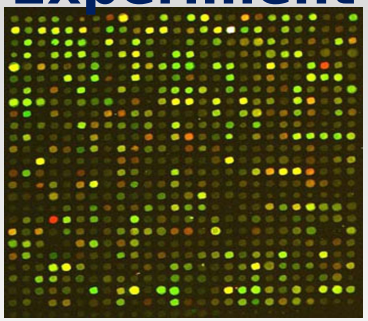
Functional annotation with Blast2GO

Why Blast2GO?

Experiment



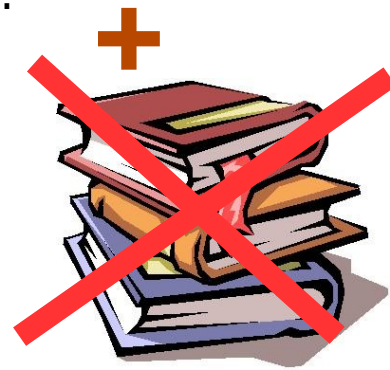
Data-Analysis

GEPAS
Gene Expression Pattern Analysis Suite

Gene-List

MNAT1
CTNNBL1
ENOX2
GTPBP1
RALY
TAGLN2
RAB3A
PPP2R5A
MAPRE1
.....
...

Functional Annotation



Functional Profiling

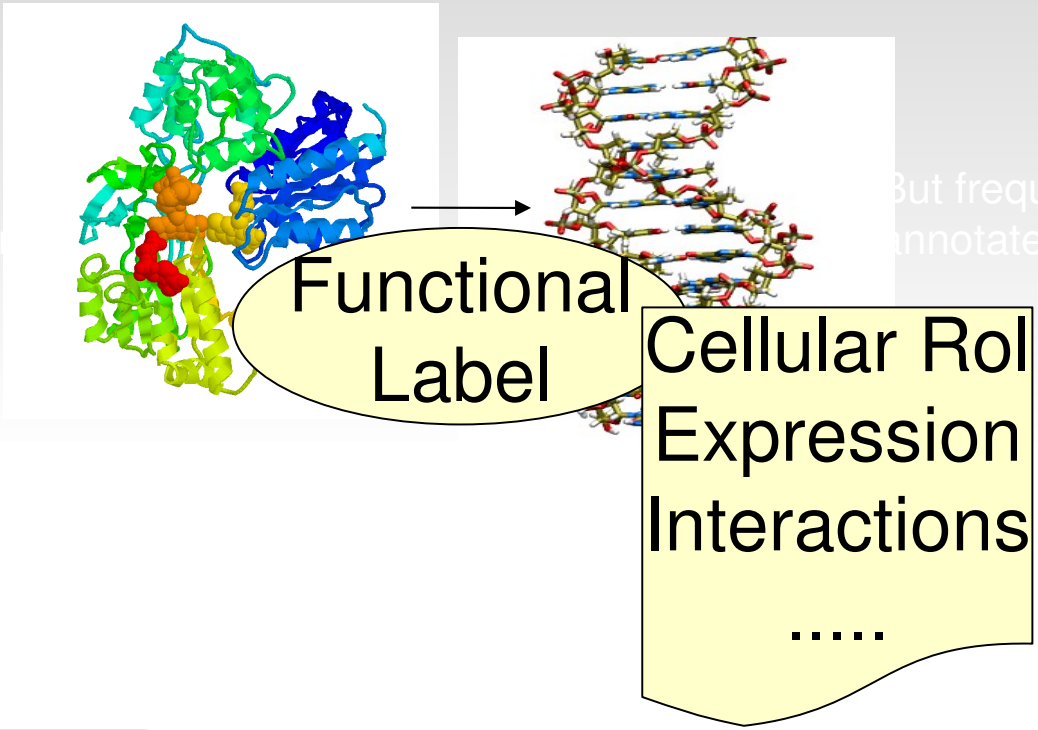
Functional interpretation

What does Blast2GO do?

Generates annotations

Visualization of funcional annotations

What is functional annotation?



Controlled Vocabulary

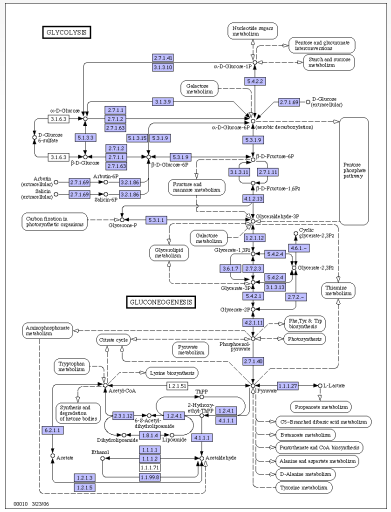
High throughput

Accessible

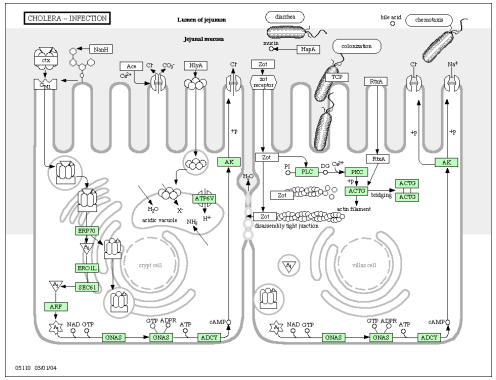
Functional Vocabularies

Molecular Function
Biological Process
Cellular Component

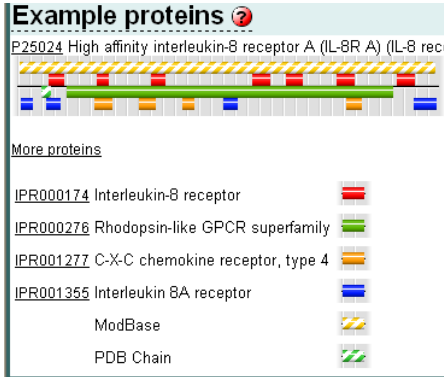
Metabolic pathways



KEGG orthologues



Functional motifs

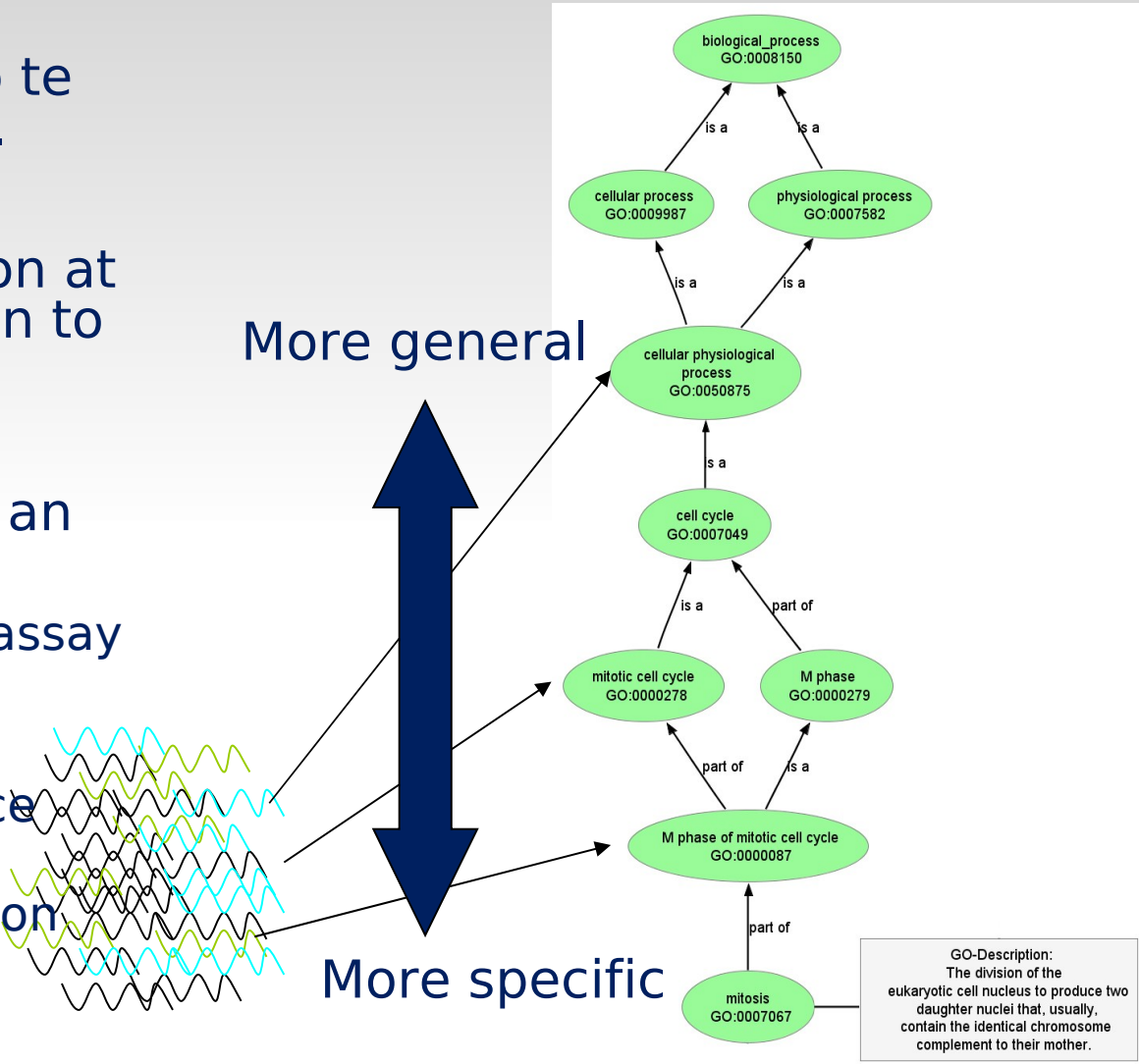


The Gene Ontology

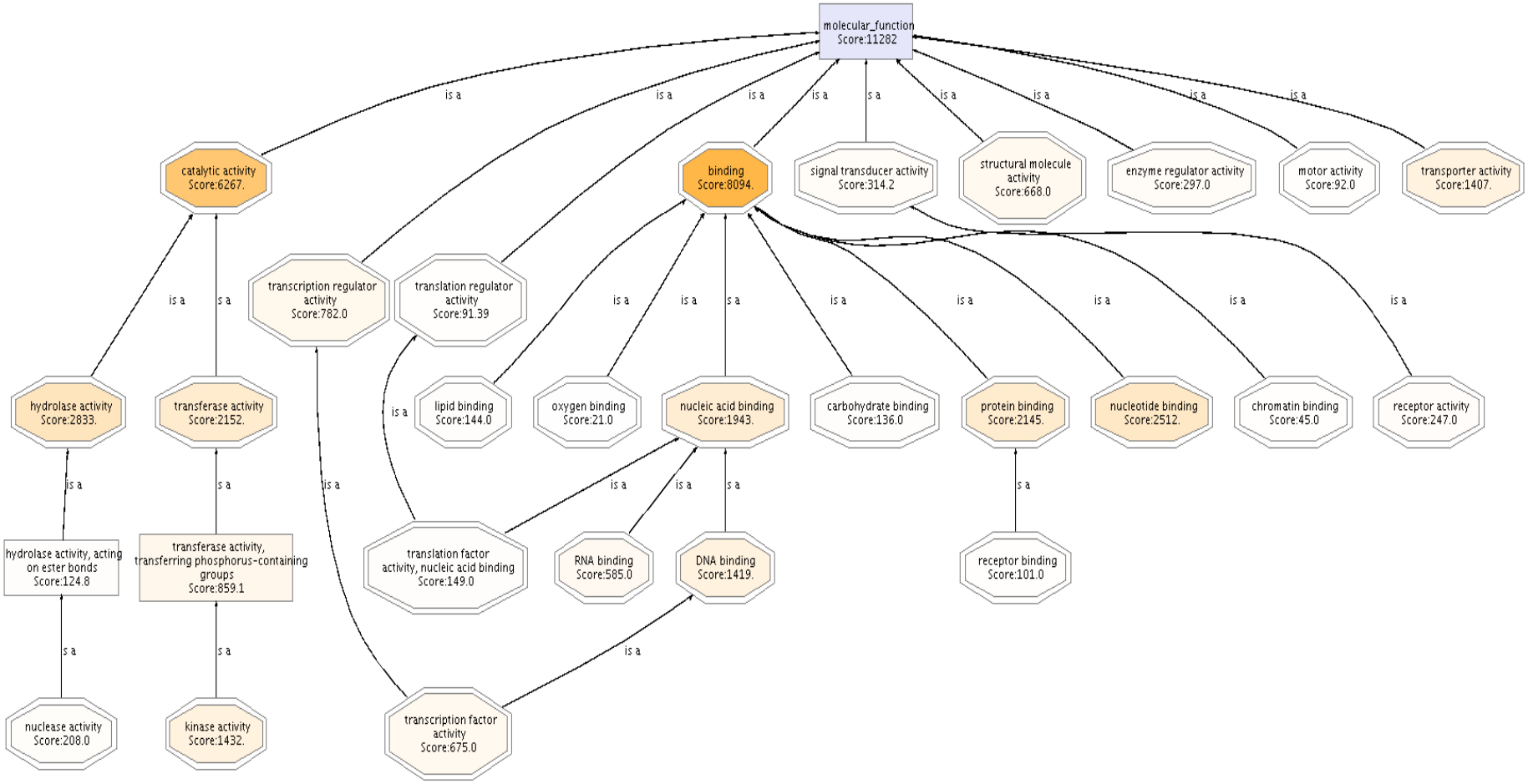
- ✓ Project developed by the Gene Ontology Consortium
- ✓ Provides a controlled vocabulary to describe gene and gene product attributes in any organism
- ✓ Includes both the development of the Ontology and the maintenance of a Database of annotations

The Ontology

- ✓ Annotations are given to te most specific (low) level.
- ✓ True path rule: annotation at a term implies annotation to all its parent terms
- ✓ Annotation is given with an Evidence Code:
  - o **IDA**: inferred by direct assay
  - o **TAS**: traceable author statement
  - o **ISS**: infered by sequence similarity
  - o **IEA**: electronic annotation
  - o ….



The GO has a DAG structure

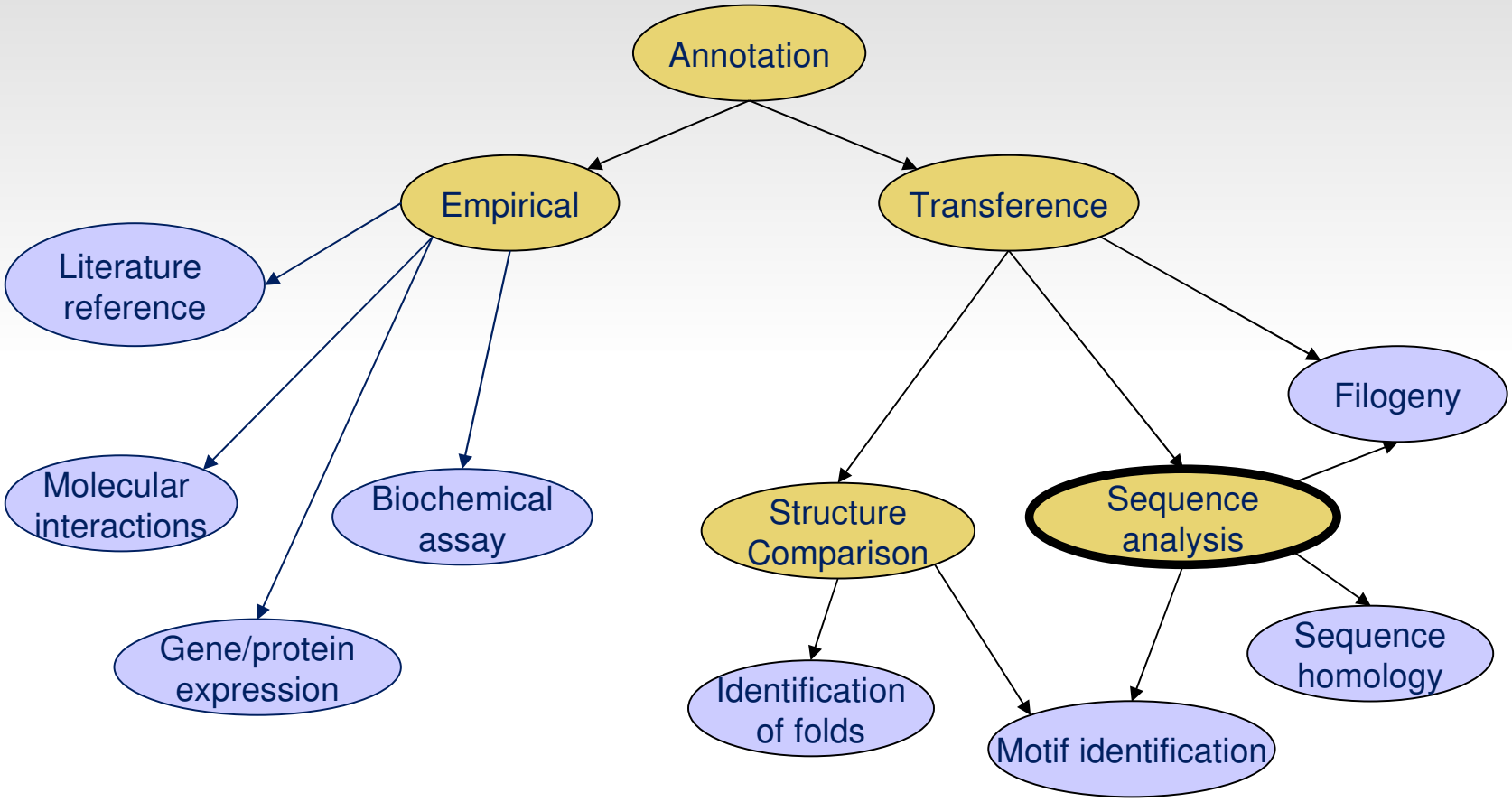


The Gene Ontology Database (GOA)

http://www.geneontology.org/GO.current.annotations.shtml

- ✓ There is a collaborating institution per organism to provide annotations
- ✓ Most of the GOA annotations come from UniProt
- ✓ Most of the annotations are electronic annotations

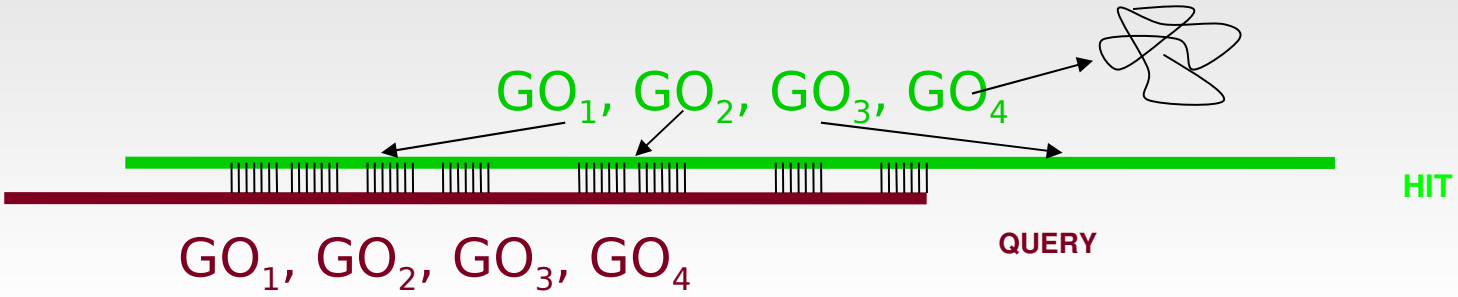
Functional assignment



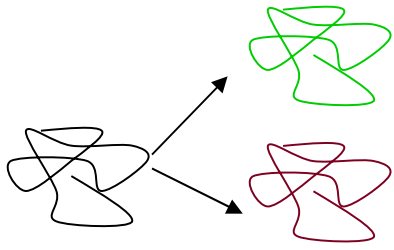
Automatic annotation

- ✓ GO annotations can be created by comparision to annotated sequences
- ✓ To achieve enough coverage, high-throughput, automatic annotation is required
- ✓ The most effective (also error prone) automatic annotation method is transfer from sequence similarity

Concerns in functional transfer by similarity



- ✓ Level of homology (~ from 40-60% is possible)
- ✓ The overlap query and hit sequences
- ✓ The domain or structure function association
- ✓ The paralog problem: genes with similar sequences might have different functional specifications
- ✓ The evidence for the original annotation
- ✓ Balance between quality and quantity: depends on the use

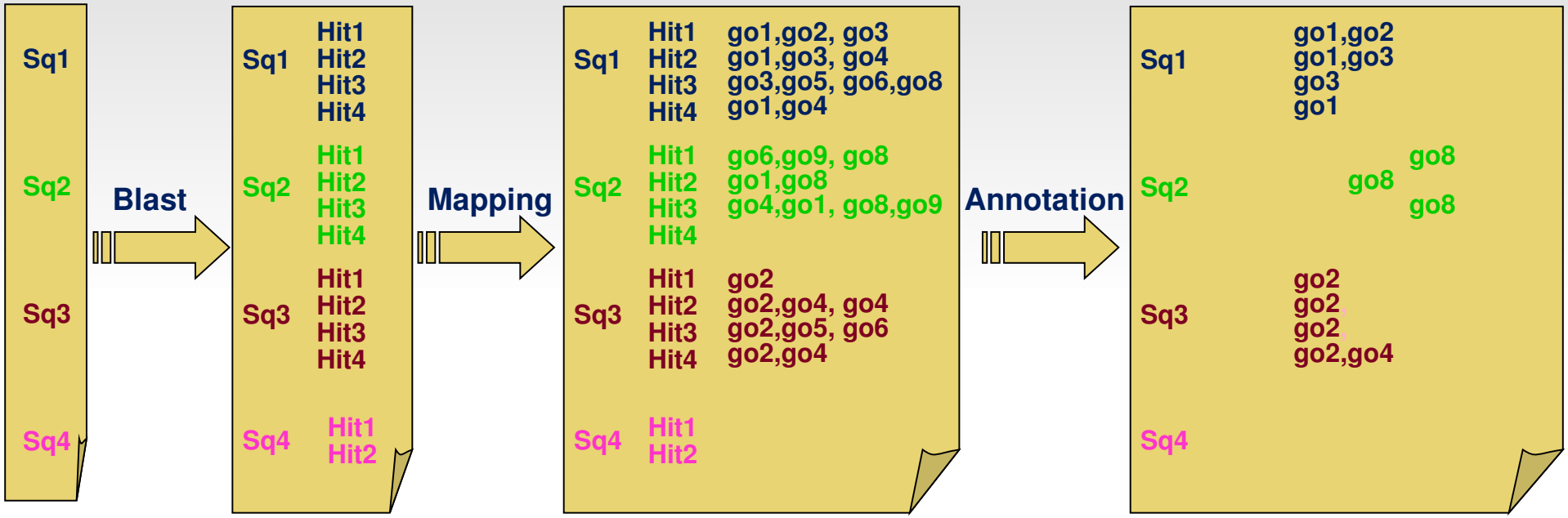


Blast2GO

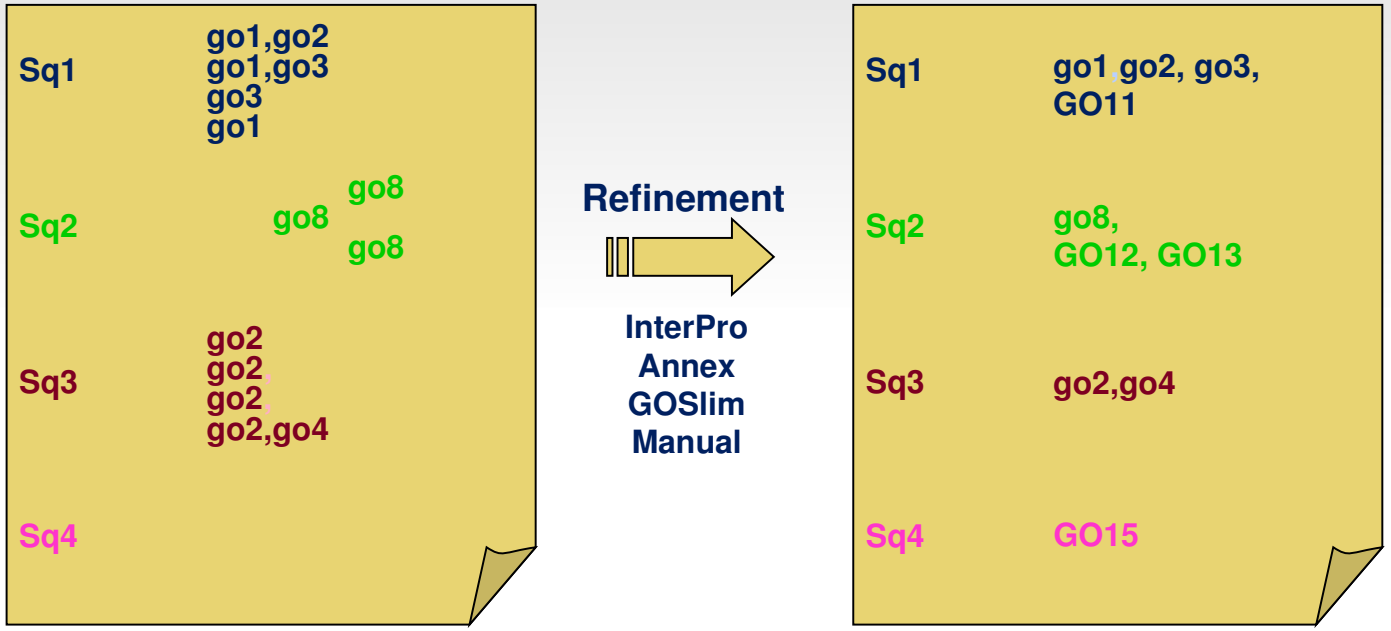
- ✓ Suite for functional annotation and data mining on functional data
  - Considerations for **annotation**
    - Simlarity
    - Length of the overlap
    - Percentage of hit sequence spanned by the overlap
    - Evidence original annotation
    - Blast hits and motif hits
    - Refinement by additional methods
  - Visualization:
    - Annotation charts
    - **Knowledge discovery on the DAG**

- ✓ Desktop Java application
- ✓ web interface @ Babelomics: **Babelomics for non-model**

Blast2GO Annotation strategy



Blast2GO Annotation Strategy



Blast2GO annotation rule

Lowest term above threshold

Similarity requirement

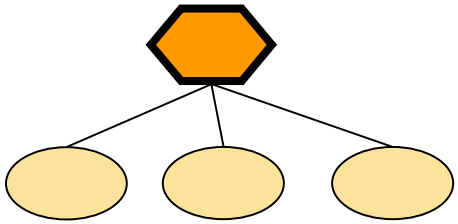
$$sim = \frac{\sum positives_{hsp}}{\sum alignmentlength_{hsp}}$$

Quality of annotation source

EC	weight
IC	1
TAS	1
IDA	1
IMP	0.9
IGI	0.9
IPI	0.9
ISS	0.8
IEP	0.8
NAS	0.7
IEA	0.7
ND	0.5
NR	0.5
RCA	0.5

Evidence codes

Possibility of abstraction



Recall vs. Precision

Lowest.node [(max.sim x ECw) + (#GO-1 x GOw) >= threshold]

Blast2GO annotation rule

Lowest.node [(max.sim x ECw) + (#GO-1 x GOw) >= threshold]

- When I have a GO with ECw = 1 and I do not allow abstraction (GOw = 0), then the Annotation Score = %similarity
- If the ECw < 1 my similarity requirement is higher to obtain the same Annotation Score
- If I allow abstraction GOw > 0, then with less similarity I can obtain the required Annotation Score at a parent node

Annotation example

Let consider Sequence Query A with the following Blast result:

Hit sequence	% similarity	#GO terms	Evidence Code
1	**60%**	GO1	**IDA**
2	**65%**	GO2	**ISS**
3	**67%**	GO3	**IEA**

GO2 and **GO3** are brother terms with parent term GO4

AS = %sim * ECw + (#GO-1) * GOw

Which GO annotations will be transferred?

Annotation example

Hit sequence	% similarity	#GO terms	Evidence Code
1	**60%**	GO1	**IDA**
2	**65%**	GO2	**ISS**
3	75%	GO3	**IEA**

GO2 and **GO3** are brother terms with parent term GO4

AS = %sim * ECw + (#GO-1) * GOw

Scenario 1

- ECw (IDA)=1; ECw(ISS) = 0.8; ECw(IEA) = 0.6 **(Evidence Code Control)**
- Annotation threshold is set to **55**
- **GOw = 0** (no contribution from children terms)

AS(GO1) = **(60 * 1)** + (1-1 * 0) = 60 > 55 --> **GO1 is transfered** to the query sequence
AS(GO2) = (65 * 0.8) + (1-1 * 0) = 52 < 55 --> GO2 is NOT transfered
AS(GO3) = (67 * 0.7) + (1-1 * 0) = 47 < 55 --> GO3 is NOT transfered
AS(GO4) = (67 * 0.7) + (2-1 * 0) = 47 < 55 --> GO4 is NOT transfered

Annotation example

Hit sequence	% similarity	#GO terms	Evidence Code
1	**60%**	GO1	**IDA**
2	**65%**	GO2	**ISS**
3	**67%**	GO3	**IEA**

GO2 and **GO3** are brother terms with parent term GO4

AS = %sim * ECw + (#GO-1) * GOw

Scenario 2

- ECw (IDA)=1; ECw(ISS) = 0.8; ECw(IEA) = 0.7 **(Evidence Code Control)**
- Annotation threshold is set to **55**
- **GOw = 10 (the children contribution is enabled)**

AS(GO1) = (60 * 1) + (1-1 * 10) = 60 > 55 --> **GO1 is transfered** to the query sequence

AS(GO2) = (65 * 0.8) + (1-1 * 10) = 52 < 55 --> GO2 is NOT transfered

AS(GO3) = (67 * 0.7) + (1-1 * 10) = 47 < 55 --> GO3 is NOT transfered

AS(GO4) = (67 * 0.7) + **(2-1 * 10)** = 57 > 55 --> **GO4 is transfered**

Annotation example

Hit sequence	% similarity	#GO terms	Evidence Code
1	**60%**	GO1	**IDA**
2	**65%**	GO2	**ISS**
3	**67%**	GO3	**IEA**

GO2 and **GO3** are brother terms with parent term GO4

AS = %sim * ECw + (#GO-1) * GOw

Scenario 3

- ECw (IDA)=1; ECw(ISS) = 0.8; ECw(IEA) = 0.7 **(Evidence Code control)**
- Annotation threshold is set to **50**
- **GOw = 10** (the children contribution is enabled)

AS(GO1) = (60 * 1) + (1-1 * 10) = 60 > 50 --> **GO1 is transfered** to the query sequence

AS(GO2) = (65 * 0.8) + (1-1 * 10) = **52 > 50** --> **GO2 is transfered** to the query sequence

AS(GO3) = (67 * 0.7) + (1-1 * 10) = 47 < 50 --> GO3 is NOT transfered

AS(GO4) = (67 * 0.7) + (2-1 * 10) = 57 > 50 --> GO4 is NOT transfered (transferred child)

Annotation example

Hit sequence	% similarity	#GO terms	Evidence Code
1	**60%**	GO1	**IDA**
2	**65%**	GO2	**ISS**
3	**67%**	GO3	**IEA**

GO2 and **GO3** are brother terms with parent term GO4

$$AS = \%sim * ECw + (\#GO-1) * GOw$$

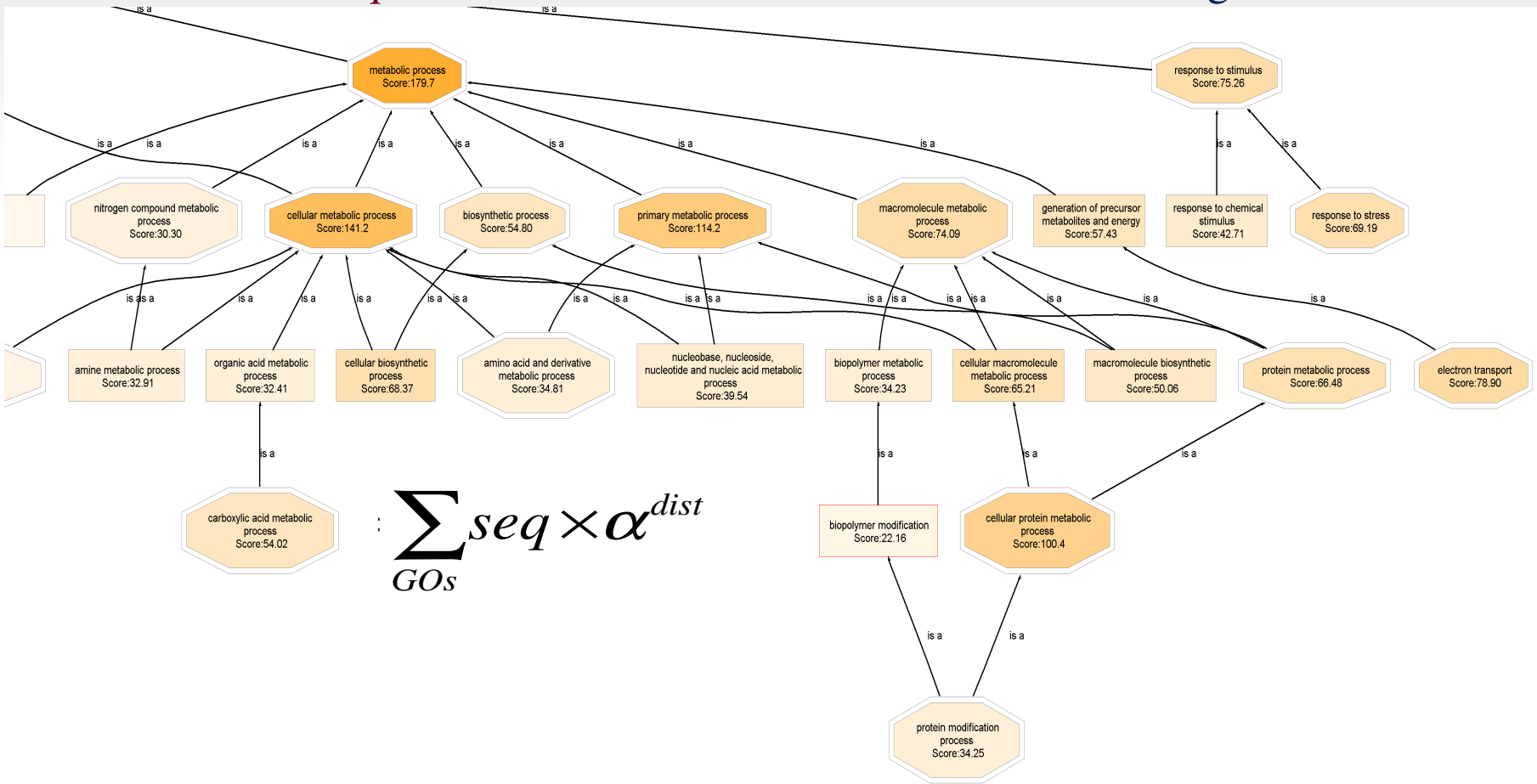
Scenario 4

- ECw (IDA)=1; ECw(ISS) = 1; ECw(IEA) = **1 (no Evidence Code control)**
- Annotation threshold is set to **55**
- **GOw = 10** (the childern contribution is enabeled)

AS(GO1) = (60 * 1) + (1-1 * 10) = 60 > 55 --> **GO1 is transfered** to the query sequence
AS(GO2) = **(65 * 1)** + (1-1 * 10) = 65 > 55 --> **GO2 is transfered**
AS(GO3) = **(67 * 1)** + (1-1 * 10) = 67 > 55 --> **GO3 is transfered**
AS(GO4) = (67 * 1) + (2-1 * 10) = 77 > 55 --> GO4 is NOT transfered (transferred child)

B2G Highlighting on the DAG
The B2G Score

- ✓ Coloring strategy to highlight regions in the DAG where the most interesting information is concentrated
- ✓ The confluence score (B2G score) keeps a balance between the number of annotated sequences at one node and the distance to the origin of annotation



$$\sum_{GOs} seq \times \alpha^{dist}$$

Hands-on