

Expanding GO annotations with text classification

Nicko Goncharoff
Reel Two, Inc.

Linking literature to GO

- GO literature annotations – reference articles that illustrate the underlying concept of a GO Term
 - Not meant to be a comprehensive list of relevant literature
- Researchers could use a resource that expands amount of GO-related literature
 - Quick understanding of functions for unfamiliar genes and proteins

Linking Medline to GO

Reel Two's Gene Ontology Knowledge Discovery System already classifies Medline according GO

GO KDS
Gene Ontology Knowledge Discovery System

Home
Help

Search for GO term Search by keyword Classify document

Gene Ontology Terms:
cell-matrix adhesion [\[remove\]](#) **Limit to:**
 2 terms with no documents [\[show\]](#)

The binding of a cell to the extracellular matrix via adhesion molecules. [\[hide\]](#)

[\[tree\]](#)
[biological_process](#) > [cell_communication](#) > [cell_adhesion](#) > [cell-matrix_adhesion](#)

Keywords:

★=Molecular Function ★=Biological Process ★=Cellular Component

Results 1 to 10 of 1199 ([Next 10 Results >>](#))
 All 1199 documents are categorized by Gene Ontology Term "**cell-matrix adhesion**". Documents are ranked by matching GO Terms.

1) The human family of Deafness/Dystonia peptide (DDP) related mitochondrial import proteins. Jin H Kendall E Freeman TC Roberts RG Vetric DL GKT Medical School, Guy's Hospital, London, SE1 9RT, United Kingdom. The gene responsible for the human genetic neurodegenerative disorder DFN-1/MTS encodes a small protein known as deafness dystonia peptide (DDP). It bears a st... [open complete abstract](#)
 Gene Ontology Terms: ★★★★★(G) heavy metal binding ★★★★★(G) mitochondrial membrane ★★★★★(G) mitochondrial inner membrane ★★★★★(G) mitochondrial inner membrane translocase complex ★★★★★(G) cell communication ★★★★★(G) cell adhesion ★★★★★(G) cell-matrix adhesion ★★★★★(G) sensory perception ★★★★★(G) hearing ★★★★★(G) zinc binding ★★★★★(G) perception of external stimulus ★★★★★(G) perception of abiotic stimulus ★★★★★(G) perception of sound ★★★★★(G) response to external stimulus ★★★★★(G) response to abiotic stimulus ★★★★★(G) inner membrane ★★★★★(G) cell ★★★★★(G) cell growth and/or maintenance ★★★★★(G) mitochondrion ★★★★★(G) intracellular ★★★★★(G) metabolism ★★★★★(G) cytoplasm ★★★★★(G) mitochondrial translocation ★★★★★(G) mitochondrion organization and biogenesis ★★★★★(G) ligand binding or carrier ★★★★★(G) cytoplasm organization and biogenesis ★★★★★(G) organelle organization and biogenesis ... [\(show all\)](#)

2) Transcript encoded on the opposite strand of the human steroid 21-hydroxylase/complement component C4 gene locus. Morel Y Bristol J Gitelman SE Miller WL Department of Pediatrics, University of California, San Francisco 94143-0434. The gene encoding human adrenal steroid 21-hydroxylase (P450c21) and its highly similar pseudogene are duplicated in tandem with the tw... [open complete abstract](#)

Fully automated

Not interactive

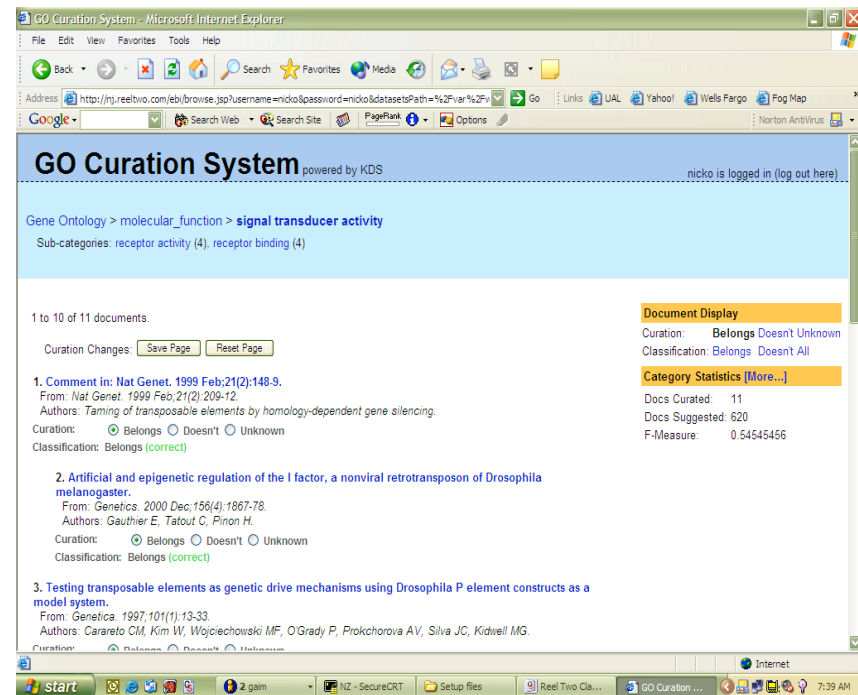
Not validated

The Next Step

- Researchers found GO KDS useful in some areas
- Reel Two expected category bias in training
- Reel Two wanted validation and user feedback
- Goal: a more accurate, user friendly research tool
- Next step: Collaboration with EBI and Flybase

Validation

- Reel Two developed an interactive version of GO KDS
- Curators from EBI and Flybase performed validation using the GO Slim taxonomy
- Validation took place over 10 days in November 2003



Results

Positive

- System performance generally quite accurate
- Top 100-200 articles were ~90-98% correct
- Poor performing categories typically improved with user input

Negative

- Certain biases found in categories with few training examples
- Accuracy often fell off by midway through predictions
- Sometimes required confirming or correcting several dozen predictions

Examples

- **cellular component > extracellular matrix (GO:0005578)**
 - Generally accurate - Some bias toward collagen.
- **molecular function > nucleotide binding (GO:0000166)**
 - High confidences correct, by ~50% many false positives. Bias to ABC transporters and ATP binding.
- **biological process > behavior (GO:0007610)**
 - High confidences incorrect due to strong bias toward HPRT.
- **biological process > cell homeostasis (GO: 0019725)**
 - No training data. Example of system learning.

Observations

- Confidence above 70-80%
- Confidence at ~40%
- Category bias
- Quick improvement
- Obsolete or changed GO terms
- Some GO terms probably too broad
- Possible bias in electronic annotation

Improving Training Data

- Training data used all go/gene_association files, including electronic GO annotations
 - May have led to bias
- Possible revised approach:
 - Filter out electronic annotations
 - Check for redundant Pubmed ID/GO term pairs
- Goal: more representative training set

Getting Results to Researchers

- Expand system to cover all of GO
- Interface for user feedback
- Interface for users' own curations
- System learns and improves – updates on weekly basis
- Could be deployed to research community via GO site by mid-2004, given funding



FlyBase

ReelTwo

Acknowledgements

Michael Ashburner, Rolf Apweiler, Daniel Barrell, Evelyn Camon, Emily Dimmer, Rebecca Foulger, Vivian Lee

Contact Information

Nicko Goncharoff
Reel Two, Inc.
2255 Van Ness Avenue, Suite 203
San Francisco, CA 94118 USA
+1-415-775-7630
nicko@reeltwo.com