Computational Molecular Biology
Group 4: Gene Ontology

Griffin Lunn

Vaibhav Deoda

Azhar Mirza

Final Presentation: The Gene Ontology Project and

BLASTING AMIGOS

Synopsis

- Introduction
- Literature review
- Program overview
- Implementation
- Test run
- Test run data analysis
- Conclusion
- Recommendations

Introduction

- Our group has been assigned to investigate the Gene Ontology project, which is a valuable tool in Bioinformatics.

- We plan to learn about the project and try to implement a novel program to help gain information from these massive databases of valuable gene data

GOALS

1) Learn about the Gene Ontology project and its place in Bioinformatics

3) Learn techniques that are useful for implementation of the above, preferably the PERL language and MySQL

3) Construct a program(‘s) that are beneficial to the Gene Ontology project

4) Take another team’s data in our class and use it as an input and generate an output that is value-added

5) Analyze the data and suggest improvements for the program

Consider the following problem: Biologists work day and night doing experiments that generate more data than ever.

How can they organize and access their data efficiency? How about for various types of data for various species? How can they integrate all this information seamlessly?

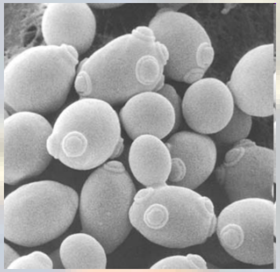
Solution: Gene Ontology

- An ontology is a relationships between various concepts inside of a domain, in this instance for molecular/cell biology.

- This is done by using a *controlled vocabulary,* which tags entries with a consistent methodology which makes data retrieval easier.

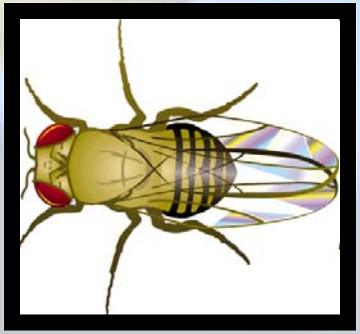
Gene Ontology Project

- Started in the late 90’s
- Combined the talents of scientists working on gene databases for yeast, fruit fly, and mouse
- Grew to cover more model organisms and eventually more organisms



+

+



= GO

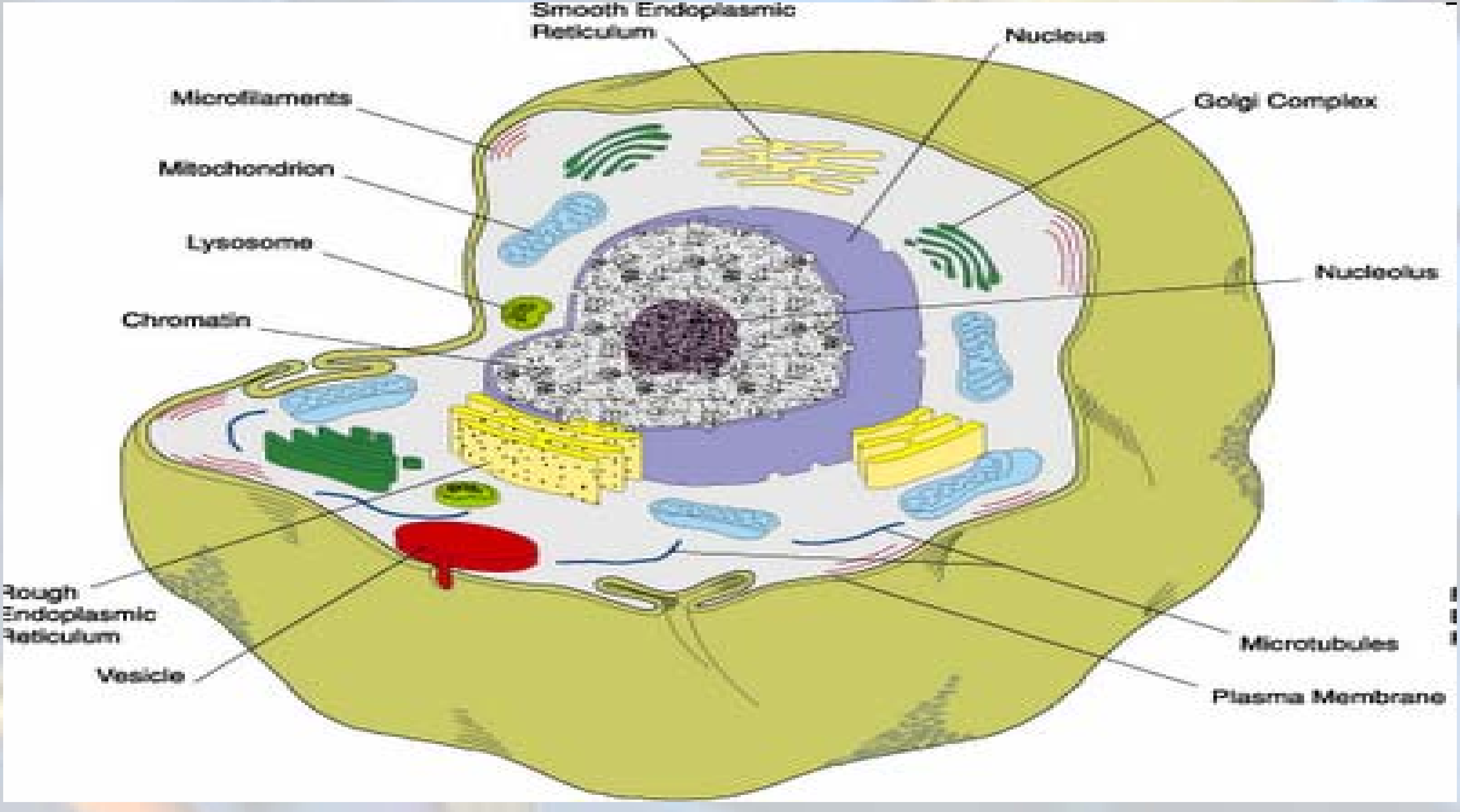
Structure of the GO project

- Made up of 3 Ontologies

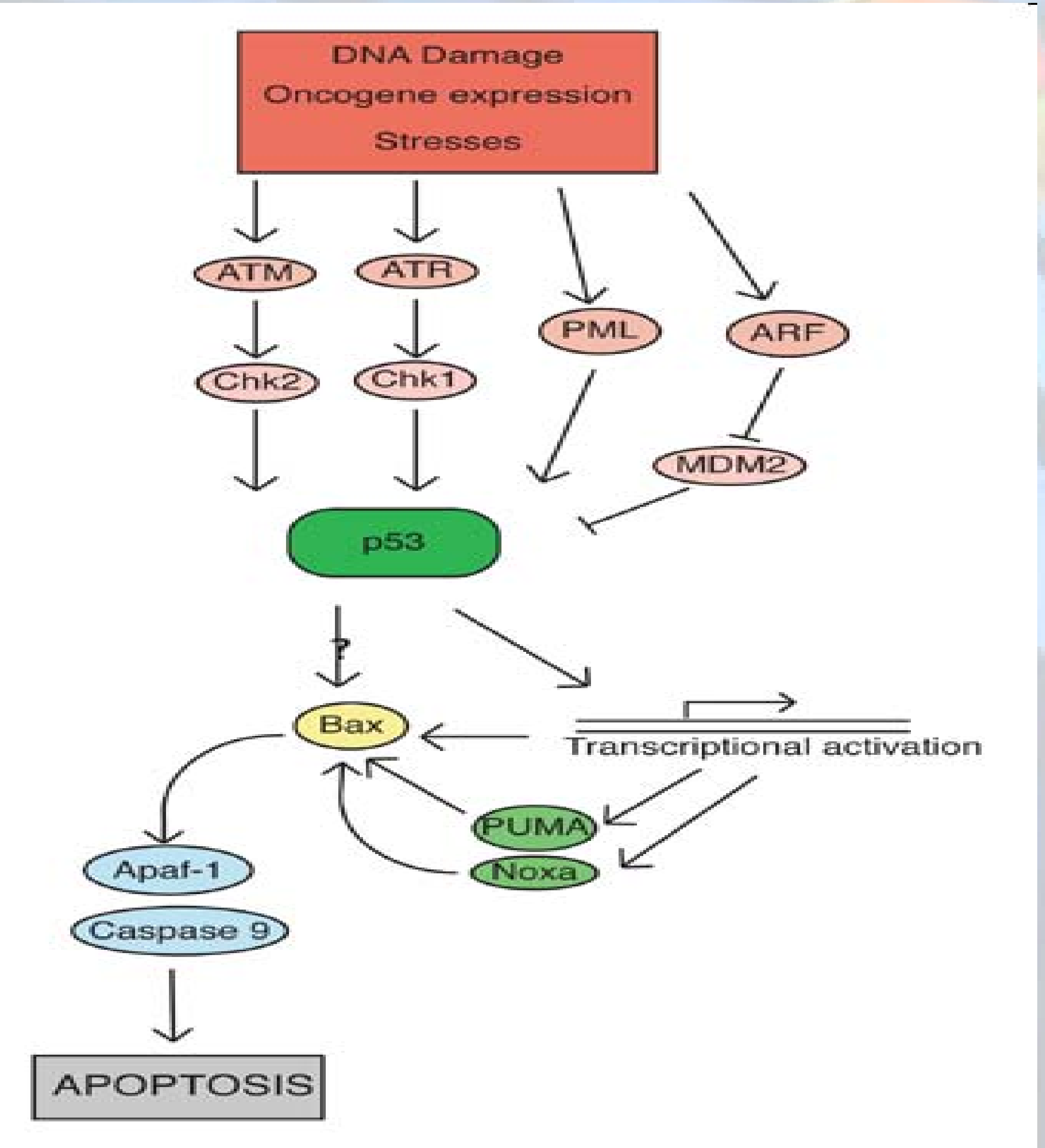
- Consists of GO terms annotated to Gene Products (proteins)

- Can be searched with AmiGO and edited with OBE-edit

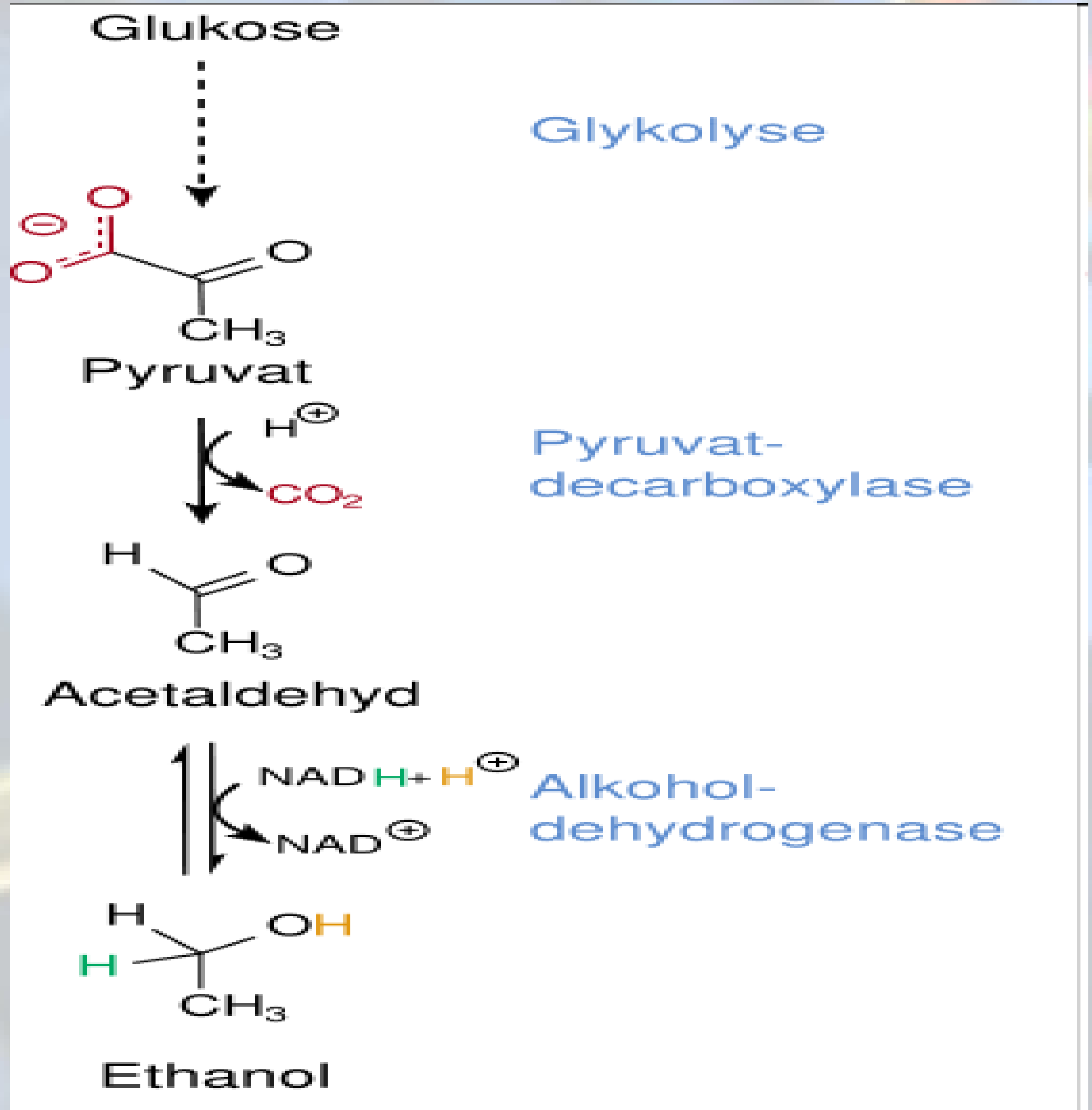
Cellular components



molecular functions



biological processes



So what does a Gene Ontology do?

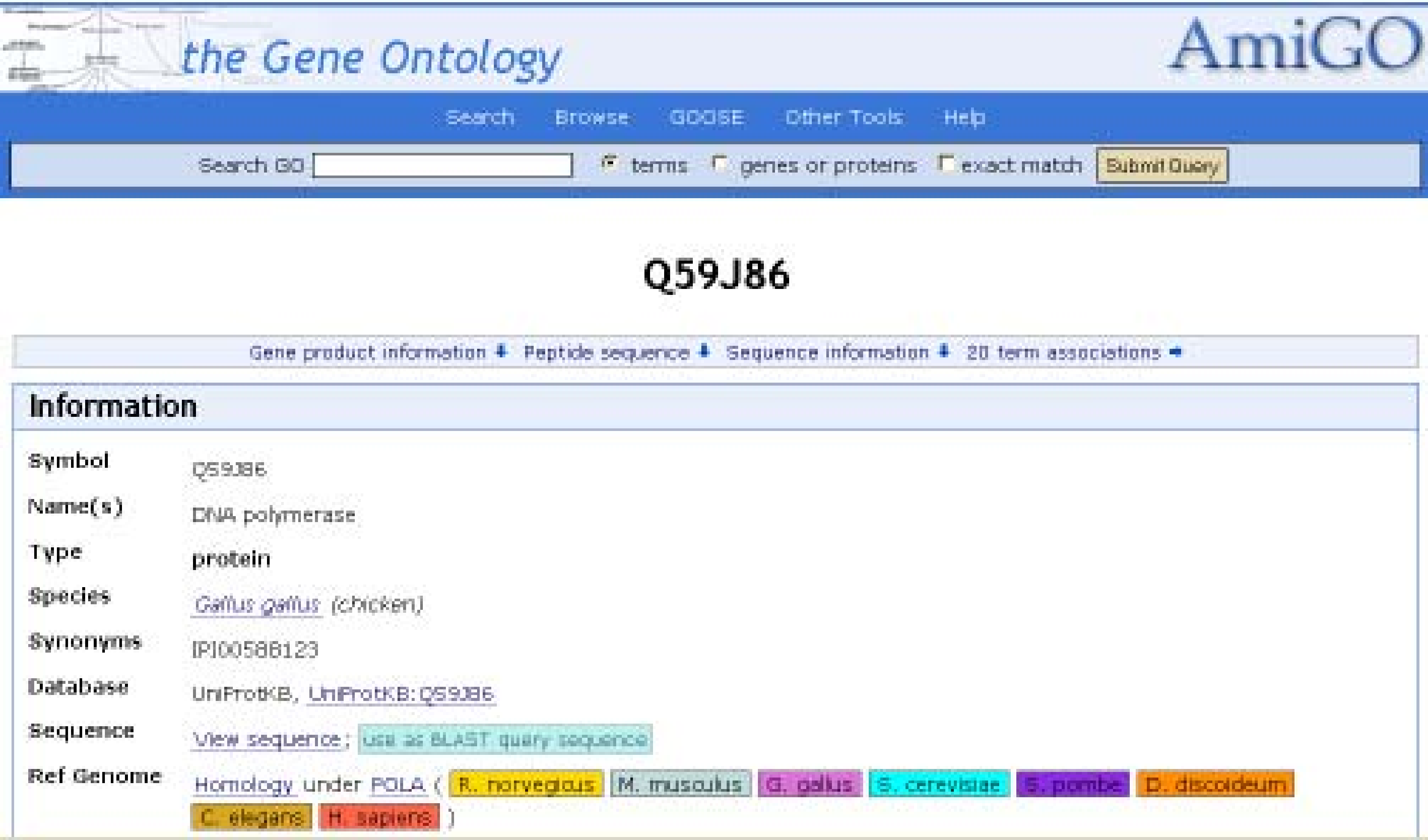
- A Gene Ontology takes a gene product (protein) and gives it a cellular context.

- For each of the three ontology's, gene products can be placed where they belong, and various keywords can be looked up to find the associated gene products.

Example of gene product data

- Look up gene “Q59J86”
- Gives:
- Name(s) “DNA polymerase”
- Type “protein”
- Species “Gallus gallus (chicken)”
- Synonyms “IPI00588123”
- Sequence
- References
- Term associations

Example from AmiGO



Q59J86

Information

Symbol	Q59J86
Name(s)	DNA polymerase
Type	protein
Species	*Gallus gallus* (chicken)
Synonyms	IPI00588123
Database	UniProtKB, UniProtKB:Q59J86
Sequence	View sequence; use as BLAST query sequence
Ref Genome	Homology under POLA (R. norvegicus M. musculus G. gallus S. cerevisiae S. pombe D. discoideum C. elegans H. sapiens)

Go Term

- A decriptive term that is used to give a gene product a cellular, molecular, or biological context

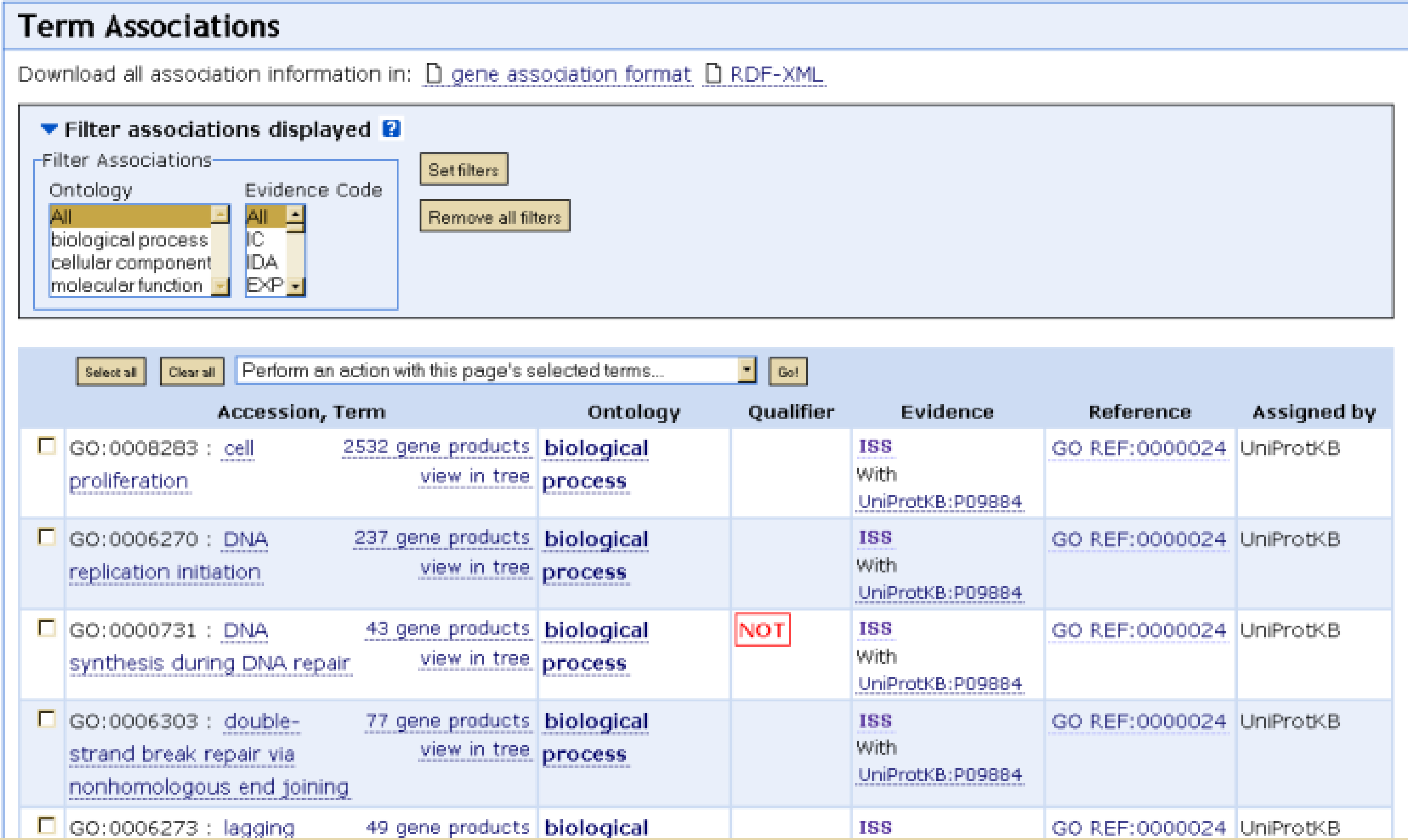
- Terms are standardized across all databases and use synonyms to bridge gaps in spelling or similar function

- Older terms can become obsolete

Anatomy of a GO term

- Term “Cell wall”
- ID number “GO:00005618
- Ontology “Cellular components”
- Definition “The rigid or semi-rigid envelope lying outside the cell membrane of plant, fungal, and most prokaryotic cells, maintaining their shape and protecting them from osmotic lysis. In plants it is made of cellulose and, often, lignin; in fungi it is composed largely of polysaccharides; in bacteria it is composed of peptidoglycan. “
- Synonyms “None”
- Lineage “shows graph”
- Gene products “1045 found”
- LINK

Example from AmiGO



Term Associations

Download all association information in: gene association format RDF-XML

▼ Filter associations displayed

Filter Associations

Ontology: All, biological process, cellular component, molecular function

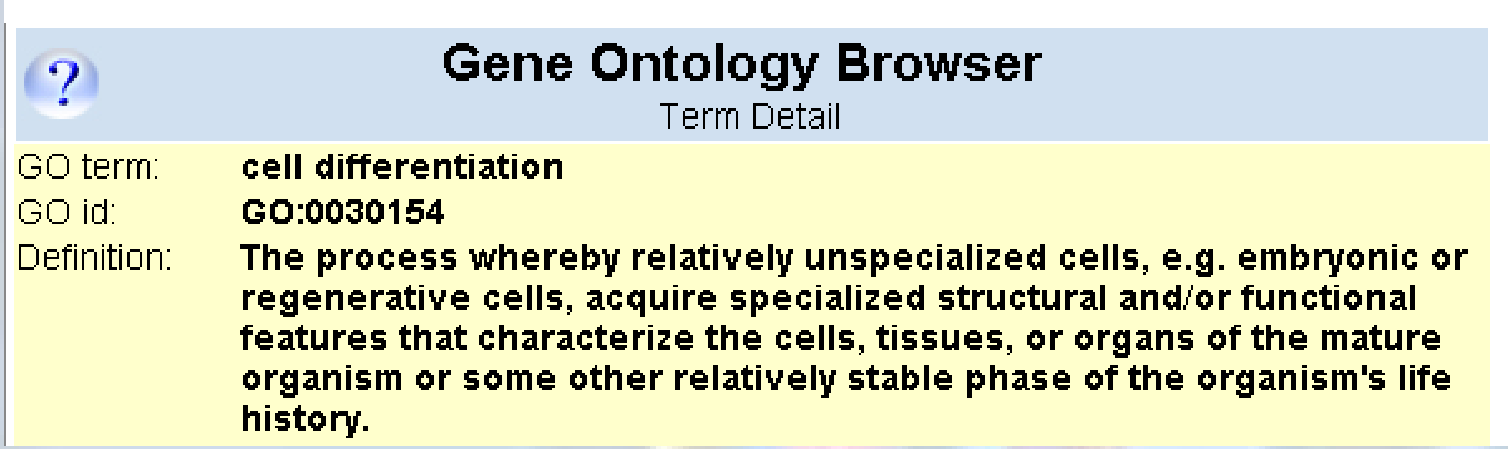
Evidence Code: All, IC, IDA, EXP

	Accession, Term		Ontology	Qualifier	Evidence	Reference	Assigned by
□	GO:0008283 : cell proliferation	2532 gene products view in tree	biological process		ISS With UniProtKB:P09884	GO REF:0000024	UniProtKB
□	GO:0006270 : DNA replication initiation	237 gene products view in tree	biological process		ISS With UniProtKB:P09884	GO REF:0000024	UniProtKB
□	GO:0000731 : DNA synthesis during DNA repair	43 gene products view in tree	biological process	NOT	ISS With UniProtKB:P09884	GO REF:0000024	UniProtKB
□	GO:0006303 : double-strand break repair via nonhomologous end joining	77 gene products view in tree	biological process		ISS With UniProtKB:P09884	GO REF:0000024	UniProtKB
□	GO:0006273 : lagging	49 gene products	biological		ISS	GO REF:0000024	UniProtKB

Term Obsoleteness

- If a term is found to be misleading or can be described with a better term, it is rendered obsolete
- The term is NOT DELETED, but is marked obsolete and a new term may be proposed

GO definitions

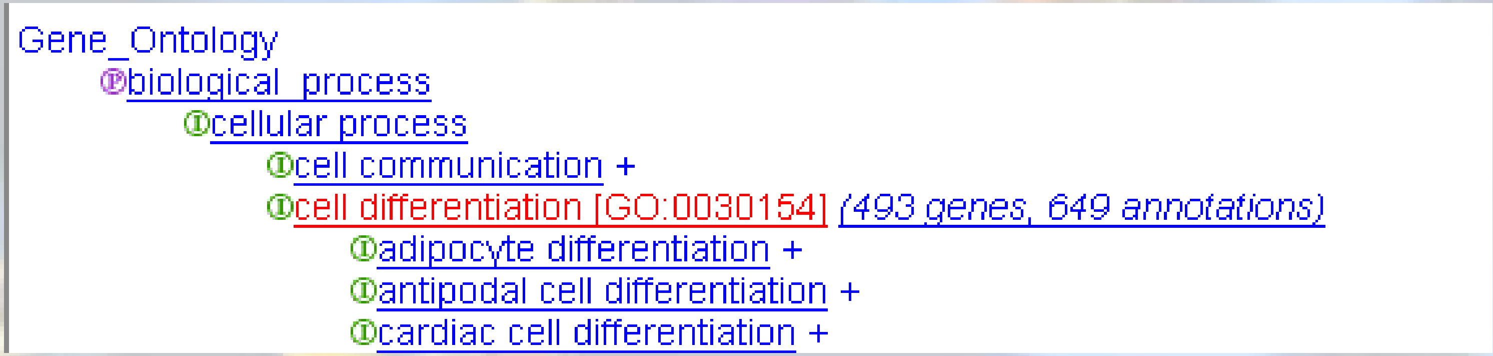


Gene Ontology Browser

Term Detail

GO term:	**cell differentiation**
GO id:	**GO:0030154**
Definition:	**The process whereby relatively unspecialized cells, e.g. embryonic or regenerative cells, acquire specialized structural and/or functional features that characterize the cells, tissues, or organs of the mature organism or some other relatively stable phase of the organism's life history.**

Written Definition, not searchable



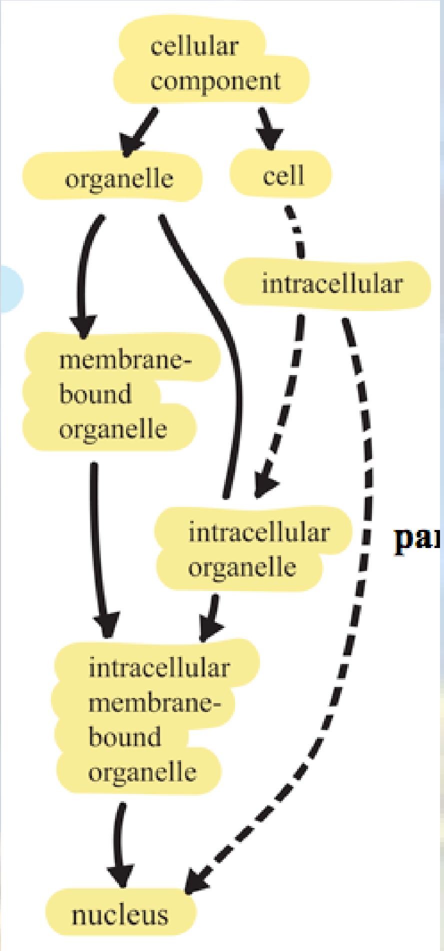
- Gene_Ontology
  - Ⓟbiological_process
    - Ⓘcellular process
      - Ⓘcell communication +
      - Ⓘcell differentiation [GO:0030154] *(493 genes, 649 annotations)*
        - Ⓘadipocyte differentiation +
        - Ⓘantipodal cell differentiation +
        - Ⓘcardiac cell differentiation +

Graph structure, searchable

Graph structure

- The ontologies are structured as directed acyclic graphs, which are graphs that do not cycle or repeat
- These are similar to hierarchies but differ in that a more specialized term (child) can be related to more than one less specialized term (parent)
- This allows annotations to one GO term to be also annotated to related GO terms connected in the graph structure

Example



Solid lines are Is_a relationships

Dotted Lines are Part_of relationships

Types of Relationships

- Is_a [i]
- Part_of [p]
- Regulates/ positively_regulates / negatively_regulates [r]

```
GO:0010467 : gene expression
[r] GO:0010468 : regulation of gene expression
---[i] GO:0045449 : regulation of transcription
[p] GO:0006350 : transcription
---[r] GO:0045449 : regulation of transcription
```

Is_a Relationships

- Simple parent-child relationship
- A is_a B means A is a subclass of B

```
GO:0043232 : intracellular non-membrane-bound organelle
[i] GO:0005694 : chromosome
---[i] GO:0000228 : nuclear chromosome
```

Part_of Relationships

- C part_of D means that whenever C is present, it is always a part of D, but C does not always have to be present.

```
[i] GO:0042597 : periplasmic space
---[p] GO:0055040 : periplasmic flagellum
```

“When a periplasmic flagellum is present, it is always part_of a periplasmic space. However, every periplasmic space does not necessarily have a periplasmic flagellum.”

Relationship Transitivity

- Is_a Transitivity:
- A nucleus must be an organelle
- Part_of Transitivity:
- All intracellular organelles must be intracelluar
- Regulation Transitivity
- If process B is regulated and is_a child of Process A, regulating process B will regulate process A

Problem

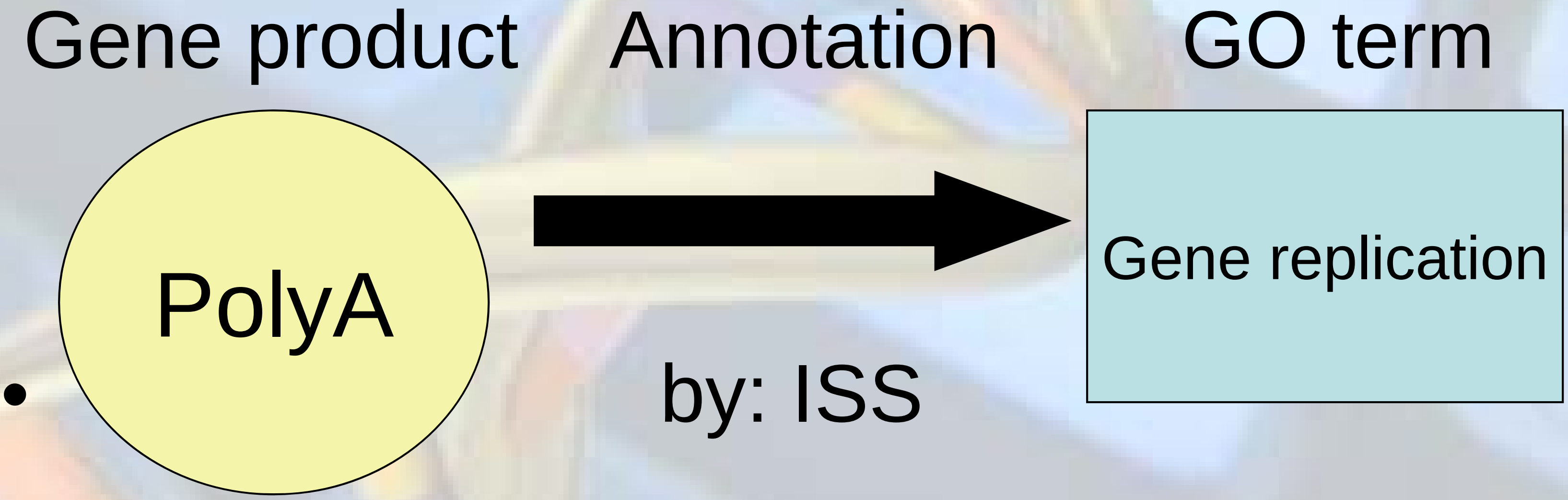
- How do we know which go terms apply for which gene products, and vice versa?

- Gene Product Go term

PolyA
Gene replication

Annotation!

- Annotating is the process of associating a gene product with a GO term



Gene product

Annotation

GO term

PolyA

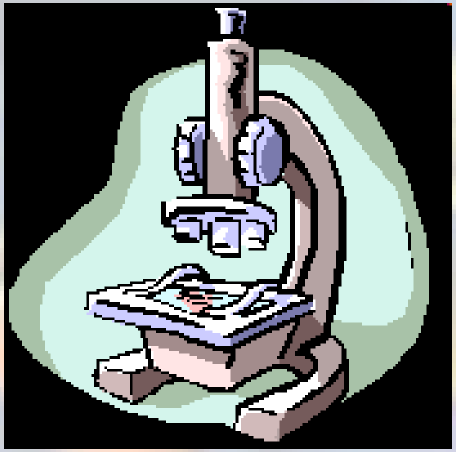
Gene replication

by: ISS

Types of Annotation

- Electronic Annotation:
- Uses computational methods like sequence simularity or genomic models to determine the GO term associations. Very fast but not especially accurate.

- Manual Annotation:



Uses primary research or review from published literature to make the annotation. Highly accurate but very labor intensive

Evidence codes:

- Experimental Evidence Codes
  - EXP: Inferred from Experiment
  - IDA: Inferred from Direct Assay
  - IPI: Inferred from Physical Interaction
  - IMP: Inferred from Mutant Phenotype
  - IGI: Inferred from Genetic Interaction
  - IEP: Inferred from Expression Pattern
- Computational Analysis Evidence Codes
  - ISS: Inferred from Sequence or Structural Similarity
  - ISO: Inferred from Sequence Orthology
  - ISA: Inferred from Sequence Alignment
  - ISM: Inferred from Sequence Model
  - IGC: Inferred from Genomic Context
  - RCA: inferred from Reviewed Computational Analysis
- Author Statement Evidence Codes
  - TAS: Traceable Author Statement
  - NAS: Non-traceable Author Statement
- Curator Statement Evidence Codes
  - IC: Inferred by Curator
  - ND: No biological Data available
- Automatically-assigned Evidence Codes
  - IEA: Inferred from Electronic Annotation

Computational Analysis Evidence Codes

- After a computer has generated annotations, they are usually checked over by a human curator for accuracy.
- If a human curator has not checked over the output data, the annotations are assigned the code IEA until they are.
- Currently, all data shown by AmiGO has been allegedly looked over by at least one human being

How is this useful?

- The Gene Ontology project is always growing with new genes discovered daily

- Annotations give these new genes a cellular context and help Scientists understand how these genes function in the grand scheme of things

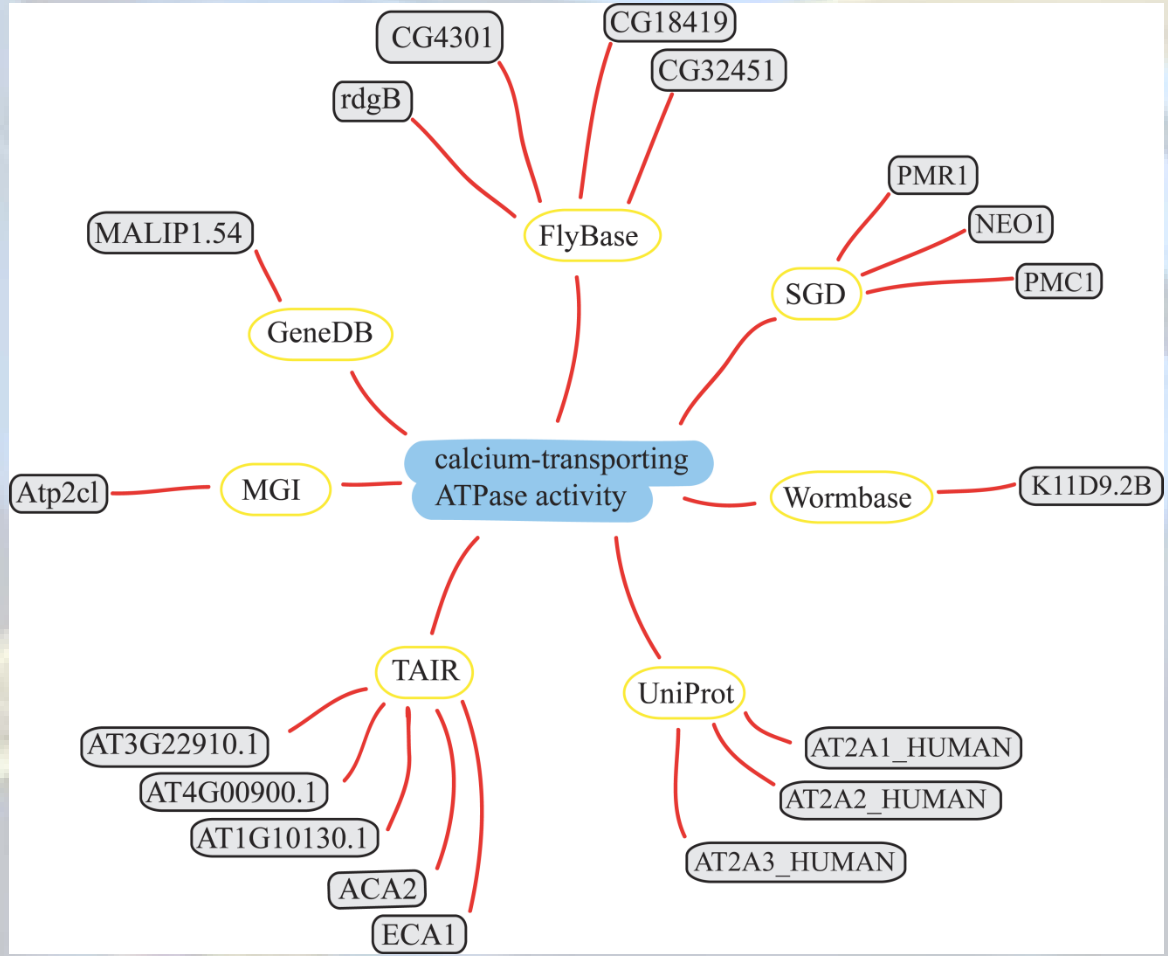
Example

- Biologist isolates genes and uses a genetic analyzer to determine the nucleotide sequence of each gene

- The biologist then uses a computer program to find a similar gene to each of the discovered genes (BLAST), and then uses another computer program (AMIGO) to find the GO terms associated with the similar gene.

- By assuming that similar gene sequences have a similar cellular context, these GO terms could be annotated to these new genes, which allows the scientist to understand what these genes do, in a very short period of time.

Gene ontology data across species



Database structure

- All 3 Gene Ontologies, Annotations, and Gene products are stored in one relational database.
- The Database is written in MySQL and is updated with various daily, weekly, and monthly builds in addition to various mirrors and stored previous builds
- The database can be accessed by AmiGO or queried remotely by various methods, or even downloaded
- The Ontology data is in OBO file format (Open Biomedical Ontologies)

Gene Ontology Tools

- The Gene Ontology Consortium itself has created tools to help create, search, and analyze its data and also supports 3rd party applications on their website

- The GOC created AmiGO and OBO-edit to read and edit the database data respectfully

- 3rd party developers have created GO browsers, annotators, and data analyzers, among other tools

AmiGO

- Browser and search tool created by the GOC to quickly search their database online.

- Currently only shows manual annotations (ones that have been reviewed by a curator and don’t have the evidence code IEA)

- Can search by gene name or go term, and provides selected gene information, sequence, term associations, and the acyclic graph data for that gene’s associations

OBO-Edit

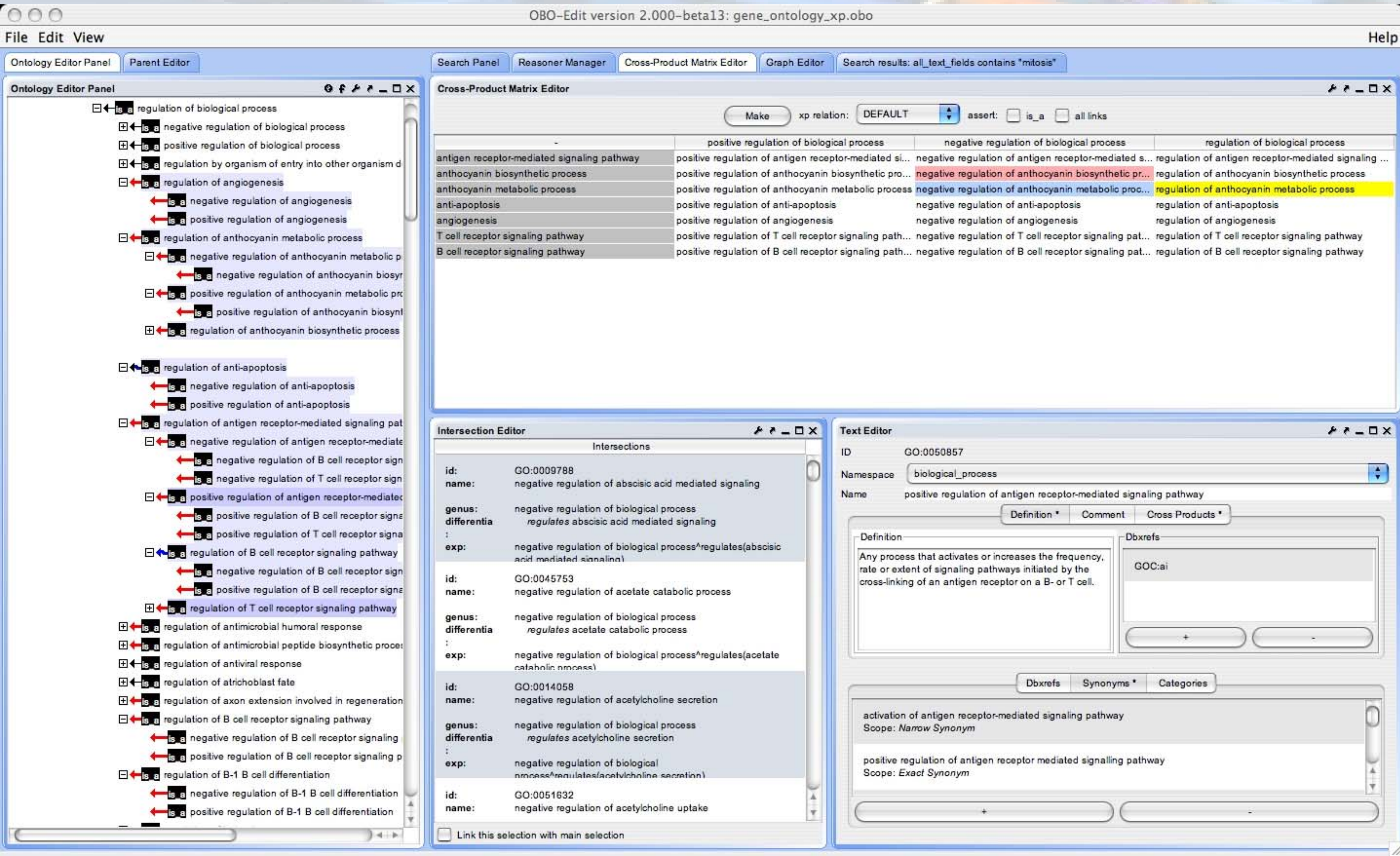
- Originally designed for the Open Biomedical Ontology by Berkeley Bioinformatics and Ontologies Project.

- Written in java and optimized for the OBO file format and works in a graph-based interface that is easy for biologists to edit and understand

- All 3 Ontologies are designed in this program, and all GO terms are given their relationships and definitions.

- Includes a reasoning engine to establish links that have not been found by the curator

OBO-edit in action



```
⊟ is a regulation of biological process
    ⊞ is a negative regulation of biological process
    ⊞ is a positive regulation of biological process
    ⊞ is a regulation by organism of entry into other organism d
    ⊟ is a regulation of angiogenesis
        is a negative regulation of angiogenesis
        is a positive regulation of angiogenesis
    ⊟ is a regulation of anthocyanin metabolic process
        ⊟ is a negative regulation of anthocyanin metabolic p
            is a negative regulation of anthocyanin biosyr
        ⊟ is a positive regulation of anthocyanin metabolic prc
            is a positive regulation of anthocyanin biosynt
        ⊞ is a regulation of anthocyanin biosynthetic process

    ⊟ is a regulation of anti-apoptosis
        is a negative regulation of anti-apoptosis
        is a positive regulation of anti-apoptosis
    ⊟ is a regulation of antigen receptor-mediated signaling pat
        ⊟ is a negative regulation of antigen receptor-mediate
            is a negative regulation of B cell receptor sign
            is a negative regulation of T cell receptor sign
        ⊟ is a positive regulation of antigen receptor-mediatec
            is a positive regulation of B cell receptor signa
            is a positive regulation of T cell receptor signa
        ⊟ is a regulation of B cell receptor signaling pathway
            is a negative regulation of B cell receptor sign
            is a positive regulation of B cell receptor signa
        ⊞ is a regulation of T cell receptor signaling pathway
    ⊞ is a regulation of antimicrobial humoral response
    ⊞ is a regulation of antimicrobial peptide biosynthetic proce:
    ⊞ is a regulation of antiviral response
    ⊞ is a regulation of atrichoblast fate
    ⊞ is a regulation of axon extension involved in regeneration
    ⊟ is a regulation of B cell receptor signaling pathway
        is a negative regulation of B cell receptor signaling
        is a positive regulation of B cell receptor signaling p
    ⊟ is a regulation of B-1 B cell differentiation
        is a negative regulation of B-1 B cell differentiation
        is a positive regulation of B-1 B cell differentiation
```

xp relation: DEFAULT assert: □ is_a □ all links

-	positive regulation of biological process	negative regulation of biological process	regulation of biological process
antigen receptor-mediated signaling pathway	positive regulation of antigen receptor-mediated si...	negative regulation of antigen receptor-mediated s...	regulation of antigen receptor-mediated signaling ...
anthocyanin biosynthetic process	positive regulation of anthocyanin biosynthetic pro...	negative regulation of anthocyanin biosynthetic pr...	regulation of anthocyanin biosynthetic process
anthocyanin metabolic process	positive regulation of anthocyanin metabolic process	negative regulation of anthocyanin metabolic proc...	regulation of anthocyanin metabolic process
anti-apoptosis	positive regulation of anti-apoptosis	negative regulation of anti-apoptosis	regulation of anti-apoptosis
angiogenesis	positive regulation of angiogenesis	negative regulation of angiogenesis	regulation of angiogenesis
T cell receptor signaling pathway	positive regulation of T cell receptor signaling path...	negative regulation of T cell receptor signaling pat...	regulation of T cell receptor signaling pathway
B cell receptor signaling pathway	positive regulation of B cell receptor signaling path...	negative regulation of B cell receptor signaling pat...	regulation of B cell receptor signaling pathway

id: GO:0009788
name: negative regulation of abscisic acid mediated signaling
genus: negative regulation of biological process
differentia: *regulates* abscisic acid mediated signaling
exp: negative regulation of biological process^regulates(abscisic acid mediated signaling)

id: GO:0045753
name: negative regulation of acetate catabolic process
genus: negative regulation of biological process
differentia: *regulates* acetate catabolic process
exp: negative regulation of biological process^regulates(acetate catabolic process)

id: GO:0014058
name: negative regulation of acetylcholine secretion
genus: negative regulation of biological process
differentia: *regulates* acetylcholine secretion
exp: negative regulation of biological process^regulates(acetylcholine secretion)

id: GO:0051632
name: negative regulation of acetylcholine uptake

□ Link this selection with main selection

ID GO:0050857
Namespace biological_process
Name positive regulation of antigen receptor-mediated signaling pathway

Definition
Any process that activates or increases the frequency, rate or extent of signaling pathways initiated by the cross-linking of an antigen receptor on a B- or T cell.

Dbxrefs
GOC:ai

activation of antigen receptor-mediated signaling pathway
Scope: *Narrow Synonym*

positive regulation of antigen receptor mediated signalling pathway
Scope: *Exact Synonym*

Gosling

- Stands for GO similarity listing using information graphs

- Is a gene product annotator that uses sequence similarity to predict GO term associations by using a rule-based decision tree.

- Is designed to handle very large data sets very quickly, yet when compared to a test data set, is more accurate than similar programs

- Currently unavailable on https://www.sapac.edu.au/gosling

BLAST

- Stands for Basic Local Alignment Search Tool

- Is a group of programs used to compare sequence data to various (user’s choice) of sequence databases

- In short, BLAST finds high-scoring segment pairs (HSP) in the sequence and compares them to other sequences using a modified Smith-Waterman algorithm

- BLAST is not as accurate as the Smith-Waterman method, but is over 50 times faster

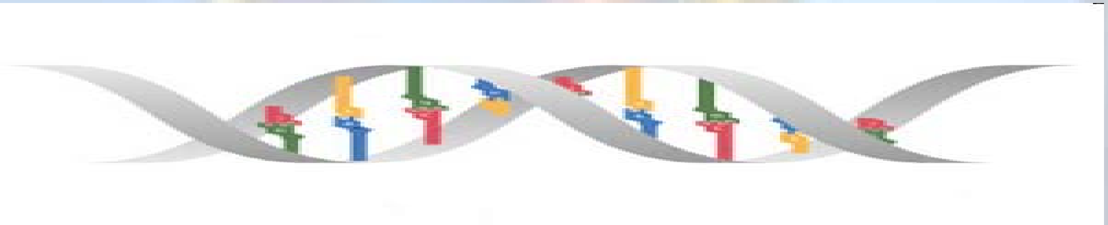
Part 2

- Program Design and Implementation

Our Project:

BLASTing AMIGO's

- Input
  - – Gene sequence data (nucleotide or AA) in FASTA format
- Output



  - Go Term, Description, Annotation in a MySql Database.

Design Goals

- Easy to use for Biologists
- Fast, results in minutes.
- Accurate, gives correct GO term associations
- Comprehensive, for each gene sequence gives many accession numbers which yields many go terms

Major Steps

- Remotely query blast and get blast output.
- Extract accession numbers from the blast output.
- Query GO database with these accession numbers and extract the associated GO terms
- Dump the output generated into a table.

❖ The Project basically integrates blast and amigo and removes a lot of manual work!

Perl

- Perl is nicknamed "the Swiss Army chainsaw of programming languages" due to its flexibility and adaptability.
- Just like C(Procedural).
- Very easy to use.

Why do Biologist use Perl ?

- Open Source.
- Most of biology works is centered around text manipulation.

Remote access to blast

Bio Perl

- Core Package
- Run Package
- Bio Perl DB package
- Network Package

Bio::Search::Hit::HitI

Output Part 1

- Lots of information from NCBI Website saved in a text file.

- Accession numbers taken out from this file.

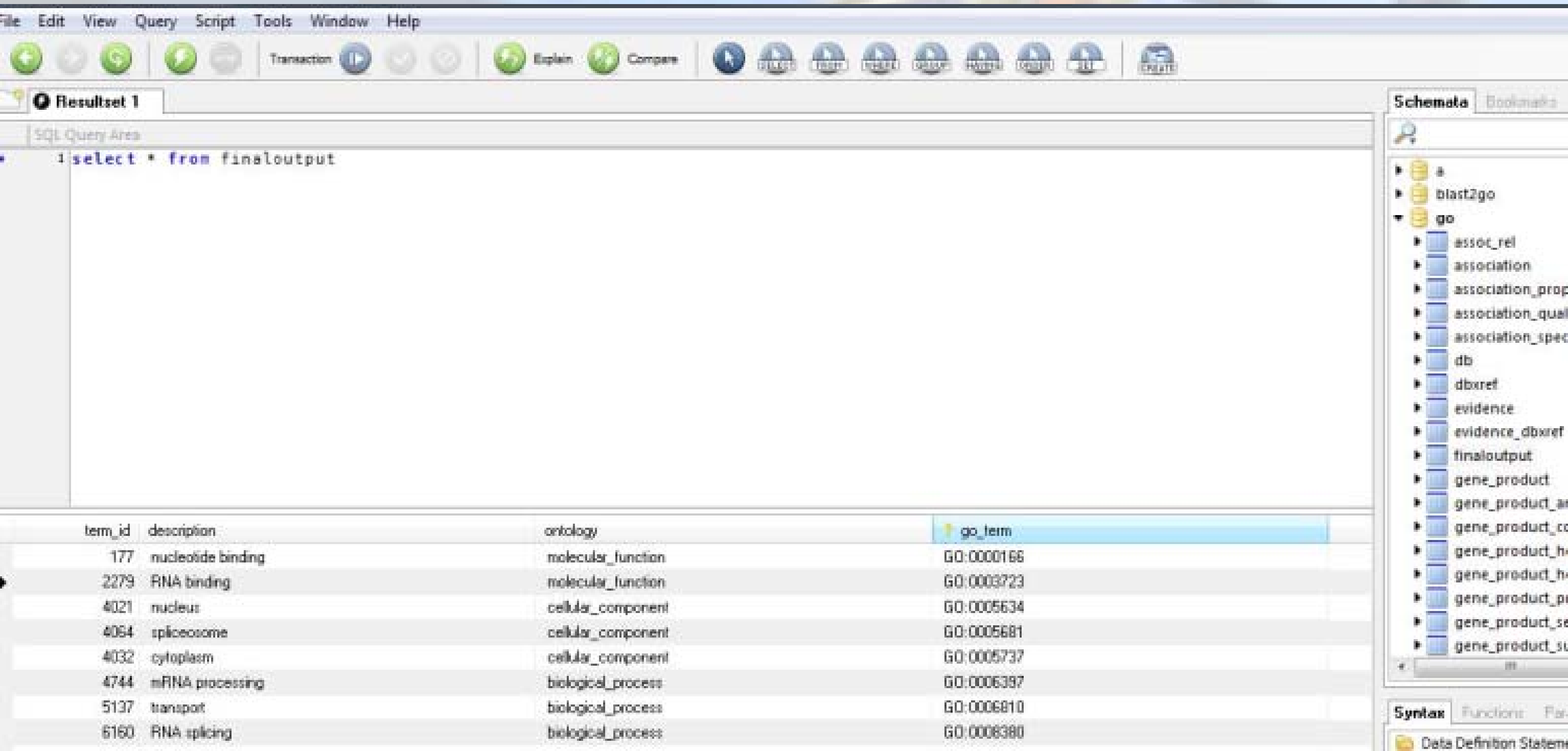
Querying GO Database

Module Used : DBI

Syntax :

```
Obj = DBI->connect(‘dbi:mysql:Dbase’,’username’,’pass’);
obj->prepare(‘query’);
obj->execute;
obj->fetchrow_array;
```

Final output

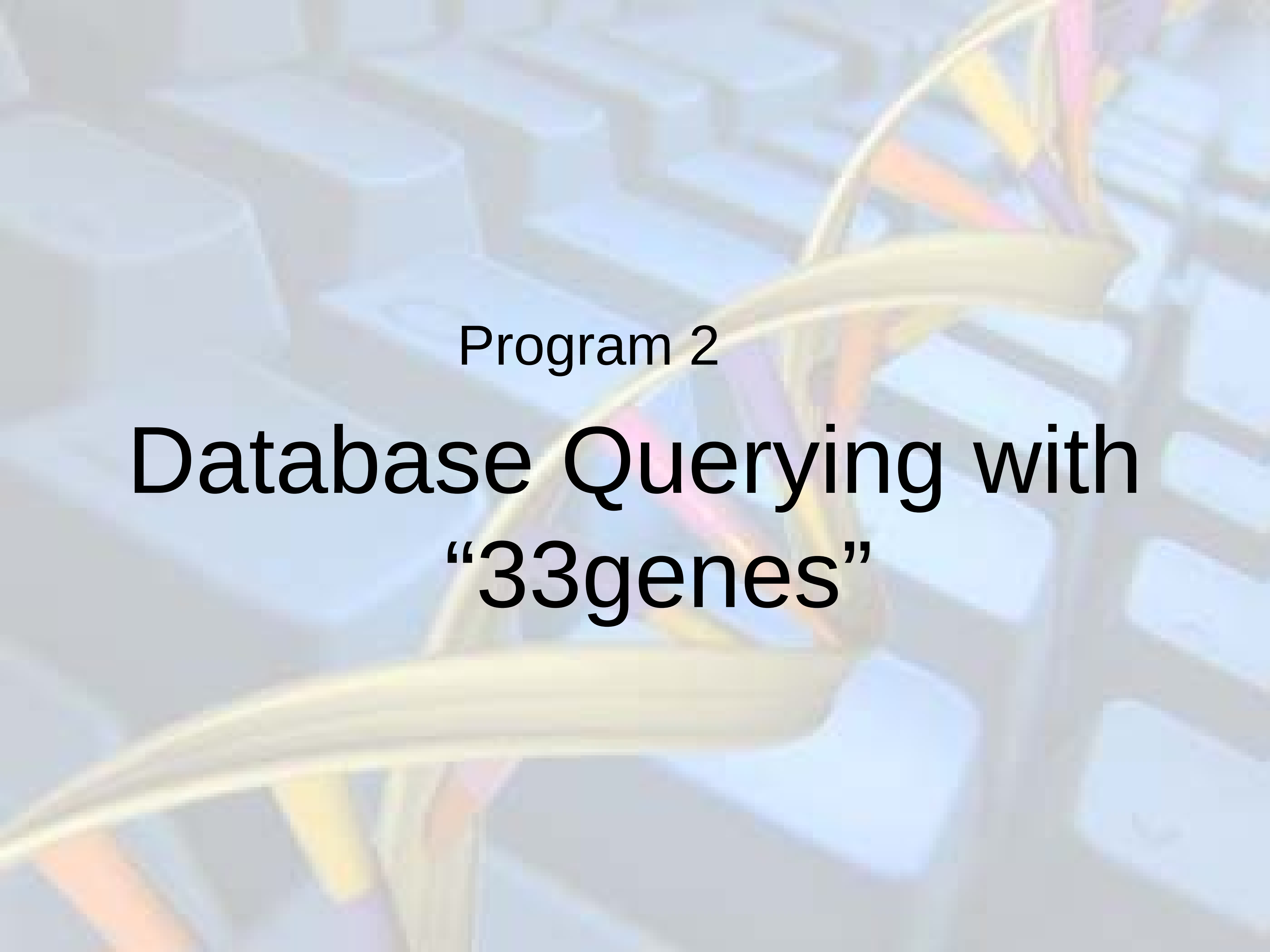


```
select * from finaloutput
```

term_id	description	ontology	go_term
177	nucleotide binding	molecular_function	GO:0000166
2279	RNA binding	molecular_function	GO:0003723
4021	nucleus	cellular_component	GO:0005634
4064	spliceosome	cellular_component	GO:0005681
4032	cytoplasm	cellular_component	GO:0005737
4744	mRNA processing	biological_process	GO:0006397
5137	transport	biological_process	GO:0006810
6160	RNA splicing	biological_process	GO:0008380

Note

- There can be some blast results with no accession numbers.
- The program does not validate input.
- They code right now runs from command prompt but can be easily enhanced to a website!
- Easily enhanced to have different control parameters.

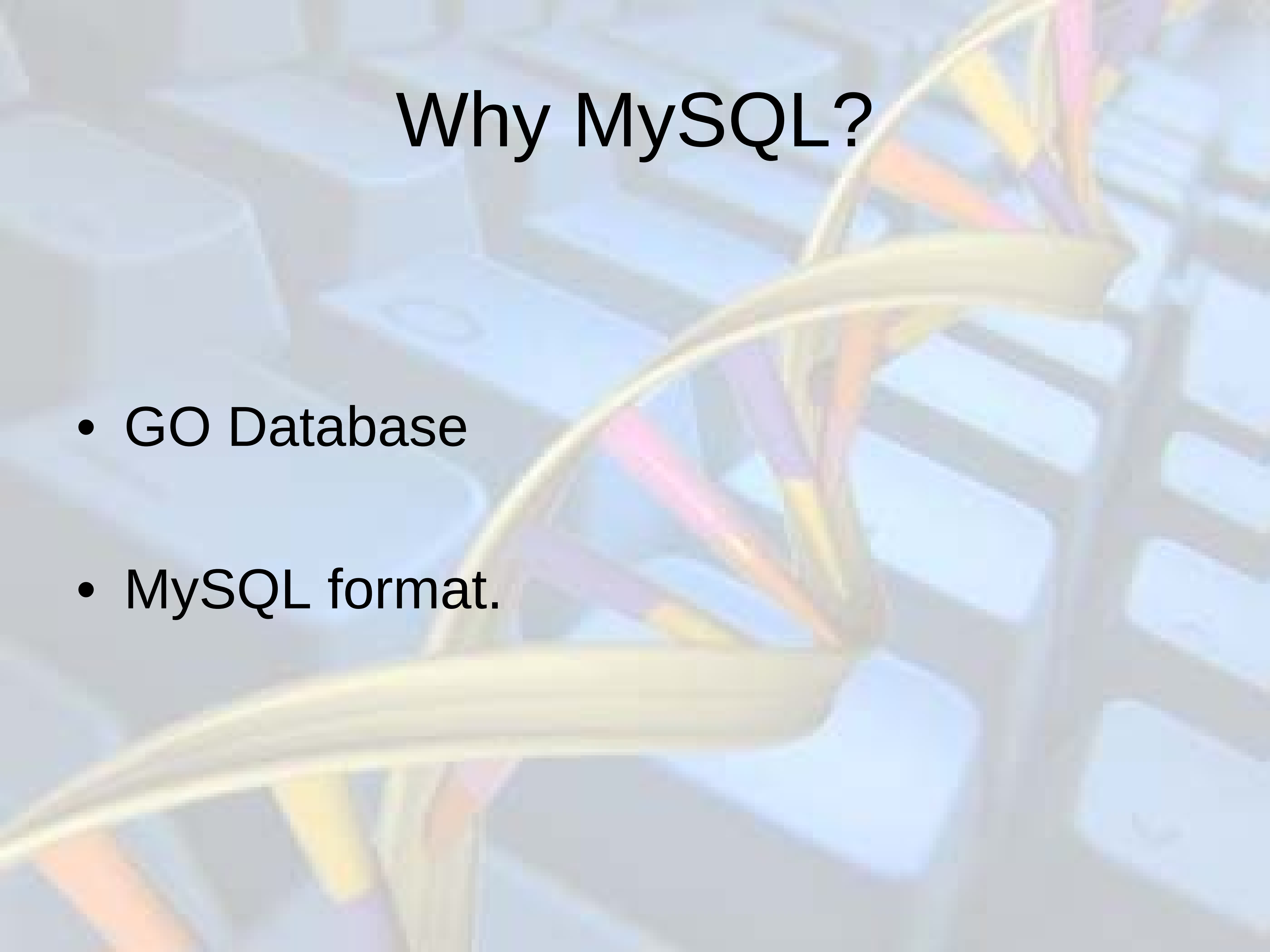


Program 2

Database Querying with “33genes”

MySQL

- Most popular open-source, free, high performance DB engine.
- Fast, reliable, scalable etc.
- Works great with PHP, Perl etc.
- Integrated with common applications



Why MySQL?

- GO Database
- MySQL format.

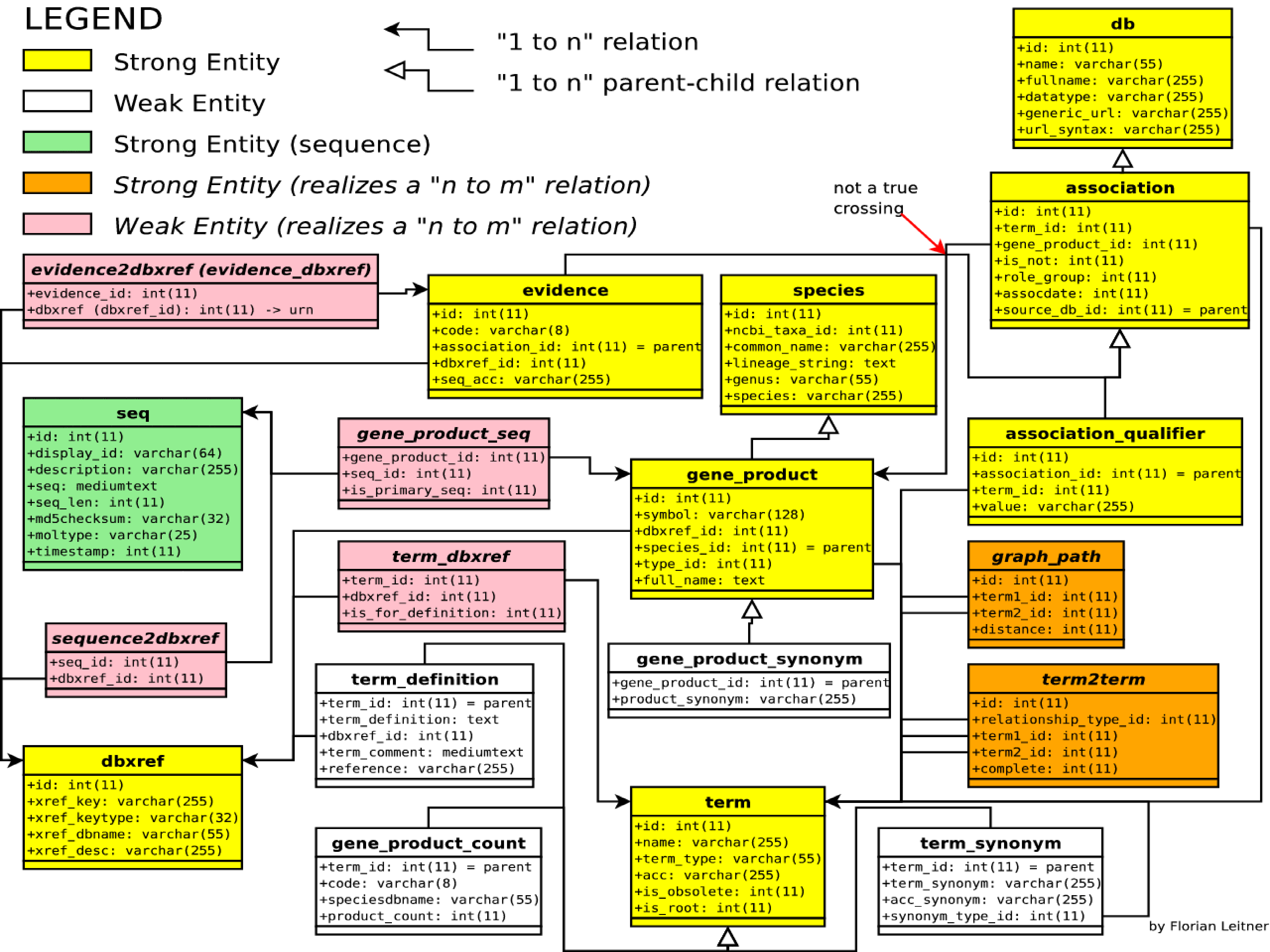
Go Database

- termdb (44 mb)

  Small database, easy to load, less terms

- assocdb (4 gb)

  Large database, difficult to load, more terms very complex.



LEGEND

🟨 Strong Entity

□ Weak Entity

🟩 Strong Entity (sequence)

🟧 *Strong Entity (realizes a "n to m" relation)*

🟥 *Weak Entity (realizes a "n to m" relation)*

"1 to n" relation

"1 to n" parent-child relation

not a true crossing

db
+id: int(11)
+name: varchar(55)
+fullname: varchar(255)
+datatype: varchar(255)
+generic_url: varchar(255)
+url_syntax: varchar(255)

association
+id: int(11)
+term_id: int(11)
+gene_product_id: int(11)
+is_not: int(11)
+role_group: int(11)
+assocdate: int(11)
+source_db_id: int(11) = parent

evidence2dbxref (evidence_dbxref)
+evidence_id: int(11)
+dbxref (dbxref_id): int(11) -> urn

evidence
+id: int(11)
+code: varchar(8)
+association_id: int(11) = parent
+dbxref_id: int(11)
+seq_acc: varchar(255)

species
+id: int(11)
+ncbi_taxa_id: int(11)
+common_name: varchar(255)
+lineage_string: text
+genus: varchar(55)
+species: varchar(255)

seq
+id: int(11)
+display_id: varchar(64)
+description: varchar(255)
+seq: mediumtext
+seq_len: int(11)
+md5checksum: varchar(32)
+moltype: varchar(25)
+timestamp: int(11)

gene_product_seq
+gene_product_id: int(11)
+seq_id: int(11)
+is_primary_seq: int(11)

gene_product
+id: int(11)
+symbol: varchar(128)
+dbxref_id: int(11)
+species_id: int(11) = parent
+type_id: int(11)
+full_name: text

association_qualifier
+id: int(11)
+association_id: int(11) = parent
+term_id: int(11)
+value: varchar(255)

term_dbxref
+term_id: int(11)
+dbxref_id: int(11)
+is_for_definition: int(11)

graph_path
+id: int(11)
+term1_id: int(11)
+term2_id: int(11)
+distance: int(11)

sequence2dbxref
+seq_id: int(11)
+dbxref_id: int(11)

gene_product_synonym
+gene_product_id: int(11) = parent
+product_synonym: varchar(255)

term_definition
+term_id: int(11) = parent
+term_definition: text
+dbxref_id: int(11)
+term_comment: mediumtext
+reference: varchar(255)

term2term
+id: int(11)
+relationship_type_id: int(11)
+term1_id: int(11)
+term2_id: int(11)
+complete: int(11)

dbxref
+id: int(11)
+xref_key: varchar(255)
+xref_keytype: varchar(32)
+xref_dbname: varchar(55)
+xref_desc: varchar(255)

gene_product_count
+term_id: int(11) = parent
+code: varchar(8)
+speciesdbname: varchar(55)
+product_count: int(11)

term
+id: int(11)
+name: varchar(255)
+term_type: varchar(55)
+acc: varchar(255)
+is_obsolete: int(11)
+is_root: int(11)

term_synonym
+term_id: int(11) = parent
+term_synonym: varchar(255)
+acc_synonym: varchar(255)
+synonym_type_id: int(11)

by Florian Leitner

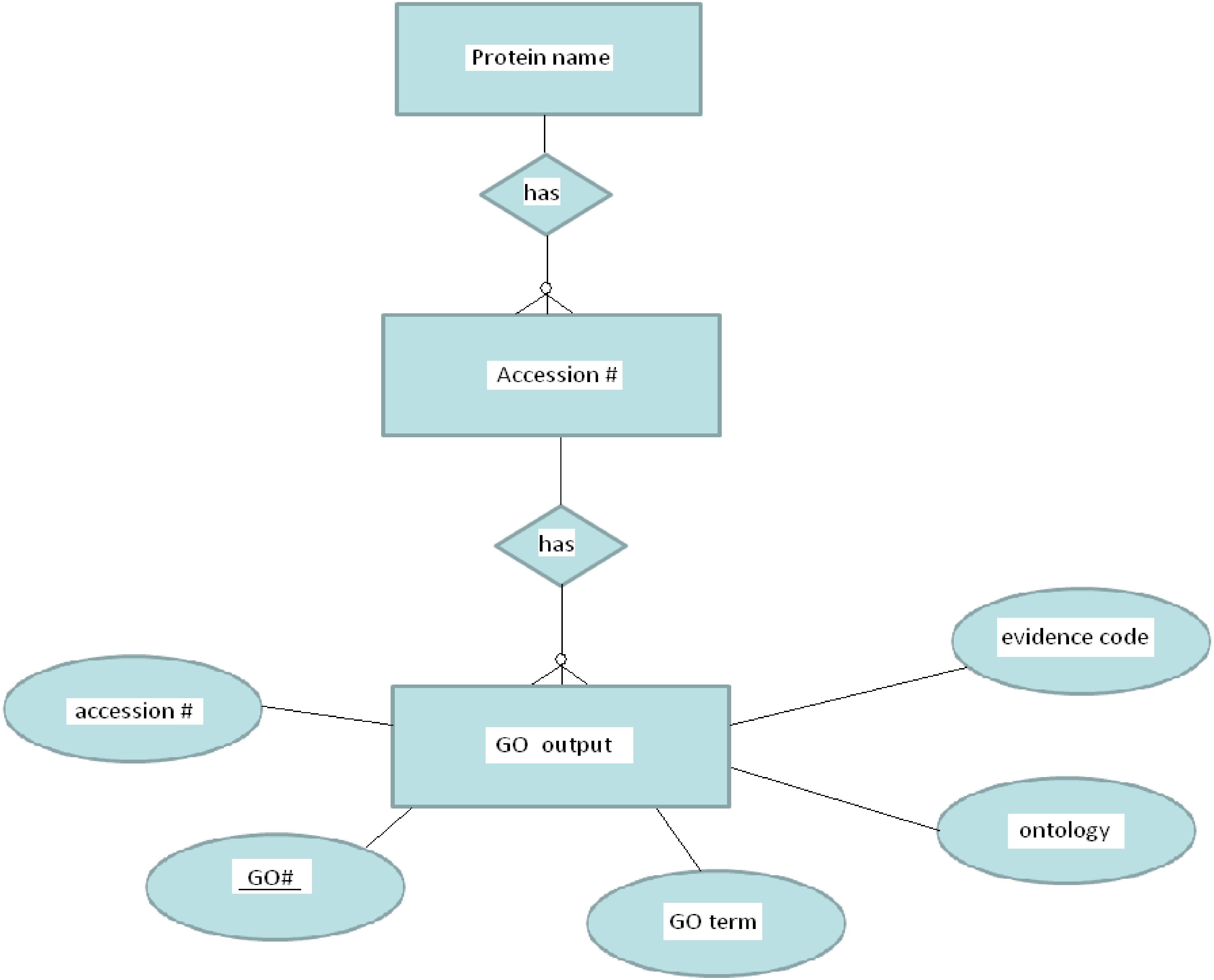
Querying GO database

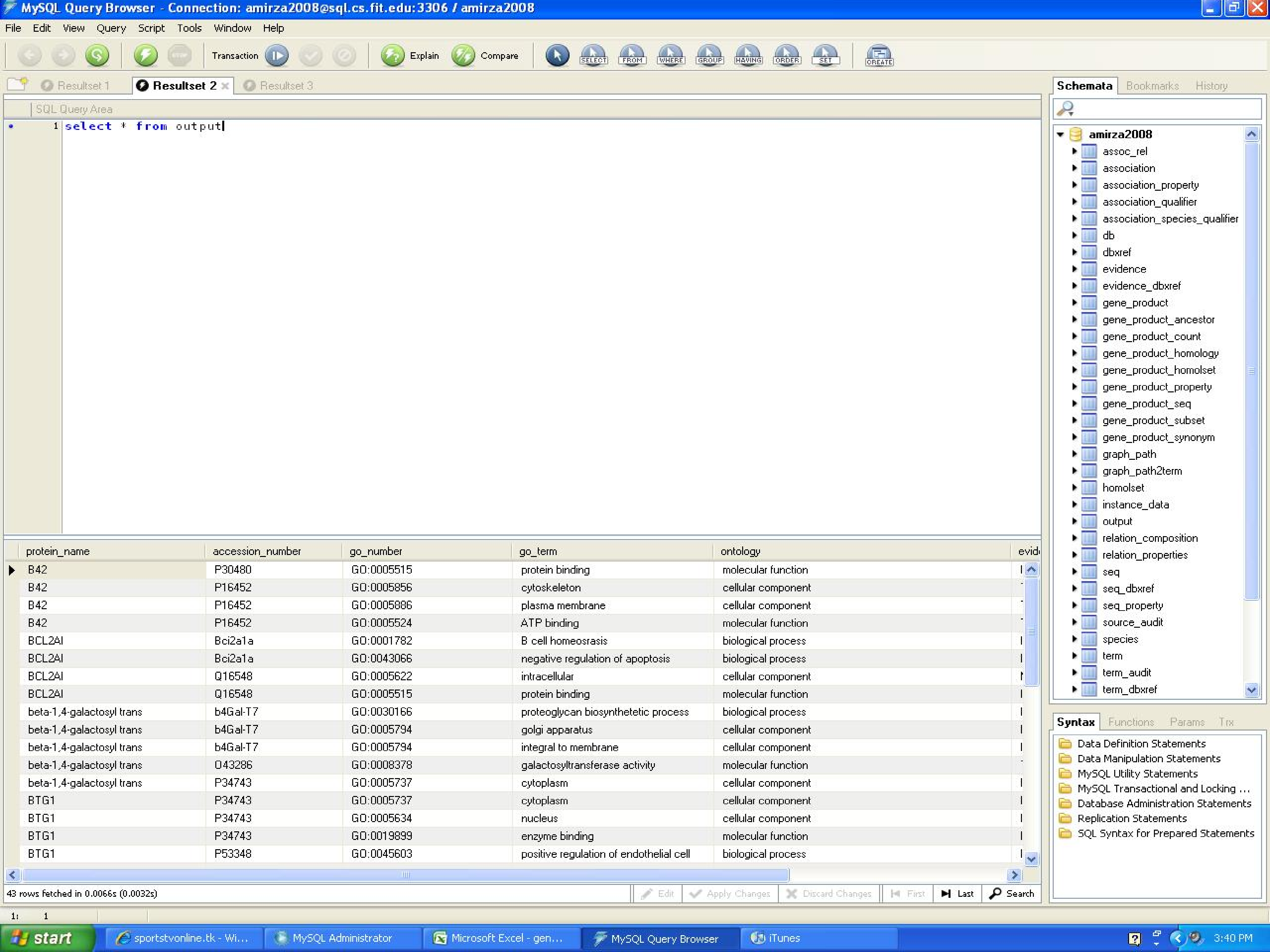
- to get GO terms-

```
select distinct `term`.`name`,`term`.`acc`,`term`.`term_type` from association,term
where `association`.`term_id` = `term`.`id`and (term_id) in
(SELECT distinct term_id FROM association,gene_product where
`gene_product`.`id`=`association`.`gene_product_id` and (`gene_product`.`id`) in
(select id from gene_product where symbol = 'CCR6'))
```

- to get evidence code-

```
(SELECT evidence.association_id FROM evidence where association_id in
(select association.id from association,gene_product where
association.gene_product_id = gene_product.id and symbol='ccr6'))
```


Protein name
has
Accession #
has
evidence code
accession #
GO output
ontology
GO#
GO term



```
select * from output
```

protein_name	accession_number	go_number	go_term	ontology	evid
B42	P30480	GO:0005515	protein binding	molecular function	I
B42	P16452	GO:0005856	cytoskeleton	cellular component	-
B42	P16452	GO:0005886	plasma membrane	cellular component	-
B42	P16452	GO:0005524	ATP binding	molecular function	-
BCL2AI	Bci2a1a	GO:0001782	B cell homeosrasis	biological process	I
BCL2AI	Bci2a1a	GO:0043066	negative regulation of apoptosis	biological process	I
BCL2AI	Q16548	GO:0005622	intracellular	cellular component	I
BCL2AI	Q16548	GO:0005515	protein binding	molecular function	I
beta-1,4-galactosyl trans	b4Gal-T7	GO:0030166	proteoglycan biosynthetic process	biological process	I
beta-1,4-galactosyl trans	b4Gal-T7	GO:0005794	golgi apparatus	cellular component	I
beta-1,4-galactosyl trans	b4Gal-T7	GO:0005794	integral to membrane	cellular component	I
beta-1,4-galactosyl trans	O43286	GO:0008378	galactosyltransferase activity	molecular function	-
beta-1,4-galactosyl trans	P34743	GO:0005737	cytoplasm	cellular component	I
BTG1	P34743	GO:0005737	cytoplasm	cellular component	I
BTG1	P34743	GO:0005634	nucleus	cellular component	I
BTG1	P34743	GO:0019899	enzyme binding	molecular function	I
BTG1	P53348	GO:0045603	positive regulation of endothelial cell	biological process	I

43 rows fetched in 0.0066s (0.0032s)

Part 4

- Discussion
- Recommendations
- Conclusion

Discussion

- BLASTing AmiGOs was able to take FASTA sequences and generate GO terms for each sequence completely automatically.

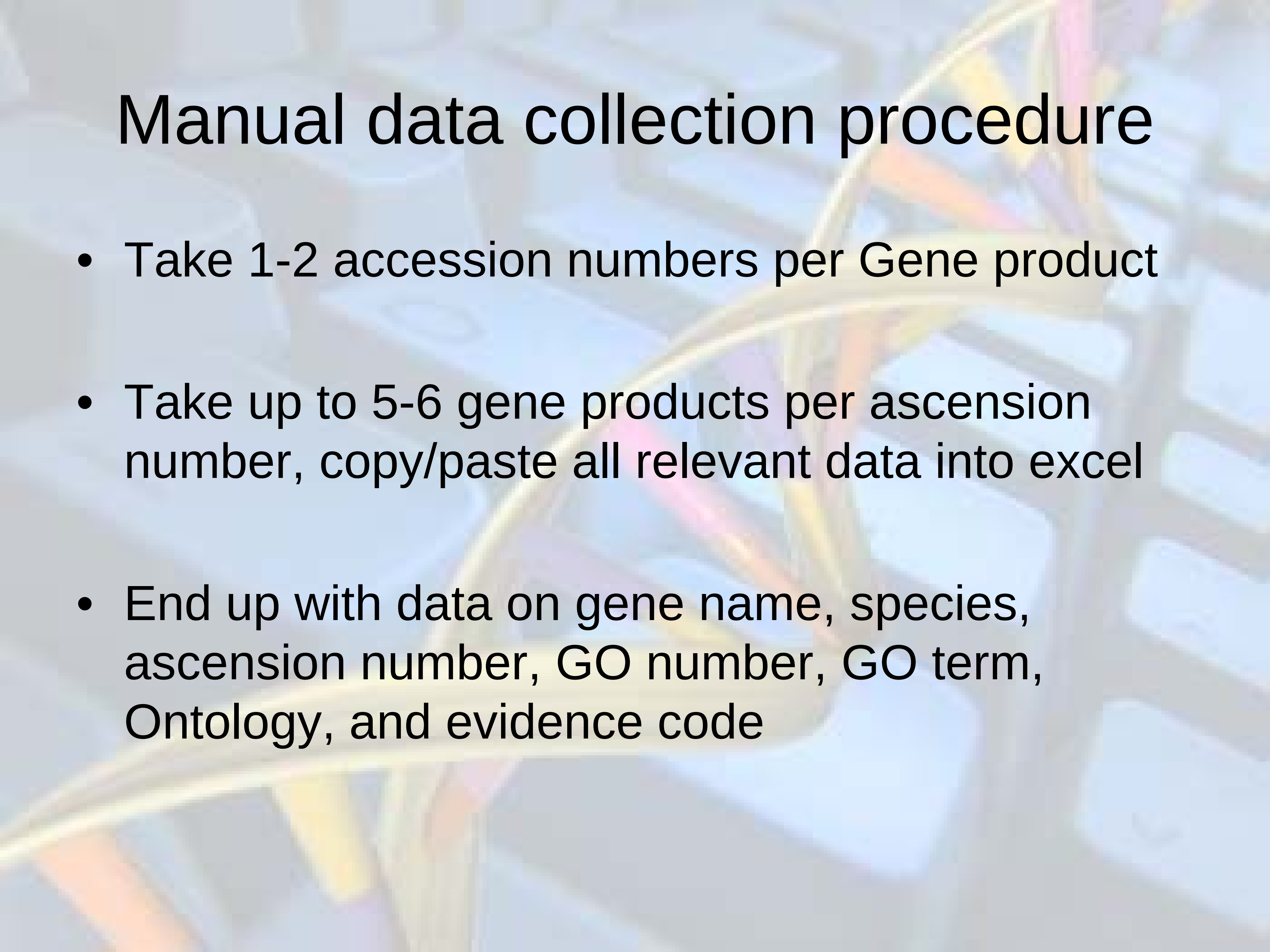
- “33” was able to take Gene products and find GO terms for them and dump them into the GO output Database.

- To give a comparison, Griffin and Azhar ran the 33 genes into AmiGO and MANUALLY extracted the GO terms and built a database (in excel)

Why Manually?

- Biologists tend to not consult computer scientists to automate data collection

- It is common for biologists to do manual data collection because hiring a computer scientist to automate it cost too much.



Manual data collection procedure

- Take 1-2 accession numbers per Gene product
- Take up to 5-6 gene products per ascension number, copy/paste all relevant data into excel
- End up with data on gene name, species, ascension number, GO number, GO term, Ontology, and evidence code

Manual data results

- Collected 155 Go terms for 32 genes with 1 gene having no hits
- Took about 4-5 hours to get a partial GO term list, estimating about 8-12 hours for a complete list
- Human error is very likely to cause atleast a few mistakes in the database

species	Assention number	Go number	Go term	ontology	evidence code
Homo sapiens	P30480	GO:0005515	protein binding	molecular function	IPI
Homo sapiens	P16452	GO:0005856	cytoskeleton	cellular component	TAS
Homo sapiens	P16452	GO:0005886	plasma membrane	cellular component	TAS
Homo sapiens	P16452	GO:0005524	ATP binding	molecular function	TAS
Mus musculus	Bcl2a1a	GO:0001782	B cell homeostasis	biological process	IDA
Mus musculus	Bcl2a1a	GO:0043066	negative regulation of apoptosis	biological process	IDA
Homo sapiens	Q16548	GO:0005622	intracellular	cellular component	NAS
Homo sapiens	Q16548	GO:0005515	protein binding	molecular function	IPI
Pan troglodytes	b4Gal-T7	GO:0030166	proteoglycan biosynthetic process	biological process	ISS
Pan troglodytes	b4Gal-T7	GO:0005794	Golgi apparatus	cellular component	IDA
Pan troglodytes	b4Gal-T7	GO:0016021	integral to membrane	cellular component	ISS
Homo sapiens	O43286	GO:0008378	galactosyltransferase activity	molecular function	TAS
Gallus gallus	P34743	GO:0005737	cytoplasm	cellular component	ISS
Gallus gallus	P34743	GO:0005634	nucleus	cellular component	ISS
Gallus gallus	P34743	GO:0019899	enzyme binding	molecular function	ISS
Bos taurus	P53348	GO:0045603	positive regulation of endothelial cell differentiation	biological process	ISS
Bos taurus	P53348	GO:0042981	regulation of apoptosis	biological process	ISS
Bos taurus	P53348	GO:0005737	cytoplasm	cellular component	ISS
Homo sapiens	P51684	GO:0006935	chemotaxis	biological process	TAS
Homo sapiens	P51684	GO:0007204	elevation of cytosolic calcium ion concentration	biological process	TAS
Homo sapiens	P51684	GO:0006959	humoral immune response	biological process	TAS
Homo sapiens	P35354	GO:0019371	cyclooxygenase pathway	biological process	NAS
Homo sapiens	P35354	GO:0008217	regulation of blood pressure	biological process	ISS
Homo sapiens	P35354	GO:0050727	regulation of inflammatory response	biological process	NAS
Homo sapiens	Q99424	GO:0008206	bile acid metabolic process	biological process	TAS
Homo sapiens	Q99424	GO:0005777	peroxisome	cellular component	NAS
Homo sapiens	Q99424	GO:0003997	acyl-CoA oxidase activity	molecular function	TAS
Homo sapiens	P09919	GO:0008284	positive regulation of cell proliferation	biological process	TAS
Homo sapiens	P09919	GO:0005737	cytoplasm	cellular component	IDA
Homo sapiens	P09919	GO:0005856	cytoskeleton	cellular component	IDA
Homo sapiens	P09919	GO:0005615	extracellular space	cellular component	TAS
Homo sapiens	Q99062	GO:0006952	defense response	biological process	TAS
Homo sapiens	Q99062	GO:0007165	signal transduction	biological process	NAS
Homo sapiens	Q99062	GO:0005887	integral to plasma membrane	cellular component	TAS
Homo sapiens	Q99062	GO:0004872	receptor activity	molecular function	TAS
xxxxxxx	xxxx	xxxxxx	xxxxxx	xxxxxxx	xxxxxx
Homo sapiens	Q16690	GO:0006470	protein amino acid dephosphorylation	biological process	TAS
Homo sapiens	Q16690	GO:0004725	protein tyrosine phosphatase activity	molecular function	TAS
Bos taurus	P42891	GO:0016486	peptide hormone processing	biological process	IDA
Bos taurus	P42891	GO:0051605	protein maturation via proteolysis	biological process	IDA
Bos taurus	P42891	GO:0042803	protein homodimerization activity	molecular function	IPI

Automatic method

- The 33 genes can have all their GO terms located in a short period of time (around 10-15 minutes)
- This method removes virtually all human error involved in collecting Go terms
- Less Sanity is lost in the process

Conclusion

- We were able to learn about the Gene Ontology project, PERL (BIOPERL), and MySQL.

- We were able to automate various portions of converting FASTA files to GO terms associations and to automate database querying to remarkably reduce human input.

- Running our automated scripts was orders of magnitude faster than doing it manually, more complete, and more accurate.

Recommendations

- Should have had better project guidelines
- More human interaction can be automated from both programs
- Scoring system for Go terms could be implemented
- Finding a way to query in parallel instead of in series
- Finding a way to Query AmiGO remotely without downloading it

References

- www.geneontology.org
- www.NCBI.gov/blast
- https://www.sapac.edu.au/gosling/
- www.cpan.org