ClusterInspector

a tool to visualize ontology based relationships

Sèverine Bérard   Laurent Tichit   **Carl Herrmann**

INRA Toulouse   IML & LIF Marseille   LGPD/IBDM Marseille

Jobim 2005 - Lyon

SimCT : Similarity Clustering Tree

a tool to visualize ontology based relationships

Sèverine Bérard

INRA Toulouse

Laurent Tichit

IML & LIF Marseille

Carl Herrmann

LGPD/IBDM Marseille

A frequent situation

large scale biological dataset

how to group together **objects**
which are **similar** ?

A frequent situation

➢ **objects ?**

- genes
- proteins
- DNA/RNA stretches
- tissues
- mutants
- ...

➢ **similar ?**

- sequence
- 3D structure
- expression pattern
- function
- localization
- phenotype
- ...

A frequent situation

➢ **objects ?**

- genes
- proteins
- DNA/RNA stretches
- tissues
- mutants
- ...

➢ **similar ?**

- sequence
- 3D structure
- expression pattern
- function
- localization
- phenotype
- ...

**similarity : often more *qualitative*
than *quantitative* ...**

ontologies

➢ features increasingly described by ontologies

- functions
- anatomical description
- developmental stages

Open Biomedical Ontologies
~ 40 ontologies
http://obo.sourceforge.net/main.html

➢ definition:
*an ontology is a **hierarchical** data structure with **terms** and **relationships***

ontologies

➢ features increasingly described by ontologies

- functions
- anatomical description
- developmental stages

→ **Open Biomedical Ontologies**
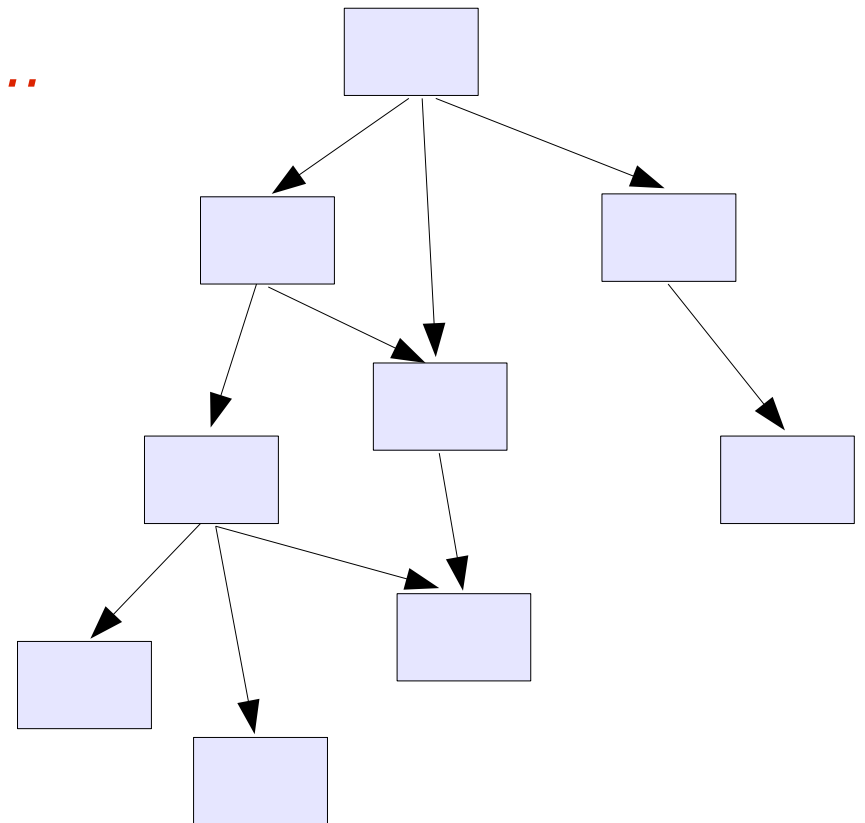~ 40 ontologies
http://obo.sourceforge.net/main.html

➢ definition:
an ontology is a **hierarchical** data structure with **terms** and **relationships**

the structure of ontologies (DAG) can be used to quantify a similarity between terms

structure of ontologies
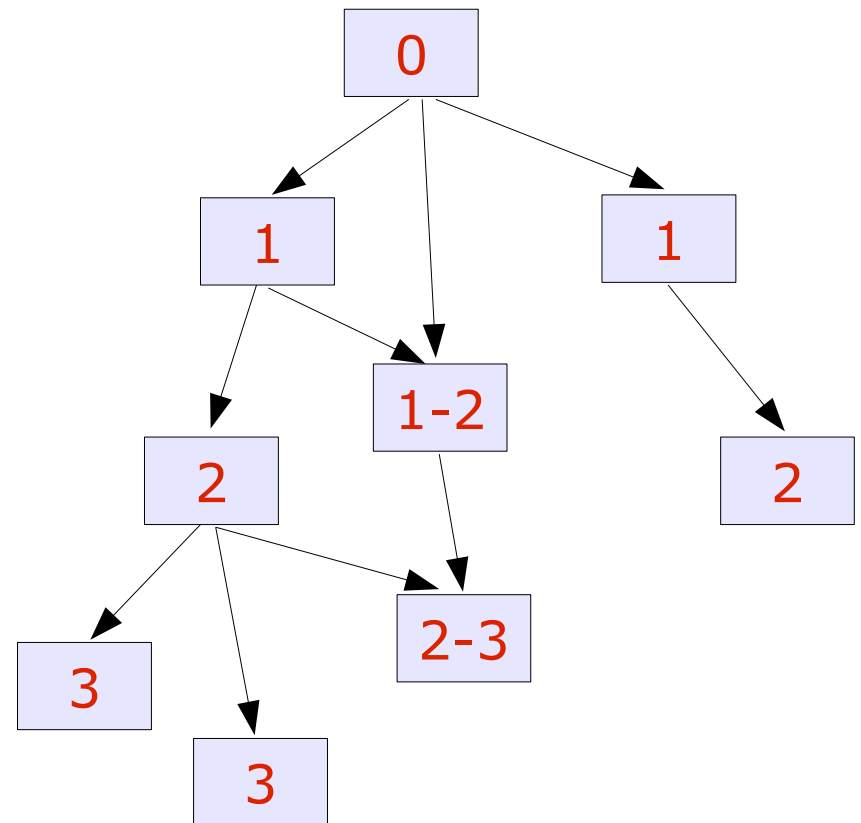
*higher terms are more **generic**…*

*…lower terms are more **specific***

structure of ontologies

*higher terms are more **generic**...*

can we quantify this ?

*...lower terms are more **specific***
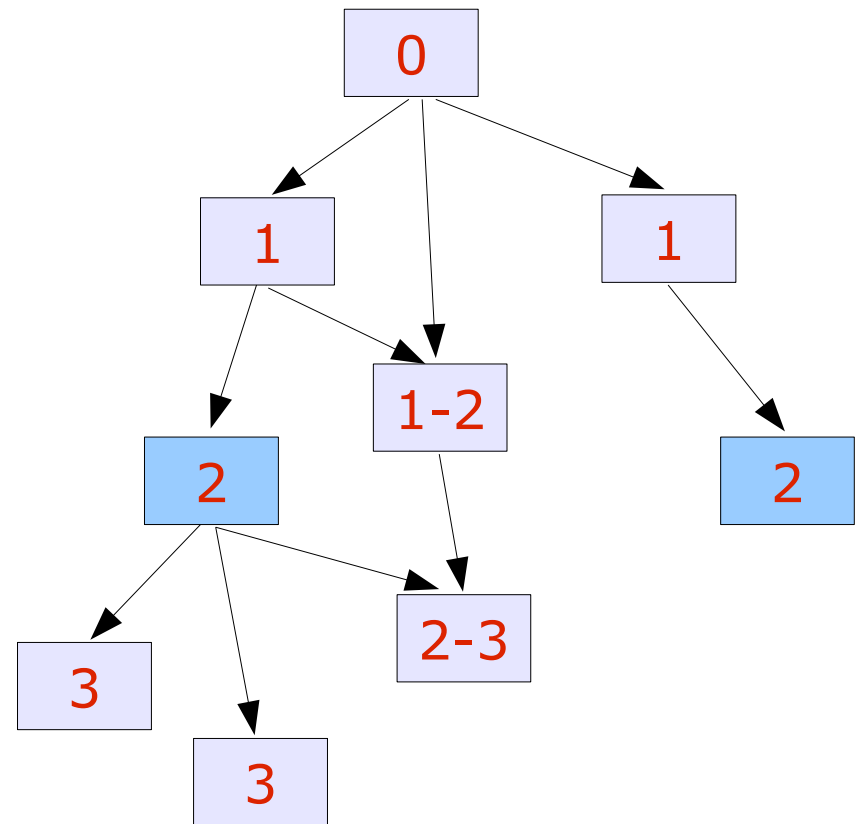
structure of ontologies

- ➤ **depth**
  - not unambigously defined

  - does not take into account the fact that some branches are longer than others

structure of ontologies

➢ **depth**

- not unambigously defined

- does not take into account the fact that some branches are longer than others
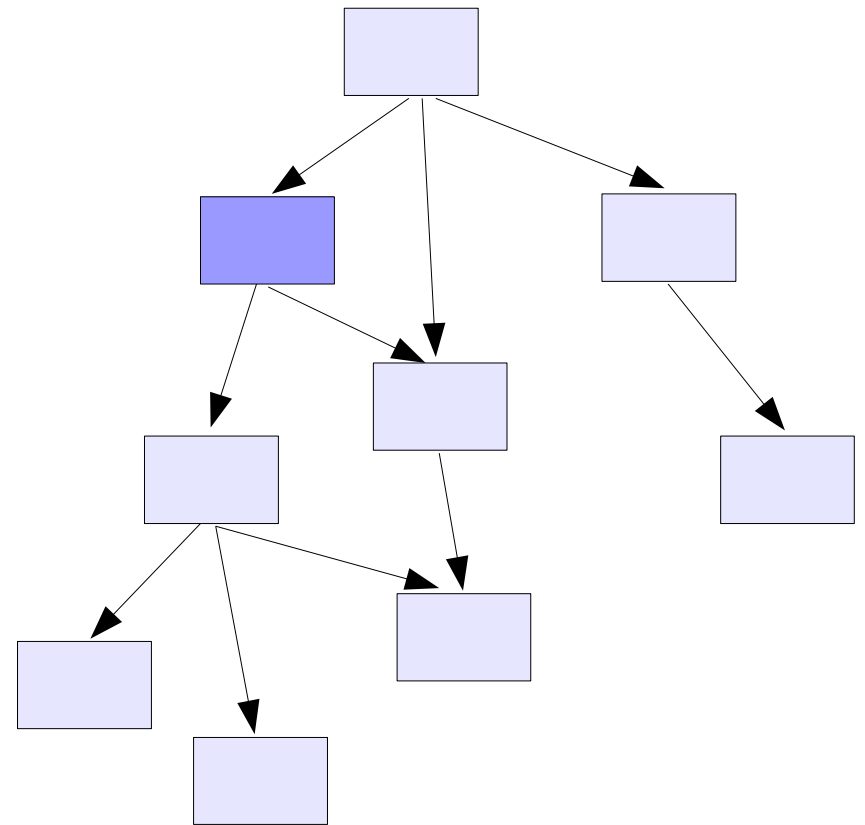
Precision & similarity

➢ we introduce our definitions :

- **precision** of term
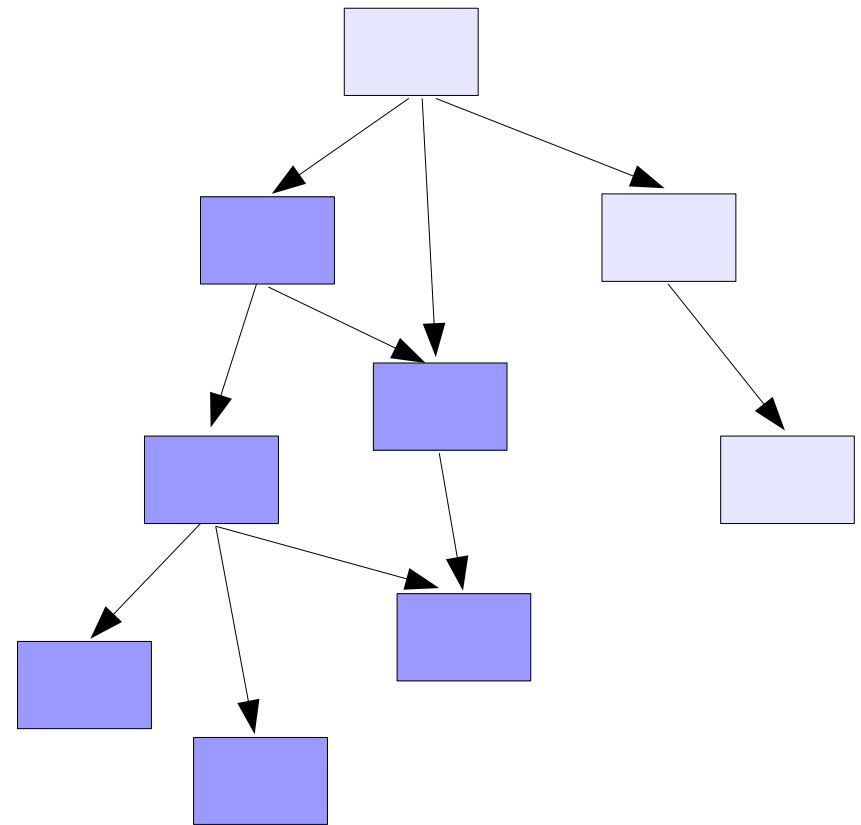
- **similarity** between two terms

Precision & similarity

➢ **precision** of a **term**

Precision & similarity

➢ **precision** of a **term**

- n = size of the **sub-DAG**

- N = size of ontology

Precision & similarity

- ➢ **precision** of a **term**

  - n = size of the **sub-DAG**
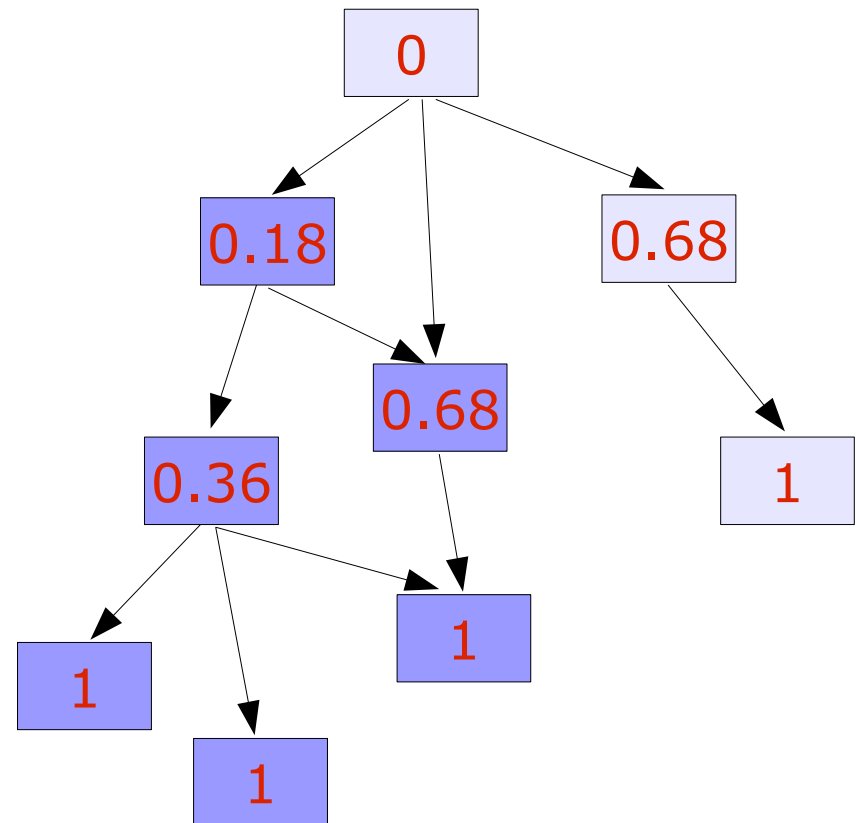
  - N = size of ontology

$$p = \frac{\log(n/N)}{\log(1/N)}$$
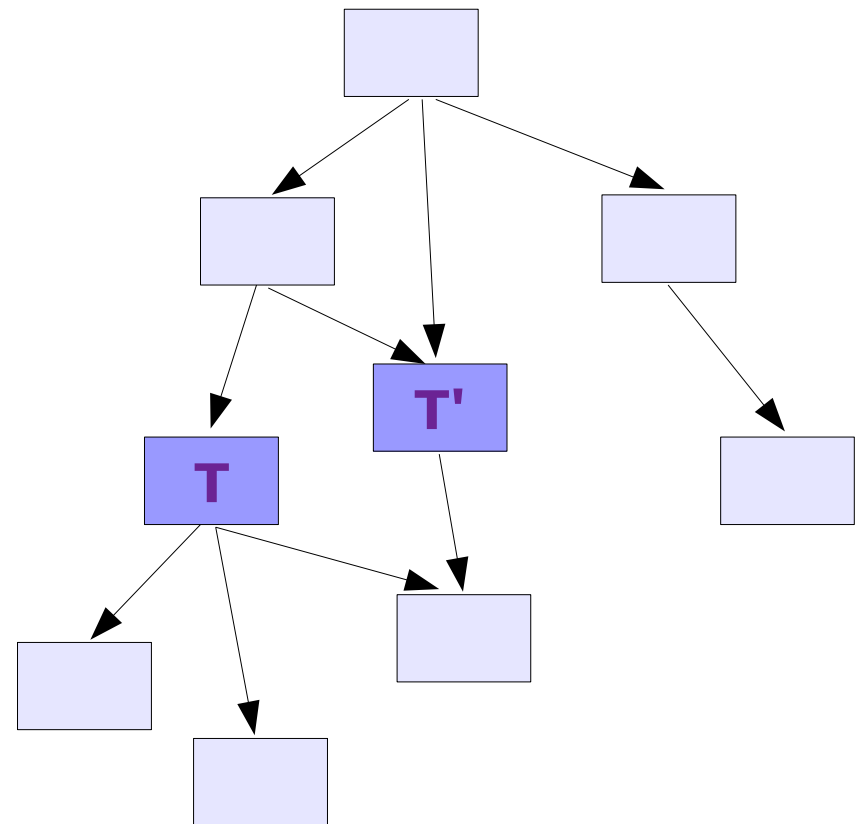
Precision & similarity

➢ **precision** of a **term**

- n = size of the **sub-DAG**

- N = size of ontology

$$p = \frac{\log(n/N)}{\log(1/N)}$$
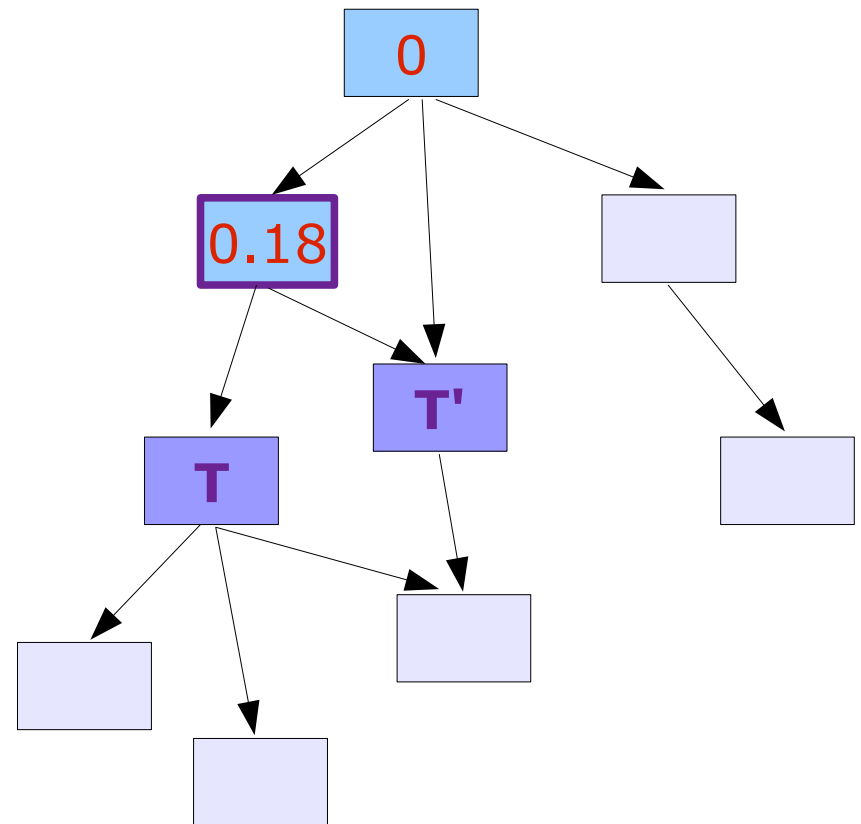
Precision & similarity

- **similarity** between 2 terms:

Precision & similarity

➢ **similarity** between 2 terms:

s(T,T')
=
precision of the most precise common ancestor

Our tool SimCT

Goal:

➢ cluster biological objects based on the **similarity between their annotations**

➢ display the relationships as a **hierarchical tree**

very general:
any type of biological objects from any organism annotated to any OBO ontology !

treating multiple annotations

one object

several annotations

dorsal

heart development
immune response
mesoderm development
regulation of transciption
...

which one shall we choose ?

treating multiple annotations

one object

several annotations

dorsal →
heart development
immune response
mesoderm development
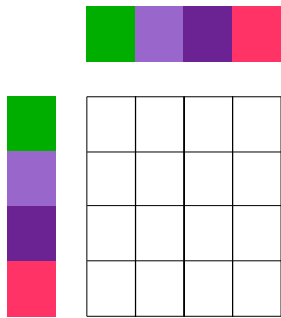regulation of transcription
...

which one shall we choose ?

we consider object | ontoID as fundamental entities (leaves)

flowchart of SimCT

object | ontoID
object | ontoID
object | ontoID
object | ontoID

pairwise
similarities
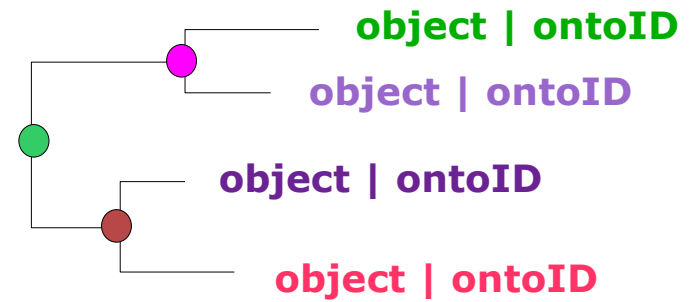
similarity matrix

hierarchical
clustering

`((a,b),c,(d,e))`

newick file

java
tree viewer

clustering tree

object | ontoID
object | ontoID
object | ontoID
object | ontoID

clustering tree

nodes annotated to **most precise common ancestor** of underlying nodes/leaves

ontoID

object | ontoID

object | ontoID

ontoID

object | ontoID

ontoID

object | ontoID
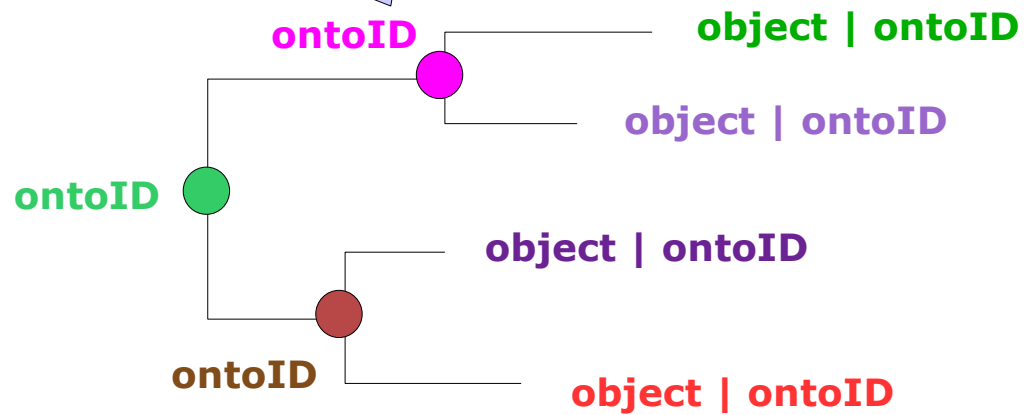
clustering tree

branch length = p(⬤) - p(⬤)

object | ontoID

object | ontoID

object | ontoID

object | ontoID

─ ⟶ **precision** +

available online
http://gin.univ-mrs.fr/SimCT

an example (1)

➢ study of the pleiotropy of a protein

  how widespread are the biological functions in which `dorsal` *is involved ?*

➢ build all `dorsal`|`annotation` combinations and cluster them...

Java Applet Window

Display ?

SRI : 0.0   +

precision : 0.227

>
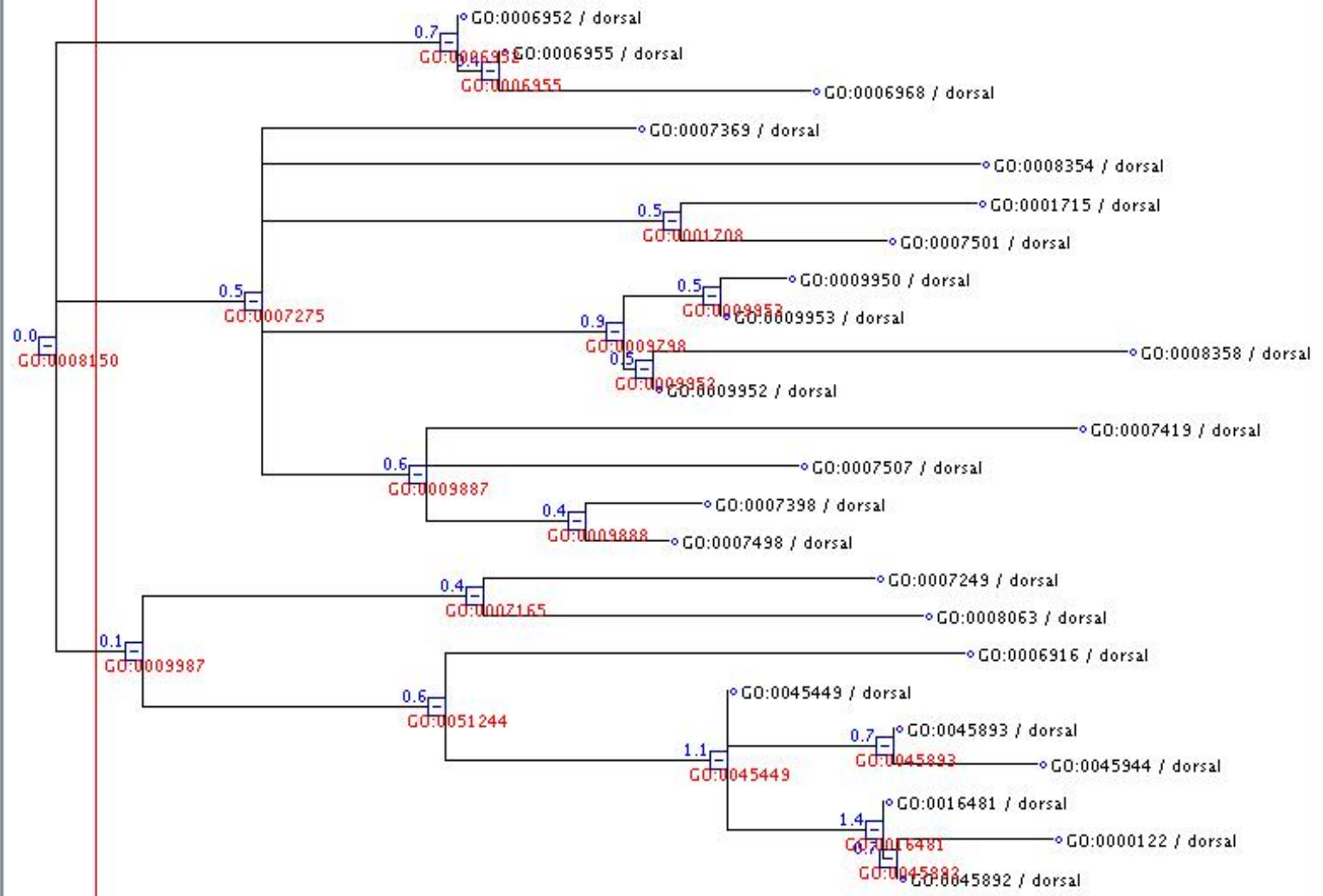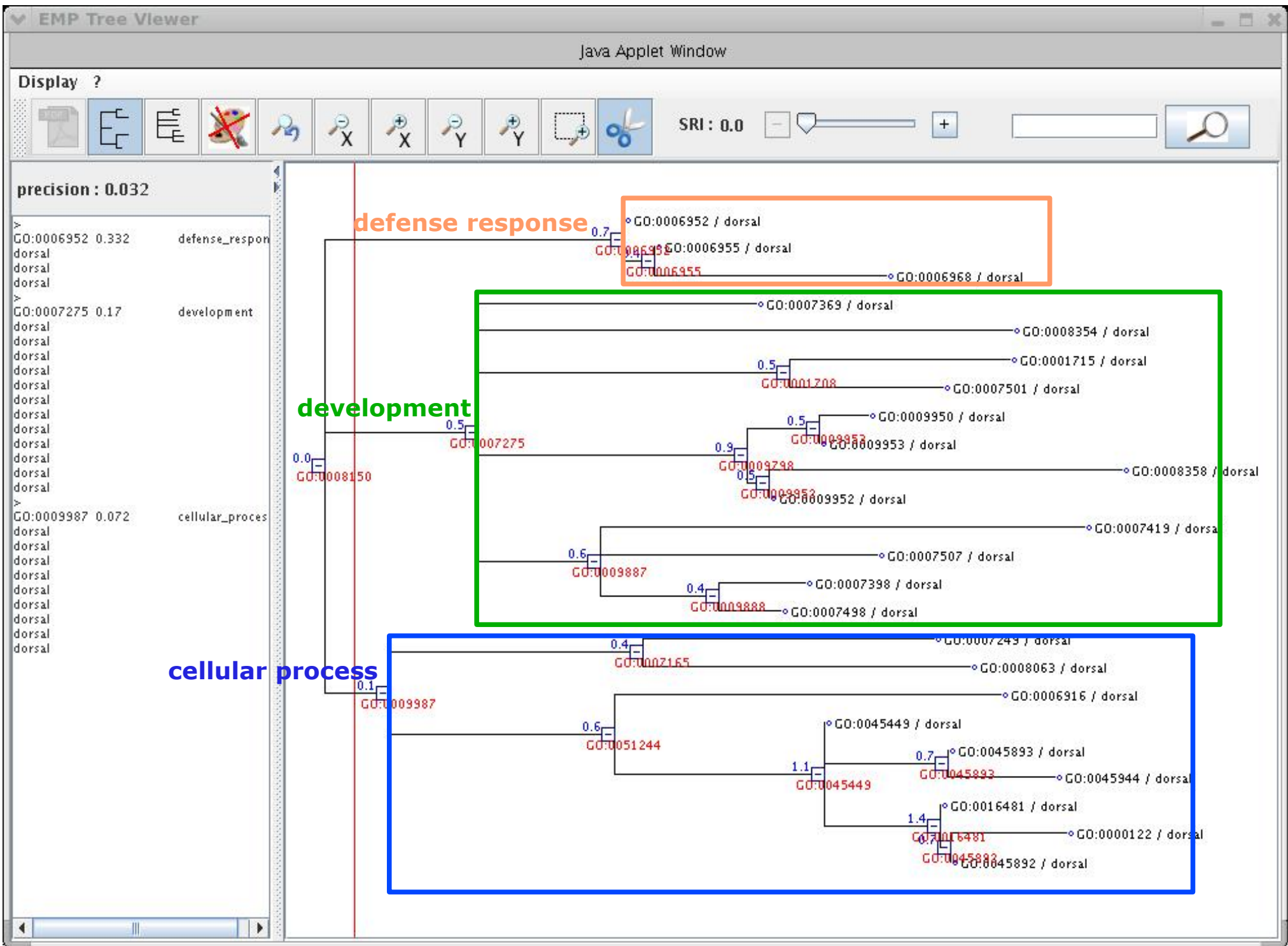GO:0006952 0.332   defense_respon
dorsal
dorsal
dorsal
>
GO:0007369 0.479
dorsal
>
GO:0008354 0.764
dorsal
>
GO:0001708 0.517   cell_fate_specifi
dorsal
dorsal
>
GO:0009798 0.469   axis_specificati
dorsal
dorsal
dorsal
dorsal
>
GO:0009887 0.307   organogenesis
dorsal
dorsal
dorsal
dorsal
>
GO:0007165 0.353   signal_transduc
dorsal
dorsal
>
GO:0051244 0.322   regulation_of_c
dorsal
dorsal
dorsal
dorsal
dorsal
dorsal
dorsal

defense response

GO:0006952 / dorsal
0.7
GO:0006955 GO:0006955 / dorsal
GO:0006955
GO:0006968 / dorsal

GO:0007369 / dorsal

GO:0008354 / dorsal

cell fate specification
0.5
GO:0001715 / dorsal
GO:0001708
GO:0007501 / dorsal

axis specification
0.5
GO:0009950 / dorsal
0.5
GO:0007275   GO:0009953
GO:0009953 / dorsal
0.9
GO:0009798
0.5
GO:0008358 / dorsal
GO:0008150
0.0
GO:0008150
GO:0009952
GO:0009952 / dorsal

organogenesis
GO:0007419 / dorsal
0.6
GO:0007507 / dorsal
GO:0009887
0.4
GO:0007398 / dorsal
GO:0009888
GO:0009888
GO:0007498 / dorsal

signal transduction
GO:0007249 / dorsal
GO:0007165
GO:0008063 / dorsal

0.1
GO:0009987
GO:0006916 / dorsal

regulation of cellular physiological processes
0.6
GO:0051244
GO:0045449 / dorsal
1.1
0.7
GO:0045893 / dorsal
GO:0045449
GO:0045893
GO:0045944 / dorsal
GO:0016481 / dorsal
1.4
GO:0016481
GO:0000122 / dorsal
GO:0045893
GO:0045892 / dorsal

an example (2)

➢ protein complexes

how to assign (a) molecular function(s) to a yeast protein complex ?
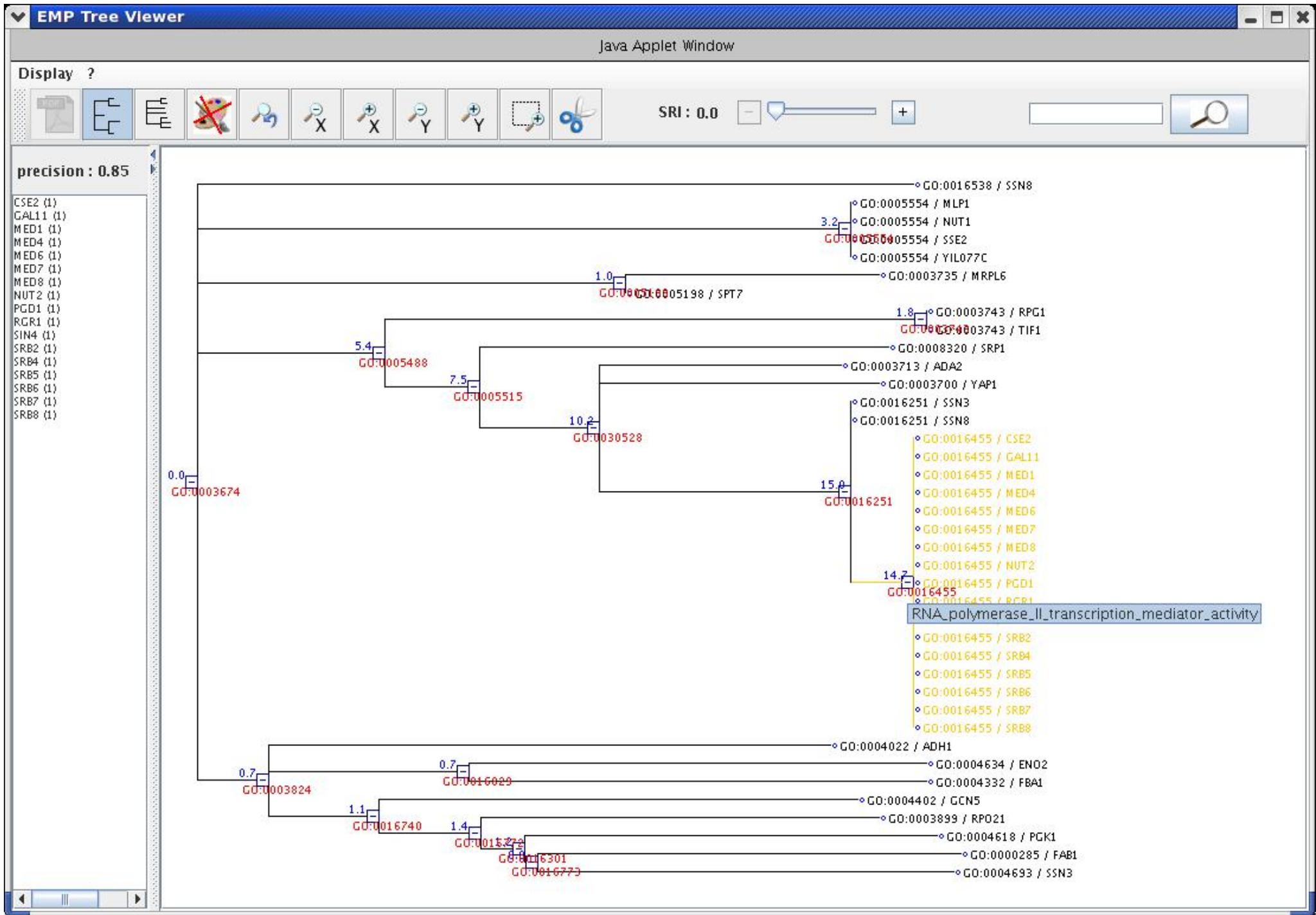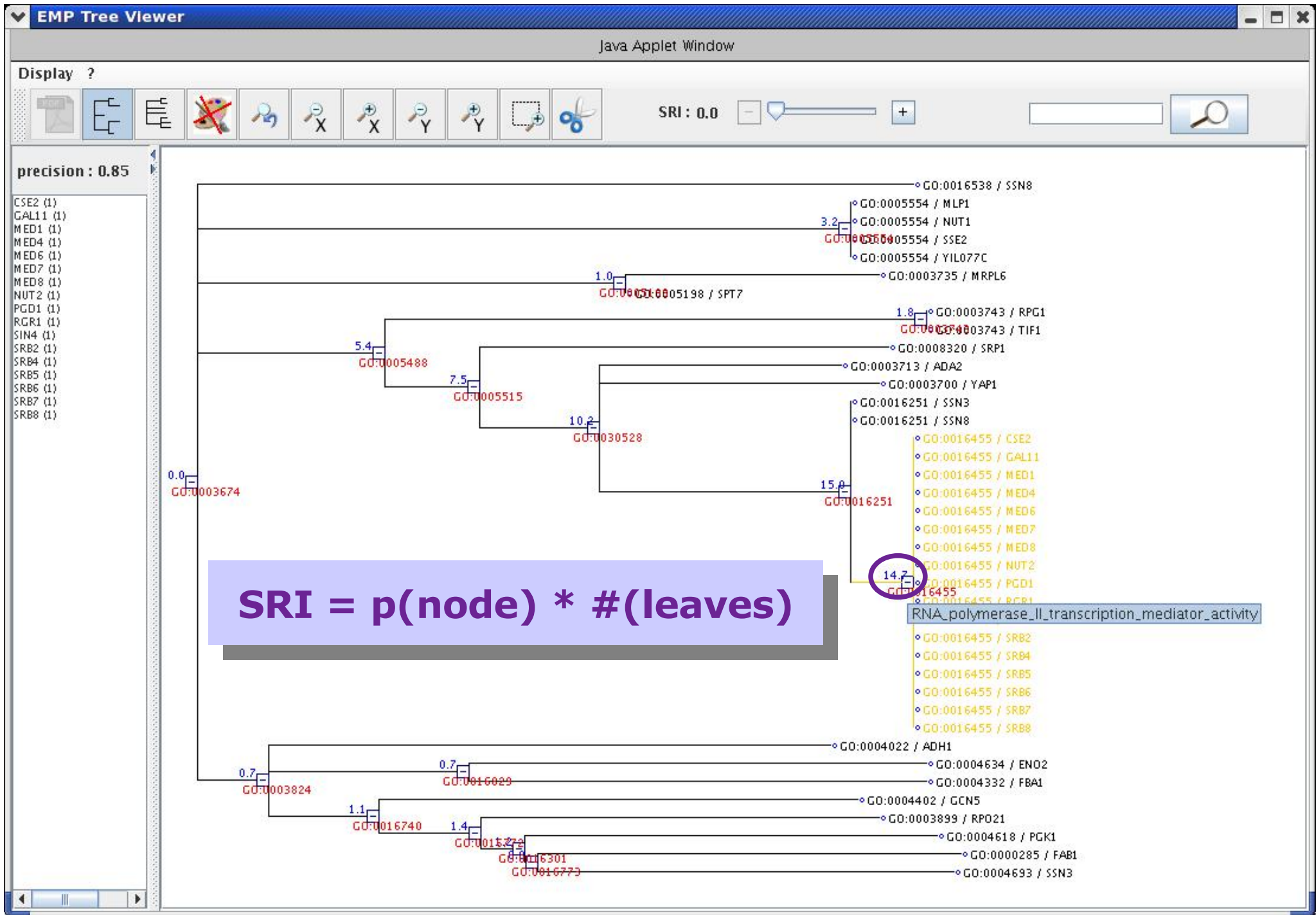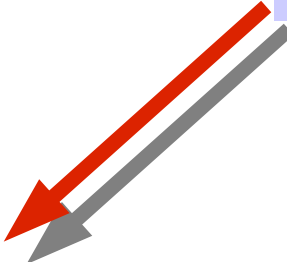
➢ build all protein|annotation and cluster them...

Java Applet Window

Display   ?

SRI : 0.0   −   +

precision : 0.85

CSE2 (1)
GAL11 (1)
MED1 (1)
MED4 (1)
MED6 (1)
MED7 (1)
MED8 (1)
NUT2 (1)
PGD1 (1)
RGR1 (1)
SIN4 (1)
SRB2 (1)
SRB4 (1)
SRB5 (1)
SRB6 (1)
SRB7 (1)
SRB8 (1)

GO:0016538 / SSN8
GO:0005554 / MLP1
3.2 GO:0005554 / NUT1
GO:0005554 GO:0005554 / SSE2
GO:0005554 / YIL077C
1.0 GO:0003735 / MRPL6
GO:0005198 GO:0005198 / SPT7
1.8 GO:0003743 / RPG1
GO:0003743 GO:0003743 / TIF1
5.4 GO:0008320 / SRP1
GO:0005488
7.5 GO:0003713 / ADA2
GO:0005515 GO:0003700 / YAP1
10.2 GO:0016251 / SSN3
GO:0030528 GO:0016251 / SSN8
GO:0016455 / CSE2
GO:0016455 / GAL11
GO:0016455 / MED1
15.0 GO:0016455 / MED4
GO:0016251 GO:0016455 / MED6
GO:0016455 / MED7
GO:0016455 / MED8
GO:0016455 / NUT2
14.2 GO:0016455 / PGD1
GO:0016455 GO:0016455 / RGR1
RNA_polymerase_II_transcription_mediator_activity
GO:0016455 / SRB2
GO:0016455 / SRB4
GO:0016455 / SRB5
GO:0016455 / SRB6
GO:0016455 / SRB7
GO:0016455 / SRB8
GO:0004022 / ADH1
0.7 0.7 GO:0004634 / ENO2
GO:0003824 GO:0016020 GO:0004332 / FBA1
GO:0004402 / GCN5
1.1 GO:0003899 / RPO21
GO:0016740 1.4 GO:0004618 / PGK1
GO:0015372 GO:0000285 / FAB1
GO:0016301 GO:0004693 / SSN3
GO:0016773

SRI = p(node) * #(leaves)

Conclusions & Outlook

2 new notions:
- **precision**
- **similarity**

clustering procedure
- visualize relationships
- annotate clusters
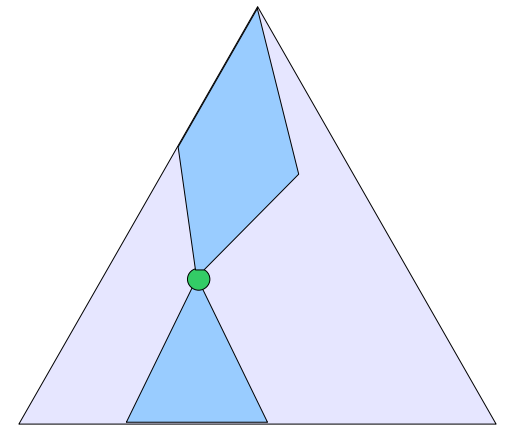- define sub-clusters

Conclusions & Outlook

2 new notions:
- **precision**
- **similarity**

clustering procedure
- visualize relationships
- annotate clusters
- define sub-clusters

slim ontologies

different definition of precision

acknowlegments

➤ *futur users for suggestions and feedback !!*

`http://gin.univ-mrs.fr/SimCT`