Annotation and the analysis of annotation terms.

Brian J. Knaus
USDA Forest Service
Pacific Northwest Research Station
Why annotate?

• Assess quality of assembly
• Characterize assembly
• Identify genes/suites of genes which are of a priori interest.
• Identify genes/suites of genes which have been experimentally determined to be of interest (i.e., significantly differentially expressed).
• Gene enrichment analysis (comparison of a set of genes of interest to a null set).
Quality of annotation

<table>
<thead>
<tr>
<th>Evidence code</th>
<th>Evidence code description</th>
<th>Source of evidence</th>
<th>Manually checked</th>
<th>Current number of annotations*</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDA</td>
<td>Inferred from direct assay</td>
<td>Experimental</td>
<td>Yes</td>
<td>71,050</td>
</tr>
<tr>
<td>IEP</td>
<td>Inferred from expression pattern</td>
<td>Experimental</td>
<td>Yes</td>
<td>4,598</td>
</tr>
<tr>
<td>IGI</td>
<td>Inferred from genetic interaction</td>
<td>Experimental</td>
<td>Yes</td>
<td>8,311</td>
</tr>
<tr>
<td>IMP</td>
<td>Inferred from mutant phenotype</td>
<td>Experimental</td>
<td>Yes</td>
<td>61,549</td>
</tr>
<tr>
<td>IPI</td>
<td>Inferred from physical interaction</td>
<td>Experimental</td>
<td>Yes</td>
<td>17,043</td>
</tr>
<tr>
<td>ISS</td>
<td>Inferred from sequence or structural similarity</td>
<td>Computational</td>
<td>Yes</td>
<td>196,643</td>
</tr>
<tr>
<td>RCA</td>
<td>Inferred from reviewed computational analysis</td>
<td>Computational</td>
<td>Yes</td>
<td>103,792</td>
</tr>
<tr>
<td>IGC</td>
<td>Inferred from genomic context</td>
<td>Computational</td>
<td>Yes</td>
<td>4</td>
</tr>
<tr>
<td>IEA</td>
<td>Inferred from electronic annotation</td>
<td>Computational</td>
<td>No</td>
<td>15,687,382</td>
</tr>
<tr>
<td>IC</td>
<td>Inferred by curator</td>
<td>Indirectly derived from experimental or computational evidence made by a curator</td>
<td>Yes</td>
<td>5,167</td>
</tr>
<tr>
<td>TAS</td>
<td>Traceable author statement</td>
<td>Indirectly derived from experimental or computational evidence made by the author of the published article</td>
<td>Yes</td>
<td>44,564</td>
</tr>
<tr>
<td>NAS</td>
<td>Non-traceable author statement</td>
<td>No <code>source of evidence</code> statement given</td>
<td>Yes</td>
<td>25,656</td>
</tr>
<tr>
<td>ND</td>
<td>No biological data available</td>
<td>No information available</td>
<td>Yes</td>
<td>132,192</td>
</tr>
<tr>
<td>NR</td>
<td>Not recorded</td>
<td>Unknown</td>
<td>Yes</td>
<td>1,185</td>
</tr>
</tbody>
</table>


Also see: http://www.geneontology.org/GO.evidence.shtml
Monogenic (Mendelian) traits

<table>
<thead>
<tr>
<th>Character</th>
<th>Dominant trait</th>
<th>Recessive trait</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seed shape</td>
<td>Spherical</td>
<td>Wrinkled</td>
</tr>
<tr>
<td>Seed color</td>
<td>Yellow</td>
<td>Green</td>
</tr>
<tr>
<td>Flower color</td>
<td>Purple</td>
<td>White</td>
</tr>
<tr>
<td>Pod shape</td>
<td>Inflated</td>
<td>Constricted</td>
</tr>
<tr>
<td>Pod color</td>
<td>Green</td>
<td>Yellow</td>
</tr>
</tbody>
</table>

- Flower position: Axial, Terminal
- Stem height: Tall, Dwarf
Quantitative traits

Tobin/Dusheck, Asking About Life, 2/e
Figure 16.6

Number of individuals

Height in inches

Copyright © 2001 by Harcourt, Inc. All rights reserved.
Ontology

A structure of concepts or entity within a domain, organized by relationships.

A structured and controlled vocabulary.
Sequence Ontology: SO

http://www.sequenceontology.org/

‘terms and relationships used to describe the features and attributes of biological sequence.’ (E.g., binding_site, exon, etc.)

File formats:
- GFF3
- GVF
- OBO flat file
Gene Ontology: GO

http://www.geneontology.org/

‘standardizing the representation of gene and gene product attributes across species and databases’

Structured and controlled vocabularies.
• Biological process
• Cellular component
• Molecular function
Gene Ontology: GO
Directed, acyclic graph

Gene Ontology: GO
Basic Local Alignment Search Tool: BLAST


Use a protein database
• blastx – nucleotide 6-frame translation-protein
• blastp – protein-protein

• $E$-values – expectation value, analogous to p-values. No standard cut-off exists (i.e., $p < 0.05$).
• Hsp-length – length of match.

**Smith-Waterman algorithm**

<table>
<thead>
<tr>
<th>Δ</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>C</th>
<th>C</th>
<th>U</th>
<th>C</th>
<th>G</th>
<th>C</th>
<th>U</th>
<th>U</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Δ</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>A</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
</tr>
<tr>
<td>A</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.7</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.7</td>
</tr>
<tr>
<td>U</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.7</td>
<td>0.3</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>1.0</td>
<td>0.0</td>
</tr>
<tr>
<td>G</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.3</td>
<td>0.0</td>
<td>0.7</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.7</td>
<td>0.7</td>
<td>1.0</td>
</tr>
<tr>
<td>C</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>2.0</td>
<td>1.3</td>
<td>0.3</td>
<td>1.0</td>
<td>0.3</td>
<td>2.0</td>
<td>0.7</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>C</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.7</td>
<td>1.0</td>
<td>3.0</td>
<td>1.7</td>
<td>1.3</td>
<td>1.0</td>
<td>1.3</td>
<td>1.7</td>
<td>0.3</td>
<td>0.0</td>
</tr>
<tr>
<td>A</td>
<td>0.0</td>
<td>0.0</td>
<td>2.0</td>
<td>0.7</td>
<td>3.0</td>
<td>1.7</td>
<td>2.7</td>
<td>1.3</td>
<td>1.0</td>
<td>0.7</td>
<td>1.0</td>
<td>1.3</td>
<td>1.3</td>
</tr>
<tr>
<td>U</td>
<td>0.0</td>
<td>0.0</td>
<td>0.7</td>
<td>1.7</td>
<td>0.3</td>
<td>1.3</td>
<td>2.7</td>
<td>2.3</td>
<td>1.0</td>
<td>0.7</td>
<td>1.7</td>
<td>2.0</td>
<td>1.0</td>
</tr>
<tr>
<td>U</td>
<td>0.0</td>
<td>0.0</td>
<td>0.3</td>
<td>0.3</td>
<td>1.3</td>
<td>1.0</td>
<td>2.3</td>
<td>2.3</td>
<td>2.0</td>
<td>0.7</td>
<td>1.7</td>
<td>2.7</td>
<td>1.7</td>
</tr>
<tr>
<td>G</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.3</td>
<td>0.0</td>
<td>1.0</td>
<td>1.0</td>
<td>2.0</td>
<td>3.0</td>
<td>2.0</td>
<td>1.7</td>
<td>1.3</td>
<td>2.3</td>
</tr>
<tr>
<td>A</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.7</td>
<td>0.7</td>
<td>1.0</td>
<td>0.7</td>
<td>2.0</td>
<td>3.0</td>
<td>1.7</td>
</tr>
<tr>
<td>C</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.7</td>
<td>1.0</td>
<td>2.0</td>
<td>0.7</td>
<td>1.7</td>
<td>1.7</td>
<td>3.0</td>
<td>2.7</td>
<td>1.3</td>
<td>1.0</td>
</tr>
<tr>
<td>G</td>
<td>0.0</td>
<td>0.0</td>
<td>0.7</td>
<td>1.0</td>
<td>0.3</td>
<td>0.7</td>
<td>1.7</td>
<td>0.3</td>
<td>2.7</td>
<td>1.7</td>
<td>2.7</td>
<td>2.3</td>
<td>1.0</td>
</tr>
<tr>
<td>G</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.7</td>
<td>0.7</td>
<td>0.3</td>
<td>0.3</td>
<td>1.3</td>
<td>1.3</td>
<td>2.3</td>
<td>1.3</td>
<td>2.3</td>
<td>2.0</td>
</tr>
</tbody>
</table>

Fig. 1. H_{ij} matrix generated from the application of eqn (1) to the sequences A-A-U-G-C-C-A-U-U-G-A-C-G-G and C-A-G-C-C-U-C-G-C-U-U-A-G. The underlined elements indicate the trackback path from the maximal element 3-30.

† Zero need not be included unless there are negative values of s(a,b).

-G-C-C-A-U-U-G-
-G-C-C--U-C-G-

Smith-Waterman algorithm

When both sequences are associated a similarity is calculated (i.e., BLOSUM62, PAM-120 for proteins).

When an indel is inferred the similarity decreased by a weighted value.

The alignment begins at the maximal value and a traceback procedure goes until an element of zero is reached.

Exhaustive search
• Finds the best alignment
• Computationally expensive

BLAST


Use a protein database
• blastx – nucleotide 6-frame translation-protein
• blastp – protein-protein

• $E$-values – expectation value, analogous to p-values.
• Hsp-length – length of match.

BLAST

A heuristic search.
Create a matrix of similarities among all residues.
Sequence segment = contiguous stretch of residues.
Maximal segment pair (MSP) = highest scoring pair of identical length segments.
Sequences above a cutoff score are considered to match.

1. Compile list of high scoring words.
2. Scan for hits.
3. Extend hits.

Alternatives to BLAST

BLAT – blast like alignment tool:
  http://genome.ucsc.edu/FAQ/FAQblat.html#blat3

HMMER: http://hmmmer.janelia.org/
What to do with BLAST

• Search a model organism
  Arabidopsis
  Rice
  Drosophila
  Mouse

• Search database of conserved sequences
  Core eukaryotic genes: CEGs
  Conserved orthologous sequences (Asterids)
  Custom database: genes shared among Arabidopsis and rice
  NCBI non-redundant nucleotide database
Blast2go

http://www.blast2go.org/home

• Blastx of NCBI
• Mapping – if matching proteins have a GO term include it.
• Annotation – based on evidence code, similarity and GO weight

## InterPro Scan

<table>
<thead>
<tr>
<th>InterPro Match</th>
<th>Description</th>
<th>Query Sequence</th>
<th>InterPro Scan Graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPR009518</td>
<td>Photosystem II PsbX</td>
<td>1</td>
<td><img src="https://www.ebi.ac.uk/Tools/pfa/iprscan/graph.png" alt="Graph" /></td>
</tr>
<tr>
<td>IPR023431</td>
<td>Photosystem II PsbX, type 1 subfamily</td>
<td>MF_01386</td>
<td><img src="https://www.ebi.ac.uk/Tools/pfa/iprscan/graph.png" alt="Graph" /></td>
</tr>
<tr>
<td>noIPR</td>
<td>unintegrated</td>
<td>tmhmm</td>
<td><img src="https://www.ebi.ac.uk/Tools/pfa/iprscan/graph.png" alt="Graph" /></td>
</tr>
</tbody>
</table>

© European Bioinformatics Institute 2006-2011. EBI is an Outstation of the European Molecular Biology Laboratory.

---

http://www.ebi.ac.uk/Tools/pfa/iprscan/
KEGG: Kyoto Encyclopedia of Genes and Genomes

http://www.genome.jp/kegg/
GO term enrichment analysis
