Functional Annotation

May 23, 2007

Rama Maiti
Functional Annotation Overview

- What is annotation
- Steps we take to annotate eukaryotic genes
- Software tools we use for functional annotation
- Steps we take to manually annotate or verify an automated annotation
What are the questions?

• How did the gene get its structure and name?
• Does it really have a function assigned to it?
• Where did this information come from?
• Is it accurate? Can you rely on it?
What is Functional Annotation?

• “To annotate” is “to make or furnish critical or explanatory notes or comments”

• For genomics the ‘notes’ are about
  • Names of the gene products
  • Functions of genes within an organism

• Elements of the functional annotation process
  • Validation of the gene structure
  • Literature search, if any is available
  • Homology / domain searches
  • Assignment of function
  • Maintenance of data availability
The Annotation Pipeline

Sequencing

Automated Annotation
- protein searches (nap)
- EST, cDNA alignments (gap2)
- custom database searches (nap/gap2)
- gene prediction algorithms
- Blastp searches
- HMM searches
- SignalP/TargetP/Interpro

Manual Annotation
- Human intervention:
  - critical evaluation of automated assignments

Manual Structural Curation

Manual Functional Curation
Manual vs. Automated Annotation

- Automated Annotation is complicated by high volumes of data derived from different methods at different centers.
- High quality annotation requires manual review and intervention.
Steps in Functional Annotation

Manual curation

Functional Assignment

Verify Structural Annotation

- BLAST
- Domains
- Motifs
- Protein families

{ Analyze results of automated annotation }

GO, EC Number, Metabolic Pathways
Steps in Functional Annotation

• Analyze the gene structure (Annotation Station or preferred gene viewer)
• Name the gene product (Manatee)
  – requires analysis of the gene product
  – gene product name is primarily homology based on different bioinformatics tools
• Assign Gene Ontology terms
  – Process
  – Function
  – Component
Homology Searching
(Tools that are available to characterize a sequence)

- **WU BLAST** [http://blast.wustl.edu/](http://blast.wustl.edu/) with links to many servers
- **Pfam profiles** (profiles, or HMMs) [http://pfam.wustl.edu/](http://pfam.wustl.edu/)
- **TIGRFAMS** (profiles, or HMMs) [http://tigrblast.tigr.org/web-hmm/](http://tigrblast.tigr.org/web-hmm/)
- **Interpro** (families) [http://www.ebi.ac.uk/InterProScan/](http://www.ebi.ac.uk/InterProScan/)
- **TmHMM** (transmembrane domain) [http://www.cbs.dtu.dk/services/TMHMM/](http://www.cbs.dtu.dk/services/TMHMM/)
- **SignalP** (signal peptide cleavage sites) [http://www.cbs.dtu.dk/services/SignalP/](http://www.cbs.dtu.dk/services/SignalP/)
- **TargetP** (subcellular location) [http://www.cbs.dtu.dk/services/TargetP/](http://www.cbs.dtu.dk/services/TargetP/)
- **Protein families and clustering**
  - **TIGR Paralogous Families** (not yet available outside of TIGR)
  - **TribeMCL** [http://www.ebi.ac.uk/research/cgg/tribe/](http://www.ebi.ac.uk/research/cgg/tribe/)
Manatee

- Manatee is a web-based gene evaluation and genome annotation tool.
- Manatee displays the current annotation for prokaryotic and eukaryotic genomes.
- Manatee is an open source software available at:
  http://sourceforge.net/projects/manatee/

make high quality functional assignments using genome analyses tools. These tools consist of, but are not limited to GO classifications, blast search data, protein families.
Verify evidence from automated annotation

- BLAST matches
- HMM
- Prosite, Interpro classifications
- Motifs
- Signal Sequence
- Target Sequence
- Transmembrane domain
- Protein families
Functional annotation

Examine the gene structure
- does it make sense with respect to the alignments?
- do you need to re-curate the gene structure?

Name the gene product
- Determine whether it is published,
- Fully characterized? Give it the Swiss-Prot name.
- Sequenced but not characterized? Look at the evidence.

Add comments to comment field
- explain reasoning for others
- add personal communication information
- make comments about function or process

In many occasions after analyzing our data and make a decision about a gene function, we may need to go back and re-examine the gene structure.
Use all possible resources...
Example:-

A protein sequence from *Trypanosoma brucei*. Our task will be to annotate this protein sequence as fully as possible, given the tools at hand.

```
>unknown_T._brucei_protein_sequence
MLRRLGVRHFRRTPLLFVGGDGIFSERY
TEIDNSNERRINALKGCGMFEDEWIA
TEKVHGANFIEGEKGMIRYAKRSGIMP
PNEHFFGYHILPELQRYITSIREMLCEK
QKKKLHVVLINGELFGGKYDHPSPKT
RKTVMVAGKPRTISAVQTDSPFQYSPDL
HFYAFDIKYKETEDGDYTTL VyDEAiEL
FQRVPGLLYARAVIRGPMKavaADVE
RFVTTIPPLVGMGNYPLTGNWAEGLVV
KHSLGMAGFDPKGPTVLKFKCTAFQE
ISTDRAQGPRVDEMNRVRRDSINRAGVQ
LPDLESIVQDPIQLEASKLLNHVCENiRL
KNVLGSKICTEPEFEKEMTPDQLATLLAK
DVLKDFLKDTEPSIVNIPLIRKDLTRYV
IFESRRLVCSQWKDLIKRQSPDFSE*
```
Verify the gene structure
NCBI BLAST


<table>
<thead>
<tr>
<th>Program</th>
<th>Database</th>
<th>Query</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLASTN</td>
<td>Nucleotide</td>
<td>Nucleotide</td>
</tr>
<tr>
<td>BLASTP</td>
<td>Protein</td>
<td>Protein</td>
</tr>
<tr>
<td>BLASTX</td>
<td>Protein</td>
<td>Nucleotide → Protein</td>
</tr>
<tr>
<td>TBLASTN</td>
<td>Nucleotide → Protein</td>
<td>Protein</td>
</tr>
<tr>
<td>TBLASTX</td>
<td>Nucleotide → Protein</td>
<td>Nucleotide → Protein</td>
</tr>
</tbody>
</table>

Read → as “translated to”

TIGR Rice Workshop
BLAST: What makes a good alignment?

It depends on what you are trying to prove!

• minimum of 30% identity, better 40% & up
  – higher for short proteins
  – score is weighted for length

• full length match
  – at least 70% of both proteins

See explanation of BLAST scores on slide 56.
Example: run NCBI BLAST

BLASTP – protein against protein

Results:

The first hit in the BLASTP output, a 100% match, is to a genome project submission, which means that the entry is not
Example: navigating BLAST output

The alignment reveals three positions with variations:

I103V (very similar, both hydrophobic) conservative

D182G (negative, hydrophilic to tiny polar) non-conservative

V364A (nonpolar, aliphatic, hydrophobic to tiny, nonpolar, aliphatic) conservative

See Glossary entry for SNP
Our sequence is 99% identical to the sequence of this Swiss-Prot entry.

Another name for this protein in the literature is ‘REL2.’

<table>
<thead>
<tr>
<th>ENTRY INFORMATION</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ENTRY NAME</td>
<td>TB48 TRYBB</td>
</tr>
<tr>
<td>ACCESSION NUMBER</td>
<td>P62264</td>
</tr>
<tr>
<td>Integrated into Swiss-Prot on</td>
<td>2004-05-10</td>
</tr>
<tr>
<td>Sequence was last modified on</td>
<td>2001-03-01 (Sequence version 1)</td>
</tr>
<tr>
<td>Annotations were last modified on</td>
<td>2006-10-31 (Entry version 24)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NAME AND ORIGIN OF THE PROTEIN</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>PROTEIN NAME</td>
<td>RNA-editing ligase TbMP48, mitochondrial precursor</td>
</tr>
<tr>
<td>Synonyms</td>
<td>EC 6.5.1.3 RNA ligase</td>
</tr>
<tr>
<td>GENE NAME</td>
<td>MP48</td>
</tr>
<tr>
<td>SOURCE ORGANISM</td>
<td>Trypanosoma brucei brucei</td>
</tr>
<tr>
<td>TAXONOMY ID</td>
<td>5702 [NCBL, NEWT]</td>
</tr>
<tr>
<td>LINEAGE</td>
<td>Eukaryota, Euglenozoa, Kinetoplastida, Trypanosomatidae, Trypanosoma</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>REFERENCES</th>
<th></th>
</tr>
</thead>
</table>
Association of two novel proteins TbMP52 and TbMP48 with the Trypanosoma brucei RNA editing complex. 

<table>
<thead>
<tr>
<th>COMMENTS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>FUNCTION</td>
<td>Part of the RNA editing complex essential for cell variability. RNA editing in kinetoplastid mitochondria inserts and deletes uridylates at multiple sites in pre-mRNAs as directed by guide RNAs</td>
</tr>
<tr>
<td>CATALYTIC ACTIVITY</td>
<td>ATP + (ribonucleotide)(n) + (ribonucleotide)(m) = AMP + diphosphate + (ribonucleotide)(n+m)</td>
</tr>
<tr>
<td>SUBCELLULAR LOCATION</td>
<td>Mitochondria</td>
</tr>
</tbody>
</table>
**ENTRY INFORMATION**

<table>
<thead>
<tr>
<th><strong>ENTRY NAME</strong></th>
<th>TB48 TRYBB</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ACCESSION NUMBER</strong></td>
<td>P92864</td>
</tr>
<tr>
<td><strong>Integrated into Swiss-Prot on</strong></td>
<td>2004-05-10</td>
</tr>
<tr>
<td><strong>Sequence was last modified on</strong></td>
<td>2001-03-01 (Sequence version 1)</td>
</tr>
<tr>
<td><strong>Annotations were last modified on</strong></td>
<td>2006-10-31 (Entry version 24)</td>
</tr>
</tbody>
</table>

**NAME AND ORIGIN OF THE PROTEIN**

| **PROTEIN NAME** | RNA-editing ligase TbMP48, mitochondrial precursor |
| **Synonyms** | EC 6.5.1.3, RNA ligase |
| **GENE NAME** | MP48 |
| **SOURCE ORGANISM** | Trypanosoma brucei brucei |
| **TAXONOMY ID** | 5702 [NCBI, NEWT] |
| **LINEAGE** | Eukaryota, Euglenozoa, Kinetoplastida, Trypanosomatidae, Trypanosoma |

**REFERENCES**


**COMMENTS**

| **FUNCTION** | Part of the RNA editing complex essential for cell viability. RNA editing in kinetoplastid mitochondria inserts and deletes uridylates at multiple sites in pre-mRNAs as directed by guide RNAs |
| **CATALYTIC ACTIVITY** | ATP + (ribonucleotide)(n) + (ribonucleotide)(m) = AMP + diphosphate + (ribonucleotide)(n+m) |
| **SUBCELLULAR LOCATION** | Mitochondria |
Pubmed

- Read abstract
- If promising, read paper to be sure protein is characterized
- If characterized, it is good evidence for naming our sequence
This is a very positive hit to the RNA ligase RNL2 family domain (TIGR02307).
 Verify HMM

Total score: 923.1
Trusted cutoff: 100.0
Gathering cutoff: 100.0
Noise cutoff: -165.0

Score is well above the trusted cutoff.

Our sequence contains an RNA ligase, Rnl2 family domain, with a very strong match. Members of this TIGRfam family ligate RNA.
Non-secretory protein

See Glossary entry for Signal Peptide

SignalP-HMM result:

SignalP-HMM prediction (euk models): unknown

# data

>unknown
Prediction: Non-secretory protein
Signal peptide probability: 0.008
Signal anchor probability: 0.009
Max cleavage site probability: 0.006 between pos. 22 and 23
The sequence contains a mitochondrial targeting peptide, mTP.

### TargetP 1.1 prediction results

<table>
<thead>
<tr>
<th>Name</th>
<th>Len</th>
<th>mTP</th>
<th>SP</th>
<th>other</th>
<th>Loc</th>
<th>RC</th>
</tr>
</thead>
<tbody>
<tr>
<td>unknown_Tb_seq</td>
<td>416</td>
<td>0.728</td>
<td>0.070</td>
<td>0.209</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>cutoff</td>
<td></td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Transmembrane domains

There are no transmembrane domains.

**TMHMM result**

HELP with output formats

- # unknown Length: 416
- # unknown Number of predicted TMHS: 0
- # unknown Exp number of AAs in TMHs: 0.00491
- # unknown Exp number, first 60 AAs: 0.00077
- # unknown Total prob of N-in: 0.00474
- unknown TMHMM2.0 outside 1 416

**TMHMM posterior probabilities for unknown**

![Graph showing TMHMM posterior probabilities for unknown protein](https://example.com/tmhmm_graph.png)
Annotation of Example Protein

**BLAST:** A protein match at Swiss-Prot is 99% identical, with 2 conservative and one non-conservative amino acid substitutions. “RNA-editing ligase TbMP48, mitochondrial precursor” is the Swiss-Prot name for this close protein match.

This mitochondrial precursor of an RNA ligase was identified as a member of a multi-protein complex that catalyzes deletion editing in vitro. It was isolated from an enriched sample of Trypanosoma brucei mitochondria by sequential ion-exchange and gel filtration chromatography, followed by glycerol gradient sedimentation. The protein was not functionally characterized, but was identified as a member of an RNA-editing complex. The complex was shown to have RNA-editing function. (PMID:11134327)

**Domain:** Our sequence contains an RNA ligase, Rnl2 family TIGRFAMs domain, with a very strong match. Members of this TIGRFam family ligate (seal breaks in) RNA.

**Signal sequence:** none

**Targeting Sequence:** It contains a mitochondrial targeting sequence.

**Under the standards of this annotation project, “RNA-editing ligase TbMP48, mitochondrial precursor,” is a suitable name.**
Evidence from homology searching

Compare sequences of unknown function to those of known function.

Shared sequence identity may imply shared function:-
• Full-length match with significant identity (>30%)
• Domains and motifs
• Binding sites
• Catalytic sites

But :
- there are occurrences where one amino acid substitution changes the function of an enzyme.
- synonymous or “silent” codon substitutions may result in functional differences.
- Mutations may result in modification or deletion of function.
- all functional assignments made by similarity should be considered tentative until confirmed by experiment.
Transitive annotation

Beware!

A is like B
B is like C
C is like D
D is NOT like A!

Take a conservative approach. Err on the side of missing homology rather than stretching weak data.
Gene Ontology

GO is…

- a **method** used to structure biological knowledge using a dynamic controlled vocabulary across organisms.
- a **database** containing a shared vocabulary of descriptive terms for the description of the molecular function, biological process and cellular component of gene products.
- The Gene Ontology Consortium™ is a **collaboration** among model genome organism databases.
Topics

• Reasons GO has been developed
• Nuts and bolts of GO
• Tools
• Searching GO
• Assigning terms
• GO Slims
The Basics

• GO is a controlled vocabulary
• GO has three aspects, or ontologies:
  – Molecular function
  – Biological process
  – Cellular component
• The 3 aspects refer to genes and gene products
The specificity of GO

There is a limit to how much information can be contained in the name of a protein. For example: “translation initiation factor 2 subunit”

GO terms assigned to this tell much more:

- GO:0003743 (MF) translation initiation factor activity
- GO:0005525 (MF) GTP binding
- GO:0006413 (BP) translational initiation
- GO:0005851 (CC) eukaryotic translation initiation factor 2B complex
The Gene Ontology is like a dictionary

Each concept has:

- a name
- a definition
- an ID number

**term**: transcription initiation

**id**: GO:0006352

**definition**: Processes involved in the assembly of the RNA polymerase complex at the promoter region of a DNA template resulting in the subsequent synthesis of RNA from that promoter.
GO terms

• A GO term, or ID, is attached to every function, process or component

• There are relationships between them

• Relationships are shown by a graph
  – Directed acyclic graph
  – Sometimes called a “tree”
GO Tools

GO tools are available at the GO Consortium:

Developed and maintained by GO:
AmiGO - Searching through terms and annotations
OBO-Edit - Editing and viewing the DAG

Many others developed independently, for:
Annotation
Gene expression/microarray data
GO Slims
AmiGO

The GO Browser
Filter results

Filter by ontology
Ontology
All
Biological Process
Cellular Component
Molecular Function

Filter Gene Product Counts
Data source
All
CGD
dictyBase
FlyBase

Set filters
Remove all filters

all : all [184843]

GO:0008150 : biological_process [139437]
  •  GO:0022610 : biological adhesion [1691]
  •  GO:0065007 : biological regulation [18316]
  •  GO:0009987 : cellular process [81676]
  •  GO:0032502 : developmental process [16502]
  •  GO:0043062 : extracellular structure organization and biogenesis [313]
  •  GO:0040007 : growth [3428]
  •  GO:0042592 : homeostatic process [1533]
  •  GO:0051179 : localization [20043]
  •  GO:0040011 : locomotion [458]
  •  GO:0051235 : maintenance of localization [200]
  •  GO:0008152 : metabolic process [53763]
  •  GO:0051704 : multi-organism process [1640]
  •  GO:0032501 : multicellular organismal process [2352]
GO information to include

Independent of interface, add:

GO ID
Evidence code
Reference
Qualifier

The date is an important part of the annotation.

In Manatee:
### Filling in the GO information

<table>
<thead>
<tr>
<th><strong>Function</strong></th>
<th><strong>Process</strong></th>
<th><strong>Component</strong></th>
<th><strong>Others</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Key code</td>
<td>Key code</td>
<td>Key code</td>
<td>TmHMM</td>
</tr>
<tr>
<td>catalytic</td>
<td></td>
<td></td>
<td>signalP</td>
</tr>
<tr>
<td>enzyme</td>
<td></td>
<td></td>
<td>targetP</td>
</tr>
<tr>
<td>metabolism</td>
<td>cytoplasm</td>
<td>integral to membrane</td>
<td></td>
</tr>
<tr>
<td></td>
<td>extracellular</td>
<td>nucleus</td>
<td></td>
</tr>
<tr>
<td></td>
<td>membrane</td>
<td>mitochondrion</td>
<td></td>
</tr>
</tbody>
</table>

### Other annotations:

- PMID: TIGR_Tba1:annot
- ISBN:
- CDB: TIGR_Tba1:
- InterPro
- UniProt/Swiss-Prot
- Pfam:
- PIR:
- UniProt/TriEMBL:
- SGD:
- SGDID:
- TAIR-gene:
- FB:
- Prosite:
- protein_id:
- GerProtEC:
- MG:
- RGD:
- UniProt:
- GDB:
Assigning a GO term

1) Read the literature, not just the abstract
2) Search for GO terms
3) Record the data
GO Annotations based on similarity

- Sequence or structure
  - Similarity to GO-annotated gene products
- Domains
- EC numbers
- Pathways
- Protein families
and many more…

Annotating by similarity

use the evidence code ‘ISS’—inferred from sequence or structural similarity.

enter the database ID of the entity used to infer similarity in the ‘With’ field.
IEA: Inferred from Electronic Annotation

IEA is used when no curator has checked the annotation to verify its accuracy.

Use when an annotation:
• is based on "hits" in sequence similarity searches, if they have not been reviewed by curators

• is transferred from database records, if not reviewed by curators

• that depend directly on computation or automated transfer of annotations from a database.
  – The actual method used (BLAST search, SwissProt keyword mapping, etc.) doesn't matter.
  – If the method is match-based, a valid database ID must be entered in the with column.
GO Slim

- cut-down versions of the ontologies
- useful summary of GO annotation
- versions of GO Sloims available
  - Eukaryotic GO slim
  - Plant GO slim
  - Yeast GO slim
Points to remember

• GO enables querying across annotations
• The GO Consortium website has documentation and lists available tools
• AmiGO is available online and as downloadable resource
• GO Slims summarize your annotation
• GO annotations are worth the trouble—they enhance the value of research
MANATEE-

- Navigation, inspection & curation of gene products
  - Gene/Gene products
  - GO Assignments

- Available at: