Functional Annotation

Trudy Torto-Alalibo
Virginia Bioinformatics Institute
June 4, 2009
What is Annotation

comments, notes, explanations, or other types of external remarks that can be attached to a document......

For genomics
functional annotation means attaching biological information to sequences
Road to Functional Annotation

- Structural Annotation
  - Searches
    - Nucleotide/Protein Databases
    - Domain/Motifs
  - Assignments
    - EC Number
    - Metabolic Pathways
    - Automated GO
    - Paralogous Families
- Manual curation
Road to Functional Annotation

- Structural Annotation
  - Searches
    - Nucleotide/Protein Databases
    - Domain/Motifs
  - Assignments
    - EC Number
    - Metabolic Pathways
    - Automated GO
    - Paralogous Families
- Manual curation
Automated Searches

• Search programs can be downloaded and run internally on Unix system
• Some programs have graphic user interfaces but normally takes limited sequences
Homology or similarity based searches

- **Local pairwise alignment** tools: look for any regions of similarity within the proteins that score well.
  - **BLAST**
    - fast

- **Global pairwise alignment** tools take two sequences and attempt to find an alignment of the two over their full lengths.
  - Needleman-Wunsch
    - finds best out of all possible alignments

- **Multiple alignments** tools try to align 3 or more proteins so that the maximal number of amino acids from each protein are matched in the alignment - this may or may not include the full length of some or all of the proteins
  - clustalW
BLAST Programs

- **Blastn**: Search a nucleotide database using a nucleotide query
- **BlastP**: Search protein database using a protein query
- **Blastx**: Search protein database using a translated nucleotide query
- **Tblastn**: Search translated nucleotide database using a protein query
- **Tblastx**: Search translated nucleotide database using a translated nucleotide query
Example of BLAST output

top row is the search protein (query) and the bottom row is the match protein (subject).
Middle row is consensus
+ indicates similar amino acids
numbers indicate amino acid position in the sequence
Domain Search

Hidden Markov Models

- Statistical models of the primary structure consensus of a sequence family
HMMer
Pfam

• Large collection of protein families represented by multiple sequence alignments and HMMs
• Analyze protein sequences for Pfam match
• Look at multiple alignments of members of the gene family
Pfam 23.0 (July 2008, 10340 families)

The Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs). More...
## Significant Pfam-A Matches

Show or hide all alignments.

<table>
<thead>
<tr>
<th>Pfam-A</th>
<th>Description</th>
<th>Entry type</th>
<th>Sequence</th>
<th>HMM</th>
<th>Bits score</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>NPP1</td>
<td>Necrosis inducing protein (NPP1)</td>
<td>Family</td>
<td>2</td>
<td>237</td>
<td>1</td>
<td>289</td>
</tr>
</tbody>
</table>
INTERPRO

- Database of protein families, domains and sites with identified in known proteins which can be applied to new protein sequences
- Collects protein families from other databases such as Pfam, UniProtKb and TIGRFAMs
- Sequence search is done with InterProScan
  - downloadable (runs faster on own server, large set)
  - GUI (limited number of sequences)
INTERPRO SEARCH

Download Software

APPLICATIONS TO RUN

- BlastProDom
- HMMtigr
- TMHMM
- FPrintScan
- ProfileScan
- ScanRegExp
- PatternScan
- HMMPIR
- HMMFam
- HMMSmart
- SuperFamily
- SignalPHMM
- HMMPanther
- Gene3D

Enter or Paste a PROTEIN Sequence in any format:

Upload a file: Browse... Submit Job Reset Help
Sample InterPro Output

A lot more information can be found on the site
Other Protein Family Resources

- Prosite
- UniProtKb
- TIGRFAMs
Subcellular localization

- Signal P: Predicts the presence and location of signal peptide and cleavage sites in organism
- TMHMM: Predicts transmembrane
- TargetP: Predicts subcellular location based on chloroplast transit peptide and mitochondrial targeting sequence
8.0 Server

Server predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms: Gram-positive, Gram-negative, prokaryotes, and eukaryotes. The method incorporates a prediction of cleavage sites and a signal peptide/non-signal peptide based on a combination of several artificial neural networks and hidden Markov models.

See history of this server. All the previous versions are available online, for comparison and reference.

But using SignalP and other protein subcellular localization prediction methods:

"Proteins in the cell using TargetP, SignalP, and related tools"


For other and supplementary information see here.

### FASTA Sequence Entry

Enter a single sequence or several sequences in FASTA format into the field below:

```
sojae protein
vagaafpvdt dasqlgvdllk kaikaenamt ftgdakdlql flakgpdvdes
psaeemkees fkwlpdehra alklvegesd dyihaltage pilsgktlttt
```

### FASTA Format Directly from Your Local Disk

Upload your sequence in FASTA format directly from your local disk.

### Method

- Neural networks
- Hidden Markov models
- Both

### Graphics

- No graphics
- GIF (inline)
- GIF (inline) and EPS (as links)

### Truncation

Truncate each sequence to max. 70 residues.

We recommend that only the N-terminal part of each protein sequence is submitted. Enter 0 (zero) to disable truncation.
Sample SignalP Output
CRN2...confirmedwithproteomics

SignalP-NN result:

SignalP-NN prediction (euk networks): Sequence

>Sequence
length = 70

# Measure  Position  Value  Cutoff  signal peptide?
max. C   23    0.474  0.32  YES
max. Y   23    0.383  0.33  YES
max. S   13    0.868  0.87  NO
mean S  1-22  0.493  0.48  YES
D 1-22  0.438  0.43  YES

# Most likely cleavage site between pos. 22 and 23: TDA-SQ
Sample SignalP Output
CRN2...confirmed with proteomics

Prediction: Signal peptide
Signal peptide probability: 0.982
Signal anchor probability: 0.000
Max cleavage site probability: 0.376 between pos. 15 and 16
Road to Functional Annotation

- Structural Annotation
  - Searches
    - Nucleotide/Protein Databases
    - Domain/Motifs
  - Assignments
    - EC Number
    - Metabolic Pathways
    - Automated GO
    - Gene Families
- Manual curation
Assign EnzymeCommission (EC) Number

http://ca.expasy.org/enzyme/

- Numerical classification for enzymes based on chemical reactions catalyzed
- For example, the tripeptide aminopeptidases have the code "EC 3.4.11.4", whose components indicate the following groups of enzymes:
  - EC 3 enzymes are hydrolases (enzymes that use water to break up some other molecule)
  - EC 3.4 are hydrolases that act on peptide bonds
  - EC 3.4.11 are those hydrolases that cleave off the amino-terminal amino acid from a polypeptide
  - EC 3.4.11.4 are those that cleave off the amino-terminal end from a tripeptide
- Given a name, EC number can be searched at
Search EC numbers
http://ca.expasy.org/enzyme/

ENZYME
Enzyme nomenclature database

ENZYME is a repository of information relative to the nomenclature of enzymes. It is primarily based on the recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB) and it describes each type of characterized enzyme for which an EC (Enzyme Commission) number has been provided [More details / References / Linking to ENZYME / Disclaimer].

Release of 26-May-2009 (4141 active entries)

Access to ENZYME

- by EC number: [] [] [] [] []  Search
- by enzyme class
- by description (official name) or alternative name(s):  Search
- by chemical compound
- by cofactor
- by search in comments lines
Road to Functional Annotation

Structural Annotation

Searches

Nucleotide/Protein Databases

Domain/Motifs

Assignments

EC Number

Metabolic Pathways

Automated GO

Gene Families

Manual curation
Metabolic Pathways

- Help improve annotation by showing missing genes in essential pathways
- Useful for comparative genomics

KEGG:
http://www.genome.jp/kegg/pathway.html

Reactome: http://www.reactome.org

Metacyc: http://www.metacyc.org

Add lots of others
Road to Functional Annotation

Structural Annotation

Searches

Nucleotide/Protein Databases

Domain/Motifs

Assignments

EC Number

Metabolic Pathways

Automated GO

Gene Families

Manual curation
The Gene Ontology provides

- A controlled /universal/uniform vocabulary/language to describe
- the attributes of a gene or gene product in any organism
- Annotations with GO terms are associated with appropriate evidence codes
- These annotations are computer searchable
GO is represented by three ontologies

- **Molecular function**: What a gene product does
- **Biological process**: Why it does what it does
- **Cellular component**: where it does what it does
Automated GO Assignments

- Transfer of GO terms based on significant BLAST hits to sequences in genomes with manually curated GO terms
- Mapping between GO terms and biological concepts
- Eg Uniprot, EC number and Reactome Pathways
Gene Families

- Use evolutionary relationships to facilitate functional annotation
- Identify possible gene structure problem
Gene Families - example

• Just by duplication or vertical descent???

• Paralogous families???
Road to Functional Annotation

Structural Annotation

Searches

Nucleotide/Protein Databases

Domain/Motifs

Assignments

EC Number

Metabolic Pathways

Gene Families

Automated GO

Manual curation
What next

- All these tools can be configured into a pipeline
  Or
  Done separately with each software

Store data for each gene
  Done differently for different database
Data Summary /Storage for genes on the VBI Microbial Database

• Look for one with most details and put
Why manual Annotation

Combine all search information and evidence
Manually look through all information
Add experimental data from literature when available
Approach conservatively

Setback
Time-consuming and more expensive.
• All these are informative but there are discrepancies that may make comparative analyses especially across taxons very difficult

• Eg idiosyncrasies.....adhesion, attachment etc
• So the Gene Ontology