Introduction to Bio-Informatics
Lecture 5

Lecture by: Ahmad R. Naghsh-Nilchi, PhD
Department of Artificial Intelligence
Faculty of Computer Engineering
University of Isfahan

In the last session, we discussed:

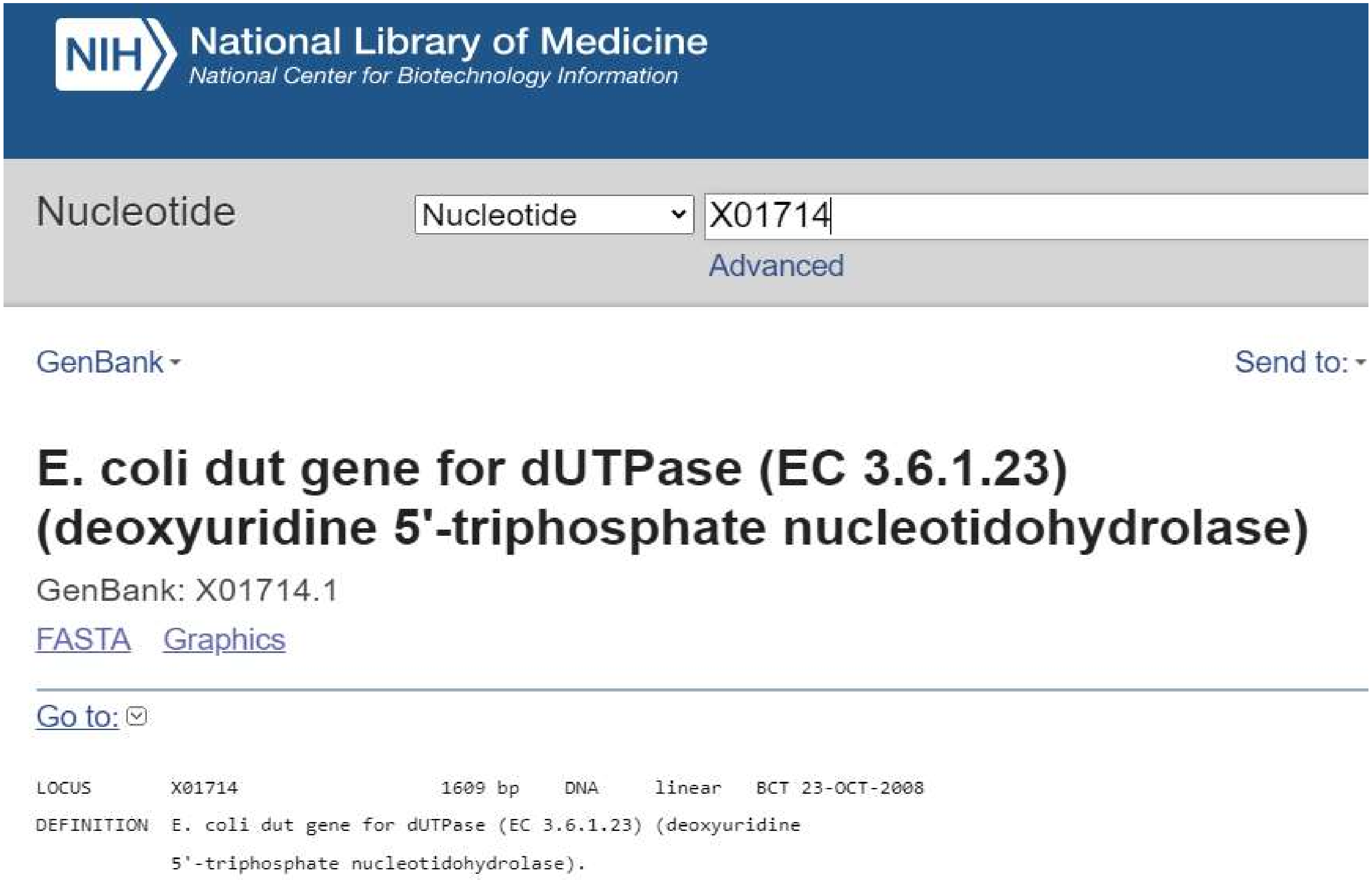
- Genes of higher Eukaryotes (animals) may span up of millions of base pairs
  - For example the human dystrophin gene (its mutation causes a dreadful disease), is 2.2 million base pairs long.
- The relationship among a gene DNA sequence, its primary transcript, the various forms of mature mRNA, and the final protein sequence can be very complex.
- A lot of database entries corresponding to partial gene sequences, or different gene-related objects:
  - such as promoter regions, mRNA, or genome fragments,
- These make their studies challenging!

Prokaryotes in GenBank

- For Prokaryotes, the limit size of the genes involved – as well as the simple (linear) relationship among the gene DNA sequence, the mRNA, the ORFs, and the final protein sequence – make all that information relatively easy to annotate and store in database records,
- That is why database entries corresponding to bacterial genes are relatively easy to read and understand.

GenBank entry for a Prokaryotic Gene

- Entry of the Escherichia coli dUTPase gene with GenBank ID of X01714.


NIH
National Library of Medicine
National Center for Biotechnology Information
Nucleotide
Nucleotide
X01714
Advanced
GenBank
Send to:

E. coli dut gene for dUTPase (EC 3.6.1.23) (deoxyuridine 5'-triphosphate nucleotidohydrolase)

GenBank: X01714.1

FASTA Graphics

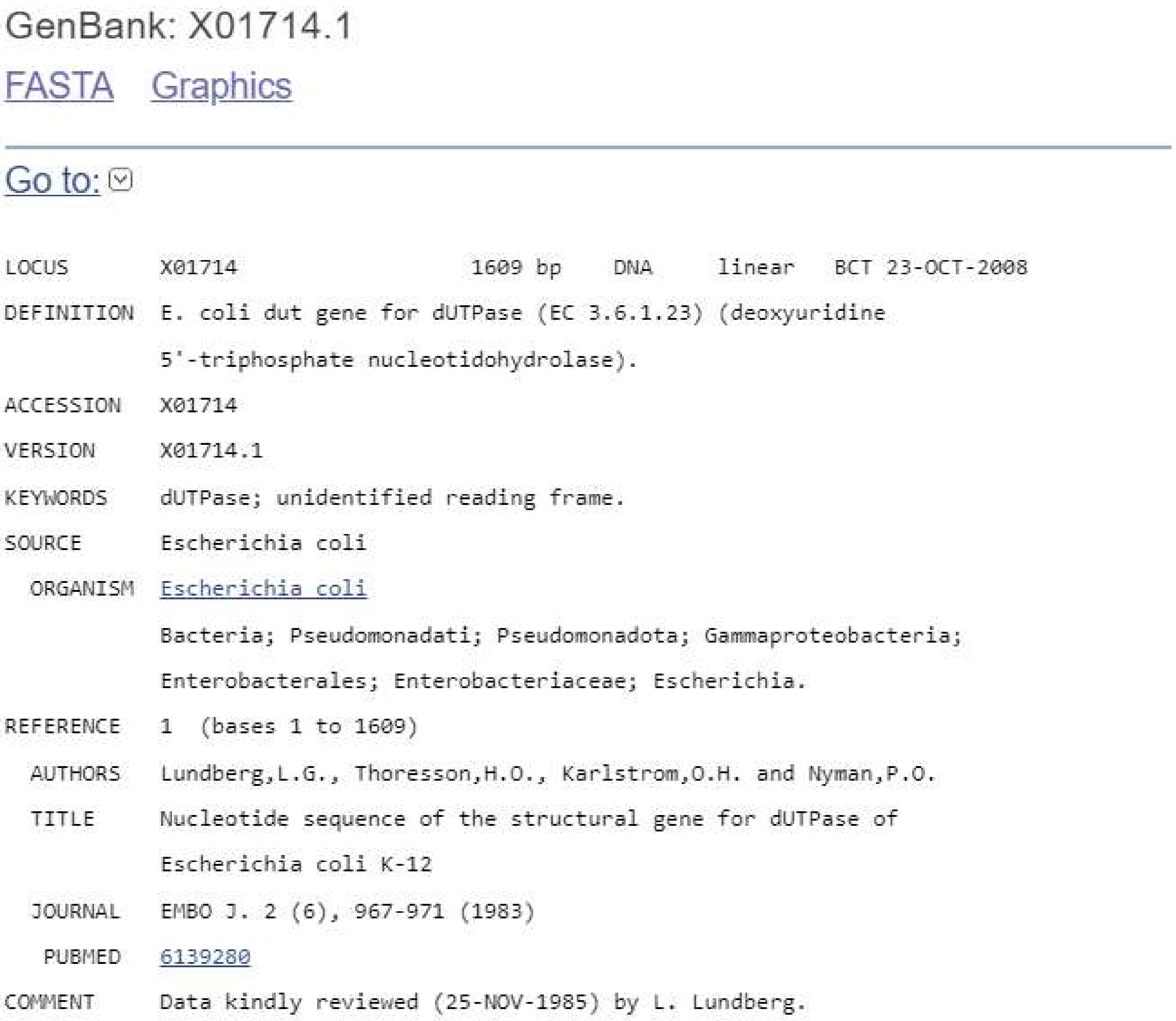
Go to:

```
LOCUS       X01714                  1609 bp    DNA     linear   BCT 23-OCT-2008
DEFINITION  E. coli dut gene for dUTPase (EC 3.6.1.23) (deoxyuridine
            5'-triphosphate nucleotidohydrolase).
```

Reading the GenBank header of a Prokaryotic entry

- **Typical keywords include:**
- **Locus:** Gives us the locus name, the size of nucleotide sequence, in base pairs, the nature of the molecule (here it is DNA), and its topology (linear or circular)
- **Definition:** A short definition of the gene corresponding to the entry sequence. Here it is the *E coli dUTPase* gene.
- **Accession:** A unique identifier within and across various databases. In this case, it is X01714.
- **Version:** Synonymous or past ID numbers

- **Keywords:** List terms that broadly characterize the entry as useful for certain database searches.
- **Source:** The common name of the relevant organism to which the sequence belongs.
- **Organism:** gives more complete identification of the organism, complete with its taxonomic classification
- **Reference:** give the credits for the sequence determination (Author, title, journal, Pubmed ID)
- **Comment:** Free-formatted text, such as acknowledgment or inf, does not fit to the other sections.



GenBank: X01714.1

FASTA Graphics

Go to: ☐

```
LOCUS       X01714                  1609 bp    DNA     linear   BCT 23-OCT-2008
DEFINITION  E. coli dut gene for dUTPase (EC 3.6.1.23) (deoxyuridine
            5'-triphosphate nucleotidohydrolase).
ACCESSION   X01714
VERSION     X01714.1
KEYWORDS    dUTPase; unidentified reading frame.
SOURCE      Escherichia coli
  ORGANISM  Escherichia coli
            Bacteria; Pseudomonadati; Pseudomonadota; Gammaproteobacteria;
            Enterobacterales; Enterobacteriaceae; Escherichia.
REFERENCE   1  (bases 1 to 1609)
  AUTHORS   Lundberg,L.G., Thoresson,H.O., Karlstrom,O.H. and Nyman,P.O.
  TITLE     Nucleotide sequence of the structural gene for dUTPase of
            Escherichia coli K-12
  JOURNAL   EMBO J. 2 (6), 967-971 (1983)
   PUBMED   6139280
COMMENT     Data kindly reviewed (25-NOV-1985) by L. Lundberg.
```

Reading the GenBank Middle Page of a Prokaryotic entry

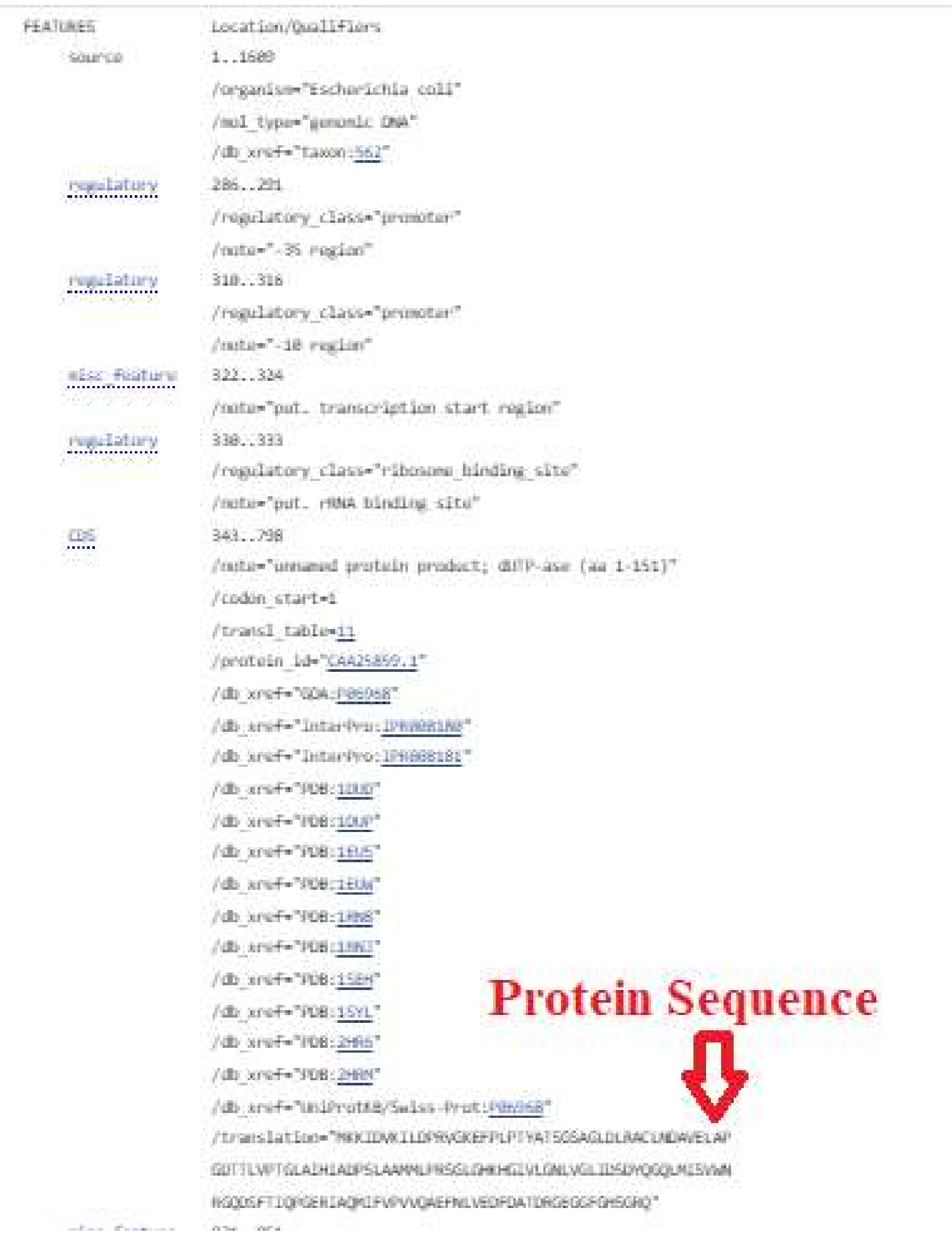
- In the middle section of the page Features Table of the entry is given.
- Describes precisely the gene regions,
- Describes the associated biological properties that have been identified in the nucleotide sequence.
- A large variety of biologically related keywords subordinated to features, in the next slide’s list

- **Sources:** Indicates the origin of specific regions of the sequence. This is useful when you want to distinguish cloning vectors from host sequences. In this case, the whole sequence comes from E *coli genomic DNA*
- **Promoter:** Shows the precise coordinates of a promoter element. In this case, a-35 region is indicated from position 286 to 291 in the nucleotide sequence,
- **Promoter:** introduces another line containing, a-10 region at position 310-316.
- **Misc feature** (miscellaneous feature): indicates the putative location of the transcription start (mRNA synthesis). In this case, this is from position 322 to 324.

- **RBS** (Ribosome Binding Site): indicates the location of the last upstream element. (Here at 330 to 333)
- **CDS** (CoDing segment): introduce a complex section that describes the gene’s open reading frame (ORF):
  - The first line indicates the coordinates of the ORF from the initial ATG to the last nucleotide of the first stop codon TAA. (Here: 343-798).
  - Each following line gives the name of a protein product, indicates the reading frame to use (here, 343 is the first base of the first codon), the genetic code to apply, and several IDs for protein sequence.
  - */translation.* The final keyword of the CDS section introduces the conceptual amino-acid sequence of the coding segment. A computer translation that uses the coordinates reading frame.
- Another **misc feature** follows the CDS section. It contains lines that point out recognized stem-loop structures and repeats (potential regulatory elements of the entry).

The Feature Section of X01714

- It is typical of a simple, well-annotated bacterial nucleotide sequence, centered on a well-identified gene.
- It, however, includes a complication: a known additional reading frame! Indicated by an additional RBS element and a second CDS section.
- This section is depicted in the next slide.



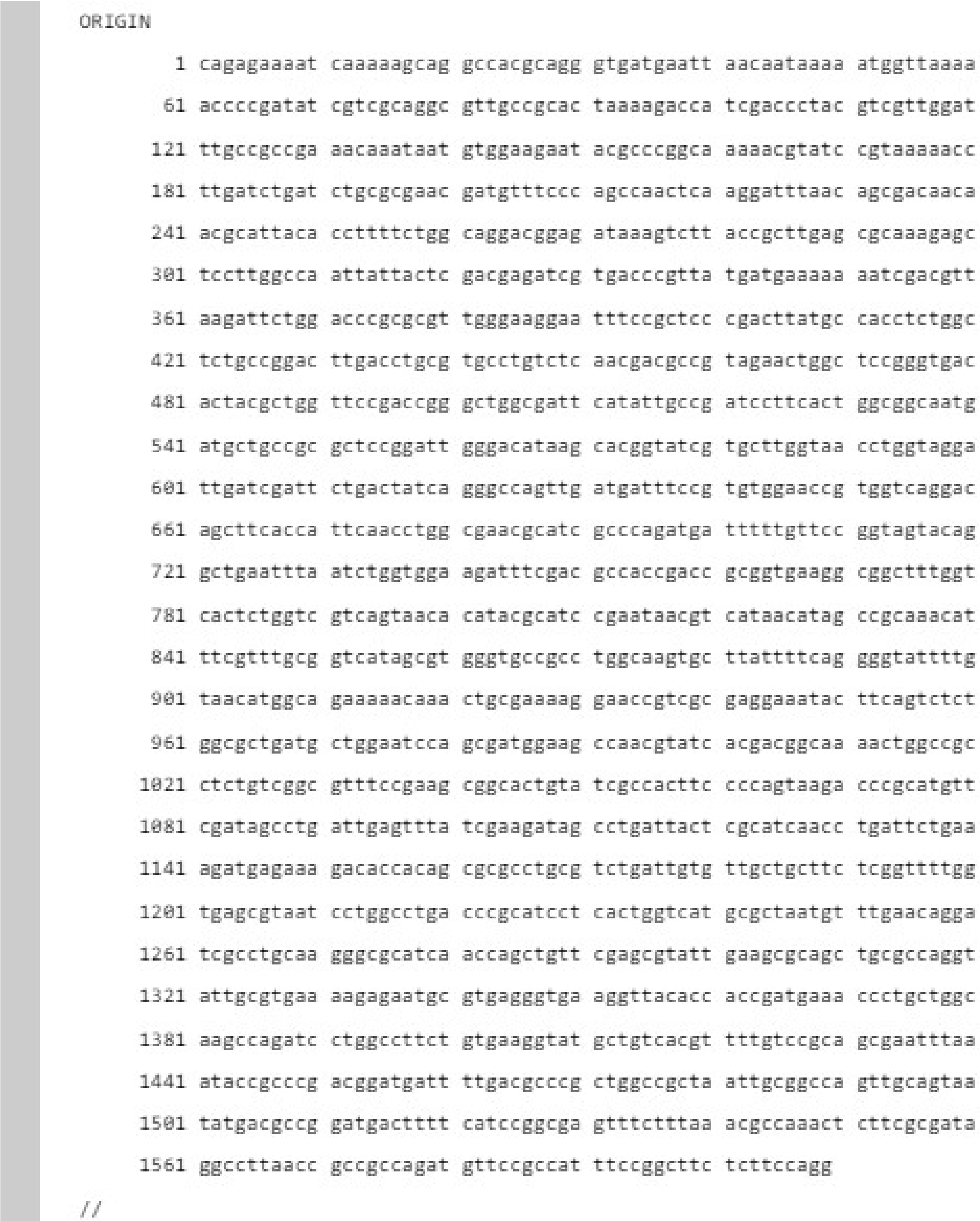
```
FEATURES             Location/Qualifiers
     source          1..1609
                     /organism="Escherichia coli"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:562"
     regulatory      286..291
                     /regulatory_class="promoter"
                     /note="-35 region"
     regulatory      310..316
                     /regulatory_class="promoter"
                     /note="-10 region"
     misc_feature    322..324
                     /note="put. transcription start region"
     regulatory      330..333
                     /regulatory_class="ribosome_binding_site"
                     /note="put. rRNA binding site"
     CDS             343..798
                     /note="unnamed protein product; dUTP-ase (aa 1-151)"
                     /codon_start=1
                     /transl_table=11
                     /protein_id="CAA25859.1"
                     /db_xref="GOA:P06968"
                     /db_xref="InterPro:IPR008180"
                     /db_xref="InterPro:IPR008181"
                     /db_xref="PDB:1DUD"
                     /db_xref="PDB:1DUP"
                     /db_xref="PDB:1EU5"
                     /db_xref="PDB:1EUW"
                     /db_xref="PDB:1RN8"
                     /db_xref="PDB:1RNJ"
                     /db_xref="PDB:1SEH"
                     /db_xref="PDB:1SYL"
                     /db_xref="PDB:2HR6"
                     /db_xref="PDB:2HRM"
                     /db_xref="UniProtKB/Swiss-Prot:P06968"
                     /translation="MKKIDVKILDPRVGKEFPLPTYATSGSAGLDLRACLNDAVELAP
                     GDTTLVPTGLAIHIADPSLAAMMLPRSGLGHKHGIVLGNLVGLIDSDYQGQLMISVWN
                     RGQDSFTIQPGERIAQMIFVPVVQAEFNLVEDFDATDRGEGGFGHSGRQ"
     misc_feature    826..851
```

Protein Sequence

⇩

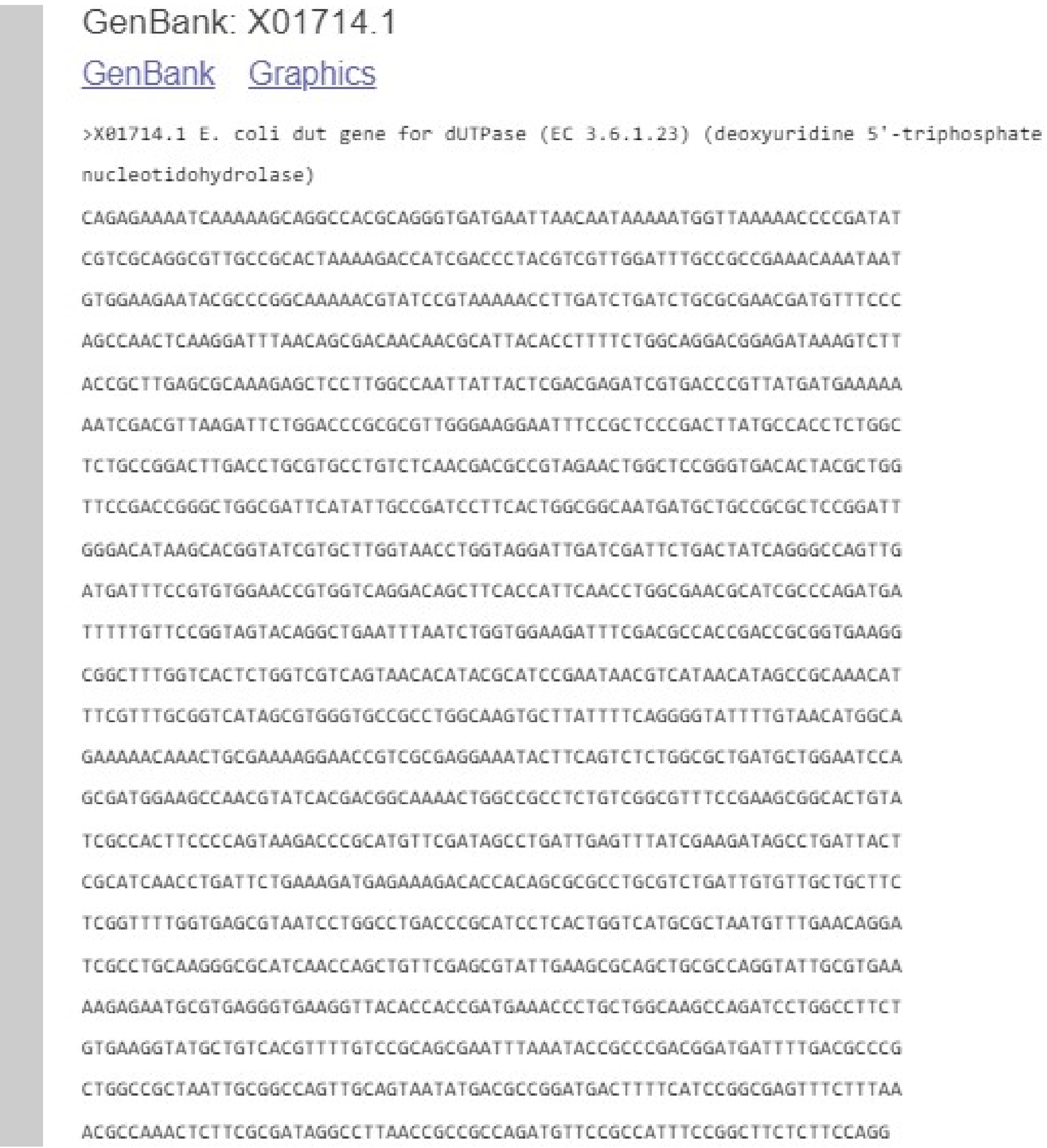
Sequence Section of Prokaryotic Entry

- The last section of GenBank entry X01714 is the nucleotide sequence section.
- It starts with the ORIGIN keyword and finishes with (//).
- Each line contains 60 nucleotides,
- Since it mixes numbers and nucleotides, you cannot directly use it as an input on most sequence analysis servers,
- Thus, you should prepare a FASTA-formatted sequence.
- To generate that, click on FASTA at the top of the page!



```
ORIGIN
        1 cagagaaaat caaaaagcag gccacgcagg gtgatgaatt aacaataaaa atggttaaaa
       61 accccgatat cgtcgcaggc gttgccgcac taaaagacca tcgaccctac gtcgttggat
      121 ttgccgccga aacaaataat gtggaagaat acgcccggca aaaacgtatc cgtaaaaacc
      181 ttgatctgat ctgcgcgaac gatgtttccc agccaactca aggatttaac agcgacaaca
      241 acgcattaca ccttttctgg caggacggag ataaagtctt accgcttgag cgcaaagagc
      301 tccttggcca attattactc gacgagatcg tgacccgtta tgatgaaaaa aatcgacgtt
      361 aagattctgg acccgcgcgt tgggaaggaa tttccgctcc cgacttatgc cacctctggc
      421 tctgccggac ttgacctgcg tgcctgtctc aacgacgccg tagaactggc tccgggtgac
      481 actacgctgg ttccgaccgg gctggcgatt catattgccg atccttcact ggcggcaatg
      541 atgctgccgc gctccggatt gggacataag cacggtatcg tgcttggtaa cctggtagga
      601 ttgatcgatt ctgactatca gggccagttg atgatttccg tgtggaaccg tggtcaggac
      661 agcttcacca ttcaacctgg cgaacgcatc gcccagatga tttttgttcc ggtagtacag
      721 gctgaattta atctggtgga agatttcgac gccaccgacc gcggtgaagg cggctttggt
      781 cactctggtc gtcagtaaca catacgcatc cgaataacgt cataacatag ccgcaaacat
      841 ttcgtttgcg gtcatagcgt gggtgccgcc tggcaagtgc ttattttcag ggtattttg
      901 taacatggca gaaaaacaaa ctgcgaaaag gaaccgtcgc gaggaaatac ttcagtctct
      961 ggcgctgatg ctggaatcca gcgatggaag ccaacgtatc acgacggcaa aactggccgc
     1021 ctctgtcggc gtttccgaag cggcactgta tcgccacttc cccagtaaga cccgcatgtt
     1081 cgatagcctg attgagttta tcgaagatag cctgattact cgcatcaacc tgattctgaa
     1141 agatgagaaa gacaccacag cgcgcctgcg tctgattgtg ttgctgcttc tcggttttgg
     1201 tgagcgtaat cctggcctga cccgcatcct cactggtcat gcgctaatgt ttgaacagga
     1261 tcgcctgcaa gggcgcatca accagctgtt cgagcgtatt gaagcgcagc tgcgccaggt
     1321 attgcgtgaa aagagaatgc gtgagggtga aggttacacc accgatgaaa ccctgctggc
     1381 aagccagatc ctggccttct gtgaaggtat gctgtcacgt tttgtccgca gcgaatttaa
     1441 ataccgcccg acggatgatt ttgacgcccg ctggccgcta attgcggcca gttgcagtaa
     1501 tatgacgccg gatgactttt catccggcga gtttctttaa acgccaaact cttcgcgata
     1561 ggccttaacc gccgccagat gttccgccat ttccggcttc tcttccagg
//
```

FASTA Formated:



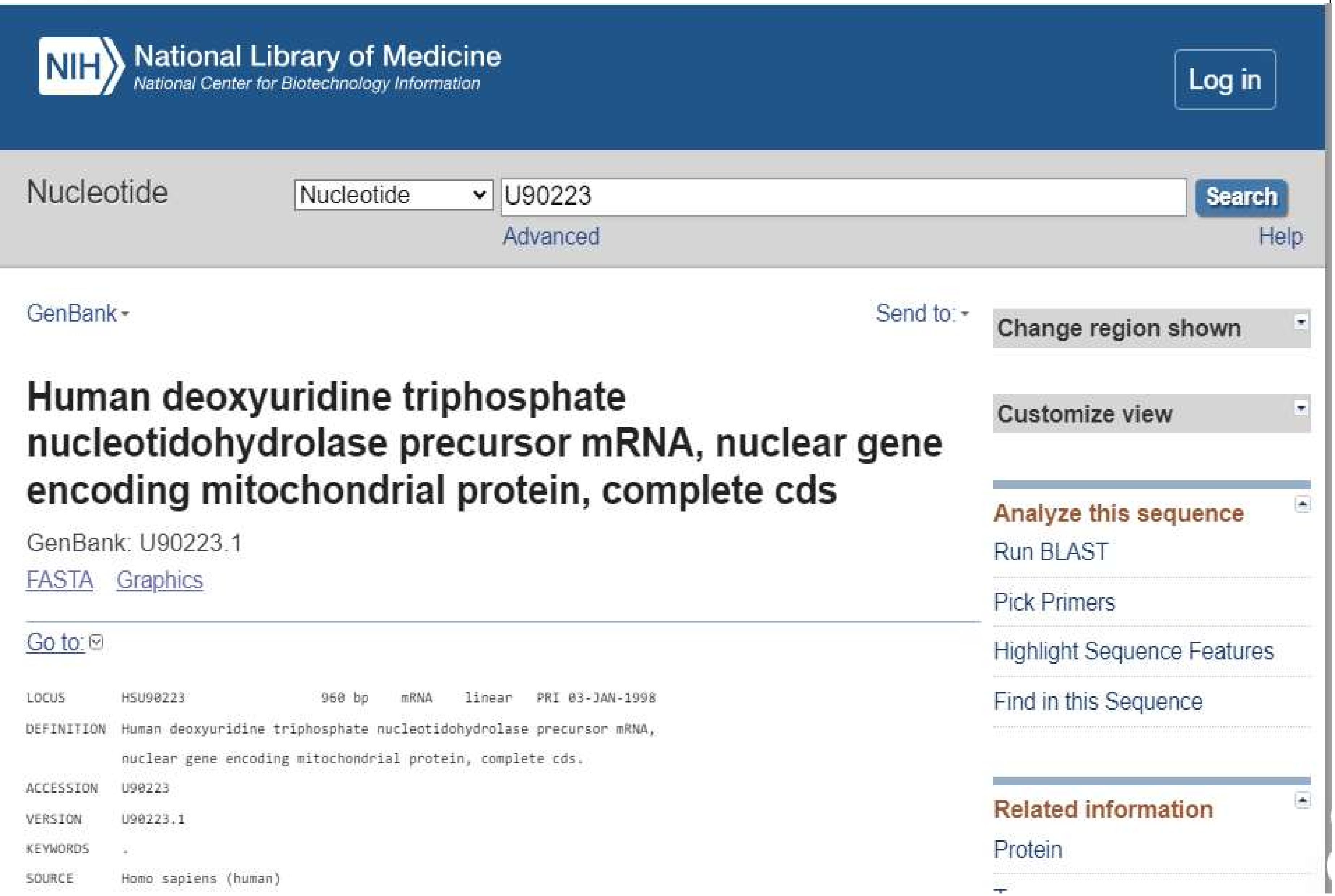
GenBank: X01714.1

GenBank Graphics

```
>X01714.1 E. coli dut gene for dUTPase (EC 3.6.1.23) (deoxyuridine 5'-triphosphate
nucleotidohydrolase)
CAGAGAAAATCAAAAAGCAGGCCACGCAGGGTGATGAATTAACAATAAAAATGGTTAAAAACCCCGATAT
CGTCGCAGGCGTTGCCGCACTAAAAGACCATCGACCCTACGTCGTTGGATTTGCCGCCGAAACAAATAAT
GTGGAAGAATACGCCCGGCAAAAACGTATCCGTAAAAACCTTGATCTGATCTGCGCGAACGATGTTTCCC
AGCCAACTCAAGGATTTAACAGCGACAACAACGCATTACACCTTTTCTGGCAGGACGGAGATAAAGTCTT
ACCGCTTGAGCGCAAAGAGCTCCTTGGCCAATTATTACTCGACGAGATCGTGACCCGTTATGATGAAAAA
AATCGACGTTAAGATTCTGGACCCGCGCGTTGGGAAGGAATTTCCGCTCCCGACTTATGCCACCTCTGGC
TCTGCCGGACTTGACCTGCGTGCCTGTCTCAACGACGCCGTAGAACTGGCTCCGGGTGACACTACGCTGG
TTCCGACCGGGCTGGCGATTCATATTGCCGATCCTTCACTGGCGGCAATGATGCTGCCGCGCTCCGGATT
GGGACATAAGCACGGTATCGTGCTTGGTAACCTGGTAGGATTGATCGATTCTGACTATCAGGGCCAGTTG
ATGATTTCCGTGTGGAACCGTGGTCAGGACAGCTTCACCATTCAACCTGGCGAACGCATCGCCCAGATGA
TTTTTGTTCCGGTAGTACAGGCTGAATTTAATCTGGTGGAAGATTTCGACGCCACCGACCGCGGTGAAGG
CGGCTTTGGTCACTCTGGTCGTCAGTAACACATACGCATCCGAATAACGTCATAACATAGCCGCAAACAT
TTCGTTTGCGGTCATAGCGTGGGTGCCGCCTGGCAAGTGCTTATTTTCAGGGGTATTTTGTAACATGGCA
GAAAAACAAACTGCGAAAAGGAACCGTCGCGAGGAAATACTTCAGTCTCTGGCGCTGATGCTGGAATCCA
GCGATGGAAGCCAACGTATCACGACGGCAAAACTGGCCGCCTCTGTCGGCGTTTCCGAAGCGGCACTGTA
TCGCCACTTCCCCAGTAAGACCCGCATGTTCGATAGCCTGATTGAGTTTATCGAAGATAGCCTGATTACT
CGCATCAACCTGATTCTGAAAGATGAGAAAGACACCACAGCGCGCCTGCGTCTGATTGTGTTGCTGCTTC
TCGGTTTTGGTGAGCGTAATCCTGGCCTGACCCGCATCCTCACTGGTCATGCGCTAATGTTTGAACAGGA
TCGCCTGCAAGGGCGCATCAACCAGCTGTTCGAGCGTATTGAAGCGCAGCTGCGCCAGGTATTGCGTGAA
AAGAGAATGCGTGAGGGTGAAGGTTACACCACCGATGAAACCCTGCTGGCAAGCCAGATCCTGGCCTTCT
GTGAAGGTATGCTGTCACGTTTTGTCCGCAGCGAATTTAAATACCGCCCGACGGATGATTTTGACGCCCG
CTGGCCGCTAATTGCGGCCAGTTGCAGTAATATGACGCCGGATGACTTTTCATCCGGCGAGTTTCTTTAA
ACGCCAAACTCTTCGCGATAGGCCTTAACCGCCGCCAGATGTTCCGCCATTTCCGGCTTCTCTTCCAGG
```

GenBank Entry of an eukaryotic mRNA

- Continue with the dUPase gene, but an eukaryotic!
- dUPase presents both prokaryotes and eukaryotes.
- Human have it too!
- A simple eukaryote dUPase gene version is U90223.
- Selecting dUPase for both types, allows us to better feel and understands the complexity differences



Human deoxyuridine triphosphate nucleotidohydrolase precursor mRNA, nuclear gene encoding mitochondrial protein, complete cds

GenBank: U90223.1

FASTA Graphics

Go to:

```
LOCUS       HSU90223                960 bp    mRNA    linear   PRI 03-JAN-1998
DEFINITION  Human deoxyuridine triphosphate nucleotidohydrolase precursor mRNA,
            nuclear gene encoding mitochondrial protein, complete cds.
ACCESSION   U90223
VERSION     U90223.1
KEYWORDS    .
SOURCE      Homo sapiens (human)
```

Reading the GenBank header of a Eukaryotic entry

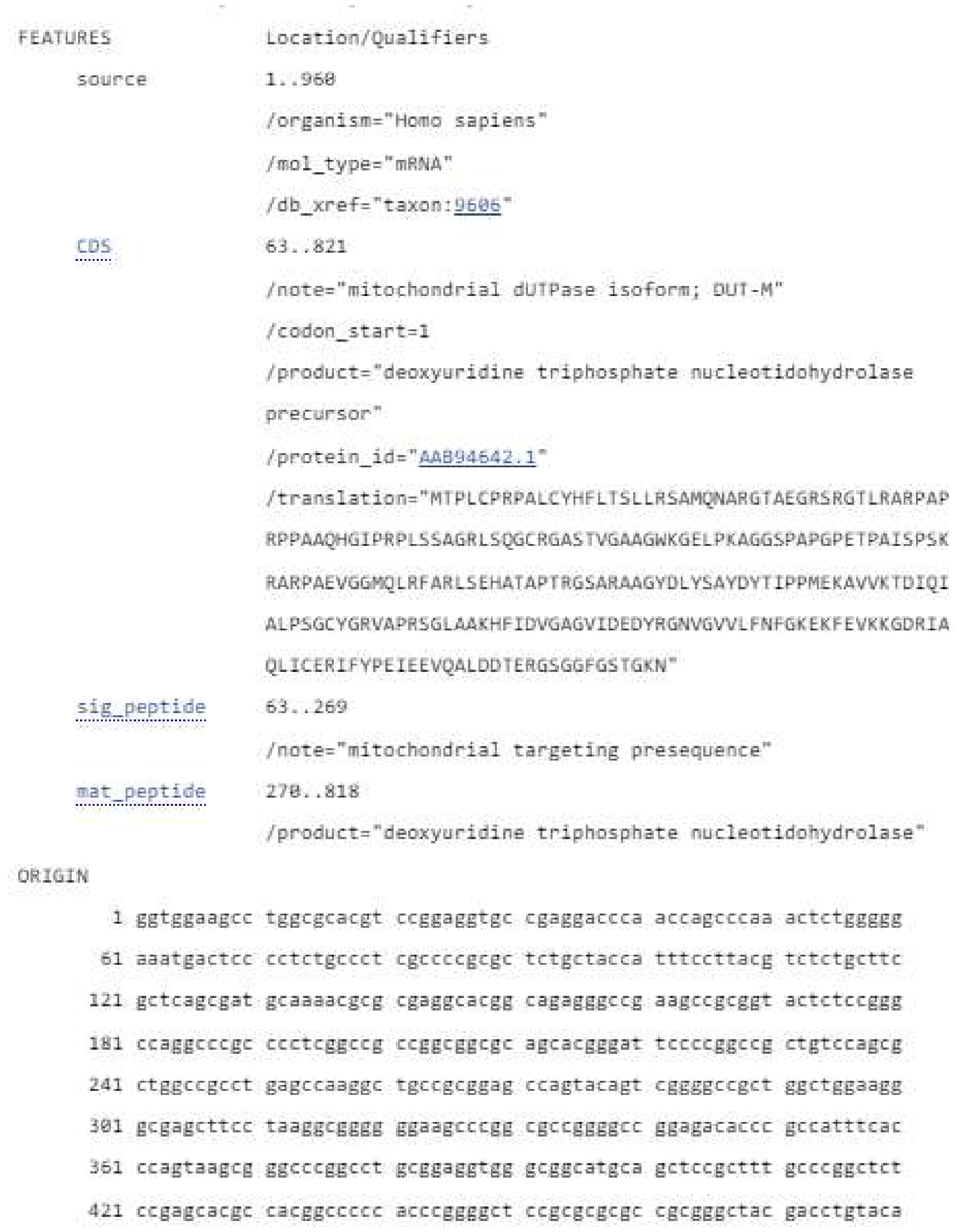
- Although it is related to a human gene, GenBank entry U90223 doesn’t look very different from entry X01714, which describes its bacterial homologue.
- The top part of the entry follows the general information keywords order: Locus, Accession, Definition, and Version!
- The Keyword line, which is supposed to list relevant and searchable terms (such as dUPase), is empty for U90223!!!)

What? How that is possible?

- Unfortunately, this is not an accident!
- It shows a common problem in sequence databases: Annotations may be incomplete!
- Information on databases may be missing or incomplete!
- A word to the wise: You should never expect GenBank (or any other sequence databases) annotations to be up-to-date!

Feature Section of U90223 Entry

- The CDS keyword indicates a coding region (63-821) sequence that corresponds to the mitochondrial form of human dUTPase.
- Mitochondria are referred to as the "powerhouses" of the cells of most eukaryotes because they generate most of the chemical energy required for cellular functions.
- Following the conceptual amino-acid translation of the ORF, the *sig peptide* keyword indicates the location of a mitochondrial targeting sequence
- The *mat peptide* keyword provides the exact boundaries of the *mature peptide* (A molecule that contains two or more amino acids)



```
FEATURES             Location/Qualifiers
     source          1..960
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
     CDS             63..821
                     /note="mitochondrial dUTPase isoform; DUT-M"
                     /codon_start=1
                     /product="deoxyuridine triphosphate nucleotidohydrolase
                     precursor"
                     /protein_id="AAB94642.1"
                     /translation="MTPLCPRPALCYHFLTSLLRSAMQNARGTAEGRSRGTLRARPAP
                     RPPAAQHGIPRPLSSAGRLSQGCRGASTVGAAGWKGELPKAGGSPAPGPETPAISPSK
                     RARPAEVGGMQLRFARLSEHATAPTRGSARAAGYDLYSAYDYTIPPMEKAVVKTDIQI
                     ALPSGCYGRVAPRSGLAAKHFIDVGAGVIDEDYRGNVGVVLFNFGKEKFEVKKGDRIA
                     QLICERIFYPEIEEVQALDDTERGSGGFGSTGKN"
     sig_peptide     63..269
                     /note="mitochondrial targeting presequence"
     mat_peptide     270..818
                     /product="deoxyuridine triphosphate nucleotidohydrolase"
ORIGIN
        1 ggtggaagcc tggcgcacgt ccggaggtgc cgaggaccca accagcccaa actctggggg
       61 aaatgactcc cctctgccct cgccccgcgc tctgctacca tttccttacg tctctgcttc
      121 gctcagcgat gcaaaacgcg cgaggcacgg cagagggccg aagccgcggt actctccggg
      181 ccaggcccgc ccctcggccg ccggcggcgc agcacgggat tccccggccg ctgtccagcg
      241 ctggccgcct gagccaaggc tgccgcggag ccagtacagt cggggccgct ggctggaagg
      301 gcgagcttcc taaggcgggg ggaagcccgg cgccggggcc ggagacaccc gccatttcac
      361 ccagtaagcg ggcccggcct gcggaggtgg gcggcatgca gctccgcttt gcccggctct
      421 ccgagcacgc cacggccccc acccggggct ccgcgcgcgc cgcgggctac gacctgtaca
```

Protein Sequence

Thank you

for your attention