SGN-6156, Lecture 11
Modeling biological networks: Gene ontology

Harri Lähdesmäki, harri.lahdesmaki@tut.fi

(Slides by Juha Kesseli 28.03.2007, updated by Harri Lähdesmäki)

Department of Signal Processing,

Tampere University of Technology

06.05.2008

Motivation

- After data analysis we commonly end up with a list of genes or gene products that are suggested as being interesting in the data in question.
  - E.g. a list of genes the expression of which should be measured to discriminate between two types of cancer.
- Experiments can be performed to validate the results but comparison with previous biological knowledge is not trivial.
  - Is there something the genes on the list have in common?
  - Do the known functions of the genes on the list match the processes known to affect the question under study?
- In particular when experimenting with different methods it would be a great benefit to have a preliminary answer to these questions automatically before consulting an expert.

Gene ontology

- The Gene Ontology (GO) project is a way of collecting biological vocabulary and information so that it can be used (e.g.) in an automated assessment of data analysis results.
- A significant improvement over using separate databases as such by enabling uniform access with the same terminology.
- GO is part of the Open Biomedical Ontologies (OBO) effort, currently collecting together over 60 different ontologies (e.g. Biological imaging methods, Drosophila development, . . .)
- The gene ontology project has two parts:
  - The ontology itself, containing a vocabulary of terms and their relationships.
  - Annotations, describing associations between the terms in the ontology and genes or gene products.

Some terms from the GO

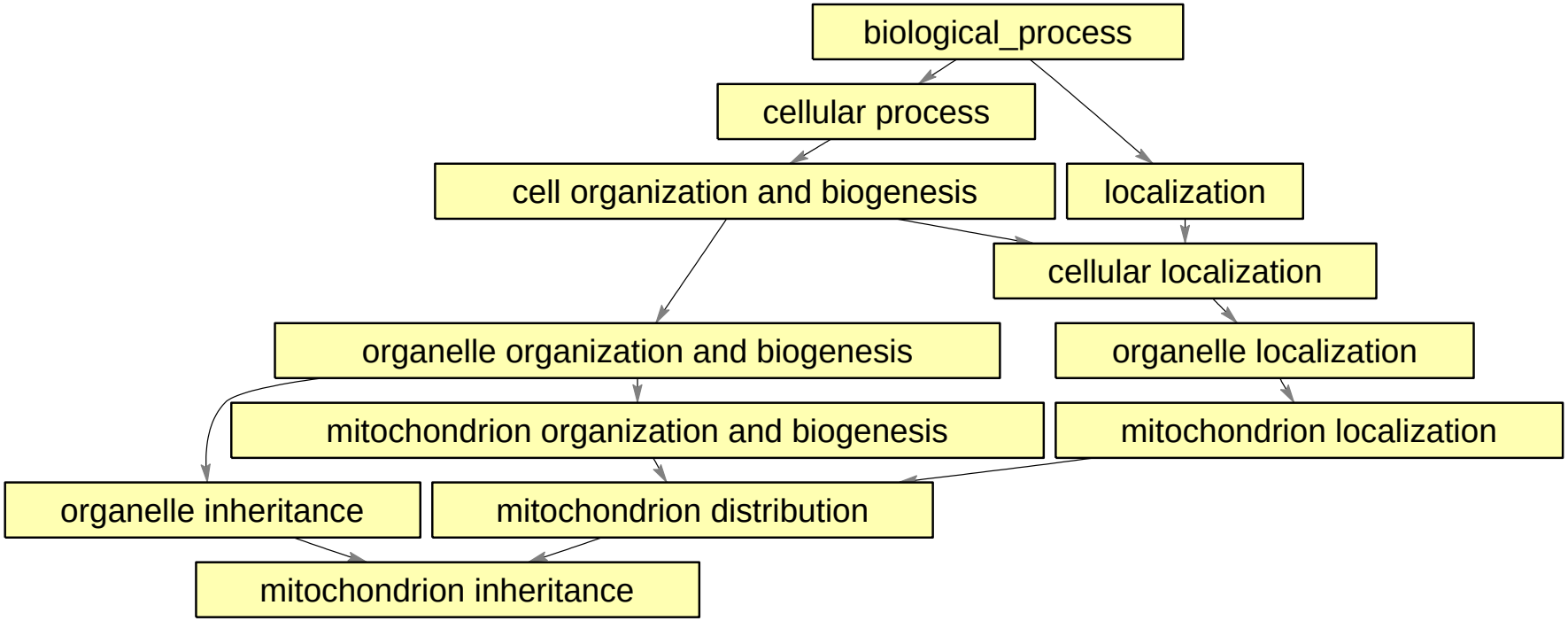

Figure 1: A part of the gene ontology containing all the ancestors of the biological process term "mitochondrion inheritance".

The three ontologies

- The ontology consists, in fact, of three distinct ontologies covering the following:
  - The molecular function of gene products
  - The role of gene products in multi-step biological processes
  - The localization of the gene products in cellular components
- Each of these contains a directed acyclic graph, each node of which is a term in the ontology.
  - Not a tree, a lower-level (more specific) node can have several parents (direct ancestors).
- The ontology is species-independent for the most part, exceptions are denoted with a sensu-tag: sporulation (sensu Bacteria), sporulation (sensu Fungi).

An example from the ontology

An example entry from GO in the OBO v1.0 file format, containing a link to two kinds of parents in GO ("is a" and "part of") and an external reference to Reactome, a database of pathways and reactions in human biology:

```
[Term]
id: GO:0000082
name: G1/S transition of mitotic cell cycle
namespace: biological_process
def: ‘‘Progression from G1 phase to S phase of the mitotic
cell cycle.’’ [GOC:mah]
xref_analog: Reactome:69206
is_a: GO:0022402 ! cell cycle process
relationship: part_of GO:0051329 ! interphase of mitotic cell
cycle*
```

Status of the GO

- The ontology is being updated rather frequently with new terms being created and old ones rendered obsolete.
  - As a result, there are terms tagged "is_obsolete: true". There should be no new annotations attached to these terms.
- The ontology can be downloaded from http://www.geneontology.org or browsed and searched online.
- There are several formats available, including ones based on XML.
- As of 05.05.2008 there are (compare to numbers from 12.02.2007):
  - 14739 (13154) biological process terms
  - 2082 (1864) cellular component terms
  - 8257 (7527) molecular function terms
  - 1160 (987) obsolete terms

Annotations

- The annotations connect genes or gene products to the ontology terms.
- A collection of annotations can be downloaded from the ontology site or e.g. accessed online from the ontology browser.
  - Note: There are a number of different sets of annotations available in addition to the ones from the ontology website.
- Each annotated item is typically associated with several terms.
  - Ideally, should be annotated to at least one process, component and function.
  - Ideally, the annotations should also be consistent (consistency is typically forced)
- Annotations are based on varied sources of information the reliability of which needs to be considered.

– The annotations contain fields describing the evidence.

An example annotation

An example from SGD (*Saccharomyces cerevisiae*) annotation file containing 6472 items associated in all three ontologies:

Database: SGD

Unique identifier in SGD: S000007287

Name of the object: 15S_RRNA

Unique index to the GO: GO:0005763 (mitochondrial small ribosomal subunit)

Reference to an article indexed in the SGD and PubMed databases ("Nucleotide sequence of the gene for the mitochondrial 15S ribosomal RNA of yeast"): SGD_REF:S000073642 |PMID:6261980

Type of evidence: ISS (inferred from sequence or structural similarity)

An example annotation (continued)

GO aspect: C (Component)

A name for the gene product in words: 15S_rRNA|15S_RRNA_2

Type of object annotated: gene

Taxonomic identifier of species encoding gene product: taxon:4932 (*Saccharomyces cerevisiae*)

The date GO annotation was defined: 20040202

Source of the annotation: SGD

- Note: As a general rule it is recommended that gene products are annotated instead of genes due to the possibility of different products and/or different functions for the same gene.

Some types of evidence

- IEA, inferred from electronic annotation
  - Automatic annotation based on sequence-similarity or existing data in databases
- ISS, inferred from sequence or structural similarity
  - E.g. sequence similarity -based results after having been reviewed for accuracy by a curator
- TAS, traceable author statement
  - The curator has read a published scientific paper and considered the annotation reliable based on this article.

Example uses of gene ontology

- Integrating proteomic information from different organisms
- Assigning functions to protein domains
- Verifying models of genetic, metabolic and product interaction networks
- Developing automated ways of deriving information about gene function from the literature
- Finding functional similarities in genes that are overexpressed or underexpressed in diseases
  - The topic of today

Use in gene list analysis

- Having obtained a list of interesting genes from an experiment we can find all the annotations for them.
- Next, from these annotations we can find out ontology terms that have more annotations in our list of genes than would be expected by chance.
  - To define "more than expected by chance" we use a statistical test described below.
  - Note that an item annotated to a term is really annotated to all the ancestors of that term as well - although not all the descendants!
- The terms of GO found now become our items of interest and can be compared with existing biological knowledge.
  - From a list of interesting genes to a list of interesting biological processes, functions, and/or components.

The hypergeometric distribution

- Consider a population of objects of two types, 1 and 2, with N items in total and M items of type 1.
- If we select, without replacement, n items from the population, the probability that exactly k objects of type 1 will be selected is given by the hypergeometric distribution with parameters N, M and n:

$$f(k, N, M, n) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}, \qquad k = 0, \ldots, n.$$

- How is this applicable in our case?

A statistical model for the annotations

- Denote with
  - N the total number of genes in our analysis
  - M the number of genes annotated with a GO term i
  - n the number of genes considered interesting ("type 1")
  - k_i the number of genes out of the above n genes that are annotated with GO term i

Testing for the enrichment of GO terms

- The probability that k_i would by chance have a value at least as large as observed in the data is given by

$$\begin{aligned} p_i &= \sum_{j=k_i}^{\min\{M_i,n\}} f(j, N, M_i, n) \\ &= 1 - \sum_{j=0}^{k_i-1} f(j, N, M_i, n) \\ &= 1 - F(k_i - 1, N, M, n), \end{aligned}$$

  where F denotes the cumulative distribution function of the hypergeometric distribution.

- The smaller this value is, the smaller the probability that the observed frequency can result by chance. In other words, the smaller the better.

Significant enrichment

- The terms with low p_i are called significant (after setting some threshold, e.g. 0.01 or 0.05).
  - For example, if there are $N = 1000$ genes out of which $M_i = 100$ are annotated with a term i. The set of interesting genes contains $n = 50$ genes. The probability that, by chance alone, at least $k_i = 10$ genes out of those 50 are annotated with term i is $p_i = 0.0274$ suggesting a significant enrichment.
  - The list of terms can be ordered according to the p-values and attention focused on the more interesting categories of the ontology.

Problems in the testing

- Since we do a number of tests (one for each term) we encounter the multiple testing problem: the significance of our results is not as high as suggested by p_i.
- Both the ontology and the annotations are updated regularly
  - Results can change overnight if you use an updated ontology, make sure you know the versions you have used in each analysis.
- The approach presented here is by no means unique.
  - When using a tool to study gene ontology enrichment, make sure you know what the software is doing.
- Gene lists produced in publications to tackle the same problem can have almost no overlap between different studies.
  - Can we really trust the results of GO enrichment analysis?

Notes on the ontology

- In practice, the results from data analysis using gene ontology are also not quite as simple to interpret as one might think:
  - Without biological knowledge it is difficult or impossible to consider the relative importance of different categories in the results (i.e. what is biologically specific and what is not).
  - Some paths between two nodes are longer than others: we should not talk about e.g. "3rd level terms of the ontology" without taking precaution.
- Automated computational GO annotations or refinements.

Summary

- The gene ontology is a useful source of biological knowledge freely available for use in automated data analysis.
- The ontology itself contains a vocabulary of terms for biological components, processes and functions.
- The annotations associate genes and/or gene products with the terms of the ontology.