Wykorzystanie R i Bioconductor do pozyskiwania, przetwarzania i wizualizacji wiedzy biologicznej zawartej w Gene Ontology (GO)

Adam Zagdański

Instytut Matematyki i Informatyki

Politechnika Wrocławska

www.im.pwr.wroc.pl/~zagdan

Wrocławski Zlot Użytkowników R

6 września 2008

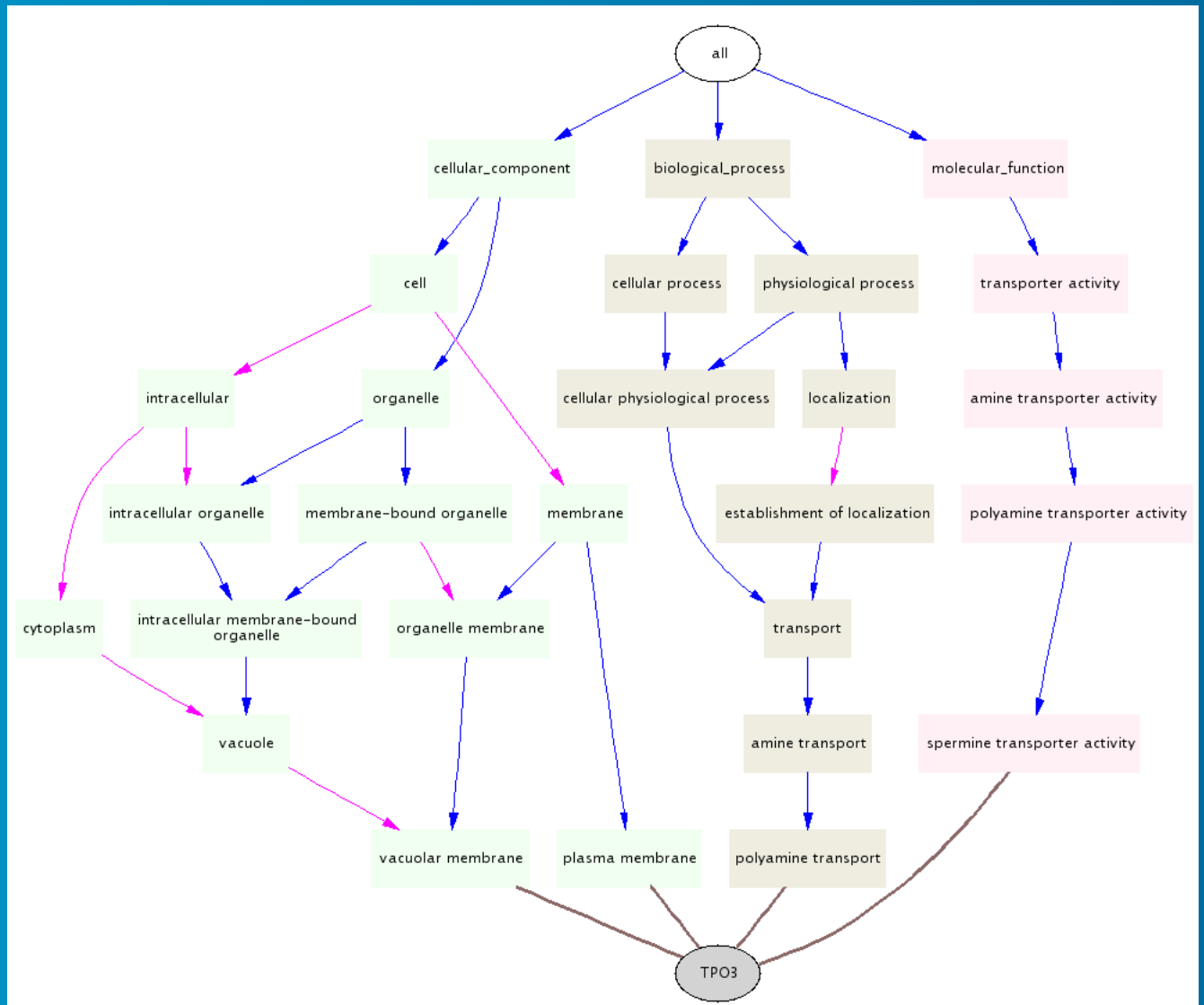
Plan prezentacji

- Gene Ontology (GO) – ogólna charakterystyka i cele projektu
- Kilka słów o projekcie Bioconductor
- Prezentacja możliwości wybranych pakietów
  - Meta-data packages
  - Software packages
- Podsumowanie

Gene Ontology (GO) – ogólna charakterystyka

- **Strona domowa: http://www.geneontology.org/**
- **GO – zbiór kategorii (GO-terms) opisujących funkcje genów w sposób niezależny od organizmu (genomu)**
- **Trzy niezależne ontologie (taksonomie)**
  - MF-molecular functions (7825 kategorii)
  - BP-biological processes (13860 kategorii)
  - CC-cellular components (1993 kategorii)
- **Hierarchiczna struktura zależności pomiędzy kategoriami (graf acykliczny skierowany)**

Gene Ontology (GO) – fragment struktury



Gene Ontology (GO) – możliwe zastosowania

- **Wspomaganie analiz przeprowadzanych na skalę całego genomu**
- **Predykcja (nieznanych) funkcji genów**
- **Ocena jakości rezultatów przeprowadzanych analiz danych eksperymentalnych, np. analizy skupień (*knowledge-based validation*)**
- **Poprawa stabilności (powtarzalności) i biologicznej istotności rezultatów analiz poprzez uwzględnienie dotychczasowej wiedzy biologicznej**

Kilka słów o Bioconductor

- **Strona domowa: http://www.bioconductor.org/**
- **Projekt rozwijany od 2001r.**
- **Aktualna wersja (stabilna): BioC 2.2**
  - **wydana 1 V 2008**
  - **zawiera 260 pakietów**
  - **zaprojektowana do pracy z R 2.7.0**
- **Oprogramowanie typu *open source* rozwijane na bazie R'a,**

  w celu wspomagania analiz i wizualizacji danych genomicznych
- **Oprócz pakietów oferujących narzędzia analityczne dostępne są także pakiety**

  zawierające tzw. meta-dane (m.in. dane związane z eksperymentami

  mikromacierzowymi, informacje przechowywane w repozytoriach biologicznych)

BIOCONDUCTOR
open source software for bioinformatics

GO
the Gene Ontology

Bioconductor – podstawowe rodzaje R-pakietów

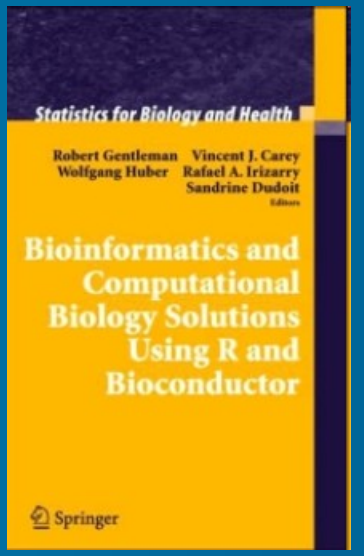
- **Software (R-pakiety wspomagające analizę danych)**
  - GOstats
  - goTools
  - Rgraphviz
- **Meta-data (m.in. adnotacje biologiczne dla GO)**
  - GO
  - YEAST
  - AnnBuilder – możliwość tworzenia własnych pakietów z meta-danymi!
- **Experiment Data**

  (m.in. wyniki eksperymentów mikromacierzowych)

Bioconductor – dodatkowe źródła informacji

- Jednym z celów projektu jest także dostarczenie wysokiej jakości dokumentacji typu *„task-oriented”* (m.in. w formie tzw. *vignettes*)
- Bioconductor Documentation http://www.bioconductor.org/docs
- The Bioconductor Project Mailing List https://stat.ethz.ch/mailman/listinfo/bioconductor
- Monografia

  R. Gentleman, V. Carey, W. Huber, R. Irizarry, S. Dudoit, *„Bioinformatics and Computational Biology Solutions Using R and Bioconductor”*, Springer, 2005



Przegląd możliwości wybranych pakietów

- **Meta-data packages**
  - **Integracja wiedzy biologicznej znajdującej się w różnych repozytoriach (m.in. Gene Ontology (GO))**
  - **Uaktualniane kwartalnie**
  - **Wybrane pakiety**
    - **GO – zawiera informacje nt. znaczenia i przyporządkowań (*annotations*) funkcji biologicznych zawartych w Gene Ontology**
    - **YEAST – przyporządkowania GO-funkcji dla genomu drożdży**
    - **AnnBuilder – możliwość stworzenia wlasnego pakietu z meta-danymi!**

Meta-data packages: Pakiet GO

- Zbudowany z wykorzystaniem pakietu AnnBuilder na bazie (publicznie dostępnych) repozytoriów
  - EntrezGene: http://ftp.ncbi.nlm.nih.gov/gene/DATA/
  - Gene Ontology: http://www.godatabase.org/dev/database/archive/latest/
- Uwaga
  - Wykorzystując AnnBuilder możemy stworzyć własny pakiet z adnotacjami biologicznymi

Meta-data packages: Pakiet GO

- **GOTERM -- Annotation of GO Identifiers to GO Terms**

  Podstawowa charakterystyka GO-funkcji (GO-terms), min.: GOID, definicja funkcji, ontologia

- **GOXXANCESTOR Annotation of GO Identifiers to their XX Ancestors**
- **GOXXCHILDREN Annotation of GO Identifiers to their XX Children**
- **GOXXOFFSPRING Annotation of GO Identifiers to their XX Offspring**
- **GOXXPARENTS Annotation of GO Identifiers to their XX Parents + information about the nature of the relationship ('is a' or 'part of'))**

Gdzie XX = [BP | CC | MF] (trzy różne ontologie występujące w GO)

Meta-data packages: Pakiet GO

Przykład 1: Informacja nt. konkretnej GO-funkcji

```
> mget("GO:0005657",env=GOTERM)

$"GO:0005657"

GOID = GO:0005657

Term = replication fork

Definition = The Y-shaped region of a replicating DNA molecule, resulting
    from the separation of the DNA strands and in which the synthesis of
    new strands takes place. Also includes associated protein complexes.

Ontology = CC
```

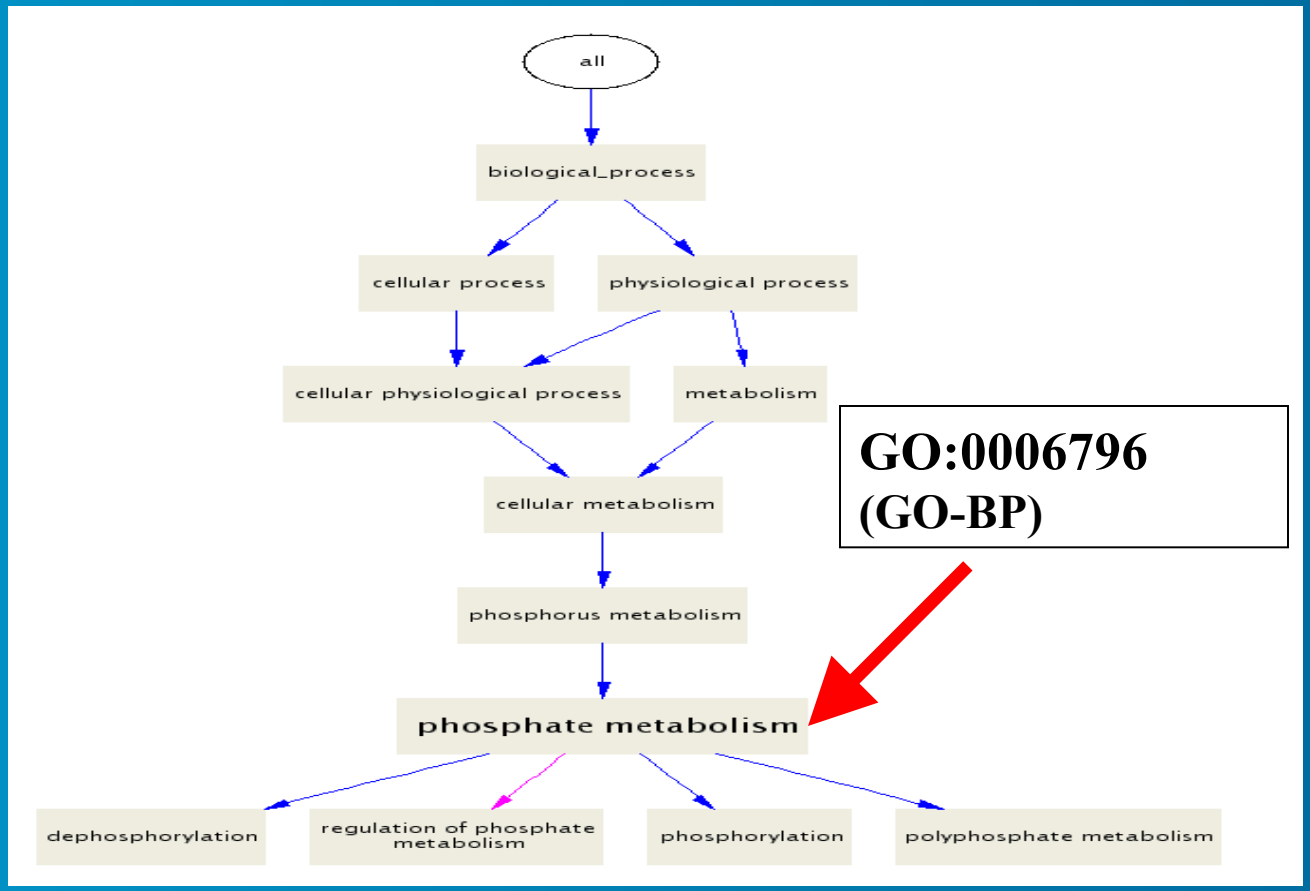
Meta-data packages: Pakiet GO

Przykład 2: Konwersja środowisko → lista

```
> as.list(GOTERM) -> GO_list

> GO_list[1:3]

$`GO:0019980`

An object of class "GOTerms"

Slot "GOID":

[1] "GO:0019980„

.................
```

Meta-data packages: Pakiet GO

Przykład 3: Lista przodków dla danej GO-funkcji



Źródło: http://mor.nlm.nih.gov/perl/gennav.pl

Meta-data packages: Pakiet GO

Przykład 3: Lista przodków dla danej GO-funkcji

```
> t1 <- "GO:0006796" #BP: phosphate metabolism

> ancestors_list <- mget(t1,GOBPANCESTOR) #list of all ancestors

> ancestors_list <- ancestors_list[[1]]          #convert to vector of
    characters

> ancestors_list

[1] "all"  "GO:0008150"  "GO:0008152" "GO:0007582" "GO:0009987"

[6] "GO:0050875"  "GO:0006793"  "GO:0044237"

> ancestors_list_GO <- mget(ancestors_list,GOTERM)
```

Meta-data packages: Pakiet GO

Przykład 3: Lista przodków dla danej GO-funkcji

```
> ancestors_list_GO <- mget(ancestors_list,GOTERM)

> sapply(ancestors_list_GO, function(x) {attr(x,"Term")})

        GO:0008150                 GO:0008152                 GO:0007582

"biological_process"          "metabolism"       "physiological process"

        GO:0009987                 GO:0050875                 GO:0006793

"cellular process„    „cellular physiological process„  „ phosphorus metabolism"
```

BIOCONDUCTOR
open source software for bioinformatics

GO
the Gene Ontology

Meta-data packages: Pakiet YEAST

- **YEASTGENENAME** Map Between Manufacturer IDs and Genes
- **YEASTGO** Map between Manufacturer IDs and Gene Ontology (GO)
- **YEASTGO2ALLPROBES** Map Between Gene Ontology (GO) Identifiers and all Manufacturer Identifiers
- **YEASTGO2PROBE** Map Between GO and Manufacturer Identifiers
- **YEASTPMID** Map between Manufacturer Identifiers and PubMed Identifiers
- **YEASTPMID2PROBE** Map between PubMed Identifiers and Manufacturer Identifiers

Meta-data packages: Pakiet YEAST

Przykład 1: Znajdź geny drożdży związane z konkretną GO-funkcją

```
> mget("GO:0005657",env=YEASTGO2PROBE)

$"GO:0005657"

    IDA       IDA       IDA              IDA       IDA       IGI       IPI

     "YCL061C" "YDR419W" "YJL090C" "YLR103C" "YLR274W" "YKL108W"

    "YKL108W"

    NAS       TAS       TAS       TAS       TAS       TAS

"YDL164C" "YBR088C" "YBR278W" "YJL090C" "YNL262W" "YPR175W"
```

Meta-data packages: Pakiet *AnnBuilder*

- **Możliwość stworzenia własnego pakietu z adnotacjami biologicznymi dla określonego zbioru genów**
- **Integracja informacji zawartej w różnych repozytoriach biologicznych**
  - GP ftp://hgdownload.cse.ucsc.edu/goldenPath/currentGenomes/Homo_sapiens/database/
  - UG ftp://ftp.ncbi.nih.gov/repository/UniGene/Homo_sapiens/Hs.data.gz
  - GO ftp://ftp.geneontology.org/pub/go/godatabase/archive/latest
  - KEGG ftp://ftp.genome.ad.jp/pub/kegg/pathway/organisms
  - YG ftp://genome-ftp.stanford.edu/pub/yeast/data_download/
  - HG ftp://ftp.ncbi.nih.gov/pub/HomoloGene/old/hmlg.ftp
  - EG ftp://ftp.ncbi.nlm.nih.gov/gene/DATA
  - IPI ftp://ftp.ebi.ac.uk/pub/databases/IPI/current/
  - YEAST ftp://ftp.yeastgenome.org/pub/yeast/sequence_similarity/domains/
  - KEGGGENOME ftp://ftp.genome.ad.jp/pub/kegg/genes/genome
  - PFAM ftp://ftp.sanger.ac.uk/pub/databases/Pfam/current_release/Pfam-A.full.gz

Meta-data packages: Pakiet *AnnBuilder*

Główna funkcja

- **ABPkgBuilder** (baseName, srcUrls, baseMapType, otherSrc, pkgName, pkgPath, organism, version, makeXML, author, fromWeb)
  - **baseName** – file containing target genes to be annotated and their corresponding mappings to GenBank accession numbers, UniGene identifiers, Image clone identifiers, or LocusLink identifiers.
    (NA are allowed)
  - **srcUrls** – vector with urls to source data files
    (Valid sources are LocusLink, UniGene, Golden Path, Gene Ontology and KEGG). Calling getSrcUrl("all", "human") will return the urls needed for building a package for human.Other valid organism names include mouse and rat

Meta-data packages: Pakiet *AnnBuilder*

Specjalna funkcja dla genomu drożdży

- **yeastPkgBuilder**(pkgName="MyYeast", pkgPath=MyDir,version = "1.1.0",author = list(author = "who",maintainer = "who@email.com"), fromWeb = TRUE)

A data package containing annotation data for MyYeast

Documentation for package `MyYeast' version 1.1.0

Help Pages

MyYeast	Bioconductor annotation data package
MyYeastALIAS	Map Open Reading Frame (ORF) Identifiers to Alias Gene Names
MyYeastCHR	Map Manufacturer IDs to Chromosomes
MyYeastCHRLENGTHS	A named vector for the length of each of the chromosomes
MyYeastCHRLOC	Map Manufacturer IDs to Chromosomal Location
MyYeastDESCRIPTION	An annotation data file that maps Open Reading Frame (ORF) identifiers to textural descriptions of the corresponding genes
MyYeastENZYME	Map Between Manufacturer IDs and Enzyme Commission (EC) Numbers
MyYeastENZYME2PROBE	Map Between Enzyme Commission Numbers and Manufacturer Identifiers
MyYeastGENENAME	Map Between Manufacturer IDs and Genes
MyYeastGO	Map between Manufacturer IDs and Gene Ontology (GO)
MyYeastGO2ALLPROBES	Map Between Gene Ontology (GO) Identifiers and all Manufacturer Identifiers
MyYeastGO2PROBE	Map Between Gene Ontology (GO) and Manufacturer Identifiers
MyYeastMAPCOUNTS	Quality control information for MyYeast
MyYeastORGANISM	The Organism for MyYeast
MyYeastPATH	Mappings between probe identifiers and KEGG pathway identifiers
MyYeastPATH2PROBE	Map between Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway identifiers and Manufacturer Identifiers
MyYeastPMID	Map between Manufacturer Identifiers and PubMed Identifiers
MyYeastPMID2PROBE	Map between PubMed Identifiers and Manufacturer Identifiers
MyYeastQC	Quality control information for MyYeast

Meta-data packages: Pakiet *AnnBuilder*

Ciekawostka

- W praktyce, informacje dotyczące genów przechowywane w różnych źródłach (repozytoriach biologicznych) mogą nie być zgodne (np.różne nazwy dla tego samego genu)
- AnnBuilder rozwiązuje tego rodzaju konflikty „metodą głosowania”, tak aby otrzymać jednoznaczne odwzorowania pomiędzy nazwami

Software packages: pakiet GOstats

- **Narzędzia umożliwiające dostęp do danych GO**
  - getEvidence() Get the Evidence codes for a set of GO terms,
  - getGOChildren() Functions to Access GO data,
  - getGOOntology() Functions to Access GO data,
  - getGOParents() Functions to Access GO data,
  - getGOTerm() Functions to Access GO data,
  - getOntology() Get GO terms for a specified ontology,
  - GOGraph() Construct the GO graph given a set of leaves
  - hasGOannote() Check for GO annotation

BIOCONDUCTOR
open source software for bioinformatics

GO
the Gene Ontology

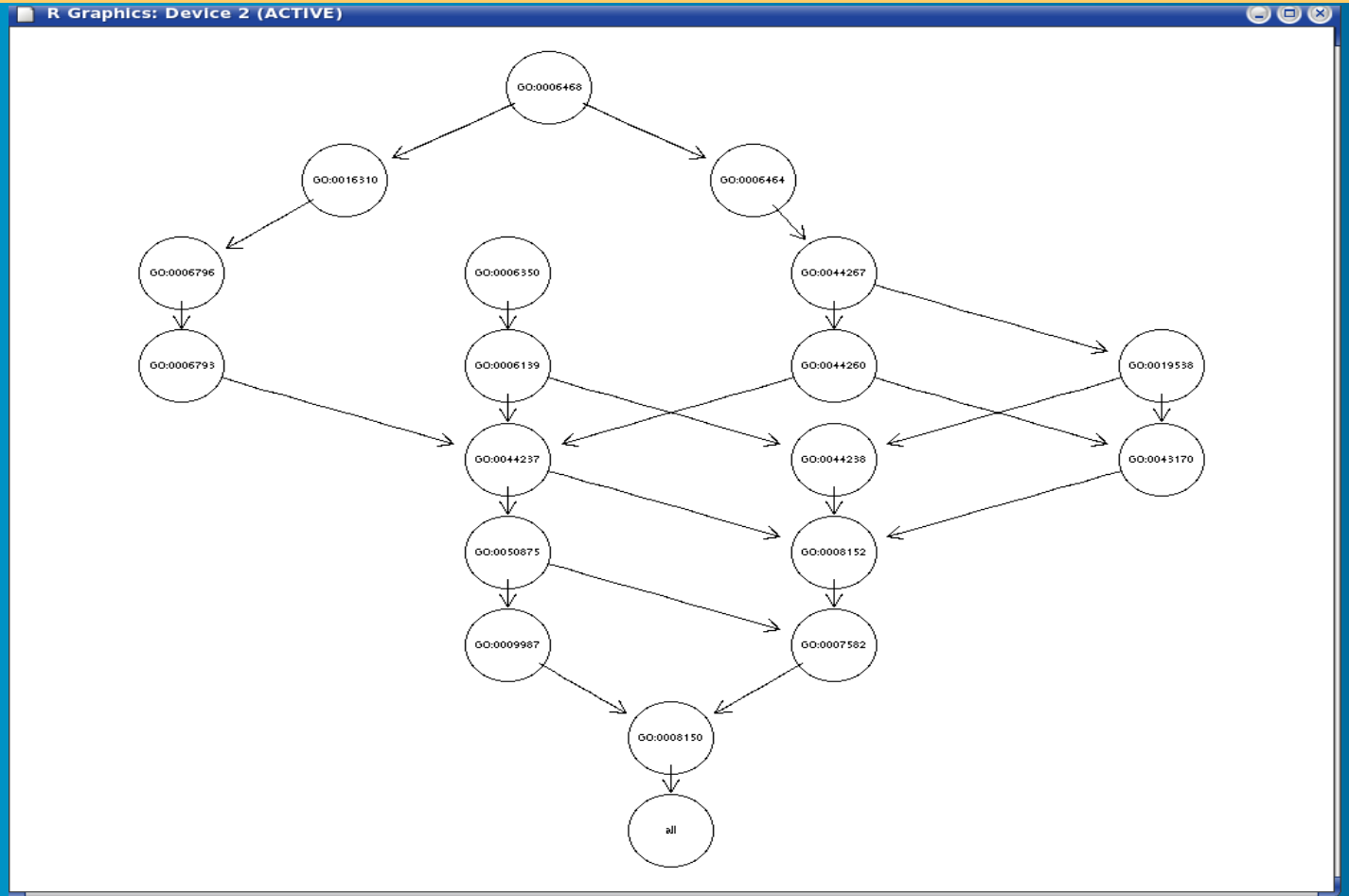
Software packages: pakiet GOstats

- **Narzędzia wspomagające analizy statystyczne**
  - **compCorrGraph()** A function to compute a correlation based graph from Gene Expression Data,
  - **shortestPath()** Shortest Path Analysis
  - **compGdist()** A function to compute the distance between pairs of nodes in a graph,
  - **GOHyperG()** Hypergeometric Tests for GO,
  - **simLL(), simLP(), simUI()** Functions to compute similarities between GO graphs and also between LocusLink IDs based on their induced GO graphs.

Software packages: pakiet Gostats

Przykład 1: Graf indukowany przez zbiór GO-funkcji

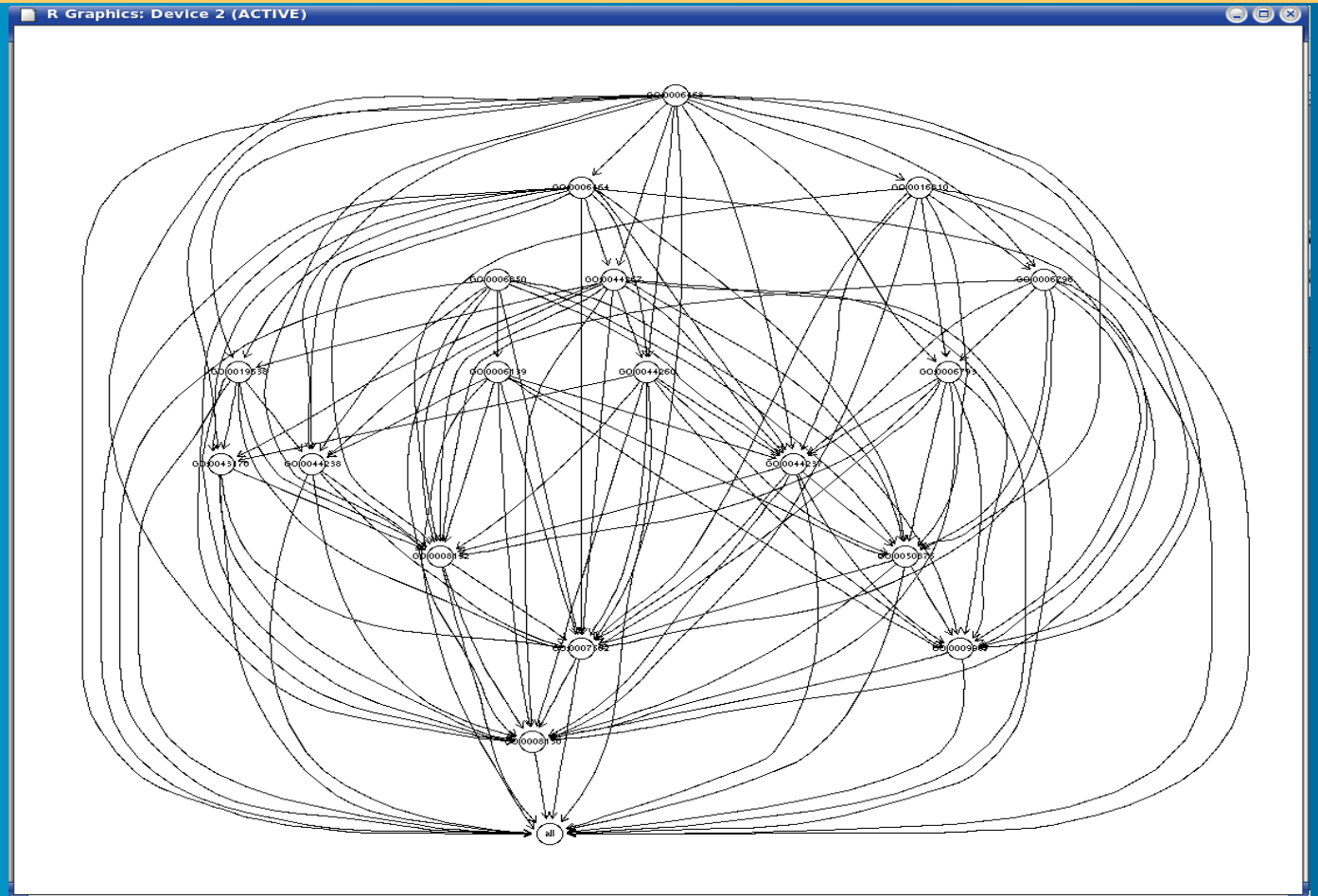
```
> GOGraph(go.terms.bp,dataenv=GOBPPARENTS)->g1
> plot(g1)
```



Software packages: pakiet Gostats

Przykład 2: Graf indukowany przez zbiór GO-funkcji

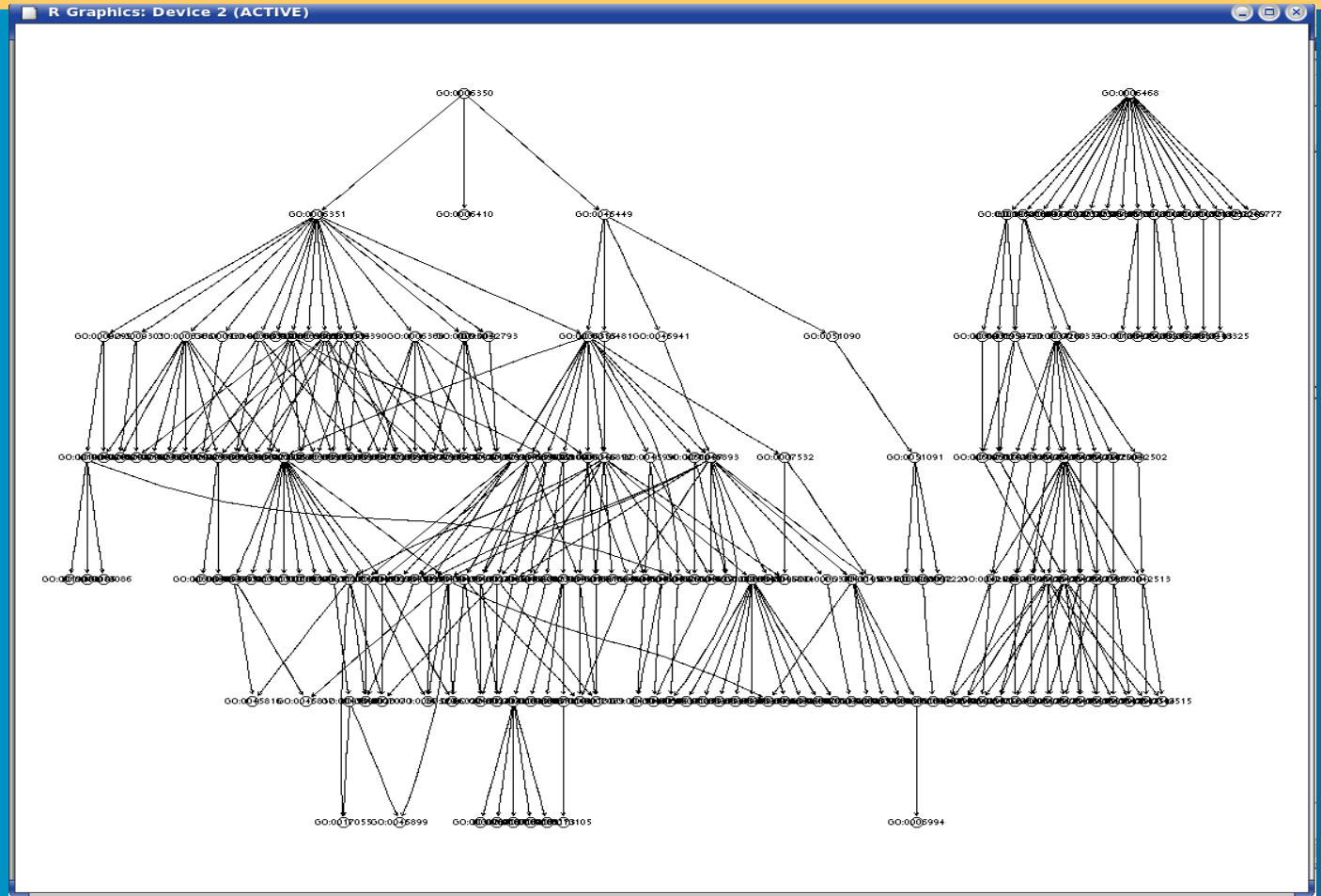
```
> GOGraph(go.terms.bp,dataenv=GOBPANCESTOR)->g2
> plot(g2)
```



Software packages: pakiet Gostats

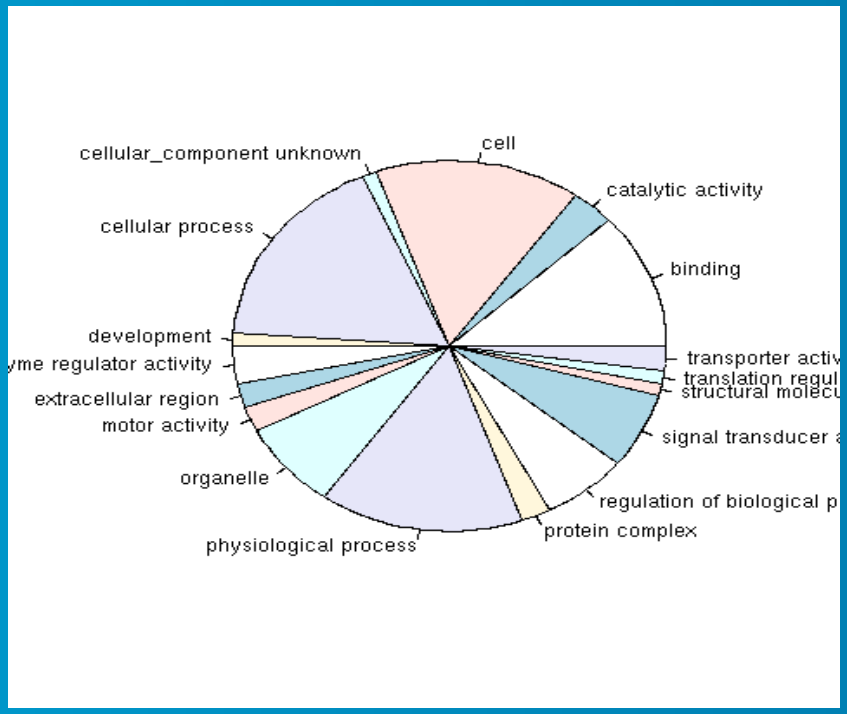
Przykład 3: Graf indukowany przez zbiór GO-funkcji

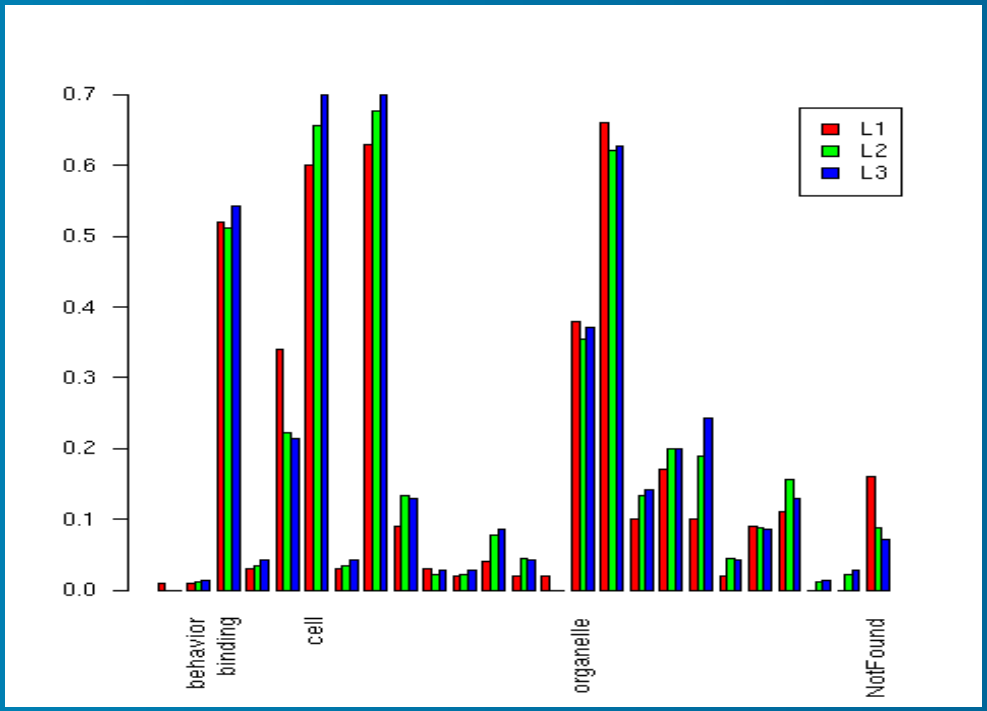
```
> GOGraph(go.terms.bp,dataenv=GOBPCHILDREN)->g3
> plot(g3)
```



Software packages: pakiet goTools

- **Graficzne porównanie przyporządkowanych GO-funkcji dla kilku zbiorów genów**





BIOCONDUCTOR
open source software for bioinformatics

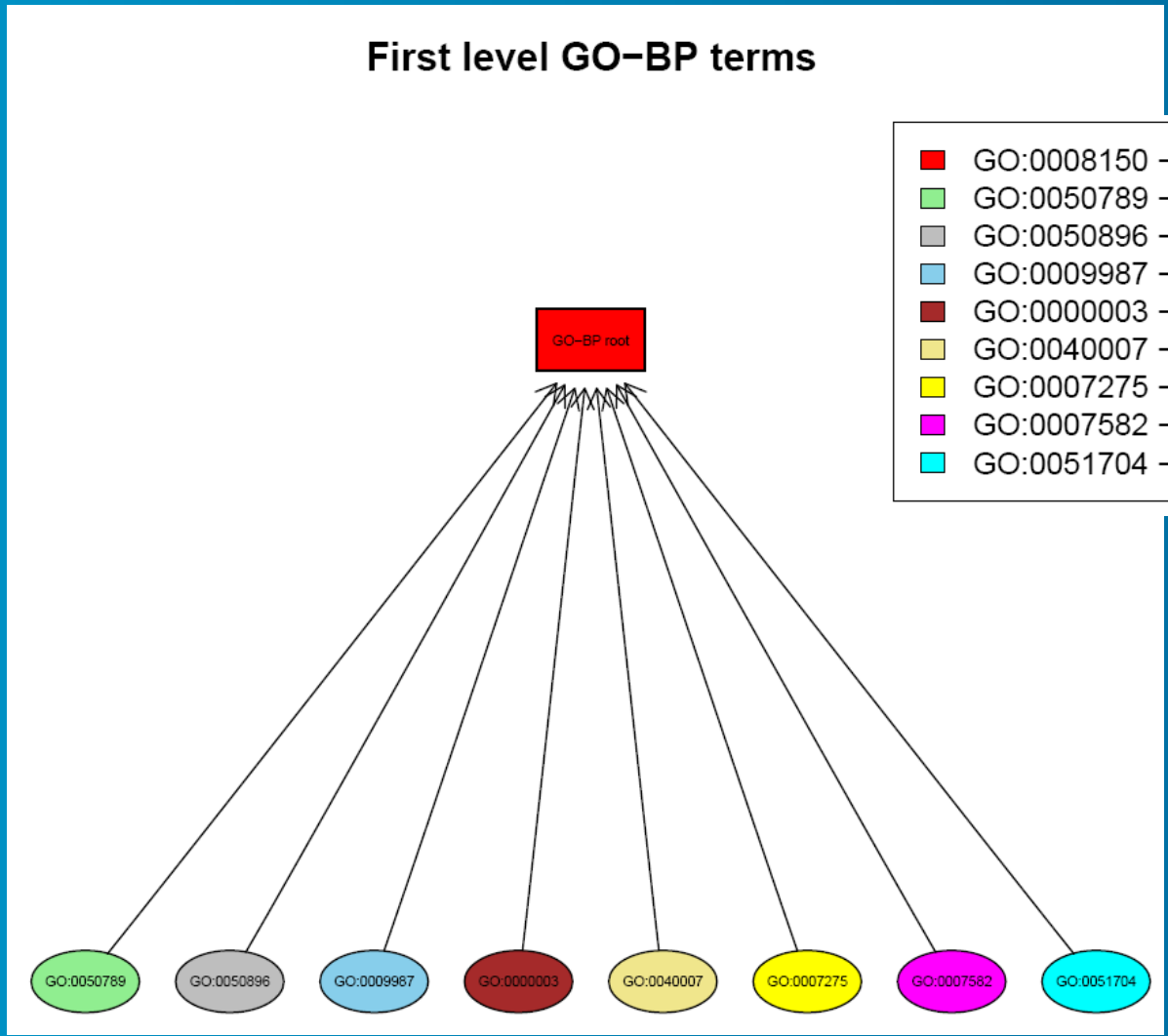
GO
the Gene Ontology

Software packages: pakiet Rgraphviz

- Interfejs R'a do programu Graphviz (Graph Visualization) http://www.graphviz.org/
- Możliwość wizualizacji struktur grafowych, w tym m.in. wizualizacja hierarchicznej struktury zależności pomiędzy GO-funkcjami
- Możliwość wizualizacji grafów stworzonych za pomocą pakietu Gostats (funkcja GOGraph(...))

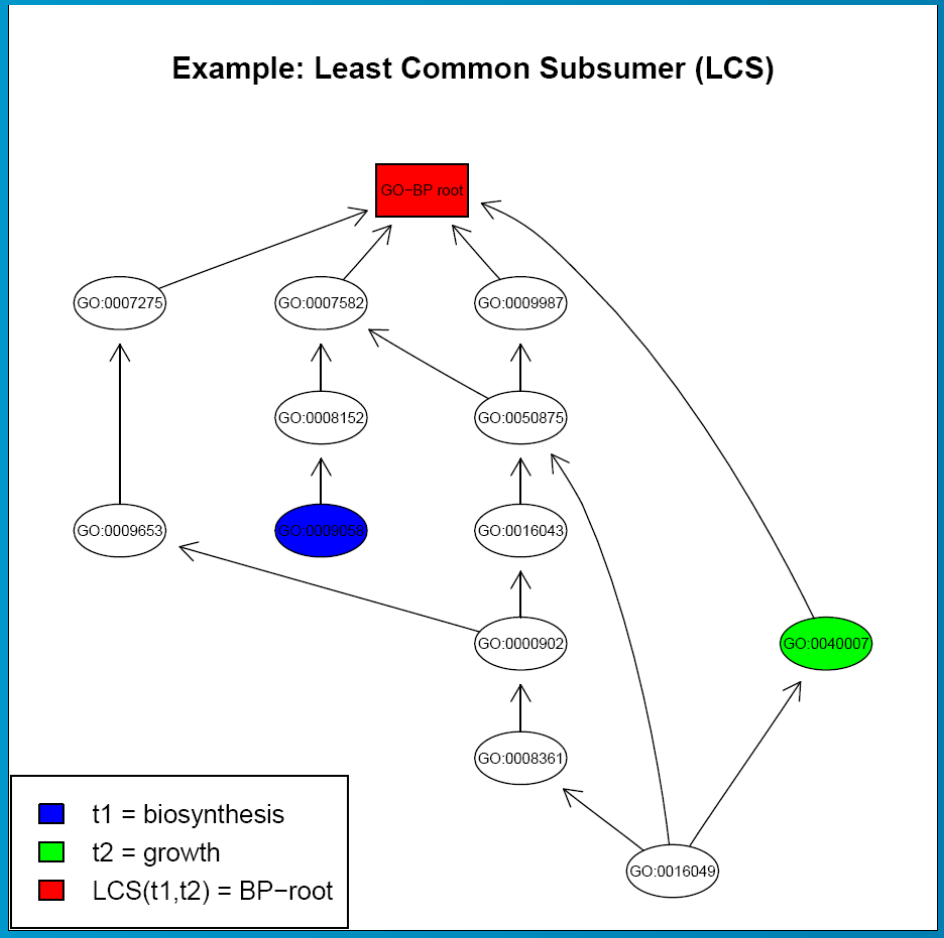
Software packages: pakiet Rgraphviz

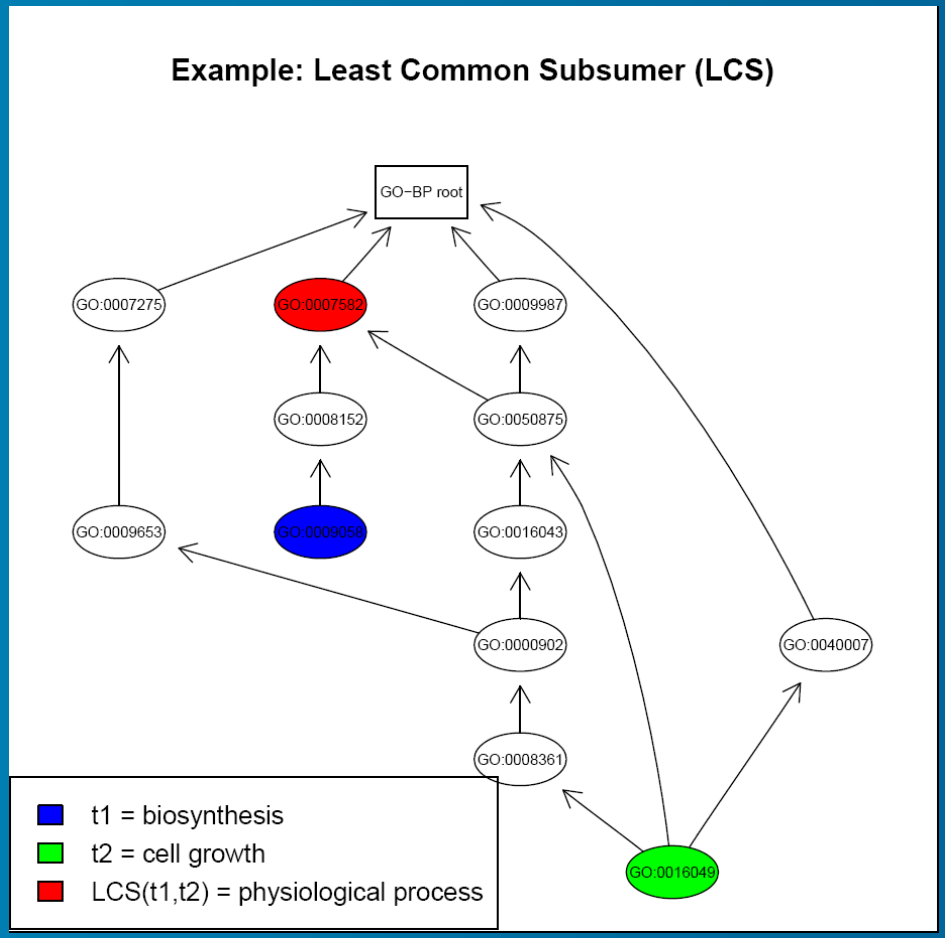
Przykład 1: GO-BP funkcje dla 1-ego poziomu



Software packages: pakiet Rgraphviz

Przykład 2: Ilustracja idei LCS





BIOCONDUCTOR
open source software for bioinformatics

GO
the Gene Ontology

Zamiast podsumowania...

- **Bioinformatyka + R ≠ Bioconductor**
- **Wiele przydatnych pakietów można znaleźć w CRAN**
  - *clusterRepro* – Reproducibility of gene expression clusters
  - *GeneTS* – Microarray Time Series and Network Analysis
  - *GOSim* – Computation of functional similarities between GO terms and gene products; GO enrichement analysis
  - *impute* – Imputation for microarray data
  - *samr* – Significance Analysis of Microarrays (SAM)
  - *sma* – Statistical Microarray Analysis
  - *pamr* – prediction analysis for microarrays (PAM)
- **Autorzy publikacji z zakresu bioinformatyki często udostępniają kody źródłowe programów w R**

Literatura

- R. Gentleman, V. Carey, W. Huber, R. Irizarry, S. Dudoit, *„Bioinformatics and Computational Biology Solutions Using R and Bioconductor"*, Springer, 2005
- A.Paquet and Y. Hwa Yang, *„Getting started with goTools package"*.
- R.Gentleman
  - *„Basic GO Usage"*,
  - *„Visualizing and Distances Using GO"*,
  - *„Using GO for Statistical Analyses"*.
- *J.Zhang, „How to use AnnBuilder"*