**DNA Sequence formats**

[Plain sequence format 1](#_Toc314165149)

[EMBL format 2](#_Toc314165150)

[FASTA format 3](#_Toc314165151)

[GCG format 4](#_Toc314165152)

[GCG-RSF (rich sequence format) 5](#_Toc314165153)

[GenBank format 6](#_Toc314165154)

[IG format 7](#_Toc314165155)

[Genomatix annotation syntax 8](#_Toc314165156)

[IUPAC nucleic acid codes 10](#_Toc314165157)

[mtDNA Codons 11](#_Toc314165158)

[Standard Amino Acids 12](#_Toc314165159)

# Plain sequence format

A sequence in plain format may contain only [IUPAC characters](http://www.genomatix.de/online_help/help/sequence_formats.html#IUPAC) 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**

# 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**

**gcgcaccccc 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 gcgcaccccc 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 gcgcaccccc 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](http://www.genomatix.de/online_help/help_eldorado/Gene2Promoter_Intro.html) or [GPD](http://www.genomatix.de/online_help/help_gpd/gpd_help.html) 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 basepairs. |
| **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. |
| **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 note, 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](http://www.genomatix.de/online_help/help_gems/FAQ_answers.html#Prom_Extraction5) is listed, separated by "/". For [Genomatix CompGen promoters](http://www.genomatix.de/online_help/help_eldorado/comparative_genomics_help.html) no transcripts are assigned, in this case the string "CompGen promoter" is denoted. |

This syntax is currently used only for sequences in the [FASTA](http://www.genomatix.de/online_help/help/sequence_formats.html#FASTA) and [GenBank](http://www.genomatix.de/online_help/help/sequence_formats.html" \l "GB) formats.

**Example (a promoter sequence in GenBank format):**

**LOCUS GXP\_170357 743 bp DNA**

**DEFINITION loc=GXL\_141619|sym=TPH2|geneid=121278|acc=GXP\_170357|**

**taxid=9606|spec=Homo sapiens|chr=12|ctg=NC\_000012|str=(+)|**

**start=70618393|end=70619135|len=743|tss=501,632|**

**homgroup=4612|promset=1|descr=tryptophan hydroxylase 2|**

**comm=GXT\_2756574/AK094614/632/gold;**

**GXT\_2799672/NM\_173353/501/bronze**

**ACCESSION GXP\_170357**

**BASE COUNT 216 a 180 c 147 g 200 t**

**ORIGIN**

**1 TTGATTACCT TATTTGATCA TTACACATTG TACGCTTGTG TCAAAATATC ACATGTGCCT**

**61 TATAAATGTG TACAACTATT AGTTATCCAT AAAAATTAAA AATTAAAAAA TCCGTAAAAT**

**121 GGTTTAAGCA TTCAGCAGTG CTGATCTTTC TTAAATTATT TTTCTAATTT TGGAAAGAAA**

**181 GCACAAAATC TTTGAATTCA CAATTGCTTA AAGACTGAGG TTAACTTGCC AGTGGCAGGC**

**241 TTGAGAGATG AGAGAACTAA CGTCAGAGGA TAGATGGTTT CTTGTACAAA TAACACCCCC**

**301 TTATGTATTG TTCTCCACCA CCCCCGCCCA AAAAGCTACT CGACCTATGA AACAAATCAC**

**361 ACTATGAGCA CAGATAACCC CAGGCTTCAG GTCTGTAATC TGACTGTGGC CATCGGCAAC**

**421 CAGAAATGAG TTTCTTTCTA ATCAGTCTTG CATCAGTCTC CAGTCATTCA TATAAAGGAG**

**481 CCCGGGGATG GGAGGATTCG CATTGCTCTT CAGCACCAGG GTTCTGGACA GCGCCCCAAG**

**541 CAGGCAGCTG ATCGCACGCC CCTTCCTCTC AATCTCCGCC AGCGCTGCTA CTGCCCCTCT**

**601 AGTACCCCCT GCTGCAGAGA AAGAATATTA CACCGGGATC CATGCAGCCA GCAATGATGA**

**661 TGTTTTCCAG TAAATACTGG GCACGGAGAG GGTTTTCCCT GGATTCAGCA GTGCCCGAAG**

**721 AGCATCAGCT ACTTGGCAGC TCA**

**//**

# 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)**

# mtDNA Codons



# Standard Amino Acids

| **Standard Amino Acids** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Amino Acid** | **3-Letter** | **1-Letter** | **Side-chain polarity** | **Side-chain charge (pH 7.4)** | [**Hydropathy index**](http://en.wikipedia.org/wiki/Hydropathy_index) | [**Absorbance**](http://en.wikipedia.org/wiki/Absorbance)**λmax(nm)** | **ε at λmax (x10−3 M−1 cm−1)** |
| [Alanine](http://en.wikipedia.org/wiki/Alanine) | ALA | A | nonpolar | neutral | 1.8 |  |  |
| [Arginine](http://en.wikipedia.org/wiki/Arginine) | ARG | R | polar | positive | −4.5 |  |  |
| [Asparagine](http://en.wikipedia.org/wiki/Asparagine) | ASN | N | polar | neutral | −3.5 |  |  |
| [Aspartic acid](http://en.wikipedia.org/wiki/Aspartic_acid) | ASP | D | polar | negative | −3.5 |  |  |
| [Cysteine](http://en.wikipedia.org/wiki/Cysteine) | CYS | C | polar | neutral | 2.5 | 250 | 0.3 |
| [Glutamic acid](http://en.wikipedia.org/wiki/Glutamic_acid) | GLU | E | polar | negative | −3.5 |  |  |
| [Glutamine](http://en.wikipedia.org/wiki/Glutamine) | GLN | Q | polar | neutral | −3.5 |  |  |
| [Glycine](http://en.wikipedia.org/wiki/Glycine) | GLY | G | nonpolar | neutral | −0.4 |  |  |
| [Histidine](http://en.wikipedia.org/wiki/Histidine) | HIS | H | polar | Positive (10%), Neutral (90%) | −3.2 | 211 | 5.9 |
| [Isoleucine](http://en.wikipedia.org/wiki/Isoleucine) | ILE | I | nonpolar | neutral | 4.5 |  |  |
| [Leucine](http://en.wikipedia.org/wiki/Leucine) | LEU | L | nonpolar | neutral | 3.8 |  |  |
| [Lysine](http://en.wikipedia.org/wiki/Lysine) | LYS | K | polar | positive | −3.9 |  |  |
| [Methionine](http://en.wikipedia.org/wiki/Methionine) | MET | M | nonpolar | neutral | 1.9 |  |  |
| [Phenylalanine](http://en.wikipedia.org/wiki/Phenylalanine) | PHE | F | nonpolar | neutral | 2.8 | 257, 206, 188 | 0.2, 9.3, 60.0 |
| [Proline](http://en.wikipedia.org/wiki/Proline) | PRO | P | nonpolar | neutral | −1.6 |  |  |
| [Serine](http://en.wikipedia.org/wiki/Serine) | SER | S | polar | neutral | −0.8 |  |  |
| [Threonine](http://en.wikipedia.org/wiki/Threonine) | THR | T | polar | neutral | −0.7 |  |  |
| [Tryptophan](http://en.wikipedia.org/wiki/Tryptophan) | TRP | W | nonpolar | neutral | −0.9 | 280, 219 | 5.6, 47.0 |
| [Tyrosine](http://en.wikipedia.org/wiki/Tyrosine) | TYR | Y | polar | neutral | −1.3 | 274, 222, 193 | 1.4, 8.0, 48.0 |
| [Valine](http://en.wikipedia.org/wiki/Valine) | VAL | V | nonpolar | neutral | 4.2 |  |  |