QUerying with Ontological Terminologies and their Annotations (QUOTA)

Yi Sun

Ontology Research Group
Miami University

OCWIC'2007
February 17,
Deer Creek Resort

Outline

- Introduction & Motivation
  - Ontology Overview
  - Basic Concepts
  - Why Query Annotations?
- Related Research
  - GeneInfoViz
  - GO Categorizer
  - eVOC
- QUOTA

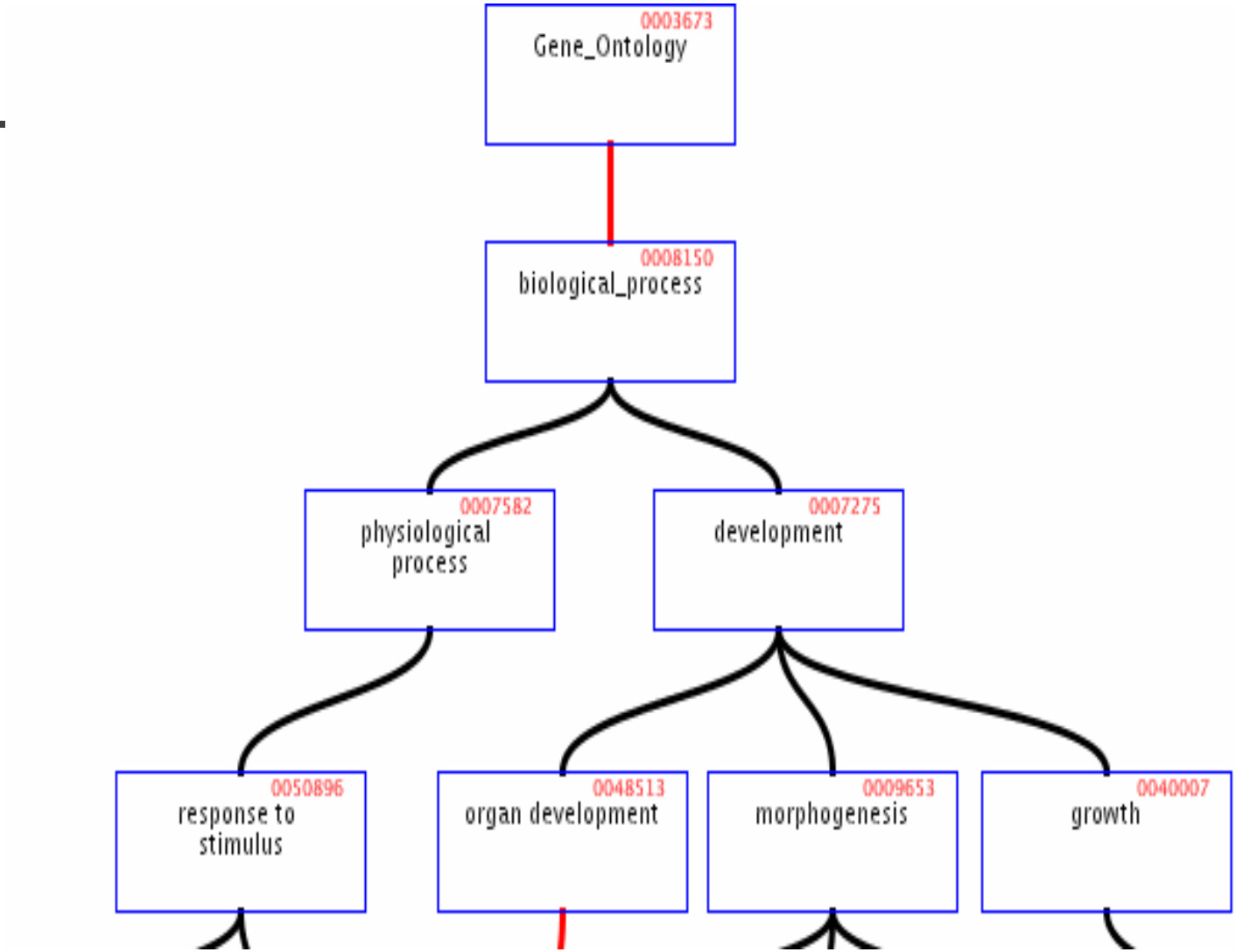
Ontology Overview

- What is Ontology?
  - Conceptualization
- What's in Ontology?
  - Concepts
  - Definitions of concepts
  - Relationships among concepts
- Why Ontology?

Gene Ontology (GO)

- Developed by the Gene Ontology Consortium (www.geneontology.org)
- GO Components
  - Cell Component (CC)
  - Molecular Function (MF)
  - Biological Process (BP)

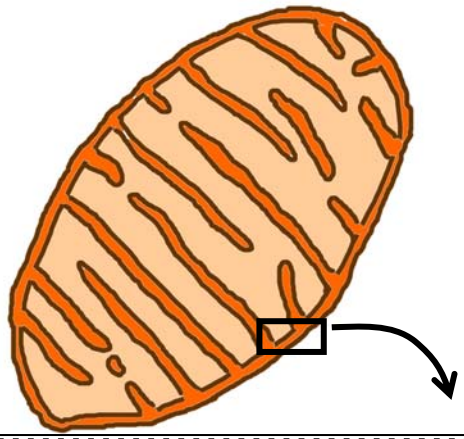
GO Example

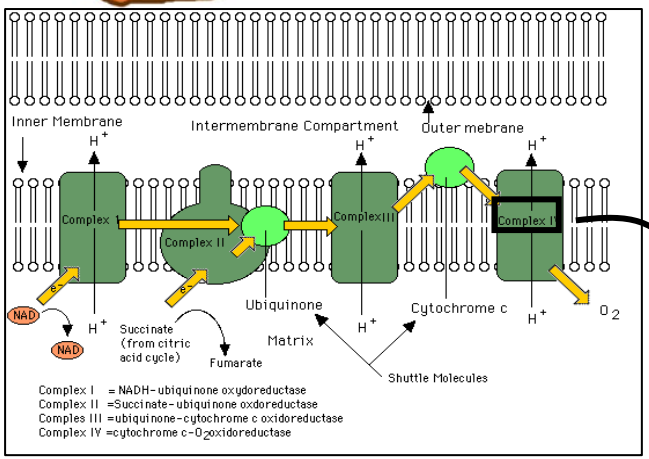

0003673
Gene_Ontology
0008150
biological_process
0007582
physiological process
0007275
development
0050896
response to stimulus
0048513
organ development
0009653
morphogenesis
0040007
growth

Annotation

- What is Annotation?
  - Association between product items and Ontology terms
- Multiple Annotations
  - e.g. in GO:
    - Gene → located in one or more cell components
    - Gene → perform one or more molecular functions
    - Gene → be active in one or more biological processes

Example: Mitochondrial P450





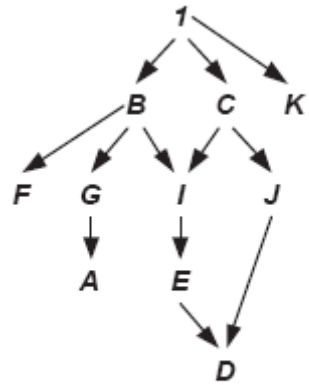
substrate + O_2 = CO_2 +H_20 product

Cellular component:
mitochondrial inner membrane
GO:0005743

Biological process:
Electron transport
GO:0006118

Molecular function:
monooxygenase activity
GO:0004497

Basic Concepts used in Research



- Interval rank
  - Improvement on Level
  - **Interval_rank (n) = (distRoot(n)+distBottom(n))/2**

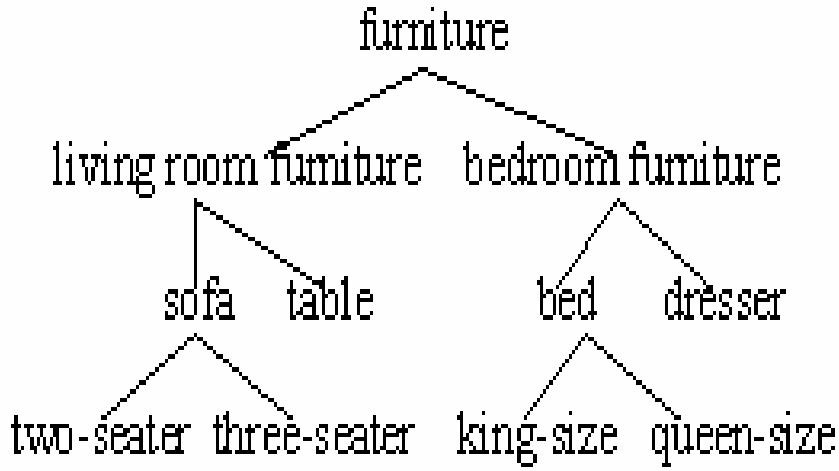
- Information Content
  - Metric proposed by Seco, Veale and Hayes (2004)
  - **IC(n) = log [(hypo(n) +1)/max]/ log (1/ max)**

hypo(n) is the number of hyponyms (descendents)

max is the maximum number of concepts in the taxonomy

- Semantic similarity

Semantic Similarity



- Use a value (similarity) to determine closeness of two objects (concepts)
- Distance Based:
  - *sim (c1, c2)* =2*len(root, c3)/(len(c1, c3) + len(c2, c3) +2* len(root, c3))

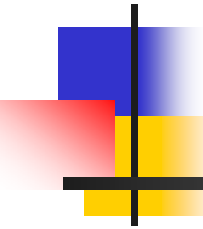
where c3 is the lowest common subsumer (LCS)

- Information Content Based:
  - *sim(c1,c2)* = 2*log IC(c3) / [log IC(c1)+log IC(c2)]

Why Query Annotations?

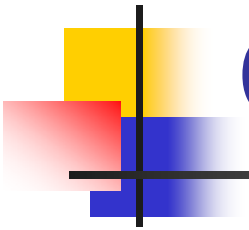
- Retrieve information and make prediction based on current information
- For example, in GO:

  Given a set of genes, what are the terms that best represent these genes, what is the relationship among these genes?



Related Research

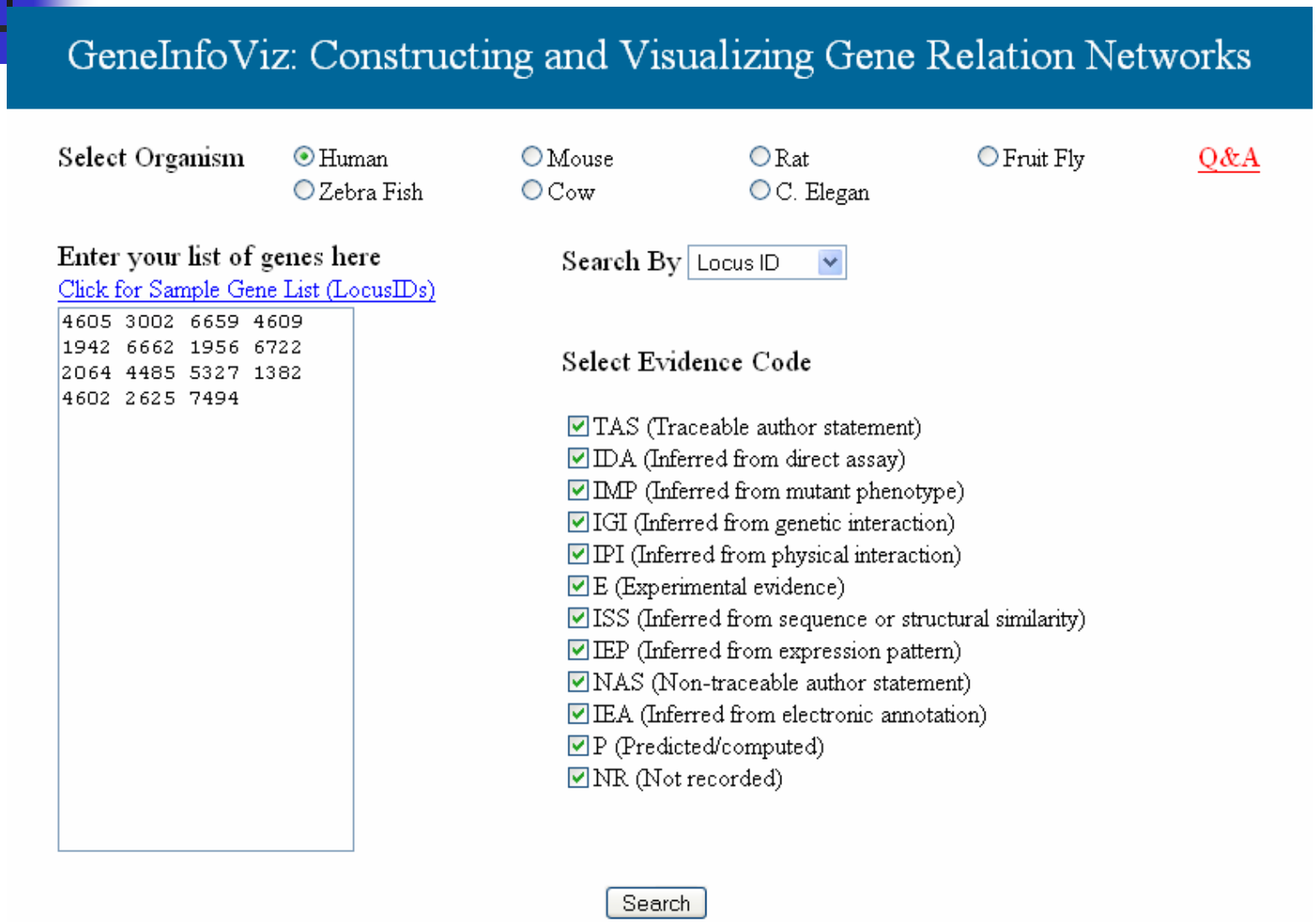
- GeneInfoViz
- GO Categorizer
  - eVOC



GeneInfoViz

- Implemented by Mi Zhou and Yan Cui, 2004
- Basic Functionality: Given a set of genes
  - 1: Display the GO terms that are annotated to these genes and highlight them in GO DAG
  - 2: Create relationship network between genes based on overlapping terms

GeneInfoViz: example



GeneInfoViz: Constructing and Visualizing Gene Relation Networks

Dynamic Visualization of Gene Relation Network Based On

Biological Process Level 1 Directed Acyclic Graph

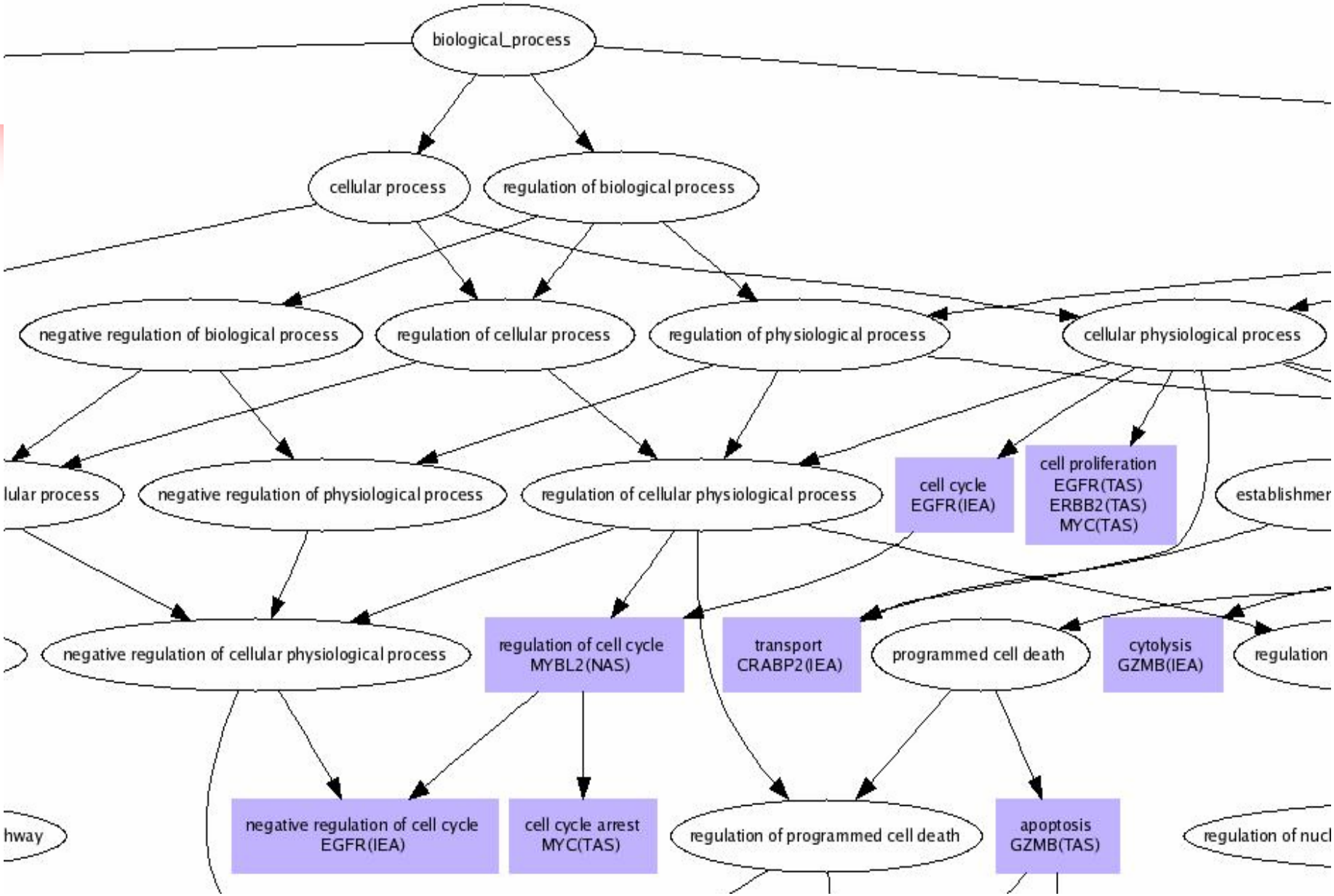
Molecular Function Level 1 Directed Acyclic Graph

Cellular Component Level 1 Directed Acyclic Graph

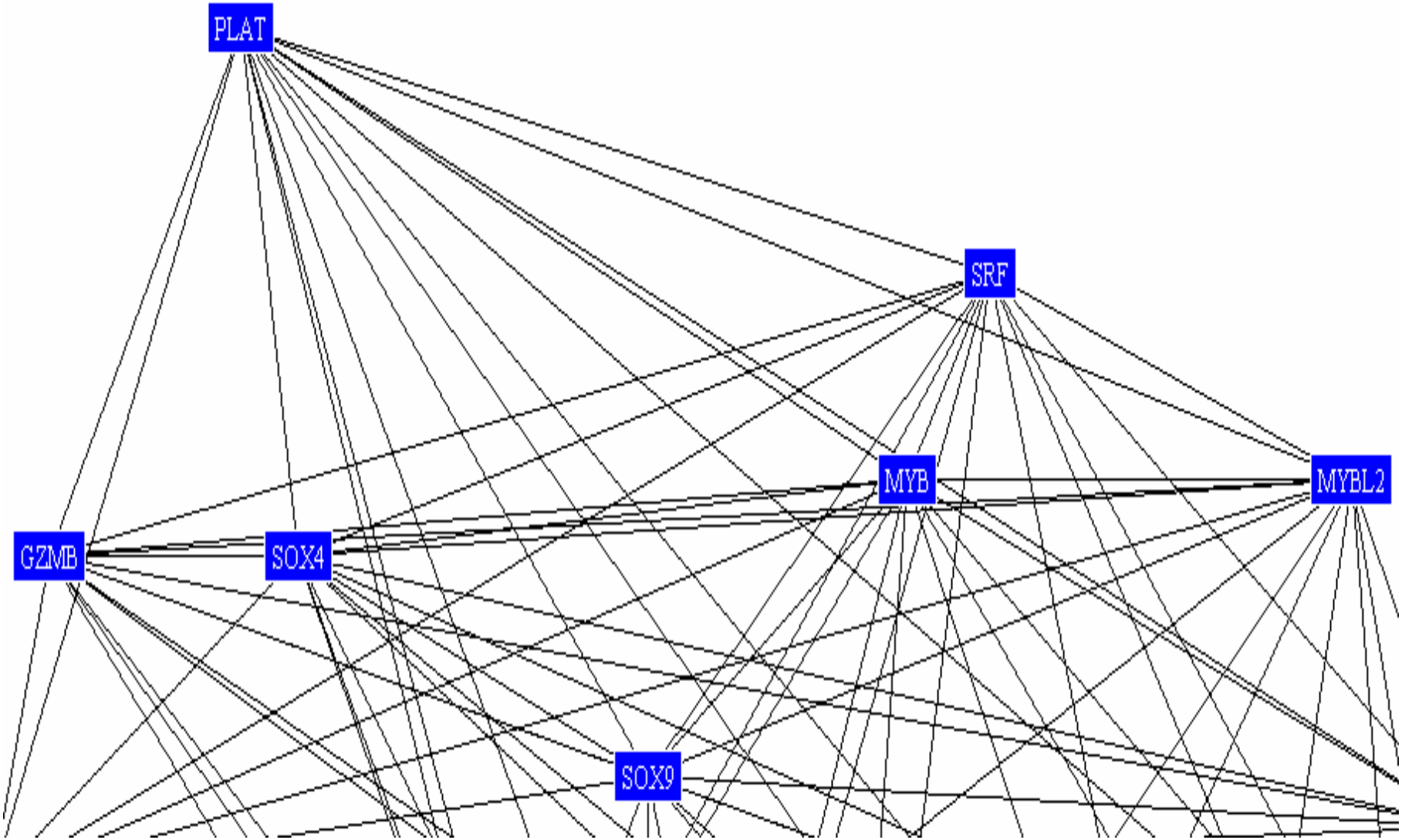
Note: Biological Process[P], Molecular Function[F], Cellular Component[C];

UniGeneID	LocusID	Symbol	Gene Name	Gene Ontology Term	Evid
Hs.405662	1382	CRABP2	cellular retinoic acid binding protein 2	lipid binding[F]	IEA
Hs.405662	1382	CRABP2	cellular retinoic acid binding protein 2	retinal binding[F]	IEA
Hs.405662	1382	CRABP2	cellular retinoic acid binding protein 2	transporter activity[F]	IEA
Hs.405662	1382	CRABP2	cellular retinoic acid binding protein 2	epidermis development[P]	TAS
Hs.405662	1382	CRABP2	cellular retinoic acid binding protein 2	regulation of transcription, DNA-dependent[P]	TAS
Hs.405662	1382	CRABP2	cellular retinoic acid binding protein 2	signal transduction[P]	TAS
Hs.405662	1382	CRABP2	cellular retinoic acid binding protein 2	transport[P]	IEA
Hs.516664	1942	EFNA1	ephrin-A1	ephrin receptor binding[F]	TAS
Hs.516664	1942	EFNA1	ephrin-A1	cell-cell signaling[P]	TAS
Hs.516664	1942	EFNA1	ephrin-A1	integral to plasma membrane[C]	TAS
Hs.516664	1942	EFNA1	ephrin-A1	membrane[C]	IEA
Hs.488293	1956	EGFR	epidermal growth factor receptor (erythroblastic l	ATP binding[F]	IEA
Hs.488293	1956	EGFR	epidermal growth factor receptor (erythroblastic l	epidermal growth factor receptor activity[F]	NR
Hs.488293	1956	EGFR	epidermal growth factor receptor (erythroblastic l	receptor activity[F]	IEA

Partial BP Branch of GO with Genes and Annotated Terms In Purple boxes



Gene relationship network based on overlap of annotating terms

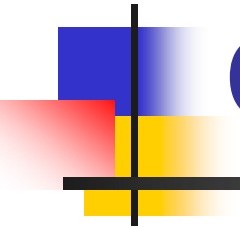


GO Categorizer

- Implemented by Cliff Joslyn and Susan Mniszewski et al in 2004
- Basic Functionality: Given a set of genes
  - Generate a ranked list of GO terms with values indicating how well they represent these genes
- Significance:
  - With this ranked list, we can easily tell if the terms best representing these genes, for example, cluster at one place or are evenly

eVOC

- Implemented by Janet Kelso et al in 2003
- Basic Functionality: Given a set of terms, return sets of genes that are annotated by each of the terms, and perform set operations based on the logical query.
- For example: Find the genes that are related to both “liver” and “cancer”
  - Set (“liver”) ∩ Set (“cancer”)



QUOTA

- Motivation & Objective
  - Querying Direction:
- In: Input to QUOTA
- Out: Output from QUOTA

1. IN: Set of Annotated Objects
OUT: Ontological Terms

- Similar to GeneInfoViz Function 1
- Generate the Annotation for each Object
- Provide various set options (AND, NOT, OR)

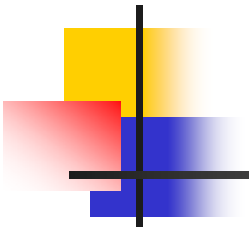
2. IN: Set of Annotated Objects OUT: Ranked Ontological Terms

- Originated from GO Categorizer
- Generate ranking score for each term
- Provide more options:
  - Interval rank
  - Information Content
  - Various Semantic Similarity measures

3. IN: Set of Annotated Objects
OUT: Similarity value matrix

- Generate similarity matrix among objects
- Based on Semantic Similarity values between each pair of annotating terms
- Various aggregation methods to integrate all SS values
- For example:
  - AO1→ t1,t3,t5; AO2→t2,t4,t6;

4. IN: Ontological Terms
OUT: Annotated Objects



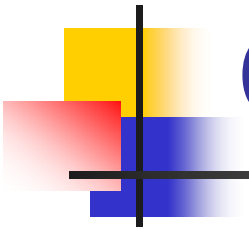
- Similar to eVOC
- Given a term, generate a set of objects annotated by this term
- Perform set operations (AND, NOT, OR)

5. IN: Two sets of Annotated Objects OUT: Semantic Similarity Value

- Input: Two sets of annotated objects
  - e.g. Two diseases, each one is potentially marked by a set of genes through some laboratory experiments
- Output: Semantic similarity between the two sets
  - e.g. A value indicating how similar two diseases are

5. IN: Two sets of Annotated Objects OUT: Semantic Similarity Value (cont…)

- Based on various Semantic Similarity measures
- Possible implementation:
  - For A={g1,g2,g3} and B ={g5,g6}
  - Calculate sim(g_i, g_j) for all pairs where g_i in A and g_j in B using various methods in Query 3
  - Use some aggregation function to determine an overall similarity between A and B



Generic Tool

- Generic
  - Works for all Ontologies and Annotation files with valid formats
- A translation program to convert Ontology into our formats has already been developed

Resources & Tools

- Gene Ontology & Annotation:
  - https: //www.geneontology.org
  - Formats:
    - GO: txt, OWL, XML
    - Annotation: XML, mySQL
- GeneInfoViz: https://www.genenet.org
- GO Categorizer: http://www.c3.lanl.gov/posoc/
- eVOC: info@egenetic.com

Thank you!