Protein databases

Henrik Nielsen

Protein databases, historical background

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 Knowledgebase (UniProtKB)
UniProt Reference Clusters (UniRef)
UniProt Archive (UniParc)
UniProt Metagenomic and Environmental Sequence Database (UniMES)

UniProt Knowledgebase Release 2013_02 (06-Feb-13)
consists of:
UniProtKB/Swiss-Prot: Annotated manually (curated) 539,165 entries
UniProtKB/TrEMBL: Computer annotated 29,769,971 entries
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
Protein sorting in eukaryotes

Different proteins belong to different compartments of the cell – and some belong outside the cell.

General annotation (Comments)

Post-translational modifications

Many proteins are modified after their synthesis in order to become active:

- **Proteolysis**: Cleavage of signal peptides, propeptides or initiating methionine
- **Glycosylation**: Particularly common in proteins on the surface of cells. Also plays a role in sorting of proteins to lysosomes
- **Phosphorylation**: Often reversible. Regulates the activity of many enzymes
More post-translational modifications

- Lipid anchors
  - (e.g. GPI anchors)
- Disulfide bonds
- Prosthetic groups
  - (e.g. metal ions)

Sequence annotation (Feature Table)

General annotation (Ontologies)
Secondary structure (Feature Table)

Protein structure

Evidence (Comments, Feature Table)
### Evidence/Confidence types

3 types of **non-experimental qualifiers** in Sequence annotation and General comment:

- **Potential**: Predicted using sequence analysis
- **Probable**: Uncertain experimental evidence
- **By similarity**: Predicted using sequence similarity

### UniProt entry, sequence(s)

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Length</th>
<th>Start</th>
<th>End</th>
<th>GenBank</th>
<th>RefSeq</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Cross-references, nucleotide sequences

<table>
<thead>
<tr>
<th>Nucleotide databases</th>
<th>GenBank</th>
<th>RefSeq</th>
</tr>
</thead>
<tbody>
<tr>
<td>BACPAC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EMBL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ensembl</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UniProt</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ensembl</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Cross-references, 3D structure

Other databases linked from UniProt (there are ~100 in total):
- Nucleotide sequences
- 3D structure
- Protein-protein interactions
- Enzymatic activities and pathways
- Gene expression (microarrays and 2D-PAGE)
- Ontologies
- Families and domains
- Organism specific databases
The genetic code

• Degenerate (redundant) but not ambiguous
  - Almost universal (deviations found in mitochondria)

Reading Frames 1

A piece of an mRNA-strand:

\[ \text{5'} \text{aug ccc aag cug aau agc gua gag ggg uuu uca uca uuu gag gac gau gua uaa} \text{3'} \]

can be divided into triplets (codons) in three ways:

1. \[ \text{aug ccc aag cug aau agc gua gag ggg uuu uca uca uuu gag gac gau gua uaa} \]
   \[ \text{M P K L N S V E G F S S F E D D V} \]

2. \[ \text{ugc cca agc uga au a gcg uag agg ggu uuu cau cau uug agg acg aug uau} \]
   \[ \text{C P S I A} \]

3. \[ \text{gcc caa gcu gaa uag cgu aga ggg guu uuc auc auu uga gga cga ugu aua} \]
   \[ \text{A Q A E R R G V F I I} \]

Each possible set of triplets is called a reading frame.

Reading Frames 2

Since there are two strands in DNA, there are six possible reading frames in a piece of DNA (three in each direction):

\[ \text{A Q A E R R G V F I I} \text{ U R C I} \]
\[ \text{C P S I A I A} \text{ R G F H H L R T M Y} \]
\[ \text{M P K L N S V E G F S S F E D D V} \text{ U} \]

A reading frame from a start codon to the first stop codon is called an open reading frame (underlined above).
Virtual Ribosome—a comprehensive DNA translation tool with support for integration of sequence feature annotation

Rasmus Winther Hovem

Center for Biological Sequence Analysis, BioCentrum—DKFZ, Technical University of Denmark, Building 345, DK-2800 Lyngby Denmark

Received January 19, 2006; Revised March 1, 2006; Accepted March 22, 2006

Abstract

Virtual Ribosome is a DNA translation tool with two areas of focus. One providing a graphical user interface for RNA gene analysis, integration, data display, and annotation of sequence features. The second area is focused on tools and solutions for integrating RNA gene analysis into next generation bioinformatics. The software is available at http://mem.bio.dtu.dk/virtual_ribosome

Introduction

Virtual Ribosome is a software package for translating DNA into RNA and displaying the information in a graphical user interface. The software is designed to be used in conjunction with other bioinformatics tools. The main function of the software is to support the