

GenBank format

```
LOCUS          CMGLOAD                      1185 bp    DNA        linear    VRT 18-APR-2005
DEFINITION    Cairina moschata (duck) gene for alpha-D globin.
ACCESSION     X01831
VERSION       X01831.1  GI:62724
KEYWORDS      alpha-globin; globin.
SOURCE        Cairina moschata (Muscovy duck)
              ORGANISM Cairina moschata
              Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
              Archosauria; Aves; Neognathae; Anseriformes; Anatidae; Cairina.
REFERENCE     1 (bases 1 to 1185)
AUTHORS       Erbil,C. and Niessing,J.
TITLE         The primary structure of the duck alpha D-globin gene: an unusual
              5' splice junction sequence
JOURNAL       EMBO J. 2 (8), 1339-1343 (1983)
PUBMED        10872328
COMMENT       Data kindly reviewed (13-NOV-1985) by J. Niessing.
```

Header

Contains unique accession ID, description, organism (taxonomy), reference publication etc. (Information that is true for ALL genes in this entry).

```
FEATURES             Location/Qualifiers
     source           1..1185
                     /organism="Cairina moschata"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:8855"
     CAAT_signal      20..24
     TATA_signal      69..73
     precursor_RNA    101..1114
                     /note="primary transcript"
     exon             101..234
                     /number=1
     CDS              join(143..234,387..591,939..1067)
                     /codon_start=1
                     /product="alpha D-globin"
                     /protein_id="CAA25966.2"
                     /db_xref="GI:4455876"
                     /db_xref="GOA:P02003"
                     /db_xref="InterPro:IPR000971"
                     /db_xref="InterPro:IPR002338"
                     /db_xref="InterPro:IPR002340"
                     /db_xref="InterPro:IPR009050"
                     /db_xref="UniProt/Swiss-Prot:P02003"
                     /translation="MLTAEDKKLIVQVWEKVAGHQEEFGSEALQRMFLAYPQTKTYFP
HFDLHPGSEQVRGHGKKVAAALGNVAVKSLDNLSQLSELSNLHAYNLRVDPVNFKLLA
QCQFQVLAHLGKDYSPEMHAAFDKFLSAVAAVLAEKYR"
     repeat_region    227..246
                     /note="direct repeat 1"
     intron           235..386
                     /number=1
     repeat_region    289..309
                     /note="direct repeat 1"
     exon             387..591
                     /number=2
     intron           592..939
                     /number=2
     exon             940..1114
                     /number=3
     polyA_signal     1095..1100
     polyA_signal     1114
```

FEATURE block

Contains a description + coordinates of individual features (e.g. genes, promoters etc.) in the DNA sequence below.

CDS: Coding sequence. The protein-coding part of a gene – notice the three intervals in the example to the left. This corresponds to three exons + two introns (in the coding part).

```
ORIGIN
1   ctgcgtggcc tcagcccctc caccctcca cgctgataag ataaggccag ggcgggagcg
61  caggggtgcta taagagctcg gccccgcggg tgtctccacc acagaaacct gtcagttgcc
121 agcctgccac gccgctgccc ccatgtgac cgccgaggac aagaagctca tcgtgcaggt
181 gtgggagaag gtggctggcc accaggagga attcggaggt gaagctctgc agaggtgtgg
241 gctgggcccc gggggcactc acaggggtgg cagcagggag caggagccct gcagcgggtg
301 tgggctggga ccagagcgc cacggggtgc gggctgagat gggcaaagca gcagggcacc
361 aaaactgact ggcctcgctc cggcaggatg ttcctcgctc acccccagac caagacctac
421 ttccccactc tcgacctgca tcccggctct gaacaggtcc gtggccatgg caagaagtgtg
481 gcggctgccc tgggcaatgc cgtgaagagc ctggacaacc tcagccaggc cctgtctgag
541 ctcagcaacc tgcatgccta caacctgcgt gttgacctg tcaacttcaa ggcaagcggg
601 gactagggtc cttgggtctg ggggtctgag ggtgtgggt gcagggtctg ggggtccagg
661 ggtctgagtt tcctggggtc tggcagtcct gggggctgag ggccagggtc ctgtgtcttt
721 ggttaccagg gtcctggggg ccagcagcca gacagcagg gctgggattg catctgggat
781 gtgggccaaga ggcctggatt gtgtttgtaa tgggagctgg gcaggggcta gggccaagggt
841 gggggactca gggcctcagg gggactcggg gggggactga gggagactca gggccatctg
901 tccggagcag ggtactaag cctgtgtttg ccttgacgt gctggcacag tgcttccagg
961 tgggtgctggc cgcacacctg gycaaaagact acagcccga gatgcatgct gcctttgaca
1021 agttcttctg cgccgtggtt gccgtgctgg ctgaaaagta cagatgagcc actgcctgca
1081 cccttgcacc ttcaataaag acaccattac cacagctctg tgtctgtgtg tgctgggact
1141 gggcatcggg ggtcccaggg agggctgggt tgcttcaca catcc
```

ORIGIN block

Contains the actual DNA sequence.

//

FASTA format

Header

Each sequence entry in a FASTA file starts with a header line, that MUST begin with ">". The first word after the ">" sign is the name of the sequence. The format allows for additional fields (e.g. comments) to be present after the name, for example:

```
>alpha-D Duck_alpha_globin
```

Important: Many pieces of bioinformatics software (also on the web) only reads the actual name (the first token) and ignore the rest. Furthermore, the name should be kept short and unique.

Sequence

The following lines are simply sequence (DNA, RNA, Protein) and ends when the next header is reached or the file ends.

>alpha-D

```
ATGCTGACCGACTCTGACAAGAAGCTGGTCCTGCAGGTGTGGGAGAAGGTGATCCGCCAC
CCAGACTGTGGAGCCGAGGCCCTGGAGAGGTGCGGGCTGAGCTTGGGGAAACCATGGGCA
AGGGGGGCGACTGGGTGGGAGCCCTACAGGGCTGCTGGGGGTGTTTCGGCTGGGGGTCAG
CACTGACCATCCCGCTCCCGCAGCTGTTACACCTACCCCCAGACCAAGACCTACTTCC
CCCCTTCGACTTGCACCATGGCTCCGACCAGGTCCGCAACCACGGCAAGAAGGTGTTGG
CCGCCTTGGGCAACGCTGTCAAGAGCCTGGGCAACCTCAGCCAAGCCCTGTCTGACCTCA
GCGACCTGCATGCCTACAACCTGCGTGTGACCCTGTCAACTTCAAGGCAGGCGGGGGAC
GGGGGTGAGGGGCCGGGGAGTTGGGGGCCAGGGACCTGGTTGGGGATCCGGGGCCATGCC
GGCGGTACTGAGCCCTGTTTTGCCTTGCAGCTGCTGGCGCAGTGCTTCCACGTGGTGCTG
GCCACACACCTGGGCAACGACTACCCCCGGAGGCACATGCTGCCTTCGACAAGTTCCTG
TCGGCTGTGTGCACCGTGCTGGCCGAGAAGTACAGATAA
```

>alpha-A

```
ATGGTGCTGTCTGCCAACGACAAGAGCAACGTGAAGGCCGTCTTCGGCAAAATCGGCGGC
CAGGCCGGTGACTTGGGTGGTGAAGCCCTGGAGAGGTATGTGGTCATCCGTCATTACCCC
ATCTCTTGTCTGTCTGTGACTCCATCCCATCTGCCCCATACTCTCCCATCCATAACTG
TCCCTGTTCTATGTGGCCCTGGCTCTGTCTCATCTGTCCCCAACTGTCCCTGATTGCCTC
TGTCCCCAGGTTGTTTCATCACCTACCCCCAGACCAAGACCTACTTCCCCACTTCGACC
TGTCACATGGCTCCGCTCAGATCAAGGGGCACGGCAAGAAGGTGGCGGAGGCACTGGTTG
AGGCTGCCAACACATCGATGACATCGCTGGTGCCCTCTCCAAGCTGAGCGACCTCCACG
CCCAAAGCTCCGTGTGGACCCCGTCAACTTCAAAGTGAGCATCTGGGAAGGGGTGACCA
GTCTGGCTCCCCTCCTGCACACACCTCTGGCTACCCCTCACCTCACCCCTTGCTCACC
ATCTCCTTTTGCCTTTCAGCTGCTGGGTCACTGCTTCCCTGGTGGTCGTGGCCGTCCACTT
CCCCTCTCTCCTGACCCCGGAGGTCCATGCTTCCCTGGACAAGTTCGTGTGTGCCGTGGG
CACCGTCCTTACTGCCAAGTACCGTTAA
```