



AUTOMATIC FUNCTIONAL ANNOTATION OF PUBCHEM BIOASSAYS

Rajarshi Guha¹, Julien Gobeill^{2,3}, Patrick Ruch³

¹ School of Informatics, Indiana University, United States of America

² Swiss-Prot research group, Swiss Institute of Bioinformatics, Geneva, Switzerland

³ University of Applied Science & University of Geneva, Switzerland

BACKGROUND and OBJECTIVES

The PubChem Bioassay collection is a set of 1293 assays that cover a wide range of compounds of different sizes (from 20 molecules to 200,000 molecules), and techniques (enzymatic, phenotypic etc.).

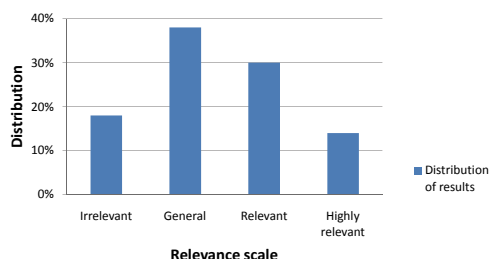
We designed an automated approach to associate each bioassays with a set of functional categories, in order to help search and navigate the assays.

DATA and METHODS

The description section of each assay record was sent to the Gene Ontology Categorizer web services. Designed to help text to GO curation, the categorizer typically outputs a ranked list of GO descriptors.

A sample (size = 30) of the resulting association was manually assessed, using a four-value scale : *highly relevant*, *relevant*, *relevant but general*, *irrelevant*.

RESULTS and NEXT STEPS



1. Relevant assignments massively overcome irrelevant associations: 82% vs 18%.

2. With the availability of formalized annotations, we plan to extend the interface to support more sophisticated similarity based searches as well as more intuitive visualizations of inter-assays relationships.

RESOURCE AVAILABILITIES

PubChem Bioassay - GO Annotations

Assay Selection [HELP]

Or select multiple assay ID's

- 580 Human H69AR Lung Tumor Ce...
- 581 Cathepsin G...
- 583 High Throughput Screening...
- 584 Promiscuous and Specific...
- 585 Promiscuous and Specific...
- 586 Dose-response cell-based...
- 587 qHTS Assay for Spectrosc...
- 588 qHTS Assay for Spectrosc...
- 589 qHTS Assay for Spectrosc...
- 590 qHTS Assay for Spectrosc...

GO Term [HELP]

GO Term Type [HELP]

- Function
- Component
- Pathway

GO Term Count [HELP]

5

Perform Query

A simple interface to the PubChem Bioassay collection, based on a local mirror of the bioassay collection augmented by GO terms predicted by GOCat (Peter Ruch). Since the GO terms were generated from an earlier version of the mirror, all the bioassays have not been annotated with GO terms.

Currently, searches based on GO terms or ID's are disabled. Also searches can only be performed using the assay ID rather than full text searches of the assay titles or descriptions.

3 557 559 561

AID 3 (NCI human tumor cell line growth inhibition assay. Data for the NCI-H226 Non-Small Cell Lung cell line)

Growth inhibition of the NCI-H226 human Non-Small Cell Lung tumor cell line is measured as a screen for anti-cancer activity. Cells are grown in 96 well plates and exposed to the test compound for 48 hours. Compounds are tested at 5 different concentrations and three endpoints are estimated from this dose response curve: GI50, concentration required for 50% inhibition of growth, TGI, the concentration requires for complete inhibition of growth, and LCS0, the concentration required for 50% reduction in cell number. These estimates are done by simple linear interpolation between the concentrations that surround the appropriate level. If a compound doesn't cause inhibition to the appropriate level, the endpoint is set to the highest concentration tested.

Term	ID	Type	Score
positive regulation of cell proliferation	0008284	p1	10000
protein amino acid binding	0005515	f1	8415
negative regulation of cell proliferation	0008285	p1	5103
negative regulation of cell growth	0030308	p1	3600
nucleus	0005634	c1	3413

Rajarshi Guha
Indiana University
Bloomington, IN 47403

The database and query interface are available at <http://goassay.rguha.net/index.html>

The GO Categorizer is available at <http://eagl.unige.ch/GOCat>