Finding Relationships Between Gene Products Using the Gene Ontology Database

Patrick Bradshaw, PhD
Bradshawp@health.missouri.edu
University of Missouri
LHNCBC, NLM, NIH

Different Methods for Finding Gene Product Similarity

- By Structural Methods
  - linear sequence homology
  - 3-dimensional comparison

- By Functional Methods (Analogy)
  - Gene Ontology annotations

Scientists Want to Find Gene Products Related Functionally

Common metabolic pathways researched:

Cell Death {neurodegenerative diseases}
Bacterial cell wall synthesis {antibiotics}
Uncontrolled Cell Cycle Progression {cancer}

Different sub-groups of genes in these pathways may be revealed using GO

Example of What Scientists May Learn from GO

- A large assembly containing 41 proteins called complex I of the respiratory chain is present in mammalian cells
  - Mutations in many of these proteins cause mitochondrial disease
- Querying GO may reveal that yeast contain a single protein with the same function that might be used in gene therapy to correct pathology associated with any of the 41 human proteins.

What is the Gene Ontology?

- A controlled vocabulary for molecular biology

- The GO terms are used as an annotation or index for gene products in other collaborating databases

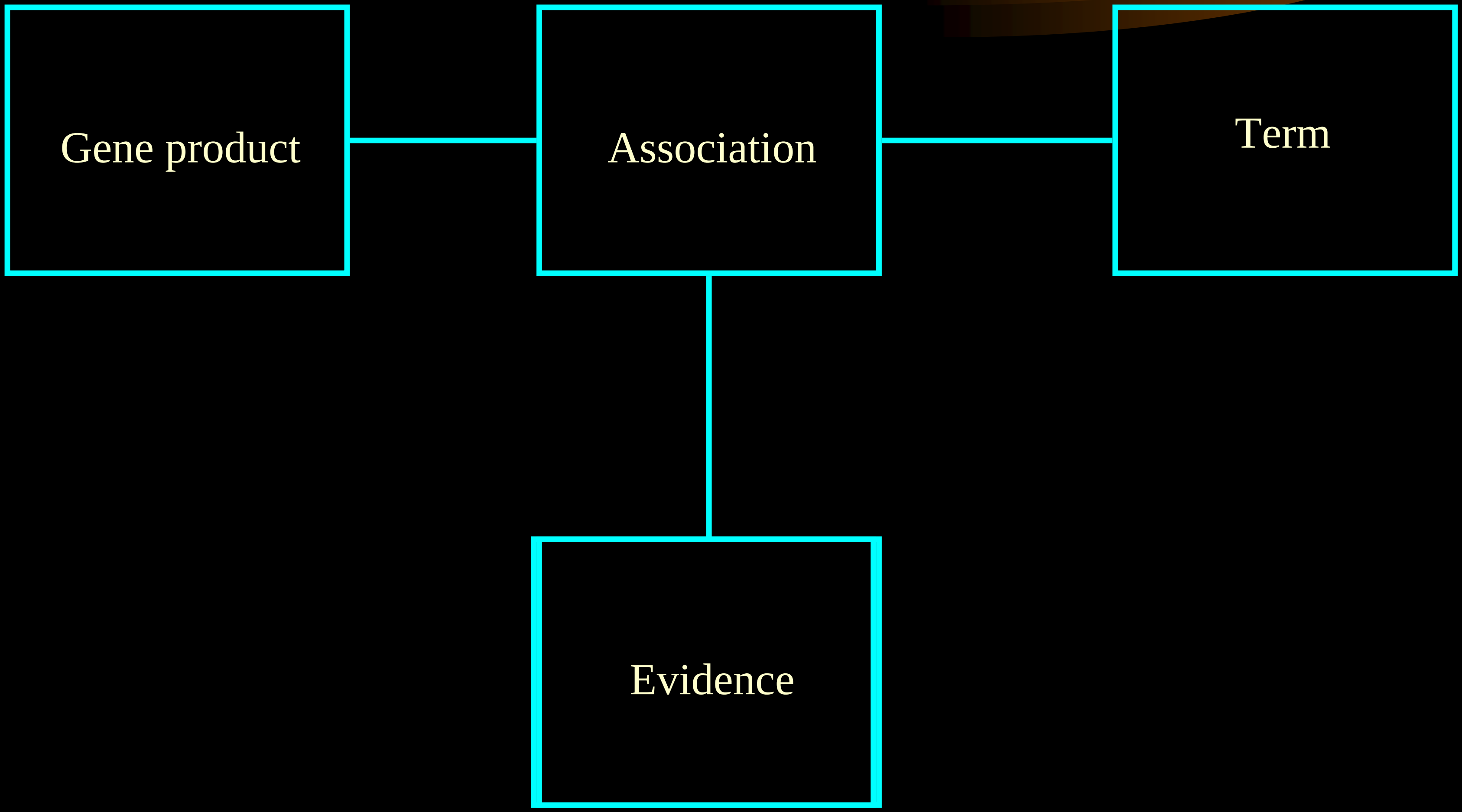
GO is Composed of 3 Branches of Concepts

- 1. Molecular Function
  i.e., Enzyme, Transporter, Trans. Factor

- 2. Cellular Component
  i.e., Nucleus, Mitochondria, ER, Golgi

- 3. Biological Process
  i.e., Transcription, Glycolysis, Cell Death

GO Database Diagram



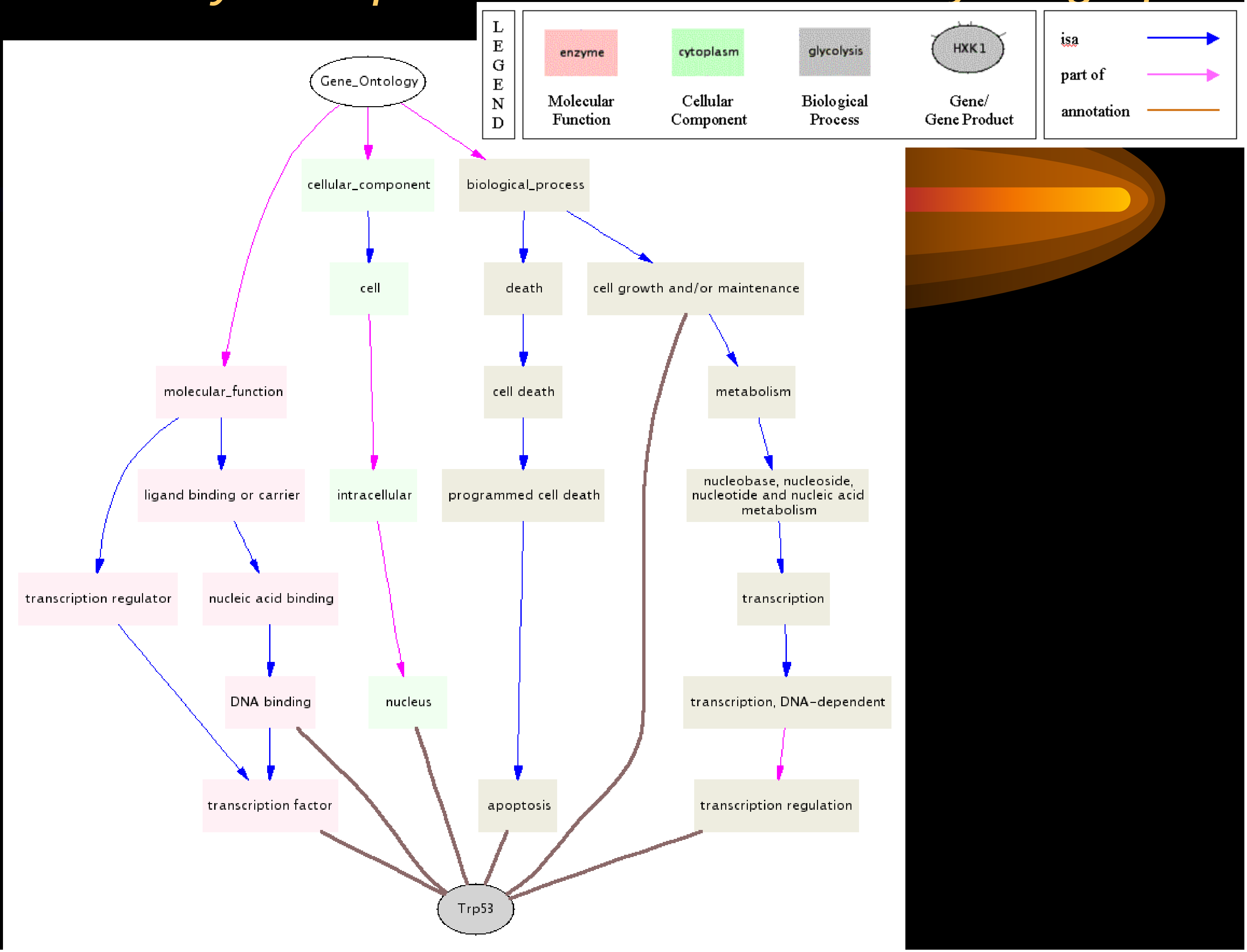
Gene product

Association

Term

Evidence

GO may be represented as directed acyclic graphs



Experimental Design: Finding Functional Identity

- For each gene in GO a list of the GO annotations (terms) was obtained.
- The genes described by exactly the same annotations were clustered.

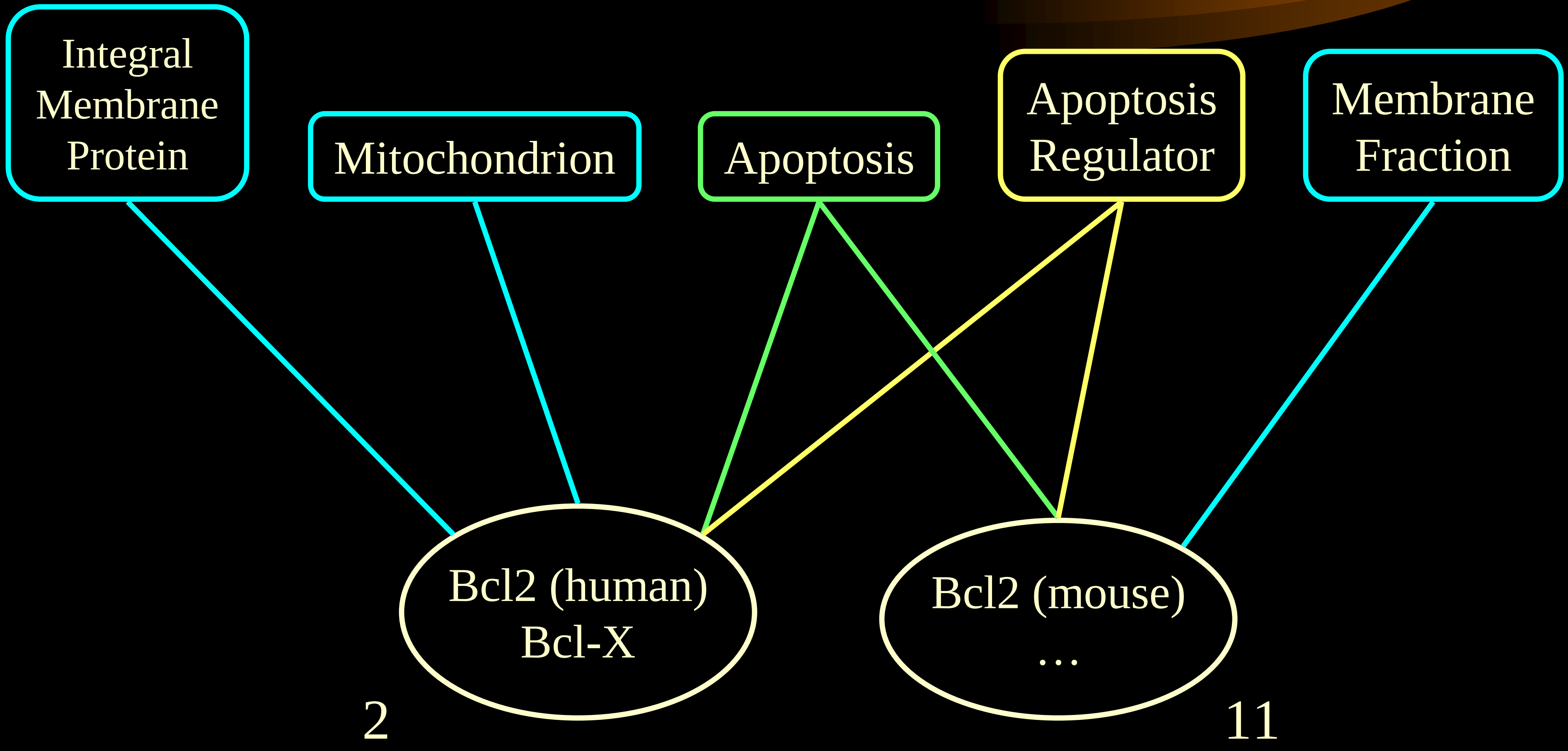
These genes should be similar based upon process, function, and localization.

Gene Comparison Using Identical Annotation is Too Limited

To Increase Cluster Size We Modified the Identity Constraints Using Different Methods

1. Only evidence from the literature was used to cluster the genes -- Traceable author statements
2. Closely related annotations were clustered together with identical ones
3. Only molecular functions, not biological processes or cellular components were used in clustering
4. Combinations of the 3 independent methods above

Gene Products with Identical Annotations are Clustered



Integral Membrane Protein

Mitochondrion

Apoptosis

Apoptosis Regulator

Membrane Fraction

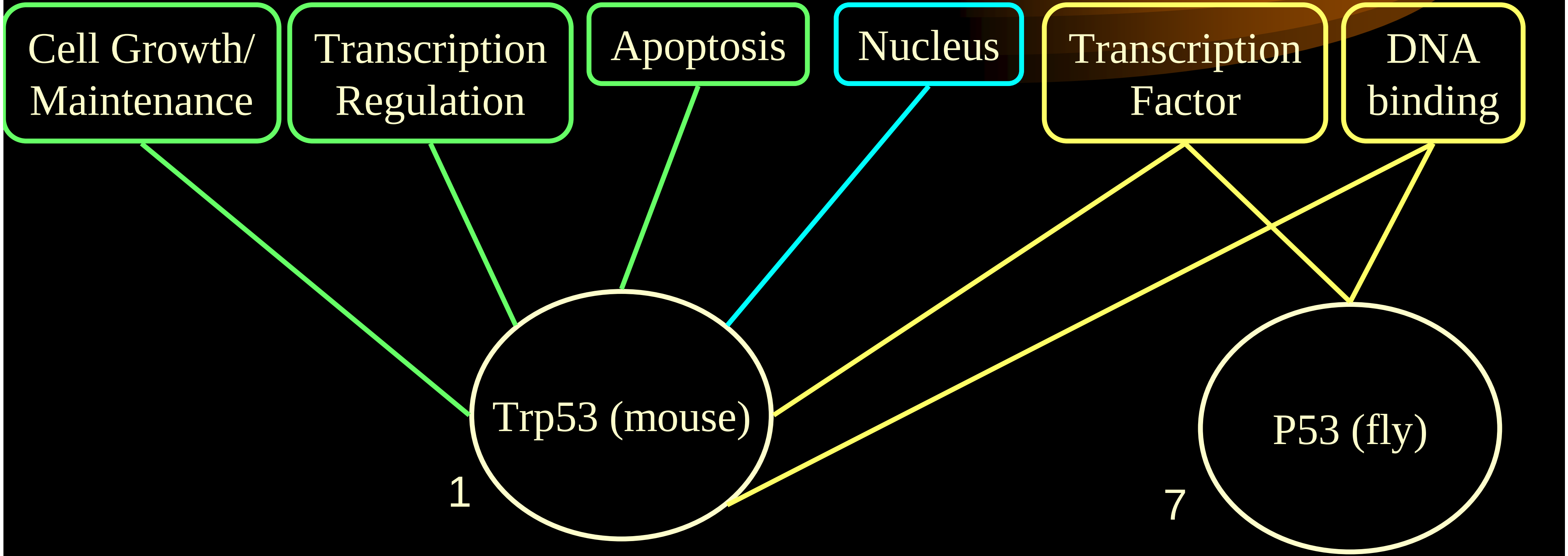
Bcl2 (human)
Bcl-X

2

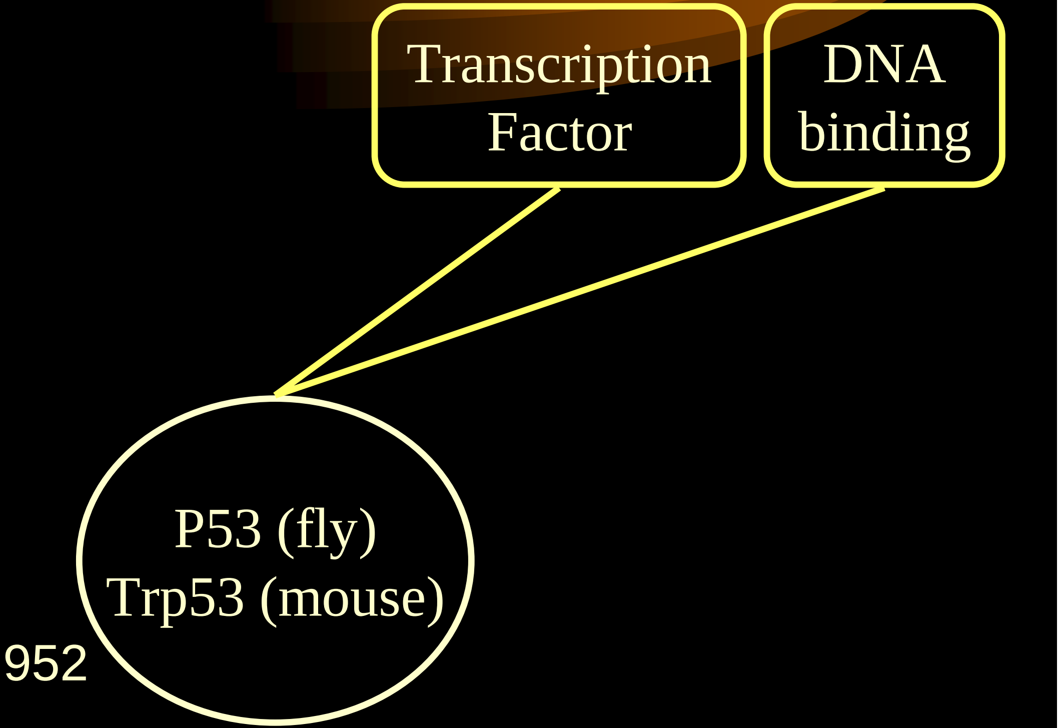
Bcl2 (mouse)
…

11

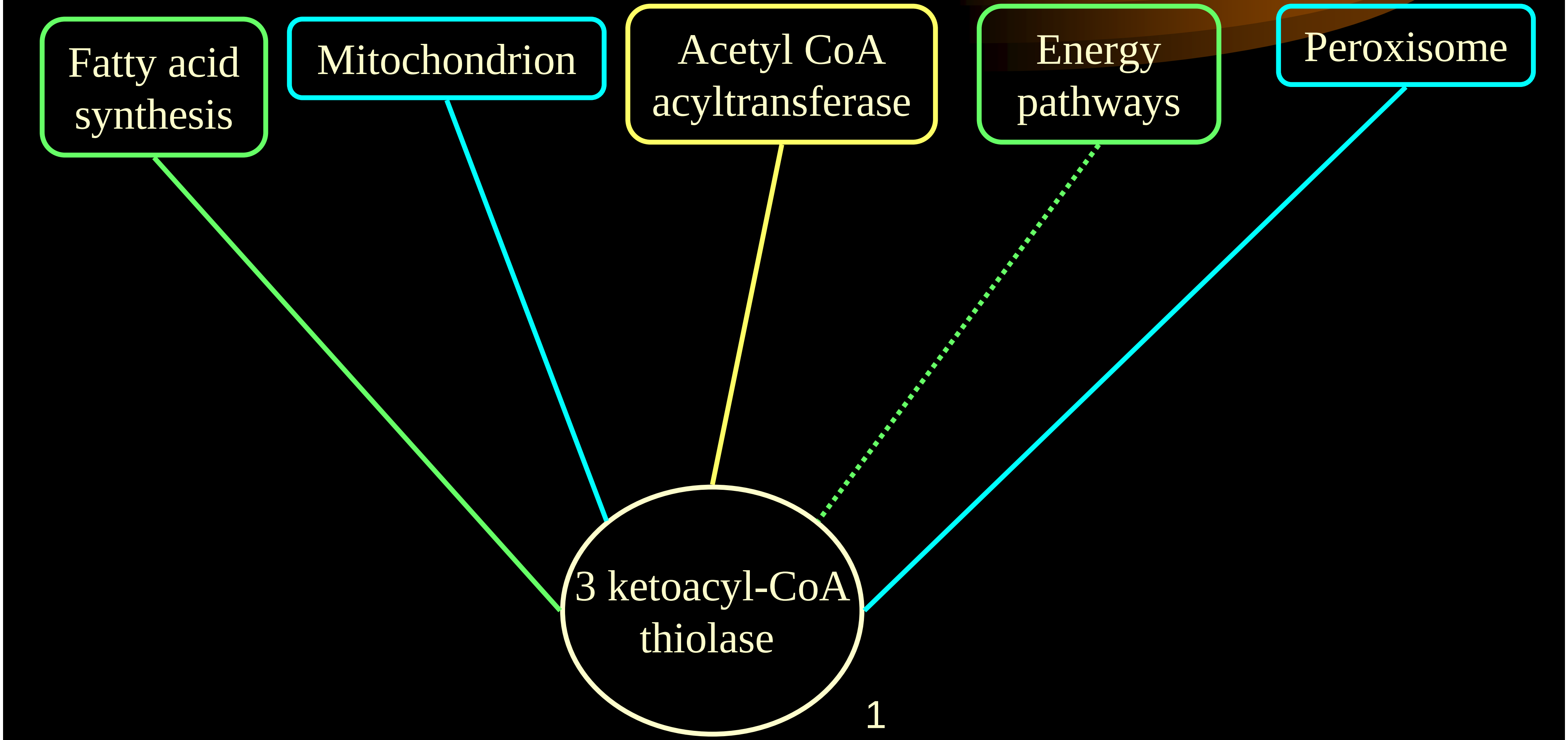
Examining Molecular Function Only



Examining Molecular Function Only: Ignoring Components and Processes



Identifying Clusters by Using Only Annotations Containing Traceable Author Statements



Fatty acid synthesis

Mitochondrion

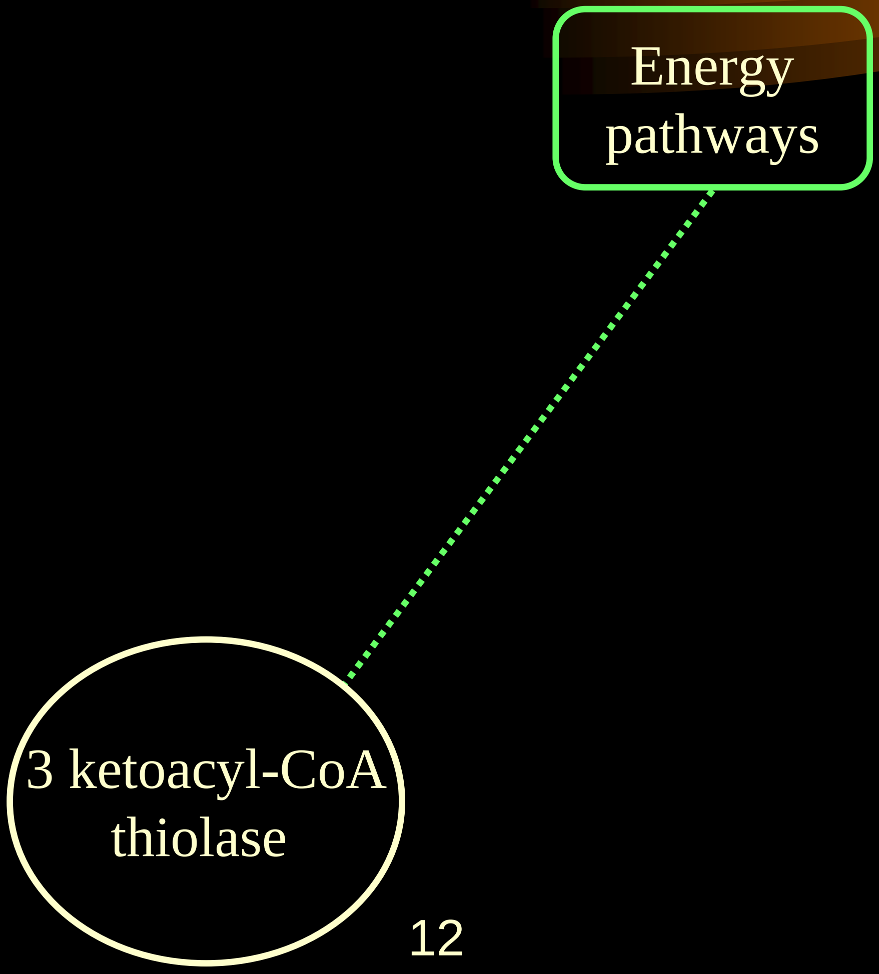
Acetyl CoA acyltransferase

Energy pathways

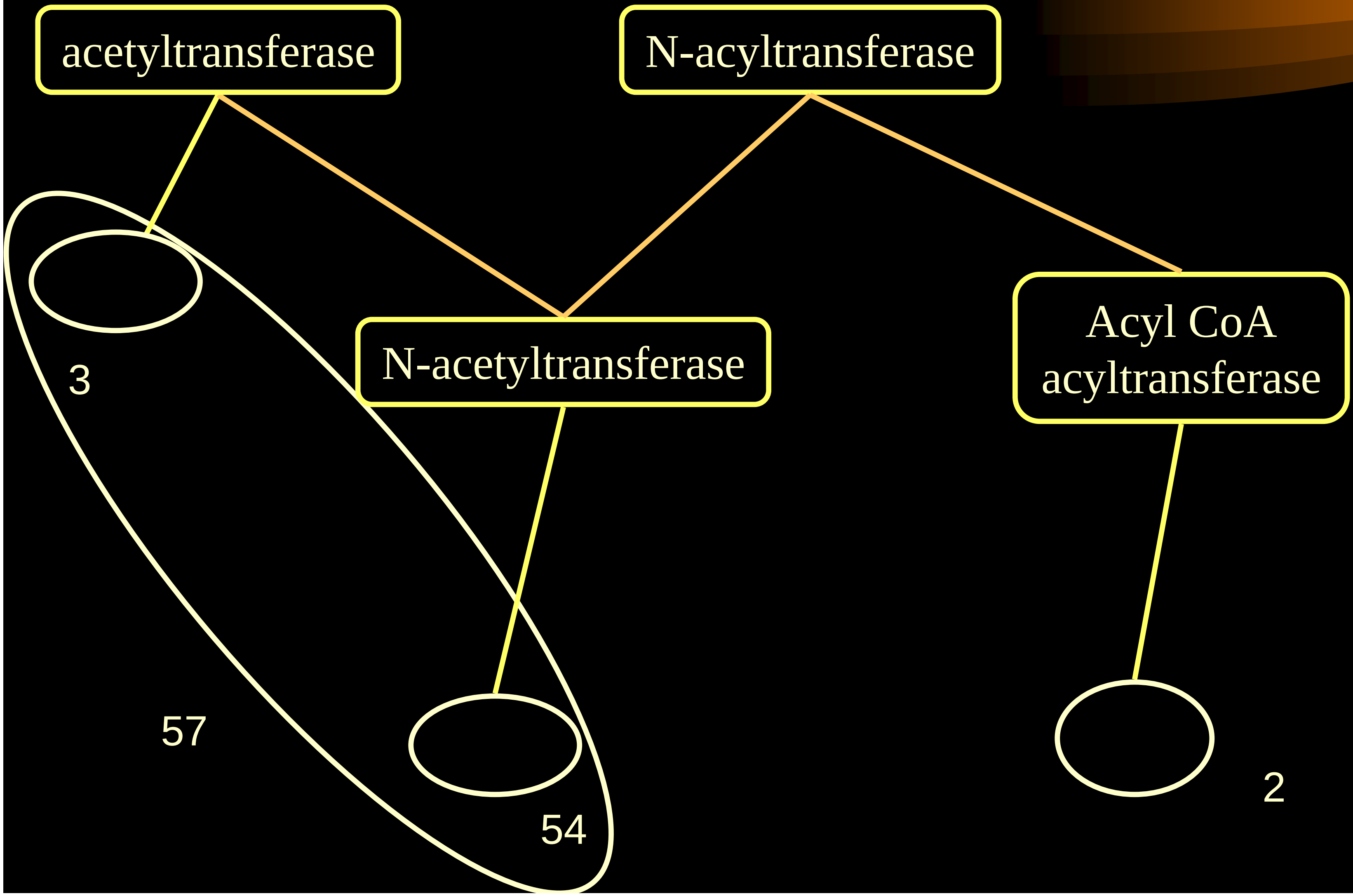
Peroxisome

3 ketoacyl-CoA thiolase

Identifying Clusters by Using Only Annotations Containing Traceable Author Statements



Clustering Genes With Closely Related Annotations



acetyltransferase

N-acyltransferase

N-acetyltransferase

Acyl CoA acyltransferase

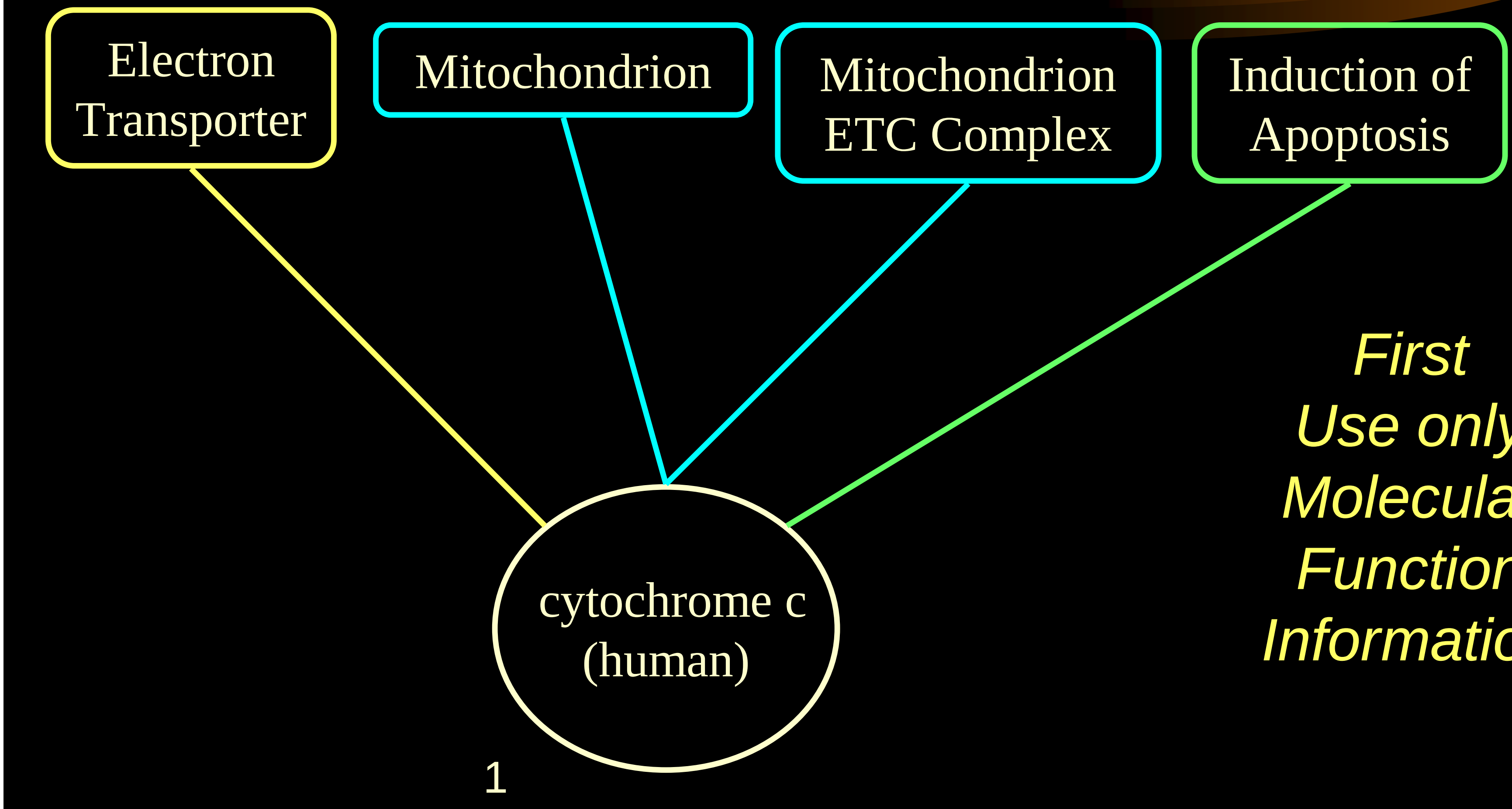
3

57

54

2

Combining Clustering Methods to Find Similar Gene Products



Electron Transporter

Mitochondrion

Mitochondrion ETC Complex

Induction of Apoptosis

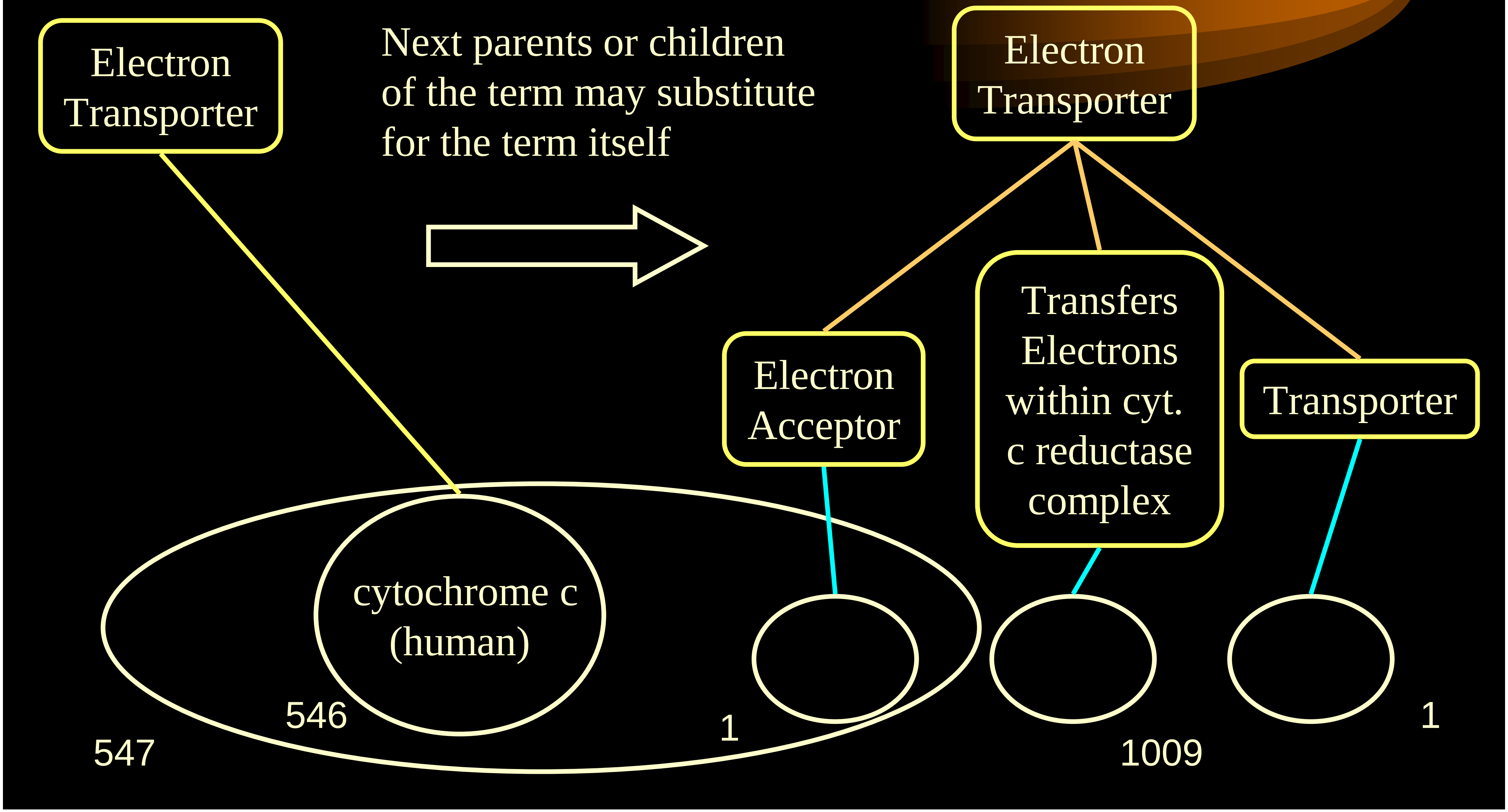
cytochrome c (human)

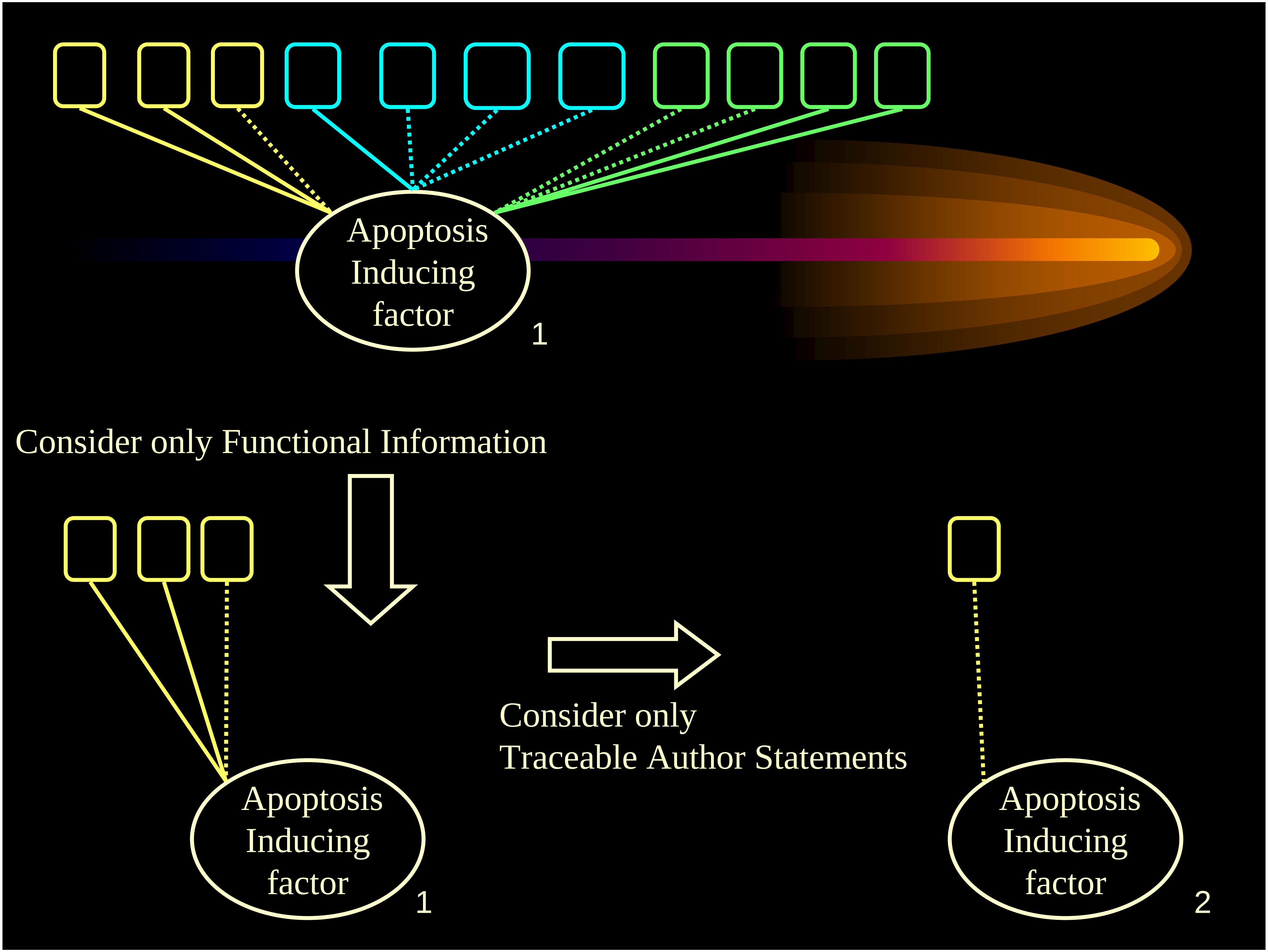
1

First
Use only
Molecular
Function
Information

Acquiring Larger Clusters of Gene Products

Next parents or children of the term may substitute for the term itself





Apoptosis Inducing factor

1

Consider only Functional Information

Apoptosis Inducing factor

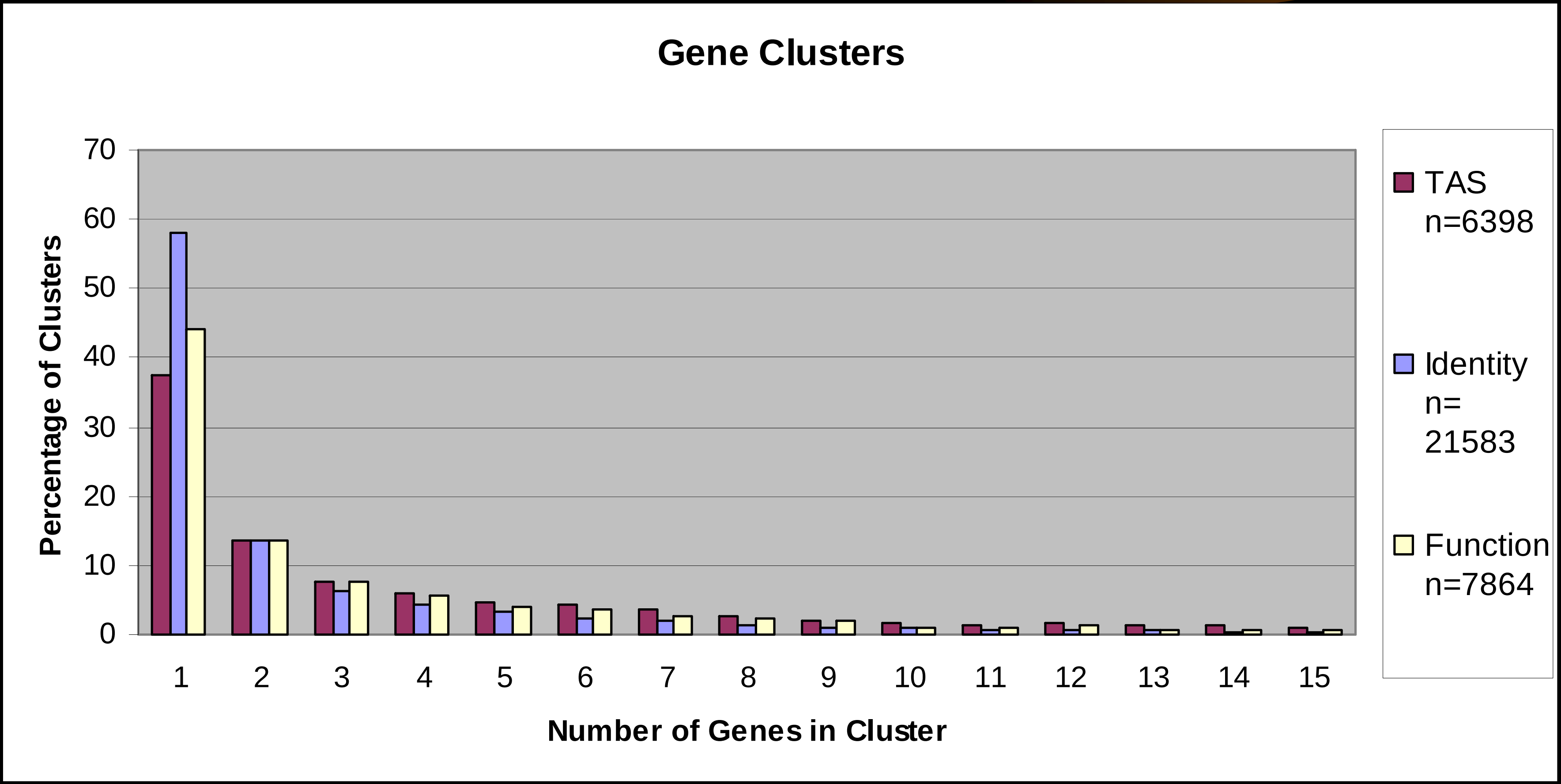
1

Consider only
Traceable Author Statements

Apoptosis Inducing factor

2

Using TAS or Function Only Reduces the % of Gene Products that Don’t Cluster with Others



Evaluation

- Does it make biological sense to group gene products in this manner?

- Examine existing databases for human manual groupings of gene products

- Compare the results with sequence homology

Possible Improvements

Instead of identical matches also include partial match of terms from one gene with identical match from the other gene (ex. A,B,C matches A,B)

Broaden the matching constraints for closely related terms to allow larger clusters

Develop or integrate the functionality into an existing web interface (GenNav) for online queries for researchers

Conclusions

- Matching identical GO terms is usually too restrictive since most genes do not even cluster with the homologous gene from another species

- Relaxing identity constraints to increase cluster size increases recall while decreasing precision

Conclusions

- Considering only molecular function in GO may form large clusters. But it may be a good way when used alone or in combination with one of the other methods for researchers to find gene products with a similar function but different molecular sequence
- More evaluation is needed

Acknowledgments- my colleagues at LHNCBC

- Olivier Bodenreider – project preceptor
- Joyce Mitchell
- Tom Rindflesch
- May Cheh
- Alexa McCray
- Chike, Laura, Larry, and Raju