Introduction to working with biological sequence databases

Lars Rønn Olsen, PhD, Associate Professor
Technical University of Denmark
Program for today

Part 1
Lecture: Biological databases: GenBank
Exercise: searching GenBank for gene sequences and information

Part 2
Lecture: BLAST
Exercise: using BLAST for sequence similarity search

Part 3
Lecture: Biological databases: UniProt
Exercise: searching UniProt for protein sequences and information
Part 1: GenBank

Slides by Henrik Nielsen, Rasmus Wernersson, Lars Rønn Olsen
Learning objectives

Part 1: GenBank

After this session you will:

• Have an overview of the content of the GenBank database

• Know how to search for sequences and sequence information using basic and advanced keyword search

• Know about the FASTA format for storing biological sequences on the computer
NGS read mapping
Cost of sequencing
Background - Nucleotide databases

- National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), USA
- Established in 1982.

**EMBL**, http://www.ebi.ac.uk/embl/
- European Bioinformatics Institute (EBI), England
- Established in 1980 by the European Molecular Biology Laboratory, Heidelberg, Germany
- Now part of **ENA**, the European Nucleotide Archive, http://www.ebi.ac.uk/ena/

**DDBJ**, http://www.ddbj.nig.ac.jp/
- National Institute of Genetics, Japan

*Together they form*
Nucleotide database growth

- Growth is roughly exponential

- But doubling time has increased from ~20 months (1990s) to ~50 months (2010)

- *The databases are public* — no restrictions on the use of the data within.

2018: ~208,000,000 sequences in GenBank alone!
Sequences on the computer: The FASTA format

>alpha-D
ATGCTGACCGACTCTGACAAGAAGCTGGTCTTGCAAGGTGTGGGAGAAGGTGATCCGCCACCCAGACTGTGGAGCCGAGGCCCTGGAGAGGTGCGGGCTGAGCTTGGGGAAACCATGGGCAAAGGGGGGCGACTGGGTGGGAGCCCTACAGGGCTGCTGGGGGTTGTTCGGCTGGGGGTCAGCACTGACCATCCCGCTCCCGCAGCTGTTCACCACCTACCCCCAGACCAAGACCTACTTCCACCACTTCGACTTGCACCATGGCTCCGACCAGGTCCGCAACCACGGCAAGAAGGTGTTGGCCGCCTTGGGCAACGCTGTCAAGAGCCTGGGCAACCTCAGCCAAGCCCTGTCTGACCTCAAGCGACCTGCATGCCTACAACCTGCGTGTCGACCCTGTCAACTTCAAGGGCGGGGACGGGGGTCAGGGGCCGGGGAGTTGGGGGCCAGGGACCTGGTTGGGGATCCGGGGCCATGCCGGCGGTACTGAGCCCTGTTTTGCCTTGCAGCTGCTGGCGCAGTGCTTCCACGTGGTGCTGGCCACACACCTGGGCAACGACTACACCCCGGAGGCACATGCTGCCTTCGACAAGTTCCTGTCCGGCTGTGTGCACCGTGCTGGCCGAGAAGTACAGATAA

>alpha-A
ATGGTGCTGTCTGCAACGACAAGAAGCAGTGAGGGCCGCTCTTCGGCAAAAAATCGGCGGCAGGCCCGGTGACCTGGGTGTGGTGAAGCCCTGGAGAGGTATGTGGTCATCCGTCATTACCCCCTCTCTTGTCTGTCTGTGACTCCATCCCATCTGCCCCCATACTCTCCCCCATCCATAACTGTCCCTTTCATATGGCCCTGGCCCTGTCTCATCTCTCCCCACCTGTCCCTGATTGCCTCTGTCCCCCAGGTTGGTCATCACCTACCCCCAGACCAAGACCTACTTCCCCCACTTCGACCTGTCCGTGTGCTGCACCCTGCCCAAAAGCTCCGTGTGGACCCCGTCAACTTCAAAGTGAGCATCTGGGAAGGGGTGACCGTCTGGCTCCCCTCTGCACACACCTCTGGCTACCCCCTCACCTCACCCCCTTGCTCACATCTCCTTTTGCCTTTCAGCTGCTGGGTCACTGCTTCCTGGTGGTCGTGGCCGTCCACTTCCCCTCTCTCCTGACCCCGGAGGTCCATGCTTCCCTGGACAAGTTCGGTGCTGTGCCGTGGGACACCGTCCTTACTGCCAAGTACCGTGTA

Reminder: Eukaryotic gene structure
GenBank format

- Originates from the GenBank database.
- Contains both a DNA sequence and annotations of features (e.g. location of genes).
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)
           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
### GenBank format - ORIGIN section

<table>
<thead>
<tr>
<th>ORIGIN</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ctgcgtggcc tcagccccctc cacccttccca cgctgataag ataaggccag ggcgggagcg</td>
</tr>
<tr>
<td>61</td>
<td>caggggtgccttaaagagctcg gccccgcggg tgtctccacc acagaaacc gtcagttgcc</td>
</tr>
<tr>
<td>121</td>
<td>agcctgccac gccgctgccg ccattgcctgac cgccgagggc aagagctca tcagctctac</td>
</tr>
<tr>
<td>181</td>
<td>tgtggagaggttgtgtgcc accagagga attcgggaagt gaagctctgc agaggtgtgcag</td>
</tr>
<tr>
<td>241</td>
<td>gctggccca ggggacttc cggccaggtg cagcctggag cggcgcagct gcagcggctt</td>
</tr>
<tr>
<td>301</td>
<td>tgggctggaga cccagagcgc cagccagtgc acggtggagt gggccagcgg ccagcgcacc</td>
</tr>
<tr>
<td>361</td>
<td>aaaactgact ggcctcgtctcc cgggagagtt ttcctcgcct cctcccagac caagacctc</td>
</tr>
<tr>
<td>421</td>
<td>tccccccactctcgacctgctca caccgctctct ggcctcgctc gaggacagtt gcgagttgag</td>
</tr>
<tr>
<td>481</td>
<td>ggccgtccccttgggaatgc cgtgaagctcg cttggcagtc gggccagcgg gcagcgcacc</td>
</tr>
<tr>
<td>541</td>
<td>ctgcatgccat cagagctgtc caccgctctct gttgcgcttg ttcctcgcct cctcccagac</td>
</tr>
<tr>
<td>601</td>
<td>gactaggggctctgggtcttg cgggcttggtt ggggagagtt ttcctcgcct cctcccagac</td>
</tr>
<tr>
<td>661</td>
<td>gctctgtgatc tctcggctc gcgctgcttg ggggagagtt ttcctcgcct cctcccagac</td>
</tr>
<tr>
<td>721</td>
<td>ggtaccagggctctggtgggg cccagagcgc cagccagtgc acggtggagt gggccagcgg</td>
</tr>
<tr>
<td>781</td>
<td>gtggccagaaggtgtgtgtgcc agaggtgtgc agaggtgtgcag</td>
</tr>
<tr>
<td>841</td>
<td>gggagaactca gggccctctcg gggagagtt ttcctcgcct cctcccagac catccctttg</td>
</tr>
<tr>
<td>901</td>
<td>tccgggagcag ggtactaag cctctgtggtt gggcaagact acagccccaga ggtgctggag</td>
</tr>
<tr>
<td>961</td>
<td>tggctgctggt ggcacactctg gggcaagact acagccccaga ggtgctggag tggaggttcc</td>
</tr>
<tr>
<td>1021</td>
<td>agttcttctgtgc ggcctgcttg gggctgcttg gggcagctga gggccagcgcactg</td>
</tr>
<tr>
<td>1081</td>
<td>ccccttgaggat tacacattac cagacagtcc tggctgcttg ggcctgcttg</td>
</tr>
<tr>
<td>1141</td>
<td>gggcatcgggg ggtcctcttg ggggtgctt gcctctttgct ggcctgcttg ggcctgcttg</td>
</tr>
</tbody>
</table>

//
GenBank format - FEATURE section

<table>
<thead>
<tr>
<th>FEATURES</th>
<th>Location/Qualifiers</th>
</tr>
</thead>
<tbody>
<tr>
<td>source</td>
<td>1..1185 /organism=&quot;Cairina moschata&quot; /mol_type=&quot;genomic DNA&quot; /db_xref=&quot;taxon:8855&quot;</td>
</tr>
<tr>
<td>CAAT_signal</td>
<td>20..24</td>
</tr>
<tr>
<td>TATA_signal</td>
<td>69..73</td>
</tr>
<tr>
<td>precursor_RNA</td>
<td>101..1114 /note=&quot;primary transcript&quot;</td>
</tr>
<tr>
<td>exon</td>
<td>101..234 /number=1</td>
</tr>
<tr>
<td>CDS</td>
<td>join(143..234,387..591,939..1067) /codon_start=1 /product=&quot;alpha D-globin&quot; /protein_id=&quot;CAA25966.2&quot; /db_xref=&quot;GI:4455876&quot; /db_xref=&quot;GOA:P02003&quot; /db_xref=&quot;InterPro:IPR000971&quot; /db_xref=&quot;InterPro:IPR002338&quot; /db_xref=&quot;InterPro:IPR002340&quot; /db_xref=&quot;InterPro:IPR009050&quot; /db_xref=&quot;UniProt/Swiss-Prot:P02003&quot; /translation=&quot;MLTAEDKKLIVQWEKVAGHQEEFGSEALQMFLAYQPTKYFP HFDLHFGSEQVRGHHKVAAALGNAVSKLDNLGQALELSNLHAYNLRVDVPVNFKLLA QCQVVLAAHLGKDYSPEMHAADFKNLSAVAAVLAEKVR&quot;</td>
</tr>
<tr>
<td>repeat_region</td>
<td>227..246 /note=&quot;direct repeat 1&quot;</td>
</tr>
<tr>
<td>intron</td>
<td>235..386 /number=1</td>
</tr>
<tr>
<td>repeat_region</td>
<td>289..309 /note=&quot;direct repeat 1&quot;</td>
</tr>
<tr>
<td>exon</td>
<td>387..591 /number=2</td>
</tr>
<tr>
<td>intron</td>
<td>592..939 /number=2</td>
</tr>
<tr>
<td>exon</td>
<td>940..1114 /number=3</td>
</tr>
<tr>
<td>polyA_signal</td>
<td>1095..1100</td>
</tr>
<tr>
<td>polyA_signal</td>
<td>1114</td>
</tr>
</tbody>
</table>
Where does the data come from?

Anyone registered with NCBI can submit data – this includes you!
Be careful! GenBank entries are not curated
Be careful! GenBank entries are not curated

<table>
<thead>
<tr>
<th>REFERENCE</th>
<th>2 (bases 1 to 11029)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUTHORS</td>
<td>Deubel,V., Malkinson,M. and Banet,C.</td>
</tr>
<tr>
<td>TITLE</td>
<td>Direct Submission</td>
</tr>
<tr>
<td>JOURNAL</td>
<td>Submitted (08-FEB-2002) CERVI, Institut Pasteur, 21 Avenue Tony Garnier, Lyon 69365, France</td>
</tr>
</tbody>
</table>

**FEATURES**

<table>
<thead>
<tr>
<th>Location/Qualifiers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1..11029</td>
</tr>
<tr>
<td>/organism=&quot;West Nile virus&quot;</td>
</tr>
<tr>
<td>/mol_type=&quot;genomic RNA&quot;</td>
</tr>
<tr>
<td>/strain=&quot;IS-98 STD1&quot;</td>
</tr>
<tr>
<td>/host=&quot;sick stork&quot;</td>
</tr>
<tr>
<td>/db_xref=&quot;taxon:11082&quot;</td>
</tr>
<tr>
<td>/note=&quot;isolated in Israel in 1998&quot;</td>
</tr>
<tr>
<td>97..10398</td>
</tr>
<tr>
<td>/codon_start=1</td>
</tr>
<tr>
<td>/product=&quot;polyprotein precursor&quot;</td>
</tr>
<tr>
<td>/protein_id=&quot;AAL87234.1&quot;</td>
</tr>
<tr>
<td>/translation=&quot;MSKKPGPGPKSRAVNMLKRGMPRVLSLIGKLKRAMSLIDGKPSRFVLALLAFRFTAIAPTRAVALDRWRGVRNKQTMKHLLSFKKELGTLATSSAINRSSHKKSKKRRGGGTGI Anyway AVMGLIASVGA VTSNFFQGKM PMTNATDVTDVITIPTAAGKNLCIVI AMDVGYMCDDTITYECPVLSAGNDPEDI DCWCTKSAVYVRYGRC TKT KTRHSRRSSR&quot;</td>
</tr>
</tbody>
</table>
Exercises: GenBank
Part 2: BLAST

Slides by Anders Gorm Pedersen, Rasmus Wernersson, Lars Rønn Olsen
Learning objectives

Part 2: BLAST

After this session you will:

• Know how sequence similarity searching works in broad strokes

• Know how to use BLAST for "reverse" searching GenBank using a sequence as the search input
BLAST
- Searching databases for sequences

BLAST (Basic Local Alignment Search Tool) is a tool to query a database for sequences similar to an input sequence.

Imagine you have sequenced a gene from an unknown sample, and you would like to know what it is. You can use BLAST in NCBI to compare your sequence to ALL the ~208,000,000 sequences in GenBank!

Sequence similarity searching

Example: you would like to know what the following sequence is:

X) AATGCCG

You have the following three sequences in your database:

A) CGTGTGATC
B) AATGCCG
C) GCTGTGAC
Sequence similarity searching

Example: you would like to know what the following sequence is:

X) AATGCCG

You have the following three sequences in your database:

A) CGTGATGTC
B) AATCCC
C) GCTGTGAC

X) AATGCCG
B) AATCCC

C) GCTGTGAC
Sequence similarity searching

Example: you would like to know what the following sequence is:

X) AATGCCG

You have the following three sequences in your database:

A) CGTGTGATC
B) AATCCCCG
C) GCTGTGAC

X) AATG−CCG
B) AATCCCCG
Sequence similarity searching

Example: you would like to know what the following sequence is:

X) AATGCCG

You have the following three sequences in your database:

A) AATCCCG
B) AATCCCCG
C) AATCC

X) AATGCCG  X) AATG−CCG  X) AATGCCG
A) AATCCCG  B) AATCCCGG  C) AAT−CC−

Which match is best?
Alignment score

There are different schemes for scoring alignments

A common one, is the use of penalties

For example, a mismatch can give a penalty of 1

Insertion of a gap could give a penalty of 5

Extension of a gap can give a penalty of 2

The lower the penalty, the better the alignment score

X) AATGCCG
A) AATCCCG
Penalty: 1

X) AATG–CCG
B) AATCCCCG
Penalty: 6

X) AATG---CCG
C) AATCCCCCG
Penalty: 8
BLAST flavors

**BLASTN**
Nucleotide query sequence
Nucleotide database

**BLASTP**
Protein query sequence
Protein database

**BLASTX**
Nucleotide query sequence
Protein database
Compares all six reading frames with the database

**TBLASTN**
Protein query sequence
Nucleotide database
"On the fly" six frame translation of database

**TBLASTX**
Nucleotide query sequence
Nucleotide database
Compares all reading frames of query with all reading frames of the database
BLASTN
- Searching GenBank for nucleotide sequences

[Image: BLASTN interface]

When is a database hit significant?

Problem:

• Even unrelated sequences can be aligned (yielding a low score)
• How do we know if a database hit is meaningful?
• When is an alignment score sufficiently high?

Solution:

• Determine the range of alignment scores you would expect to get for random reasons (i.e., when aligning unrelated sequences).
• Compare actual scores to the distribution of random scores.
• Is the real score much higher than you’d expect by chance?
Distribution of random alignment scores

Software simulation
Significance of alignment score expressed as E-value

Searching a database of *unrelated* sequences results in scores following an extreme value distribution.

The exact shape and location of the distribution depends on the exact nature of the database and the query sequence.

**E-value:** the number of *random hits* to expect for any given score.

Want E-values below 1 (the lower the better).
Significance of alignment score expressed as E-value

**E-value / Expect-value:**
Number of unrelated hits with an equal or better alignment score to expect due to strictly stochastic reasons.

**Example:**
Alignment score = 110
E-value = 8.7

Alignment score = 135
E-value = 0.0001
Exercises: BLAST
Part 3: UniProt

Slides by Henrik Nielsen, Lars Rønn Olsen
Part 3: UniProt

After this session you will:

- Have an overview of the content of the UniProt database
- Know the difference between the UniProt/Swiss-Prot and UniProt/TrEMBL databases
- Know how to search for sequences and sequence information using basic and advanced keyword search
Protein databases, historical background

- **Swiss-Prot**, http://www.expasy.org/sprot/
  - Established in 1986 in Switzerland
  - ExPASy (Expert Protein Analysis System)
  - Swiss Institute of Bioinformatics (SIB) and European Bioinformatics Institute (EBI)

- **PIR**, http://pir.georgetown.edu/
  - Established in 1984
  - National Biomedical Research Foundation, Georgetown University, USA

- *In 2002 merged into:*
- **UniProt**, http://www.uniprot.org/
  - A collaboration between SIB, EBI and Georgetown University.
UniProt

- UniProt Knowledgebase (UniProtKB)
- UniProt Reference Clusters (UniRef)
- UniProt Archive (UniParc)

- UniProt Knowledgebase Release 2018_07 (18-July-18) consists of:
  - UniProtKB/Swiss-Prot: Annotated manually *(curated)*
    - 557,992 entries
  - UniProtKB/TrEMBL: Computer annotated
    - 120,243,849 entries
Levels of curation

GenBank / EMBL / DDBJ:
• Entries created & maintained by individual contributors
• No check for redundancy

Swiss-Prot:
• Entries created & maintained by staff
• Better standards compliance

TrEMBL:
• Entries created by automatic translation of EMBL sequences & annotations
Growth of UniProt

TrEMBL

Swiss-Prot
Content of UniProt Knowledgebase

• **Amino acid sequences**

• Functional and structural annotations
  - Function / activity
  - Secondary structure
  - Subcellular location
  - Mutations, phenotypes
  - Post-translational modifications

• Origin
  - organism: Species, subspecies; classification
  - tissue

• References

• Cross references
Amino acid sequences

From where do you get amino acid sequences?

• Translation of nucleotide sequences (GenBank/EMBL/DDBJ)
• Direct amino acid sequencing: Edman degradation
• Mass spectrometry
• 3D-structures
Content of UniProt Knowledgebase

• Amino acid sequences
• Functional and structural annotations
  - Function / activity
  - Secondary structure
  - Subcellular location
  - Mutations, phenotypes
  - Post-translational modifications
• Origin
  - organism: Species, subspecies; classification
  - tissue
• References
• Cross references
Protein structure

Primary structure: Amino acid sequence

Secondary structure:
"Backbone" hydrogen bonding
Alpha helix / Beta sheet / Turn

Tertiary structure: Fold, 3D coordinates

Quaternary structure: subunits
Content of UniProt Knowledgebase

• Amino acid sequences
• Functional and structural annotations
  - Function / activity
  - Secondary structure
  - **Subcellular location**
  - Mutations, phenotypes
  - Post-translational modifications
• Origin
  - organism: Species, subspecies; classification
  - tissue
• References
• Cross references
Subcellular location / protein sorting

Various proteins belong to different *compartments* of the cell – some even belong *outside* the cell.
Content of UniProt Knowledgebase

- Amino acid sequences
- Functional and structural annotations
  - Function / activity
  - Secondary structure
  - Subcellular location
  - Mutations, phenotypes
  - Post-translational modifications
- Origin
  - organism: Species, subspecies; classification
  - tissue
- References
- Cross references
Many proteins are modified after they have been synthesized in order to become functional.

**Proteolysis:** Cleavage of *signal peptides, propeptides or initiator methionine.*

**Glycosylation:** Especially common on the *cell surface.* Plays a role in sorting of proteins to *lysosomes.*

**Phosphorylation:** Often *reversible.* Regulates the *activity* of many enzymes.
Inhibitor of serum proteases. Its primary target is elastase, but it also has a moderate affinity for plasmin and thrombin. Irreversibly inhibits trypsin, chymotrypsin and plasminogen activator. The aberrant form inhibits insulin-induced NO synthesis in platelets, decreases coagulation time and has proteolytic activity against insulin and plasmin.

Short peptide from AAT: reversible chymotrypsin inhibitor. It also inhibits elastase, but not trypsin. Its major physiological function is the protection of the lower respiratory tract against proteolytic destruction by human leukocyte elastase (HLE).

**Sites**

<table>
<thead>
<tr>
<th>Feature key</th>
<th>Position(s)</th>
<th>Length</th>
<th>Description</th>
<th>Graphical view</th>
<th>Feature identifier</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site^1</td>
<td>382-383</td>
<td>2</td>
<td>Reactive bond</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**GO - Molecular function**^1

- Glycoprotein binding # Source: Uniprot
- Protease binding # Source: Uniprot

**GO - Biological process**^1

- Acute phase response # Source: Uniprot
- Negative regulation of endopeptidase activity # Source: GO_Central
- Platelet degranulation # Source: Reactome

Complete GO annotation...

**Keywords - Molecular function**^1

Protease inhibitor, Serine protease inhibitor

**Keywords - Biological process**^1

Acute phase, Blood coagulation, Hemostasis
**UniProt entry, formatted view**

**Entry name (ID)**

**Accession #**
Entry names and accession numbers

• Entry name (UniProt ID / GenBank LOCUS)
  • Provides a mnemonic identifier for a database entry. One and only one name per entry.

• Accession #
  • Provides a stable identifier for a database entry (does not change across database versions). One or more accession numbers per entry.
# Entry information

## Entry name
A1AT_HUMAN

## Accession
Primary (citable) accession number: **P01009**
Secondary accession number(s): A6PX14, B2RDQ8, Q0PVP5, Q13672, Q53XM8, Q5U0M1, Q7M4R2, Q86U18, Q86U19, Q96BF9, Q96ES1, Q9P1P0, Q9UCE6, Q9UCM3

## Entry history
- Integrated into UniProtKB/Swiss-Prot: July 21, 1986
- Last sequence update: October 1, 1996
- Last modified: February 4, 2015

This is version 213 of the entry and version 3 of the sequence. [Complete history]

## Entry status
Reviewed (UniProtKB/Swiss-Prot)

## Annotation program
Chordata Protein Annotation Program

## Disclaimer
Any medical or genetic information present in this entry is provided for research, educational and informational purposes only. It is not in any way intended to be used as a substitute for professional medical advice, diagnosis, treatment or care.
Inhibitor of serine proteases. Its primary target is elastase, but it also has a moderate affinity for plasmin and thrombin. Irreversibly inhibits trypsin, chymotrypsin and plasminogen activator. The aberrant form inhibits insulin-induced NO synthesis in platelets, decreases coagulation time and has proteolytic activity against insulin and plasmin.

Short peptide from AAT: reversible chymotrypsin inhibitor. It also inhibits elastase, but not trypsin. Its major physiological function is the protection of the lower respiratory tract against proteolytic destruction by human leukocyte elastase (HLE).
### Names & Taxonomy

| Protein names | **Recommended name:** Alpha-1-antitrypsin  
|---------------|------------------------------------------------------------------------------------------|
| Alternative name(s): | • Alpha-1 protease inhibitor  
| | • Alpha-1-antiproteinase  
| | • Serpin A1  
| Cleaved into the following chain: | • Short peptide from AAT  
| | • Short name: SPAAAT  
| Gene names | Name: SERPINA1  
| | Synonyms: AAT, P1  
| | ORF Names: PRO0684, PRO2209  
| Organism | Homo sapiens (Human)  
| Taxonomic identifier | 9606 [NCBI]  
| Taxonomic lineage | Eukaryota > Metazoa > Chordata > Craniata > Vertebrata > Euteleostomi > Mammalia > Eutheria > Euarchontoglires > Primates > Haplorrhini > Catarhini > Hominidae > Homo }
Gene Ontology (GO)

GO - Molecular function:
- glycoprotein binding [Source: UniProtKB]
- identical protein binding [Source: IntAct]
- protease binding [Source: UniProtKB]
- serine-type endopeptidase inhibitor activity [Source: UniProtKB]

GO - Biological process:
- acute-phase response [Source: UniProtKB-KW]
- blood coagulation [Source: Reactome]
- negative regulation of endopeptidase activity [Source: GO_Central]
- platelet activation [Source: Reactome]
- platelet degranulation [Source: Reactome]
- regulation of proteolysis [Source: GO_Central]
<table>
<thead>
<tr>
<th>Feature key</th>
<th>Position(s)</th>
<th>Length</th>
<th>Description</th>
<th>Graphical view</th>
<th>Feature identifier</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turn(^1)</td>
<td>48 - 50</td>
<td>3</td>
<td>Combined sources</td>
<td><a href="#">Graphical view</a></td>
<td><a href="#">Identifier</a></td>
<td><a href="#">Actions</a></td>
</tr>
<tr>
<td>Helix(^1)</td>
<td>51 - 68</td>
<td>18</td>
<td>Combined sources</td>
<td><a href="#">Graphical view</a></td>
<td><a href="#">Identifier</a></td>
<td><a href="#">Actions</a></td>
</tr>
<tr>
<td>Beta strand(^1)</td>
<td>70 - 72</td>
<td>3</td>
<td>Combined sources</td>
<td><a href="#">Graphical view</a></td>
<td><a href="#">Identifier</a></td>
<td><a href="#">Actions</a></td>
</tr>
<tr>
<td>Beta strand(^1)</td>
<td>74 - 76</td>
<td>3</td>
<td>Combined sources</td>
<td><a href="#">Graphical view</a></td>
<td><a href="#">Identifier</a></td>
<td><a href="#">Actions</a></td>
</tr>
<tr>
<td>Helix(^1)</td>
<td>78 - 89</td>
<td>12</td>
<td>Combined sources</td>
<td><a href="#">Graphical view</a></td>
<td><a href="#">Identifier</a></td>
<td><a href="#">Actions</a></td>
</tr>
<tr>
<td>Helix(^1)</td>
<td>94 - 103</td>
<td>10</td>
<td>Combined sources</td>
<td><a href="#">Graphical view</a></td>
<td><a href="#">Identifier</a></td>
<td><a href="#">Actions</a></td>
</tr>
<tr>
<td>Turn(^1)</td>
<td>108 - 110</td>
<td>3</td>
<td>Combined sources</td>
<td><a href="#">Graphical view</a></td>
<td><a href="#">Identifier</a></td>
<td><a href="#">Actions</a></td>
</tr>
<tr>
<td>Helix(^1)</td>
<td>113 - 127</td>
<td>15</td>
<td>Combined sources</td>
<td><a href="#">Graphical view</a></td>
<td><a href="#">Identifier</a></td>
<td><a href="#">Actions</a></td>
</tr>
<tr>
<td>Beta strand(^1)</td>
<td>135 - 145</td>
<td>11</td>
<td>Combined sources</td>
<td><a href="#">Graphical view</a></td>
<td><a href="#">Identifier</a></td>
<td><a href="#">Actions</a></td>
</tr>
<tr>
<td>Helix(^1)</td>
<td>152 - 162</td>
<td>11</td>
<td>Combined sources</td>
<td><a href="#">Graphical view</a></td>
<td><a href="#">Identifier</a></td>
<td><a href="#">Actions</a></td>
</tr>
<tr>
<td>Beta strand(^1)</td>
<td>164 - 169</td>
<td>6</td>
<td>Combined sources</td>
<td><a href="#">Graphical view</a></td>
<td><a href="#">Identifier</a></td>
<td><a href="#">Actions</a></td>
</tr>
<tr>
<td>Helix(^1)</td>
<td>174 - 188</td>
<td>15</td>
<td>Combined sources</td>
<td><a href="#">Graphical view</a></td>
<td><a href="#">Identifier</a></td>
<td><a href="#">Actions</a></td>
</tr>
<tr>
<td>Turn(^1)</td>
<td>189 - 191</td>
<td>3</td>
<td>Combined sources</td>
<td><a href="#">Graphical view</a></td>
<td><a href="#">Identifier</a></td>
<td><a href="#">Actions</a></td>
</tr>
<tr>
<td>Beta strand(^1)</td>
<td>206 - 220</td>
<td>15</td>
<td>Combined sources</td>
<td><a href="#">Graphical view</a></td>
<td><a href="#">Identifier</a></td>
<td><a href="#">Actions</a></td>
</tr>
<tr>
<td>Helix(^1)</td>
<td>224 - 226</td>
<td>3</td>
<td>Combined sources</td>
<td><a href="#">Graphical view</a></td>
<td><a href="#">Identifier</a></td>
<td><a href="#">Actions</a></td>
</tr>
<tr>
<td>Beta strand(^1)</td>
<td>228 - 235</td>
<td>8</td>
<td>Combined sources</td>
<td><a href="#">Graphical view</a></td>
<td><a href="#">Identifier</a></td>
<td><a href="#">Actions</a></td>
</tr>
<tr>
<td>Beta strand(^1)</td>
<td>238 - 256</td>
<td>19</td>
<td>Combined sources</td>
<td><a href="#">Graphical view</a></td>
<td><a href="#">Identifier</a></td>
<td><a href="#">Actions</a></td>
</tr>
</tbody>
</table>
### Evidence (Comments, Feature Table)

**Experimental:**

<table>
<thead>
<tr>
<th>Feature key</th>
<th>Position(s)</th>
<th>Length</th>
<th>Description</th>
</tr>
</thead>
</table>
| Signal peptide | 1 – 24      | 24     | Manual assertion based on experiment in [1]

---

"Characterization of a 54 kDa, alpha 1-antitrypsin-like protein isolated from ascitic fluid of an endometrial cancer patient."

Tanaka N., Sekiya S., Takamizawa H., Kato N., Moriyama Y., Fujimura S.

Jpn. J. Cancer Res. 82:693-700(1991) [PubMed] [Europe PMC] [Abstract]

**Cited for:** PROTEIN SEQUENCE OF 25-39, FUNCTION.

**Predicted:**

<table>
<thead>
<tr>
<th>Feature key</th>
<th>Position(s)</th>
<th>Length</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signal peptide</td>
<td>1 – 30</td>
<td>30</td>
<td>Sequence Analysis</td>
</tr>
</tbody>
</table>

**By similarity:**

<table>
<thead>
<tr>
<th>Feature key</th>
<th>Position(s)</th>
<th>Length</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disulfide bond</td>
<td>41 ↔ 77</td>
<td></td>
<td>By similarity</td>
</tr>
</tbody>
</table>
Evidence types in UniProt

See also http://www.uniprot.org/help/evidences
UniProt entry, sequence(s)
Cross-references, nucleotide sequences

<table>
<thead>
<tr>
<th>Select the link destinations:</th>
<th>K01396 mRNA. Translation: AAB59375.1.</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMBL</td>
<td>Genomic DNA. Translation: AAB59495.1.</td>
</tr>
<tr>
<td>GenBank</td>
<td>X01683 mRNA. Translation: CAA25838.1.</td>
</tr>
<tr>
<td>DDBJ</td>
<td>M11465 mRNA. Translation: AAA51546.1.</td>
</tr>
<tr>
<td></td>
<td>J02619 Genomic DNA. Translation: AAA51547.1.</td>
</tr>
<tr>
<td></td>
<td>DQ682455 mRNA. Translation: ABG73380.1.</td>
</tr>
<tr>
<td></td>
<td>AM048838 Genomic DNA. Translation: CAJ15161.1.</td>
</tr>
<tr>
<td></td>
<td>AF113676 mRNA. Translation: AAF29581.1.</td>
</tr>
<tr>
<td></td>
<td>AF130068 mRNA. Translation: AAG35496.1.</td>
</tr>
<tr>
<td></td>
<td>BX161449 mRNA. Translation: CAD61914.1.</td>
</tr>
<tr>
<td></td>
<td>BX247968 mRNA. Translation: CAD62306.1.</td>
</tr>
<tr>
<td></td>
<td>BX248002 mRNA. Translation: CAD62334.1. Different initiation.</td>
</tr>
<tr>
<td></td>
<td>BX248257 mRNA. Translation: CAD62585.1. Different initiation.</td>
</tr>
<tr>
<td></td>
<td>AK315637 mRNA. Translation: BAG38005.1.</td>
</tr>
</tbody>
</table>

[Image of nucleotide sequence with arrow annotation]
Exercises: UniProt