genomatix

Overview of Help-Pages

DNA Sequence formats

[Plain] [FASTQ] [EMBL] [FASTA] [GCG] [GenBank] [IG] [IUPAC]
[How Genomatix represents sequence annotation]

Plain sequence format

A sequence in plain format may contain only IUPAC characters and spaces (no numbers!).

Note: A file in plain sequence format may only contain **one** sequence, while most other formats accept several sequences in one file.

An example sequence in plain format is:

```
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCC
CCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGC
CTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGCCGGGCCCCTCATAGGAGAGG
AAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCGCCGGGACAGAATGCC
CTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGAATGCTCACGCAAG
TTTAATTACAGACCTGAA
```

FASTQ format

A sequence file in FASTQ format can contain several sequences.
FASTQ is a text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores. It is mainly used for storing the output of high-throughput sequencing instruments.
A FASTQ file usually uses four lines per sequence.

1. a '@' character, followed by a sequence identifier and an optional description
2. the raw sequence letters.
3. a '+' character, optionally followed by the same sequence identifier (and any description)
4. quality values for the sequence in Line 2

An example sequence in FASTQ format is:

```
@SEQUENCE_ID
GTGGAAGTTCTTAGGGCATGGCAAAGAGTCAGAATTTGAC
+
FAFFADEDGDBGEGGBCGGHE>EEBA@@=
```

For a detailed decription please see the Wikipedia entry.

EMBL format

A sequence file in EMBL format can contain several sequences.
One sequence entry starts with an identifier line ("ID"), followed by further annotation lines. The start of the sequence is marked by a line starting with "SQ" and the end of the sequence is marked by two slashes ("//").

An example sequence in EMBL format is:

```
ID   AB000263 standard; RNA; PRI; 368 BP.
XX
AC   AB000263;
XX
DE   Homo sapiens mRNA for prepro cortistatin like peptide, complete cds.
XX
SQ   Sequence 368 BP;
     acaagatgcc attgtccccc ggcctcctgc tgctgctgct ctccggggcc acggccaccg        60
     ctgccctgcc cctggagggt ggccccaccg gccgagacag cgagcatatg caggaagcgg       120
     caggaataag gaaaagcagc ctcctgactt tcctcgcttg gtggtttgag tggacctccc       180
     aggccagtgc cgggcccctc ataggagagg aagctcggga ggtggccagg cggcaggaag       240
     gcgcacccca ccagcaatcc gcgcgccggg acagaatgcc ctgcaggaac ttcttctgga       300
     agaccttctc ctcctgcaaa taaaacctca cccatgaatg ctcacgcaag tttaattaca       360
     gacctgaa                                                               368
//
```

FASTA format

A sequence file in FASTA format can contain several sequences.
Each sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line must begin with a greater-than (">") symbol in the first column.

An example sequence in FASTA format is:

```
>AB000263 |acc=AB000263|descr=Homo sapiens mRNA for prepro cortistatin like peptide, complete cds.|len=368
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCC
CCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGC
CTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGCCGGGCCCCTCATAGGAGAGG
AAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCGCCGGGACAGAATGCC
CTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGAATGCTCACGCAAG
TTTAATTACAGACCTGAA
```

GCG format

A sequence file in GCG format contains exactly one sequence, begins with annotation lines and the start of the sequence is marked by a line ending with two dot ("..") characters. This line also contains the sequence identifier, the sequence length and a checksum. This format should only be used if the file was created with the GCG package.

An example sequence in GCG format is:

```
ID   AB000263 standard; RNA; PRI; 368 BP.
XX
AC   AB000263;
XX
DE   Homo sapiens mRNA for prepro cortistatin like peptide, complete cds.
XX
SQ   Sequence 368 BP;
AB000263  Length: 368  Check: 4514  ..
       1  acaagatgcc attgtccccc ggcctcctgc tgctgctgct ctccggggcc acggccaccg
      61  ctgccctgcc cctggagggt ggccccaccg gccgagacag cgagcatatg caggaagcgg
     121  caggaataag gaaaagcagc ctcctgactt tcctcgcttg gtggtttgag tggacctccc
     181  aggccagtgc cgggcccctc ataggagagg aagctcggga ggtggccagg cggcaggaag
     241  gcgcacccca ccagcaatcc gcgcgccggg acagaatgcc ctgcaggaac ttcttctgga
     301  agaccttctc ctcctgcaaa taaaacctca cccatgaatg ctcacgcaag tttaattaca
     361  gacctgaa
```

GCG-RSF (rich sequence format)

The new GCG-RSF can contain several sequences in one file. This format should only be used if the file was created with the GCG package.

GenBank format

A sequence file in GenBank format can contain several sequences.
One sequence in GenBank format starts with a line containing the word LOCUS and a number of annotation lines. The start of the sequence is marked by a line containing "ORIGIN" and the end of the sequence is marked by two slashes ("//").

An example sequence in GenBank format is:

```
LOCUS       AB000263                368 bp    mRNA    linear   PRI 05-FEB-1999
DEFINITION  Homo sapiens mRNA for prepro cortistatin like peptide, complete
            cds.
ACCESSION   AB000263
ORIGIN
        1 acaagatgcc attgtccccc ggcctcctgc tgctgctgct ctccggggcc acggccaccg
       61 ctgccctgcc cctggagggt ggccccaccg gccgagacag cgagcatatg caggaagcgg
      121 caggaataag gaaaagcagc ctcctgactt tcctcgcttg gtggtttgag tggacctccc
      181 aggccagtgc cgggcccctc ataggagagg aagctcggga ggtggccagg cggcaggaag
      241 gcgcacccca ccagcaatcc gcgcgccggg acagaatgcc ctgcaggaac ttcttctgga
      301 agaccttctc ctcctgcaaa taaaacctca cccatgaatg ctcacgcaag tttaattaca
      361 gacctgaa
//
```

IG format

A sequence file in IG format can contain several sequences, each consisting of a number of comment lines that must begin with a semicolon (";"), a line with the sequence name (it may not contain spaces!) and the sequence itself terminated with the termination character '1' for linear or '2' for circular sequences.

An example sequence in IG format is:

```
; comment
; comment
AB000263
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCC
CCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGC
CTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGCCGGGCCCCTCATAGGAGAGG
AAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCGCCGGGACAGAATGCC
CTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGAATGCTCACGCAAG
TTTAATTACAGACCTGAA1
```

Genomatix annotation syntax

Some Genomatix tools, e.g. Gene2Promoter or GPD allow the extraction of sequences. Genomatix uses the following syntax to annotate sequence information: each information item is denoted by a keyword, followed by a "=" and the value. These information items are separated by a pipe symbol "|". The keywords are the following:

loc	The **Genomatix Locus Id**, consisting of the string "GXL_" followed by a number.
sym	The **gene symbol**. This can be a (comma-separated) list.
geneid	The **NCBI Gene Id**. This can be a (comma-separated) list.
acc	A **unique identifier** for the sequence. E.g. for Genomatix promoter regions, the Genomatix Promoter Id is listed in this field.
taxid	The organism's **Taxon Id**
spec	The **organism name**
chr	The **chromosome** within the organism.
ctg	The **NCBI contig** within the chromosome.
str	**Strand**, (+) for sense, (-) for antisense strand.
start	**Start position** of the sequence (relative to the contig).
end	**End position** of the sequence (relative to the contig).
len	**Length** of the sequence in base pairs.
tss	A (comma-separated list of) **UTR-start/TSS position(s)**. If there are several TSS/UTR-starts, this means that several transcripts share the same promoter (e.g. when they are splice variants). The positions are relative to the promoter region.
probe	A (comma-separated list of) **Affymetrix Probe Id(s)**.
unigene	A (comma-separated list of) **UniGene Cluster Id(s)**.
homgroup	An identifier (a number) for the **homology group** (available for promoter sequences only). Orthologously related sequences have the same value in this field.
promset	If the sequence is a promoter region, the **promoter set** is denoted here.
eldorado	The **ElDorado version** from which the sequence has been extracted.
descr	The **gene description**. If several genes (i.e. NCBI gene ids) are associated with the sequence, the descriptions for all of the genes are listed, separated by ";"
comm	A **comment** field, used for additional annotation. For promoter sequences, this field contains information about the transcripts associated with the promoter. For each transcript the Genomatix Transcript Id, accession number, TSS position and quality is listed, separated by "/". For Genomatix CompGen promoters no transcripts are assigned, in this case the string "CompGen promoter" is denoted.

This syntax is currently used only for sequences in the FASTA and GenBank formats.

Example (a promoter sequence in GenBank format):

```
LOCUS       GXP_4405072(PAX6/human)    1105 bp    DNA
DEFINITION  loc=GXL_141121|sym=PAX6|geneid=5080|acc=GXP_4405072|
            taxid=9606|spec=Homo sapiens|chr=11|ctg=NC_000011|str=(-)|
            start=31806821|end=31807925|len=1105|tss=1001,1005|
            homgroup=-|promset=-|eldorado=E32R1605|descr=paired box 6|
            comm=GXT_25635656/ENST00000455099/1005/gold;
            GXT_27757207/NM_001310159/1001/bronze
ACCESSION   GXP_4405072
BASE COUNT    229 a  239 c  313 g  324 t
ORIGIN
        1 GACTTTTTTT TTTTTTCCTT TGGGAAAGGT AGGGAGGTGT TCGTACGGGA GCAGCCTCGG
       61 GGACCCCTGC ACTGGGTCAG GGCTTATGAA GCTAGAAGCG TCCCTCTGTT CCCTTTGTGA
      121 GTTGGTGGGT TGTTGGTACA TTTGGTTGGA AGCTGTGTTG CTGGTTAGGG AGACTCGGTT
      181 TTGCTCCTTG GGTTCGAGGA AAGCTGGAGA ATAGAAGCCA TTGTTTGCCG TCTGTCGGCT
      241 TTGTCGACCA CGCTCACCCC CTCCTGTTCG TACTTTTTAA AGCAGTGAGG CGAGGTAGAC
      301 AGGGTGTGTC ACAGTACAGT TAAAGGGGTG AAGATCTAAA CGCCAAAAGA GAAGTTAATC
      361 ACAATAAGTG AGGTTTGGGA TAAAAAGTTG GGCTTGCCCC TTTCAAAGTC CCAGAAAGCT
      421 GGGAGGTAGA TGGAGAGGGG GCCATTGGGA AGTTTTTTTG GTGTAGGGAG AGGAGTAGAA
      481 GATAAAGGGT AAGCAGAGTG TTGGGTTCTG GGGGTCTTGT GAAGTTCCTT AAGGAAGGAG
      541 GGAGTGTGGC CCTGCAGCCC TCCCAAACTG CTCCAGCCTA TGCTCTCCGG CACCAGGAAG
      601 TTCCAAGGTT CCCTTCCCCT GGTCTCCAAA CTTCAGGTAT TCCTCTCCCC TCACACCCCT
      661 TCAACCTCAG CTCTTGGCCT CTACTCCTTA CTCCACTGTT CCTCCTGTTT CCCCCTTCCC
      721 CTTTTCCTGG TTCTTTATAT TTTTGCAAAG TGGGATCCGA ACTTGCTAGA TTTTCCAATT
      781 CTCCCAAGCC AGACCAGAGC AGCCTCTTTT AAAGGATGGA GACTTCTGTG GCAGATGCCG
      841 CTGAAAATGT GGGTGTAATG CTGGGACTTA GAGTTTGATG ACAGTTTGAC TGAGCCCTAG
      901 ATGCATGTGT TTTTCCTGAG AGTGAGGCTC AGAGAGCCCA TGGACGTATG CTGTTGAACC
```

```
      961 ACAGCTTGAT ATACCTTTTT CTCCTTCTGT TTTGTCTTAG GGGGAAGACT TTAACTAGGG
     1021 GCGCGCAGAT GTGTGAGGCC TTTTATTGTG AGAGTGGACA GACATCCGAG ATTTCAGGCA
     1081 AGTTCTGTGG TGGCTGCTTT GGGCT
//
```

IUPAC nucleic acid codes

To represent ambiguity in DNA sequences the following letters can be used (following the rules of the *International Union of Pure and Applied Chemistry* (IUPAC)):

```
A = adenine
C = cytosine
G = guanine
T = thymine
U = uracil
R = G A (purine)
Y = T C (pyrimidine)
K = G T (keto)
M = A C (amino)
S = G C
W = A T
B = G T C
D = G A T
H = A C T
V = G C A
N = A G C T (any)
```