)
* [EMBL](http://www.ebi.ac.uk/help/formats.html#EMBL) [download](http://www.ebi.ac.uk/help/formats/EMBL.txt)
* [GCG/MSF](http://www.ebi.ac.uk/help/formats.html#MSF) [download](http://www.ebi.ac.uk/help/formats/GCG_MSF.txt)
* [GDE](http://www.ebi.ac.uk/help/formats.html#gde) [download](http://www.ebi.ac.uk/help/formats/GDE.txt)
* [Genebank](http://www.ebi.ac.uk/help/formats.html#Genebank) [download](http://www.ebi.ac.uk/help/formats/Genebank.txt)
* [FASTA (Pearson)](http://www.ebi.ac.uk/help/formats.html#fasta)[download](http://www.ebi.ac.uk/help/formats/Fasta.txt)
* [NBRF/PIR](http://www.ebi.ac.uk/help/formats.html#pir)[download](http://www.ebi.ac.uk/help/formats/NBRF_PIR.txt)
* [PDB format](http://www.ebi.ac.uk/help/formats.html#pdb) [download](http://www.ebi.ac.uk/help/formats/pdb.txt)
* [Pfam/Stockholm format](http://www.ebi.ac.uk/help/formats.html#pfam) [download](http://www.ebi.ac.uk/help/formats/pfam.txt)
* [Phylip](http://www.ebi.ac.uk/help/formats.html#Phylip) [download](http://www.ebi.ac.uk/help/formats/Phylip.txt)
* [Raw](http://www.ebi.ac.uk/help/formats.html#raw)[download](http://www.ebi.ac.uk/help/formats/Raw.txt)
* [RSF](http://www.ebi.ac.uk/help/formats.html#rsf)[download](http://www.ebi.ac.uk/help/formats/rich.txt)
* [UniProtKB/Swiss-Prot](http://www.ebi.ac.uk/help/formats.html#swiss) [download](http://www.ebi.ac.uk/help/formats/swiss.txt)

Click [here](http://emboss.sourceforge.net/docs/themes/SequenceFormats.html) to see a complete list of sequence formats supported by EMBOSS applications.

**ALN/ClustalW2 format:**

ALN format was originated in the alignment program ClustalW2. The file starts with word "CLUSTAL" and then some information about which clustal program was run and the version of clustal used.  
e.g. "CLUSTAL W (2.1) multiple sequence alignment"  
The type of clustal program is "W" and the version is 2.1.  
The alignment is written in blocks of 60 residues.   
Every block starts with the sequence names, obtained from the input sequence, and a count of the total number of residues is shown at the end of the line.   
The information about which residues match is shown below each block of residues:

"\*" means that the residues or nucleotides in that column are identical in all sequences in the alignment.  
":" means that conserved substitutions have been observed.  
"." means that semi-conserved substitutions are observed.  
  
An example is shown below.

CLUSTAL W 2.1 multiple sequence alignment

FOSB\_MOUSE ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGS 60

FOSB\_HUMAN ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPVVDPYDMPGTSYSTPGMSGYSSGGASGS 60

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*.\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*:\*.\*\*:\*\*\*\*\*\*

[Top](http://www.ebi.ac.uk/help/formats.html#top)

**AMPS Block file format:**

The first part of a block-file contains the identifier codes of the sequences that are to follow. Each code is prefixed by the > symbol, codes must not contain spaces. e.g.  
>HAHU  
>Trypsin  
>A0046  
>Seq1   
  
  
etc.   
  
The number of ">" symbols is read in the beginning of the file until a \* symbol is found. The \* signals the beginning of the multiple alignment which is stored VERTICALLY, thus columns are individual sequences, whilst rows are aligned positions. The \* symbol must lie over the first sequence. A further star in the same column signals the end of the alignment. Software then uses the number of ">" symbols at the beginning of the file to work out how many columns to read from the \* position. It is therefore important that the only ">" symbols in the file are those that define the identifiers, and the only symbols are those defining the start and end of the multiple alinnment. A simple, small block-file is shown below. 

>Seq\_1  
>A0231  
>HAHU  
>Four\_Alpha  
>Globin  
>GLobin\_C  
\*  
ARNDLQ  
AAAAAA  
PPPPPP  
PP PPP  
WW WWW  
LLLLLL  
IIVVLL  
\*

[Top](http://www.ebi.ac.uk/help/formats.html#top)

**Codata Format:**The first line starts with the text ENTRY". The end of a sequence is delineated by "///". The "SEQUENCE" line specifies the beginning of the sequence lines (starting on the next line), and no sequence is assumed to appear in the entry if the "SEQUENCE" line is missing.

ENTRY IXI\_234

SEQUENCE

5 10 15 20 25 30

1 T S P A S I R P P A G P S S R P A M V S S R R T R P S P P G

31 P R R P T G R P C C S A A P R R P Q A T G G W K T C S G T C

61 T T S T S T R H R G R S G W S A R T T T A A C L R A S R K S

91 M R A A C S R S A G S R P N R F A P T L M S S C I T S T T G

121 P P A W A G D R S H E

///

ENTRY IXI\_235

SEQUENCE

5 10 15 20 25 30

1 T S P A S I R P P A G P S S R - - - - - - - - - R P S P P G

31 P R R P T G R P C C S A A P R R P Q A T G G W K T C S G T C

61 T T S T S T R H R G R S G W - - - - - - - - - - R A S R K S

91 M R A A C S R S A G S R P N R F A P T L M S S C I T S T T G

121 P P A W A G D R S H E

///

ENTRY IXI\_236

SEQUENCE

5 10 15 20 25 30

1 T S P A S I R P P A G P S S R P A M V S S R - - R P S P P P

31 P R R P P G R P C C S A A P P R P Q A T G G W K T C S G T C

61 T T S T S T R H R G R S G W S A R T T T A A C L R A S R K S

91 M R A A C S R - - G S R P P R F A P P L M S S C I T S T T G

121 P P P P A G D R S H E

///

ENTRY IXI\_237

SEQUENCE

5 10 15 20 25 30

1 T S P A S L R P P A G P S S R P A M V S S R R - R P S P P G

31 P R R P T - - - - C S A A P R R P Q A T G G Y K T C S G T C

61 T T S T S T R H R G R S G Y S A R T T T A A C L R A S R K S

91 M R A A C S R - - G S R P N R F A P T L M S S C L T S T T G

121 P P A Y A G D R S H E

///

[Top](http://www.ebi.ac.uk/help/formats.html#top)

**EMBL Format:**

The EMBL entries(as below) in the database are structured so as to be usable by human readers as well as by computer programs. Each entry in the database is composed of lines. Different types of lines, each with its own format, which are used to record the various types of data which make up the entry. Some entries will not contain all of the line types, and some line types occur many times in a single entry. As noted, each entry begins with an identification line (ID) and ends with a terminator line (//). Consult the EMBL [user manual](http://www.ebi.ac.uk/embl/Documentation/User_manual/usrman.html) for a more comprehensive guide. 

* The [ID (**ID**entification line) line](http://www.ebi.ac.uk/help/formats.html#ID) is always the first line of an entry. The general form of the ID line is:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Term** | **Primary accession number** | **Sequence Version Number** | **dataclass** | **molecule** | **Data Class** | **Taxonomic division** | **sequencelength (Base Pairs)** |
| e.g. | X14897 | SV 1 | linear | mRNA | STD | MUS | 4145 BP |

* The [XX line](http://www.ebi.ac.uk/help/formats.html#XX) contains no data or comments. It is used instead of blank lines to avoid confusion with the sequence data lines.
* The [AC (**A**ccession **N**umber) line](http://www.ebi.ac.uk/help/formats.html#AC) lists the accession numbers associated with this entry.
* The [DT (**D**a**T**e) line](http://www.ebi.ac.uk/help/formats.html#DT) shows the date/release number of creation, date/release number of the last modification of the entry and the version number.
* The [DE (**DE**scription) lines](http://www.ebi.ac.uk/help/formats.html#DE) contain general descriptive information about the sequence stored.
* The [KW (**K**ey**W**ord) lines](http://www.ebi.ac.uk/help/formats.html#KW) provide information which can be used to generate cross-reference indexes of the sequence entries based on functional, structural, or other categories deemed important. The keywords chosen for each entry serve as a subject reference for the sequence, and will be expanded as work with the database continues. Often several KW lines are necessary for a single entry.
* The [OS (**O**rganism **S**pecies) line](http://www.ebi.ac.uk/help/formats.html#OS) specifies the preferred scientific name of the organism which was the source of the stored sequence.
* The [OC (**O**rganism **C**lassification) lines](http://www.ebi.ac.uk/help/formats.html#OC) contain the taxonomic classification of the source organism.
* The [RN (**R**eference **N**umber) line](http://www.ebi.ac.uk/help/formats.html#RN) gives a unique number to each reference citation within an entry.
* The RC (**R**eference **C**omment) line type is an optional line type which appears if the reference has a comment.
* The [RP (**R**eference **P**osition) line type](http://www.ebi.ac.uk/help/formats.html#RP) is an optional line type which appears if one or more contiguous base spans of the presented sequence can be attributed to the reference in question.
* The [RX (**R**eference **Cross**-reference)](http://www.ebi.ac.uk/help/formats.html#RX) line type is an optional line type which contains a cross-reference to an external citation or abstract database.
* The [RA (**R**eference **A**uthor) lines](http://www.ebi.ac.uk/help/formats.html#RA) list the authors of the paper (or other work) cited.
* The [RT (**R**eference **T**itle) lines](http://www.ebi.ac.uk/help/formats.html#RT) give the title of the paper (or other work).
* The [RL (**R**eference **L**ocation) line](http://www.ebi.ac.uk/help/formats.html#RL) contains the conventional citation information for the reference.
* The PE (**P**rotein **E**xistance) line describes the evidence evidence for the existence of a protein.
* The [DR (**D**atabase Cross-**R**eference) line](http://www.ebi.ac.uk/help/formats.html#DR) cross-references other databases which contain information related to the entry in which the DR line appears.
* The [CC lines](http://www.ebi.ac.uk/help/formats.html#CC) are free text comments about the entry, and may be used to convey any sort of information thought to be useful.
* The [FH (**F**eature **H**eader) lines](http://www.ebi.ac.uk/help/formats.html#FH) are present only to improve readability of an entry when it is printed or displayed on a terminal screen. The lines contain no data and may be ignored by computer programs.
* The [FT (Feature Table) lines](http://www.ebi.ac.uk/help/formats.html#FT) provide a mechanism for the annotation of the sequence data. Regions or sites in the sequence which are of interest are listed in the table.

A complete and definitive description of the feature table is given [here](http://www.ebi.ac.uk/embl/Documentation/FT_definitions/feature_table.html).

* The [SQ (**S**e**Q**uence header) line](http://www.ebi.ac.uk/help/formats.html#SQ) marks the beginning of the sequence data and gives a summary of its content.
* The [sequence data lines](http://www.ebi.ac.uk/help/formats.html#SD) has lines of code starting with two blanks. The sequence is written 60 bases per line, in groups of 10 bases separated by a blank character, beginning in position 6 of the line. The direction listed is always 5' to 3'
* The [// (terminator) line](http://www.ebi.ac.uk/help/formats.html#TERM) also contains no data or comments. It designates the end of an entry.

ID X14897; SV 1; linear; mRNA; STD; MUS; 4145 BP.

XX

AC X14897;

XX

DT 23-NOV-1989 (Rel. 21, Created)

DT 18-APR-2005 (Rel. 83, Last updated, Version 3)

XX

DE Mouse fosB mRNA

XX

KW fos cellular oncogene; fosB oncogene; oncogene.

XX

OS Mus musculus (house mouse)

OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;

OC Eutheria; Euarchontoglires; Glires; Rodentia; Sciurognathi; Muroidea;

OC Muridae; Murinae; Mus.

XX

RN [1]

RP 1-4145

RX PUBMED; [2498083](http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-id+1g68n1UQZZg+%5bmedline-PMID:2498083%5d+-e).

RA Zerial M., Toschi L., Ryseck R.P., Schuermann M., Mueller R., Bravo R.;

RT "The product of a novel growth factor activated gene, fos B, interacts with

RT JUN proteins enhancing their DNA binding activity";

RL EMBO J. 8(3):805-813(1989).

XX

DR TRANSFAC; [T00291](http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-id+1g68n1UQZZg+%5btfsite-AccNumber:T00291%5d+-e); T00291.

XX

CC clone=AC113-1; cell line=NIH3T3;

XX

FH Key Location/Qualifiers

FH

FT [source](http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-e+-id+1g68n1UQZZg+%5bEMBL_features-id:X14897_1%5d) 1..4145

FT /organism="Mus musculus"

FT /mol\_type="mRNA"

FT /db\_xref="[taxon:10090](http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-id+1g68n1UQZZg+%5btaxonomy-ID:10090%5d+-e)"

FT [CDS](http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-e+-id+1g68n1UQZZg+%5bEMBL_features-id:X14897_2%5d) 1202..2218

FT /note="fosB protein (AA 1-338)"

FT /db\_xref="[GOA:P13346](http://www.ebi.ac.uk/QuickGO/QuickGO?query=P13346&mode=search&entry=&querytype=protein&showcontext=false)"

FT /db\_xref="[InterPro:IPR000837](http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-id+1g68n1UQZZg+%5binterpro-AccNumber:IPR000837%5d+-e)"

FT /db\_xref="[InterPro:IPR004827](http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-id+1g68n1UQZZg+%5binterpro-AccNumber:IPR004827%5d+-e)"

FT /db\_xref="[InterPro:IPR008917](http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-id+1g68n1UQZZg+%5binterpro-AccNumber:IPR008917%5d+-e)"

FT /db\_xref="[InterPro:IPR011700](http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-id+1g68n1UQZZg+%5binterpro-AccNumber:IPR011700%5d+-e)"

FT /db\_xref="[UniProtKB/Swiss-Prot:P13346](http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-id+1g68n1UQZZg+%5Bswissprot-AccNumber:P13346%5D+-e)"

FT /protein\_id="[CAA33026.1](http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-id+1g68n1UQZZg+-e+%5bUNIPARC-refdbsv:CAA33026.1%5d%3eparent)"

FT /translation="MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECA

FT GLGEMPGSFVPTVTAITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSY

FT STPGLSAYSTGGASGSGGPSTSTTTSGPVSARPARARPRRPREETLTPEEEEKRRVRRE

FT RNKLAAAKCRNRRRELTDRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGC

FT KIPYEEGPGPGPLAEVRDLPGSTSAKEDGFGWLLPPPPPPPLPFQSSRDAPPNLTASLF

FT THSEVQVLGDPFPVVSPSYTSSFVLTCPEVSAFAGAQRTSGSEQPSDPLNSPSLLAL"

XX

SQ Sequence 4145 BP; 960 A; 1186 C; 1007 G; 991 T; 1 other;

ataaattctt attttgacac tcaccaaaat agtcacctgg aaaacccgct ttttgtgaca 60

aagtacagaa ggcttggtca catttaaatc actgagaact agagagaaat actatcgcaa 120

actgtaatag acattacatc cataaaagtt tccccagtcc ttattgtaat attgcacagt 180

gcaattgcta catggcaaac tagtgtagca tagaagtcaa agcaaaaaca aaccaaagaa 240

aggagccaca agagtaaaac tgttcaacag ttaatagttc aaactaagcc attgaatcta 300

tcattgggat cgttaaaatg aatcttccta caccttgcag tgtatgattt aacttttaca 360

gaacacaagc caagtttaaa atcagcagta gagatattaa aatgaaaagg tttgctaata 420

gagtaacatt aaataccctg aaggaaaaaa aacctaaata tcaaaataac tgattaaaat 480

tcacttgcaa attagcacac gaatatgcaa cttggaaatc atgcagtgtt ttatttaaga 540

aaacataaaa caaaactatt aaaatagttt tagagggggt aaaatccagg tcctctgcca 600

ggatgctaaa attagacttc aggggaattt tgaagtcttc aattttgaaa cctattaaaa 660

agcccatgat tacagttaat taagagcagt gcacgcaaca gtgacacgcc tttagagagc 720

attactgtgt atgaacatgt tggctgctac cagccacagt caatttaaca aggctgctca 780

gtcatgaact taatacagag agagcacgcc taggcagcaa gcacagcttg ctgggccact 840

ttcctccctg tcgtgacaca atcaatccgt gtacttggtg tatctgaagc gcacgctgca 900

ccgcggcact gcccggcggg tttctgggcg gggagcgatc cccgcgtcgc cccccgtgaa 960

accgacagag cctggacttt caggaggtac agcggcggtc tgaaggggat ctgggatctt 1020

gcagagggaa cttgcatcga aacttgggca gttctccgaa ccggagacta agcttccccg 1080

agcagcgcac tttggagacg tgtccggtct actccggact cgcatctcat tccactcggc 1140

catagccttg gcttcccggc gacctcagcg tggtcacagg ggcccccctg tgcccaggga 1200

aatgtttcaa gcttttcccg gagactacga ctccggctcc cggtgtagct catcaccctc 1260

cgccgagtct cagtacctgt cttcggtgga ctccttcggc agtccaccca ccgccgccgc 1320

ctcccaggag tgcgccggtc tcggggaaat gcccggctcc ttcgtgccaa cggtcaccgc 1380

aatcacaacc agccaggatc ttcagtggct cgtgcaaccc accctcatct cttccatggc 1440

ccagtcccag gggcagccac tggcctccca gcctccagct gttgaccctt atgacatgcc 1500

aggaaccagc tactcaaccc caggcctgag tgcctacagc actggcgggg caagcggaag 1560

tggtgggcct tcaaccagca caaccaccag tggacctgtg tctgcccgtc cagccagagc 1620

caggcctaga agaccccgag aagagacact taccccagaa gaagaagaaa agcgaagggt 1680

tcgcagagag cggaacaagc tggctgcagc taagtgcagg aaccgtcgga gggagctgac 1740

agatcgactt caggcggaaa ctgatcagct tgaagaggaa aaggcagagc tggagtcgga 1800

gatcgccgag ctgcaaaaag agaaggaacg cctggagttt gtcctggtgg cccacaaacc 1860

gggctgcaag atcccctacg aagaggggcc ggggccaggc ccgctggccg aggtgagaga 1920

tttgccaggg tcaacatccg ctaaggaaga cggcttcggc tggctgctgc cgccccctcc 1980

accacccccc ctgcccttcc agagcagccg agacgcaccc cccaacctga cggcttctct 2040

ctttacacac agtgaagttc aagtcctcgg cgaccccttc cccgttgtta gcccttcgta 2100

cacttcctcg tttgtcctca cctgcccgga ggtctccgcg ttcgccggcg cccaacgcac 2160

cagcggcagc gagcagccgt ccgacccgct gaactcgccc tcccttcttg ctctgtaaac 2220

tctttagaca aacaaaacaa acaaacccgc aaggaacaag gaggaggaag atgaggagga 2280

gaggggagga agcagtccgg gggtgtgtgt gtggaccctt tgactcttct gtctgaccac 2340

ctgccgcctc tgccatcgga catgacggaa ggacctcctt tgtgttttgt gctccgtctc 2400

tggttttctg tgccccggcg agaccggaga gctggtgact ttggggacag ggggtggggc 2460

ggggatggac acccctcctg catatctttg tcctgttact tcaacccaac ttctggggat 2520

agatggctgg ctgggtgggt agggtggggt gcaacgccca cctttggcgt cttgcgtgag 2580

gctggagggg aaagggtgct gagtgtgggg tgcagggtgg gttgaggtcg agctggcatg 2640

cacctccaga gagacccaac gaggaaatga cagcaccgtc ctgtccttct tttcccccac 2700

ccacccatcc accctcaagg gtgcagggtg accaagatag ctctgttttg ctccctcggg 2760

ccttagctga ttaacttaac atttccaaga ggttacaacc tcctcctgga cgaattgagc 2820

ccccgactga gggaagtcga tgcccccttt gggagtctgc taaccccact tcccgctgat 2880

tccaaaatgt gaacccctat ctgactgctc agtctttccc tcctgggaaa actggctcag 2940

gttggatttt tttcctcgtc tgctacagag ccccctccca actcaggccc gctcccaccc 3000

ctgtgcagta ttatgctatg tccctctcac cctcaccccc accccaggcg cccttggccg 3060

tcctcgttgg gccttactgg ttttgggcag cagggggcgc tgcgacgccc atcttgctgg 3120

agcgctttat actgtgaatg agtggtcgga ttgctgggtg cgccggatgg gattgacccc 3180

cagccctcca aaactttccc tgggcctccc cttcttccac ttgcttcctc cctccccttg 3240

acagggagtt agactcgaaa ggatgaccac gacgcatccc ggtggccttc ttgctcaggc 3300

cccagacttt ttctctttaa gtccttcgcc ttccccagcc taggacgcca acttctcccc 3360

accctgggag ccccgcatcc tctcacagag gtcgaggcaa ttttcagaga agttttcagg 3420

gctgaggctt tggctcccct atcctcgata tttgaatccc caaatatttt tggactagca 3480

tacttaagag ggggctgagt tcccactatc ccactccatc caattccttc agtcccaaag 3540

acgagttctg tcccttccct ccagctttca cctcgtgaga atcccacgag tcagatttct 3600

attttttaat attggggaga tgggccctac cgcccgtccc ccgtgctgca tggaacattc 3660

cataccctgt cctgggccct aggttccaaa cctaatccca aaccccaccc ccagctattt 3720

atccctttcc tggttcccaa aaagcactta tatctattat gtataaataa atatattata 3780

tatgagtgtg cgtgtgtgtg cgtgtgcgtg cgtgcgtgcg tgcgtgcgag cttccttgtt 3840

ttcaagtgtg ctgtggagtt caaaatcgct tctggggatt tgagtcagac tttctggctg 3900

tccctttttg tcaccttttt gttgttgtct cggctcctct ggctgttgga gacagtcccg 3960

gcctctccct ttatcctttc tcaagtctgt ctcgctcaga ccacttccaa catgtctcca 4020

ctctcaatga ctctgatctc cggtntgtct gttaattctg gatttgtcgg ggacatgcaa 4080

ttttacttct gtaagtaagt gtgactgggt ggtagatttt ttacaatcta tatcgttgag 4140

aattc 4145

//

[Top](http://www.ebi.ac.uk/help/formats.html#top)

**FASTA Format:**

* This format contains a single header line providing the sequence name, and optionally a description, followed by lines of sequence data.
* Sequences in FASTA formatted files are preceded by a line starting with a " >" symbol.
* The first word on this line is the name of the sequence. The rest of the line is a description of the sequence.

Term Entry Name Molecule Type Gene Name Sequence Length

e.g. FOSB\_MOUSE Protein fosB 338 bp

* The remaining lines contain the sequence itself, usually formated to 60 characters per line.
* Depending on the application blank lines in a FASTA file are ignored or treated as terminating the sequence
* Depending on the application spaces or other non-sequence symbols (dashes, underscores, periods) in a sequence are either ignored or treated as gaps.
* FASTA files containing multiple sequences are just the same, with one sequence listed right after another. This format is accepted for many multiple sequence alignment programs.

>FOSB\_MOUSE Protein fosB. 338 bp

MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA

ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGS

GGPSTSTTTSGPVSARPARARPRRPREETLTPEEEEKRRVRRERNKLAAAKCRNRRRELT

DRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEGPGPGPLAEVRD

LPGSTSAKEDGFGWLLPPPPPPPLPFQSSRDAPPNLTASLFTHSEVQVLGDPFPVVSPSY

TSSFVLTCPEVSAFAGAQRTSGSEQPSDPLNSPSLLAL

[Top](http://www.ebi.ac.uk/help/formats.html#top)

**GCG/MSF Format**

* The file may begin with as many lines of comment or description as required.
* The comments are terminated with a line starting with two slashes.
* The first mandatory line that is recognised as part of the MSF file is the line containing the text "MSF:", this line also includes the sequence length, type and date plus an internal check sum value.
* The next line is a mandatory blank line inserted before the sequence names.
* There then follows one line per sequence describing the sequence name, length, checksum and a weight value. Only one name per line is allowed; the qualifier "Name: " is followed by the sequence name. Names are restricted to 10 characters or less. Extra characters, between the sequence names and "Len: " are acceptable if they contain no blank characters. Another blank line is added followed by a line starting with two slashes "//" , this indicates the end of the name list.
* There then follows another blank line.
* Sequences are interleaved on separate lines with gaps represented by periods. Each sequence line starts with the sequence name which is separated from the aligned sequence residues by white space.

MSF: 510 Type: P Check: 7736 ..

Name: ACHE\_BOVIN oo Len: 510 Check: 7842 Weight: 16.0

Name: ACHE\_HUMAN oo Len: 510 Check: 8553 Weight: 17.8

Name: ACHE\_MOUSE oo Len: 510 Check: 229 Weight: 12.5

Name: ACHE\_RAT oo Len: 510 Check: 8410 Weight: 14.2

Name: ACHE\_XENLA oo Len: 510 Check: 2702 Weight: 39.2

//

ACHE\_BOVIN MAGALLCALL LLQLLGRGEG KNEELRLYHY LFDTYDPGRR PVQEPEDTVT

ACHE\_HUMAN MARAPLGVLL LLGLLGRGVG KNEELRLYHH LFNNYDPGSR PVREPEDTVT

ACHE\_MOUSE MAGALLGALL LLTLFGRSQG KNEELSLYHH LFDNYDPECR PVRRPEDTVT

ACHE\_RAT MTMALLGTLL LLALFGRSQG KNEELSLYHH LFDNYDPECR PVRRPEDTVT

ACHE\_XENLA MESGVRILSL LILLHNSLAS ESEESRLIKH LFTSYDQKAR PSKGLDDVVP

ACHE\_BOVIN ISLKVTLTNL ISLNEKEETL TTSVWIGIDW QDYRLNYSKG DFGGVETLRV

ACHE\_HUMAN ISLKVTLTNL ISLNEKEETL TTSVWIGIDW QDYRLNYSKD DFGGIETLRV

ACHE\_MOUSE ITLKVTLTNL ISLNEKEETL TTSVWIGIDW HDYRLNYSKD DFAGVGILRV

ACHE\_RAT ITLKVTLTNL ISLNEKEETL TTSVWIGIEW QDYRLNFSKD DFAGVEILRV

ACHE\_XENLA VTLKLTLTNL IDLNEKEETL TTNVWVQIAW NDDRLVWNVT DYGGIGFVPV

[Top](http://www.ebi.ac.uk/help/formats.html#top)

**GDE Format:**  
  
GDE format is a tagged field format used for storing all available information about a sequence. The format matches very closely the GDE internal structures for sequence data. The format consists of text records starting and ending with braces ('{}'). Between the open and close braces are several tagged field lines specifying different pieces of information about a given sequence. The tag values can be wrapped with double quote characters ('""') as needed. If quotes are not used, the first white space delimited string is taken as the value.Any fields that are not specified are assumed to be the default values. Offsets can be negative as well as positive. Genbank entries written out in this format will have all (") converted to ('), and all ({}) converted to ([]) to avoid confusion in the parser. Leading and trailing gaps are removed prior to writing each sequence. This format is deliberately verbose in order to be simple to duplicate. 

{

name "Short name for sequence"   
longname "Long (more descriptive) name for sequence"   
sequence-ID "Unique ID number"   
creation-date "mm/dd/yy hh:mm:ss"   
direction [-1|1]   
strandedness [1|2]   
type [DNA|RNA||PROTEIN|TEXT|MASK]   
offset (-999999,999999)   
group-ID (0,999)   
creator "Author's name"   
descrip "Verbose description"   
comments "Lines of comments that can be fairly arbitrary text about a   
sequence. Return characters are allowed, but no internal double quotes   
or brace characters. Remember to close with a double quote"   
sequence "gctagctagctagctagctcttagctgtagtcgtagctgatgctagct   
gatgctagctagctagctagctgatcgatgctagctgatcgtagctgacg   
gactgatgctagctagctagctagctgtctagtgtcgtagtgcttattgc"

}

[Top](http://www.ebi.ac.uk/help/formats.html#top)

**Genebank Format:**

GenBank is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences. Although there is daily exchange of information with the [EMBL Nucleotide Sequence Database](http://www.ebi.ac.uk/help/formats.html#embl), it has it's own sequence format shown below. Each GenBank entry includes a concise description of the sequence, the scientific name and taxonomy of the source organism, and a table of features that identifies coding regions and other sites of biological significance, such as transcription units, sites of mutations or modifications, and repeats. Protein translations for coding regions are included in the feature table. Bibliographic references are included along with a link to the Medline unique identifier for all published sequences. Each sequence entry is composed of lines. Different types of lines, each with their own format, are used to record the various data that make up the entry. 

* LOCUS: Short name for this sequence (Maximum of 32 characters).
* DEFINITION: Definition of sequence (Maximum of 80 characters).
* ACCESSION: accession number of the entry.
* VERSION: Version of the entry.
* DBSOURCE: Shows the source, the date of creation and last modification of the database entry.
* KEYWORDS: Keywords for the entry.
* AUTHORS: Authors for the work.
* TITLE: Title of the publication.
* JOURNAL: Journal reference for the entry.
* MEDLINE: Medline ID.
* COMMENT: Lines of comments.
* SOURCE ORGANISM: The organism from which the sequence was derived.
* ORGANISM: Full name of organism (Maximum of 80 characters).
* AUTHORS: Authors of this sequence (Maximum of 80 characters).
* ACCESSION: ID Number for this sequence (Maximum of 80 characters).
* FEATURES: Features of the sequence.
* ORIGIN: Beginning of sequence data.
* // End of sequence data.

LOCUS MMFOSB 4145 bp mRNA linear ROD 12-SEP-1993

DEFINITION Mouse fosB mRNA.

ACCESSION X14897

VERSION X14897.1 GI:50991

KEYWORDS fos cellular oncogene; fosB oncogene; oncogene.

SOURCE Mus musculus.

ORGANISM [Mus musculus](http://www.ncbi.nlm.nih.gov/htbin-post/Taxonomy/wgetorg?id=10090" \t "_mainwindow)

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;

Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.

REFERENCE 1 (bases 1 to 4145)

AUTHORS Zerial,M., Toschi,L., Ryseck,R.P., Schuermann,M., Muller,R. and

Bravo,R.

TITLE The product of a novel growth factor activated gene, fos B,

interacts with JUN proteins enhancing their DNA binding activity

JOURNAL EMBO J. 8 (3), 805-813 (1989)

MEDLINE [89251612](http://www.ncbi.nlm.nih.gov/entrez/utils/qmap.cgi?uid=89251612&form=6&db=m&Dopt=r)

PUBMED [2498083](http://www.ncbi.nlm.nih.gov/entrez/utils/qmap.cgi?uid=2498083&form=6&db=m&Dopt=r)

COMMENT clone=AC113-1; cell line=NIH3T3.

FEATURES Location/Qualifiers

source 1..4145

/organism="Mus musculus"

/db\_xref="taxon:10090"

[CDS](http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?val=50991&itemid=1&view=gbwithparts) 1202..2218

/note="fosB protein (AA 1-338)"

/codon\_start=1

/protein\_id="[CAA33026.1](http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?val=CAA33026.1)"

/db\_xref="GI:50992"

/db\_xref="MGD:[95575](http://www.informatics.jax.org/searches/accession_report.cgi?id=MGI:95575" \t "_mainwindow)"

/db\_xref="SWISS-PROT:P13346"

/translation="MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQEC

AGLGEMPGSFVPTVTAITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGT

SYSTPGLSAYSTGGASGSGGPSTSTTTSGPVSARPARARPRRPREETLTPEEEEKRRV

RRERNKLAAAKCRNRRRELTDRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAH

KPGCKIPYEEGPGPGPLAEVRDLPGSTSAKEDGFGWLLPPPPPPPLPFQSSRDAPPNL

TASLFTHSEVQVLGDPFPVVSPSYTSSFVLTCPEVSAFAGAQRTSGSEQPSDPLNSPS

LLAL"

BASE COUNT 960 a 1186 c 1007 g 991 t 1 others

ORIGIN

1 ataaattctt attttgacac tcaccaaaat agtcacctgg aaaacccgct ttttgtgaca

61 aagtacagaa ggcttggtca catttaaatc actgagaact agagagaaat actatcgcaa

121 actgtaatag acattacatc cataaaagtt tccccagtcc ttattgtaat attgcacagt

181 gcaattgcta catggcaaac tagtgtagca tagaagtcaa agcaaaaaca aaccaaagaa

241 aggagccaca agagtaaaac tgttcaacag ttaatagttc aaactaagcc attgaatcta

301 tcattgggat cgttaaaatg aatcttccta caccttgcag tgtatgattt aacttttaca

361 gaacacaagc caagtttaaa atcagcagta gagatattaa aatgaaaagg tttgctaata

421 gagtaacatt aaataccctg aaggaaaaaa aacctaaata tcaaaataac tgattaaaat

481 tcacttgcaa attagcacac gaatatgcaa cttggaaatc atgcagtgtt ttatttaaga

541 aaacataaaa caaaactatt aaaatagttt tagagggggt aaaatccagg tcctctgcca

601 ggatgctaaa attagacttc aggggaattt tgaagtcttc aattttgaaa cctattaaaa

661 agcccatgat tacagttaat taagagcagt gcacgcaaca gtgacacgcc tttagagagc

721 attactgtgt atgaacatgt tggctgctac cagccacagt caatttaaca aggctgctca

781 gtcatgaact taatacagag agagcacgcc taggcagcaa gcacagcttg ctgggccact

841 ttcctccctg tcgtgacaca atcaatccgt gtacttggtg tatctgaagc gcacgctgca

901 ccgcggcact gcccggcggg tttctgggcg gggagcgatc cccgcgtcgc cccccgtgaa

961 accgacagag cctggacttt caggaggtac agcggcggtc tgaaggggat ctgggatctt

1021 gcagagggaa cttgcatcga aacttgggca gttctccgaa ccggagacta agcttccccg

1081 agcagcgcac tttggagacg tgtccggtct actccggact cgcatctcat tccactcggc

1141 catagccttg gcttcccggc gacctcagcg tggtcacagg ggcccccctg tgcccaggga

1201 aatgtttcaa gcttttcccg gagactacga ctccggctcc cggtgtagct catcaccctc

1261 cgccgagtct cagtacctgt cttcggtgga ctccttcggc agtccaccca ccgccgccgc

1321 ctcccaggag tgcgccggtc tcggggaaat gcccggctcc ttcgtgccaa cggtcaccgc

1381 aatcacaacc agccaggatc ttcagtggct cgtgcaaccc accctcatct cttccatggc

1441 c

[Top](http://www.ebi.ac.uk/help/formats.html#top)

**NBRF/PIR Format:**

* A sequence in PIR format consists of:
  1. One line starting with
     1. a "**>**" (greater-than) sign, followed by
     2. a two-letter code describing the sequence type (P1, F1, DL, DC, RL, RC, or XX), followed by
     3. a semicolon, followed by
     4. the sequence identification code (the database ID-code).
  2. One line containing a textual description of the sequence.
  3. One or more lines containing the sequence itself. The end of the sequence is marked by a "**\***" (asterisk) character.
* A file in PIR format may comprise more than one sequence.

|  |  |
| --- | --- |
| **Sequence type** | **Code** |
| Protein (complete) | P1 |
| Protein (fragment) | F1 |
| DNA (linear) | DL |
| DNA (circular) | DC |
| RNA (linear) | RL |
| RNA (circular) | RC |
| tRNA | N3 |
| other functional RNA | N1 |

>P1;CRAB\_ANAPL

ALPHA CRYSTALLIN B CHAIN (ALPHA(B)-CRYSTALLIN).

MDITIHNPLI RRPLFSWLAP SRIFDQIFGE HLQESELLPA SPSLSPFLMR

SPIFRMPSWL ETGLSEMRLE KDKFSVNLDV KHFSPEELKV KVLGDMVEIH

GKHEERQDEH GFIAREFNRK YRIPADVDPL TITSSLSLDG VLTVSAPRKQ

SDVPERSIPI TREEKPAIAG AQRK\*

[Top](http://www.ebi.ac.uk/help/formats.html#top)

**PDB Format:**

**Basic Notions of the Format Description**

**Character Set**

Only non-control ASCII characters, as well as the space and end-of-line indicator, appear in a PDB coordinate entry file. Namely:

abcdefghijklmnopqrstuvwxyzABCDEFGHIJKLMNOPQRSTUVWXYZ

1234567890

` - = [ ] \ ; ' , . / ~ ! @ # $ % ^ & \* ( ) \_ + { } | : " < > ?

the space, and end-of-line. The end-of-line indicator is system-specific. Unix uses a line feed character; other systems may use a carriage return followed by a line feed.

**Special Characters**

Greek letters are spelled out, i.e., alpha, beta, gamma, etc.

Bullets are represented as (DOT).

Right arrow is represented as -->.

Left arrow is represented as <--.

Superscripts are initiated and terminated by double equal signs, e.g., S==2+==.

Subscripts are initiated and terminated by single equal signs, e.g., F=c=.

If "=" is surrounded by at least one space on each side, then it is assumed to be an equal sign, e.g., 2 + 4 = 6.

Commas, colons, and semi-colons are used as list delimiters in records which have one of the following data types:

List

SList

Specification List

Specification

If a comma, colon, or semi-colon is used in any context other than as a delimiting character, then the character must be escaped, i.e., immediately preceded by a backslash, "\". Examples of this use are found in line 4 of each of the following:

COMPND MOL\_ID: 1;

COMPND 2 MOLECULE: GLUTATHIONE SYNTHETASE;

COMPND 3 CHAIN: NULL;

COMPND 4 SYNONYM: GAMMA-L-GLUTAMYL-L-CYSTEINE\:GLYCINE LIGASE

COMPND 5 (ADP-FORMING);

COMPND 6 EC: 6.3.2.3;

COMPND 7 ENGINEERED: YES

COMPND MOL\_ID: 1;

COMPND 2 MOLECULE: S-ADENOSYLMETHIONINE SYNTHETASE;

COMPND 3 CHAIN: A, B;

COMPND 4 SYNONYM: MAT, ATP\:L-METHIONINE S-ADENOSYLTRANSFERASE;

COMPND 5 EC: 2.5.1.6;

COMPND 6 ENGINEERED: YES;

COMPND 7 BIOLOGICAL\_UNIT: TETRAMER;

COMPND 8 OTHER\_DETAILS: TETRAGONAL MODIFICATIONs

[Top](http://www.ebi.ac.uk/help/formats.html#top)

**Pfam/Stockholm Format:**

The "[Pfam/Stockholm](http://www.cgr.ki.se/cgb/groups/sonnhammer/Stockholm.html)" format is a system for marking up features in a multiple alignment. These mark-up annotations are preceded by a 'magic' label, of which there are four types.   
  
**Header:**  
The first line in the file must contain a format and version identifier, currently:

# STOCKHOLM 1.0  
  
**The sequence alignment:**  
  
< seqname> <aligned sequence>   
< seqname> <aligned sequence>   
< seqname> <aligned sequence>   
.  
.  
//

<seqname> stands for "sequence name", typically in the form "name/start-end" or just "name".  
The "//" line indicates the end of the alignment.  
Sequence letters may include any characters except whitespace. Gaps may be indicated by "." or "-".  
Wrap-around alignments are allowed in principle, mainly for historical reasons, but are not used in e.g. Pfam. Wrapped alignments are discouraged since they are much harder to parse.  
  
**The alignment mark-up:**  
  
Mark-up lines may include any characters except whitespace. Use underscore ("\_") instead of space.

#=GF <feature> <Generic per-File annotation, free text>  
#=GC <feature> <Generic per-Column annotation, exactly 1 char per column>  
#=GS <seqname> <feature> <Generic per-Sequence annotation, free text>  
#=GR <seqname> <feature> <Generic per-Sequence AND per-Column markup, exactly 1 char per column>  
  
**Example:**

# STOCKHOLM 1.0  
#=GF ID CBS  
#=GF AC PF00571  
#=GF DE CBS domain  
#=GF AU Bateman A  
#=GF CC CBS domains are small intracellular modules mostly found   
#=GF CC in 2 or four copies within a protein.   
#=GF SQ 67  
#=GS O31698/18-71 AC O31698  
#=GS O83071/192-246 AC O83071  
#=GS O83071/259-312 AC O83071  
#=GS O31698/88-139 AC O31698  
#=GS O31698/88-139 OS Bacillus subtilis  
O83071/192-246 MTCRAQLIAVPRASSLAE..AIACAQKM....RVSRVPVYERS  
#=GR O83071/192-246 SA 999887756453524252..55152525....36463774777  
O83071/259-312 MQHVSAPVFVFECTRLAY..VQHKLRAH....SRAVAIVLDEY  
#=GR O83071/259-312 SS CCCCCHHHHHHHHHHHHH..EEEEEEEE....EEEEEEEEEEE  
O31698/18-71 MIEADKVAHVQVGNNLEH..ALLVLTKT....GYTAIPVLDPS  
#=GR O31698/18-71 SS CCCHHHHHHHHHHHHHHH..EEEEEEEE....EEEEEEEEHHH  
O31698/88-139 EVMLTDIPRLHINDPIMK..GFGMVINN......GFVCVENDE  
#=GR O31698/88-139 SS CCCCCCCHHHHHHHHHHH..HEEEEEEE....EEEEEEEEEEH  
#=GC SS\_cons CCCCCHHHHHHHHHHHHH..EEEEEEEE....EEEEEEEEEEH  
O31699/88-139 EVMLTDIPRLHINDPIMK..GFGMVINN......GFVCVENDE  
#=GR O31699/88-139 AS \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
#=GR\_O31699/88-139\_IN \_\_\_\_\_\_\_\_\_\_\_\_1\_\_\_\_\_\_\_\_\_\_\_\_\_\_2\_\_\_\_\_\_\_\_\_\_0\_\_\_\_  
//

**Phylip Format:**

* The first line of the input file contains the number of species, the number of sequences and their length (in characters)separated by blanks.
* The next line contains the sequence name, followed by the sequence in blocks of 10 characters.

1 338 I   
FOSB\_MOUSE MFQAFPGDYD SGSRCSSSPS AESQYLSSVD SFGSPPTAAA SQECAGLGEM  
 PGSFVPTVTA ITTSQDLQWL VQPTLISSMA QSQGQPLASQ PPAVDPYDMP  
 GTSYSTPGLS AYSTGGASGS GGPSTSTTTS GPVSARPARA RPRRPREETL  
 TPEEEEKRRV RRERNKLAAA KCRNRRRELT DRLQAETDQL EEEKAELESE  
 IAELQKEKER LEFVLVAHKP GCKIPYEEGP GPGPLAEVRD LPGSTSAKED  
 GFGWLLPPPP PPPLPFQSSR DAPPNLTASL FTHSEVQVLG DPFPVVSPSY  
 TSSFVLTCPE VSAFAGAQRT SGSEQPSDPL NSPSLLAL

[Top](http://www.ebi.ac.uk/help/formats.html#top)

**Raw Format:**

Like text/plain format except that it removes any white space or digits, accepts only alphabetic characters and rejects anything else. This means that it is safer to use this format that plain format. If you have digits and spaces or TAB characters, these are removed and ignored. If you have other non-alphabetic characters (for example, punctuation characters), then the sequence will be rejected as erroneous. 

ataaattcttattttgacactcaccaaaatagtcacctggaaaacccgctttttgtgaca

aagtacagaaggcttggtcacatttaaatcactgagaactagagagaaatactatcgcaa

actgtaatagacattacatccataaaagtttccccagtccttattgtaatattgcacagt

gcaattgctacatggcaaactagtgtagcatagaagtcaaagcaaaaacaaaccaaagaa

aggagccacaagagtaaaactgttcaacagttaatagttcaaactaagccattgaatcta

tcattgggatcgttaaaatgaatcttcctacaccttgcagtgtatgatttaacttttaca

[Top](http://www.ebi.ac.uk/help/formats.html#top)

**RSF Format:**

RSF means rich sequence format and it is created by the Editor in SeqLab. The format is recognised by the word !!RICH\_SEQUENCE at the beginning of  
the file. It contains one or more sequences that may or may not be related. In addition to the sequence data, each sequence can be annotated with descriptive sequence information such as:

* Creator/author of the sequence
* Sequence weight
* Creation date
* One-line description of the sequence
* Offset, or the number of leading gaps in a sequence that is part of an alignment or fragment assembly project Known sequence features

!!RICH\_SEQUENCE 1.0

..

{

name chkhba

type DNA

longname chkhba

checksum 980

creation-date 4/15/98 16:42:47

strand 1

sequence

ACACAGAGGTGCAACCATGGTGCTGTCCGCTGCTGACAAGAACAACGTCAAGGGCATCTT

CACCAAAATCGCCGGCCATGCTGAGGAGTATGGCGCCGAGACCTTGGAAAGGATGTTCAC

CACCTACCCCCCAACCAAGACCTACTTCCCCCACTTCGATCTGTCACACGGCTCCGCTCA

...

}

{

name davagl

type DNA

longname davagl

checksum 7399

creation-date 4/15/98 16:42:47

strand 1

sequence

GTGCTCTCGGATGCTGACAAGACTCACGTGAAAGCCATCTGGGGTAAGGTGGGAGGCCAC

GCCGGTGCCTACGCAGCTGAAGCTCTTGCCAGAACCTTCCTCTCCTTCCCCACTACCAAA

...

}

[Top](http://www.ebi.ac.uk/help/formats.html#top)

**Macsim Format:**

<!-- This is the Document Type Definition (DTD) for Macsim. -->

<!-- A DTD for describing Multiple Alignments of Complete Sequences -->

<!-- and Information Mining -->

<!-- This DTD was created by Julie Thompson (julie@igbmc.u-strasbg.fr) -->

<!-- Institut de Genetique et de Biologie Moleculaire et Cellulaire, -->

<!-- Strasbourg, France. -->

<!-- Email the above address for corrections and suggestions. -->

<!-- This DTD's DISTRIBUTION and USE is UNLIMITED under the condition -->

<!-- that its entire content remains intact. -->

<!-- THIS DTD AND DOCUMENTATION IS PROVIDED 'AS IS,' AND COPYRIGHT -->

<!-- HOLDERS MAKE NO REPRESENTATIONS OR WARRANTIES, EXPRESS OR IMPLIED,-->

<!-- INCLUDING BUT NOT LIMITED TO, WARRANTIES OF MERCHANTABILITY OR -->

<!-- FITNESS FOR ANY PARTICULAR PURPOSE OR THAT THE USE OF THE DTD -->

<!-- OR DOCUMENTATION WILL NOT INFRINGE ANY THIRD PARTY PATENTS, -->

<!-- COPYRIGHTS, TRADEMARKS OR OTHER RIGHTS. -->

<!-- COPYRIGHT HOLDERS WILL NOT BE LIABLE FOR ANY DIRECT, INDIRECT, -->

<!-- SPECIAL OR CONSEQUENTIAL DAMAGES ARISING OUT OF ANY USE OF THE -->

<!-- DTD OR DOCUMENTATION. -->

<!-- The name and trademarks of the copyright holder may NOT be used in-->

<!-- advertising or publicity pertaining to the DTD without -->

<!-- specific, written prior permission. Title to copyright in this -->

<!-- DTD and any associated documentation will at all times remain -->

<!-- with copyright holders. -->

<!-- Version 1.1 : -->

<!-- Version 1.2 : 2005/01/04 Julie added taxid -->

<!-- Version 1.3 : 2005/01/17 Raymond changed aln-txt to freetext and -->

<!-- : add owner+type in freetext and consensus -->

<!-- Version 1.4 : 2005/03/16 Raymond added ? to fscore? -->

<!-- Version 1.5 : 2005/03/29 Julie added sense -1 0 1 -->

<!-- Version 1.6 : 2006/07/11 Julie added surface accessibility and -->

<!-- residue contact list -->

<!ELEMENT macsim (alignment)>

<!ELEMENT

alignment (aln-name, aln-score?,

aln-note?,

(sequence | freetext | consensus | column-score | surface-accessibility)+)

>

<!ELEMENT aln-name (#PCDATA)>

<!ELEMENT aln-score (#PCDATA)>

<!ELEMENT aln-note (#PCDATA)>

<!-- owner signification : 0 for all, 1-n for group, seq-name for sequence -->

<!ELEMENT

freetext (freetext-name,

freetext-owner,

freetext-type,

freetext-data)

>

<!ELEMENT

consensus (cons-name,

cons-owner,

cons-type,

cons-data)

>

<!ELEMENT

column-score (colsco-name,

colsco-owner,

colsco-type,

colsco-data)

>

<!ELEMENT

surface-accessibility (suracc-name,

suracc-owner,

suracc-type,

suracc-data)

>

<!ELEMENT freetext-name (#PCDATA)>

<!ELEMENT freetext-owner (#PCDATA)>

<!ELEMENT freetext-type (#PCDATA)>

<!ELEMENT freetext-data (#PCDATA)>

<!ELEMENT cons-name (#PCDATA)>

<!ELEMENT cons-owner (#PCDATA)>

<!ELEMENT cons-type (#PCDATA)>

<!ELEMENT cons-data (#PCDATA)>

<!ELEMENT colsco-name (#PCDATA)>

<!ELEMENT colsco-owner (#PCDATA)>

<!ELEMENT colsco-type (#PCDATA)>

<!ELEMENT colsco-data (#PCDATA)>

<!ELEMENT suracc-name (#PCDATA)>

<!ELEMENT suracc-owner (#PCDATA)>

<!ELEMENT suracc-type (#PCDATA)>

<!ELEMENT suracc-data (#PCDATA)>

<!-- A sequence must minimally have a name with a type  
 attribute and some sequence data, the info is optional -->

<!ELEMENT sequence (seq-name,

seq-info?,

seq-data)>

<!ATTLIST sequence seq-type (Protein | DNA | PDB) #REQUIRED>

<!ELEMENT seq-name (#PCDATA)>

<!ELEMENT seq-data (#PCDATA)>

<!-- The info section can contain any of the following, in any order -->

<!ELEMENT seq-info (accession | nid | definition | organism | taxid | lifedomain | ec  
 | hydrophobicity | fragment | keywordlist | complex | pub | ftable | residue-contact-list |   
 dbxreflist | length | weight | group | cksum | score | sense | status)+>

<!ELEMENT accession (#PCDATA)>

<!ELEMENT nid (#PCDATA)>

<!ELEMENT definition (#PCDATA)>

<!ELEMENT organism (#PCDATA)>

<!ELEMENT taxid (#PCDATA)>

<!ELEMENT lifedomain (#PCDATA)>

<!ELEMENT ec (#PCDATA)>

<!ELEMENT hydrophobicity (#PCDATA)>

<!ELEMENT fragment EMPTY>

<!ATTLIST fragment status (Yes | No) "No">

<!ELEMENT keywordlist (keyword+)>

<!ELEMENT keyword (#PCDATA)>

<!ELEMENT complex (#PCDATA)>

<!ELEMENT pub (pubxref | authors | journal | other | title)\*>

<!ELEMENT authors (#PCDATA)>

<!ELEMENT journal (#PCDATA)>

<!ELEMENT other (#PCDATA)>

<!ELEMENT pubxref (#PCDATA)>

<!ELEMENT title (#PCDATA)>

<!ELEMENT ftable (fitem+)>

<!ELEMENT fitem (ftype,

fstart,

fstop,

fcolor,

fscore?,

fnote?)>

<!ATTLIST fitem status (Confirmed | Predicted) "Confirmed">

<!ELEMENT ftype (#PCDATA)>

<!ELEMENT fstart (#PCDATA)>

<!ELEMENT fstop (#PCDATA)>

<!ELEMENT fcolor (#PCDATA)>

<!ELEMENT fscore (#PCDATA)>

<!ELEMENT fnote (#PCDATA)>

<!ELEMENT residue-contact-list (contact-residue1 , residue-contact+)>

<!ELEMENT residue-contact (contact-residue2,

contact-distance?,

contact-note?)>

<!ELEMENT contact-residue2 (#PCDATA)>

<!ELEMENT contact-distance (#PCDATA)>

<!ELEMENT contact-note (#PCDATA)>

<!ELEMENT dbxreflist (#PCDATA)>

<!ELEMENT dbxref (dbname,

dbid,

dbnote?,

dbnumber?)>

<!ELEMENT dbname (#PCDATA)>

<!ELEMENT dbid (#PCDATA)>

<!ELEMENT dbnote (#PCDATA)>

<!ELEMENT dbnumber (#PCDATA)>

<!ELEMENT length (#PCDATA)>

<!ELEMENT weight (#PCDATA)>

<!ELEMENT group (#PCDATA)>

<!ELEMENT cksum (#PCDATA)>

<!ELEMENT score (#PCDATA)>

<!ELEMENT sense (#PCDATA)>

<!ELEMENT status (#PCDATA)>

[Top](http://www.ebi.ac.uk/help/formats.html#top)

**UniProtKB/Swiss-Prot Format:**  
  
UniProtKB/Swiss-Prot is an annotated protein sequence database. The UniProtKB/Swiss-Prot protein knowledgebase consists of sequence entries. Sequence entries are composed of different line types, each with their own format. For standardisation purposes the format of UniProtKB/Swiss-Prot follows as closely as possible that of the EMBL Nucleotide Sequence Database. The UniProtKB/Swiss-Prot user manual is available[here](http://ca.expasy.org/sprot/userman.html). The entries in the UniProtKB/Swiss-Prot database are structured so as to be usable by human readers as well as by computer programs. The explanations, descriptions, classifications and other comments are in ordinary English. Wherever possible, symbols familiar to biochemists, protein chemists and molecular biologists are used. Each sequence entry is composed of lines. Different types of lines, each with their own format, are used to record the various data that make up the entry. 

* The [ID (**ID**entification) line](http://www.ebi.ac.uk/help/formats.html#SID) is always the first line of an entry. The general form of the ID line is:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Term** | **ID** | **ENTRY\_NAME** | **STATUS** | **SEQUENCE\_LENGTH.** |
| e.g. | ID | FOSB\_MOUSE | Reviewed | 338 AA |

* + **Entry name:** The first item on the ID line is the entry name of the sequence. This name is a useful means of identifying a sequence. The entry name consists of up to ten uppercase alphanumeric characters.
  + **Status:** To distinguish the fully annotated entries in the Swiss-Prot section of the UniProt Knowledgebase from the computer-annotated entries in the TrEMBL section, the 'status' of each entry is indicated in the first (ID) line of each entry. The two defined classes are:
    - Reviewed   
      Entries that have been manually reviewed and annotated by UniProtKB curators (Swiss-Prot section of the UniProt Knowledgebase).
    - Unreviewed  
      Computer-annotated entries that have not been reviewed by UniProtKB curators (TrEMBL section of the UniProt Knowledgebase).
  + **Length of the molecule:**The sequence length in amino acids.
* The [AC (**AC**cession number) line](http://www.ebi.ac.uk/help/formats.html#SAC) lists the accession number(s) associated with an entry.
* The [DT (**D**a**T**e) lines](http://www.ebi.ac.uk/help/formats.html#SDT) shows the date of creation and last modification of the database entry.
* The [DE (**DE**scription) lines](http://www.ebi.ac.uk/help/formats.html#SDE) contain general descriptive information about the sequence stored.
* The [GN (**G**ene **N**ame) line](http://www.ebi.ac.uk/help/formats.html#SGN) contains the name(s) of the gene(s) that code for the stored protein sequence.
* The [OS (Organism Species) line](http://www.ebi.ac.uk/help/formats.html#SOS) specifies the organism(s) which was (were) the source of the stored sequence.
* The OG (**O**r**G**anelle) line indicates if the gene coding for a protein originates from the mitochondria, the chloroplast, a cyanelle, or a plasmid.
* The PR (**PR**oject) line shows the International Nucleotide Sequence Database Collaboration (INSDC) Project Identifier that has been assigned to the entry.
* The [OC (**O**rganism **C**lassification) lines](http://www.ebi.ac.uk/help/formats.html#SOC) contain the taxonomic classification of the source organism.
* The [OX (**O**rganism taxonomy **Cross**-Reference) line](http://www.ebi.ac.uk/help/formats.html#SOX) is used to indicate the identifier to a specific organism in a taxonomic database.
* The [RN (**R**eference **N**umber) line](http://www.ebi.ac.uk/help/formats.html#SRN) gives a sequential number to each reference citation in an entry.
* The [RP (Reference Position) line](http://www.ebi.ac.uk/help/formats.html#srp) describes the extent of the work carried out by the authors of the reference cited.
* The RC (**R**eference **C**omment) lines are optional lines which are used to store comments relevant to the reference cited.
* The [RX (**R**eference **Cross**-Reference) line](http://www.ebi.ac.uk/help/formats.html#SRX) is an optional line which is used to indicate the identifier assigned to a specific reference in a bibliographic database.
* The [RA (**R**eference **A**uthor) lines](http://www.ebi.ac.uk/help/formats.html#SRA) list the authors of the paper (or other work) cited.
* The [RT (**R**eference **T**itle) lines](http://www.ebi.ac.uk/help/formats.html#SRT) give the title of the paper (or other work) cited.
* The [RL (**R**eference **L**ocation) lines](http://www.ebi.ac.uk/help/formats.html#SRL) contain the conventional citation information for the reference.
* The [CC lines](http://www.ebi.ac.uk/help/formats.html#SCC) are free text comments on the entry, and are used to convey any useful information.
* The [DR (**D**atabase cross-**R**eference) lines](http://www.ebi.ac.uk/help/formats.html#SDR) are used as pointers to information related to UniProtKB/Swiss-Prot entries and found in other data collections.
* The KW (**K**ey**W**ord) lines provide information that can be used to generate indexes of the sequence entries based on functional, structural, or other categories.
* The [FT (**F**eature **T**able) lines](http://www.ebi.ac.uk/help/formats.html#SFT)provide a precise but simple means for the annotation of the sequence data. The table describes regions or sites of interest in the sequence. In general the feature table lists posttranslational modifications, binding sites, enzyme active sites, local secondary structure or other characteristics reported in the cited references.
* The [SQ (**S**e**Q**uence header) line](http://www.ebi.ac.uk/help/formats.html#SSQ) marks the beginning of the sequence data and gives a quick summary of its content.
* The [sequence data line](http://www.ebi.ac.uk/help/formats.html#SSD) has a line code consisting of two blanks rather than the two-letter codes used until now. The sequence counts 60 amino acids per line, in groups of 10 amino acids, beginning in position 6 of the line.
* The [// (terminator) line](http://www.ebi.ac.uk/help/formats.html#STERM) contains no data or comments and designates the end of an entry.

ID FOSB\_MOUSE Reviewed; 338 AA.

AC P13346;

DT 01-JAN-1990, integrated into UniProtKB/Swiss-Prot.

DT 01-JAN-1990, sequence version 1.

DT 20-FEB-2007, entry version 54.

DE Protein fosB.

GN Name=Fosb;

OS Mus musculus (Mouse).

OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;

OC Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Sciurognathi;

OC Muroidea; Muridae; Murinae; Mus.

OX NCBI\_TaxID=[10090](http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-id+newid+%5btaxonomy-ID:10090%5d+-e);

RN [1]

RP NUCLEOTIDE SEQUENCE [MRNA].

RX MEDLINE=[89251612](http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-id+newid+%5bmed2pub-id:89251612%5d%3Emedline+-view+MedlineRef); PubMed=[2498083](http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-id+newid+%5bmedline-PMID:2498083%5d+-e);

RA Zerial M., Toschi L., Ryseck R.-P., Schuermann M., Mueller R.,

RA Bravo R.;

RT "The product of a novel growth factor activated gene, fos B, interacts

RT with JUN proteins enhancing their DNA binding activity.";

RL EMBO J. 8:805-813(1989).

RN [2]

RP NUCLEOTIDE SEQUENCE [GENOMIC DNA].

RX MEDLINE=[92158623](http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-id+newid+%5bmed2pub-id:92158623%5d%3Emedline+-view+MedlineRef); PubMed=[1741260](http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-id+newid+%5bmedline-PMID:1741260%5d+-e); DOI=[10.1093/nar/20.2.343](http://dx.doi.org/10.1093/nar/20.2.343);

RA Lazo P.S., Dorfman K., Noguchi T., Mattei M.-G., Bravo R.;

RT "Structure and mapping of the fosB gene. FosB downregulates the

RT activity of the fosB promoter.";

RL Nucleic Acids Res. 20:343-350(1992).

CC -!- FUNCTION: FosB interacts with Jun proteins enhancing their DNA

CC binding activity.

CC -!- SUBUNIT: Heterodimer (By similarity).

CC -!- SUBCELLULAR LOCATION: Nucleus.

CC -!- INDUCTION: By growth factors.

CC -!- SIMILARITY: Belongs to the bZIP family. Fos subfamily.

CC -!- SIMILARITY: Contains 1 bZIP domain.

CC -----------------------------------------------------------------------

CC Copyrighted by the UniProt Consortium, see http://www.uniprot.org/terms

CC Distributed under the Creative Commons Attribution-NoDerivs License

CC -----------------------------------------------------------------------

DR EMBL; [X14897](http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-id+newid+%5bembl-AccNumber:X14897%5d+-e); [CAA33026](http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-e+%5bEMBL_features-prd:CAA33026%5d).1; -; mRNA.

DR EMBL; [AF093624](http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-id+newid+%5bembl-AccNumber:AF093624%5d+-e); [AAD13196](http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-e+%5bEMBL_features-prd:AAD13196%5d).1; -; Genomic\_DNA.

DR PIR; [S35477](http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-id+newid+%5bpir-AccNumber:S35477%5d+-e); TVMSFB.

DR UniGene; [Mm.248335](http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-id+newid+%5bunigene-ID:Mm.248335%5d+-e); -.

DR HSSP; P01100; [1FOS](http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-id+newid+%5bhssp-ID:1FOS%5d+-e).

DR SMR; [P13346](http://swissmodel.expasy.org/repository/smr.php?sptr_ac=P13346); 157-215.

DR DIP; DIP:1067N; -.

DR TRANSFAC; [T00291](http://www.biobase.de/cgi-bin/get.cgi?database=transfac&ac=T00291); -.

DR Ensembl; [ENSMUSG00000003545](http://www.ensembl.org/Mus_musculus/geneview?gene=ENSMUSG00000003545); Mus musculus.

DR KEGG; [mmu:14282](http://www.genome.ad.jp/dbget-bin/www_bget?mmu:14282); -.

DR MGI; [MGI:95575](http://www.informatics.jax.org/searches/accession_report.cgi?id=MGI:95575); Fosb.

DR ArrayExpress; [P13346](http://www.ebi.ac.uk/aedw/dwd?uniprot=P13346); -.

DR GermOnline; [ENSMUSG00000003545](http://www.germonline.org/Mus_musculus/geneview?gene=ENSMUSG00000003545); Mus musculus.

DR InterPro; [IPR011700](http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-id+newid+%5binterpro-AccNumber:IPR011700%5d+-e); bZIP\_2.

DR InterPro; [IPR008917](http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-id+newid+%5binterpro-AccNumber:IPR008917%5d+-e); Euk\_TF\_DNA\_bd.

DR InterPro; [IPR000837](http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-id+newid+%5binterpro-AccNumber:IPR000837%5d+-e); Leuzip\_Fos.

DR InterPro; [IPR004827](http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-id+newid+%5binterpro-AccNumber:IPR004827%5d+-e); TF\_bZIP.

DR Pfam; [PF07716](http://www.sanger.ac.uk/cgi-bin/Pfam/getacc?PF07716); [bZIP\_2](http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-id+newid+%5bpfama-ID:bZIP_2%5d+-e); 1.

DR PRINTS; [PR00042](http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-id+newid+%5bprints-AccNumber:PR00042%5d+-e); LEUZIPPRFOS.

DR SMART; [SM00338](http://smart.embl-heidelberg.de/smart/do_annotation.pl?BLAST=DUMMY&DOMAIN=SM00338); BRLZ; 1.

DR PROSITE; [PS50217](http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-id+newid+%5bprosite-AccNumber:PS50217%5d+-e); BZIP; 1.

DR PROSITE; [PS00036](http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-id+newid+%5bprosite-AccNumber:PS00036%5d+-e); BZIP\_BASIC; 1.

KW [DNA-binding](http://www.expasy.org/cgi-bin/get-entries?KW=DNA-binding); [Nuclear protein](http://www.expasy.org/cgi-bin/get-entries?KW=Nuclear+protein).

FT [CHAIN](http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-e+-id+newid+%5BUNIPROT_features-id:FOSB_MOUSE_1%5D) 1 338 Protein fosB.

FT /FTId=[PRO\_0000076477](http://www.expasy.org/cgi-bin/get-sprot-variant.pl?PRO_0000076477).

FT [DOMAIN](http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-e+-id+newid+%5bUNIPROT_features-id:FOSB_MOUSE_2%5d) 183 211 Leucine-zipper.

FT [DNA\_BIND](http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-e+-id+newid+%5bUNIPROT_features-id:FOSB_MOUSE_3%5d) 161 179 Basic motif.

SQ SEQUENCE 338 AA; 35977 MW; E9D031A4BEAE48EC CRC64;

MFQAFPGDYD SGSRCSSSPS AESQYLSSVD SFGSPPTAAA SQECAGLGEM PGSFVPTVTA

ITTSQDLQWL VQPTLISSMA QSQGQPLASQ PPAVDPYDMP GTSYSTPGLS AYSTGGASGS

GGPSTSTTTS GPVSARPARA RPRRPREETL TPEEEEKRRV RRERNKLAAA KCRNRRRELT

DRLQAETDQL EEEKAELESE IAELQKEKER LEFVLVAHKP GCKIPYEEGP GPGPLAEVRD

LPGSTSAKED GFGWLLPPPP PPPLPFQSSR DAPPNLTASL FTHSEVQVLG DPFPVVSPSY

TSSFVLTCPE VSAFAGAQRT SGSEQPSDPL NSPSLLAL

//

[Top](http://www.ebi.ac.uk/help/formats.html#top)

**Known biosequence format Extensions**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **ID** | **Name** | **Read** | **Write** | **Int'leaf** | **Document** | **Content-type** | **Suffix** |
| 1 | IG|Stanford | yes | yes | -- | -- | biosequence/ig | .ig |
| 2 | GenBank|GB | yes | yes | -- | yes | biosequence/genbank | .gb |
| 3 | NBRF | yes | yes | -- | -- | biosequence/nbrf | .nbrf |
| 4 | EMBL | yes | yes | -- | yes | biosequence/embl | .embl |
| 5 | GCG | yes | yes | -- | -- | biosequence/gcg | .gcg |
| 6 | DNAStrider | yes | yes | -- | -- | biosequence/strider | .strider |
| 7 | Fitch | -- | -- | -- | -- | biosequence/fitch | .fitch |
| 8 | Pearson|FASTA | yes | yes | -- | -- | biosequence/fasta | .fasta |
| 9 | Zuker | -- | -- | -- | -- | biosequence/zuker | .zuker |
| 10 | Olsen | -- | -- | yes | -- | biosequence/olsen | .olsen |
| 11 | Phylip3.2 | yes | yes | yes | -- | biosequence/phylip2 | .phylip2 |
| 12 | Phylip|Phylip4 | yes | yes | yes | -- | biosequence/phylip | .phylip |
| 13 | Plain|Raw | yes | yes | -- | -- | biosequence/plain | .seq |
| 14 | PIR|CODATA | yes | yes | -- | -- | biosequence/codata | .pir |
| 15 | MSF | yes | yes | yes | -- | biosequence/msf | .msf |
| 16 | PAUP|NEXUS | yes | yes | yes | -- | biosequence/nexus | .nexus |
| 17 | Pretty | -- | yes | yes | -- | biosequence/pretty | .pretty |
| 18 | XML | yes | yes | -- | yes | biosequence/xml | .xml |
| 19 | BLAST | yes | -- | yes | -- | biosequence/blast | .blast |
| 20 | SCF | yes | -- | -- | -- | biosequence/scf | .scf |
| 21 | ASN.1 | -- | -- | -- | -- | biosequence/asn1 | .asn |

[Top](http://www.ebi.ac.uk/help/formats.html#top)