

Automated Enrichment of BioOntologies

Catia Pesquita, Francisco M. Couto, Mário J. Silva
XLDB, Lasige, Department of Informatics, University of Lisbon, Portugal
e-mail: cpesquita@xldb.di.fc.ul.pt

What?

Add new terms and relationships to BioOntologies using computational techniques.

Why?

Maintenance and development of BioOntologies requires experts' time and effort. Automated enrichment would help both curators and end-users.

Where?

Proposal will be tested in the Gene Ontology, the most popular BioOntology. It provides a schema for the description of gene products characteristics in a cellular context.

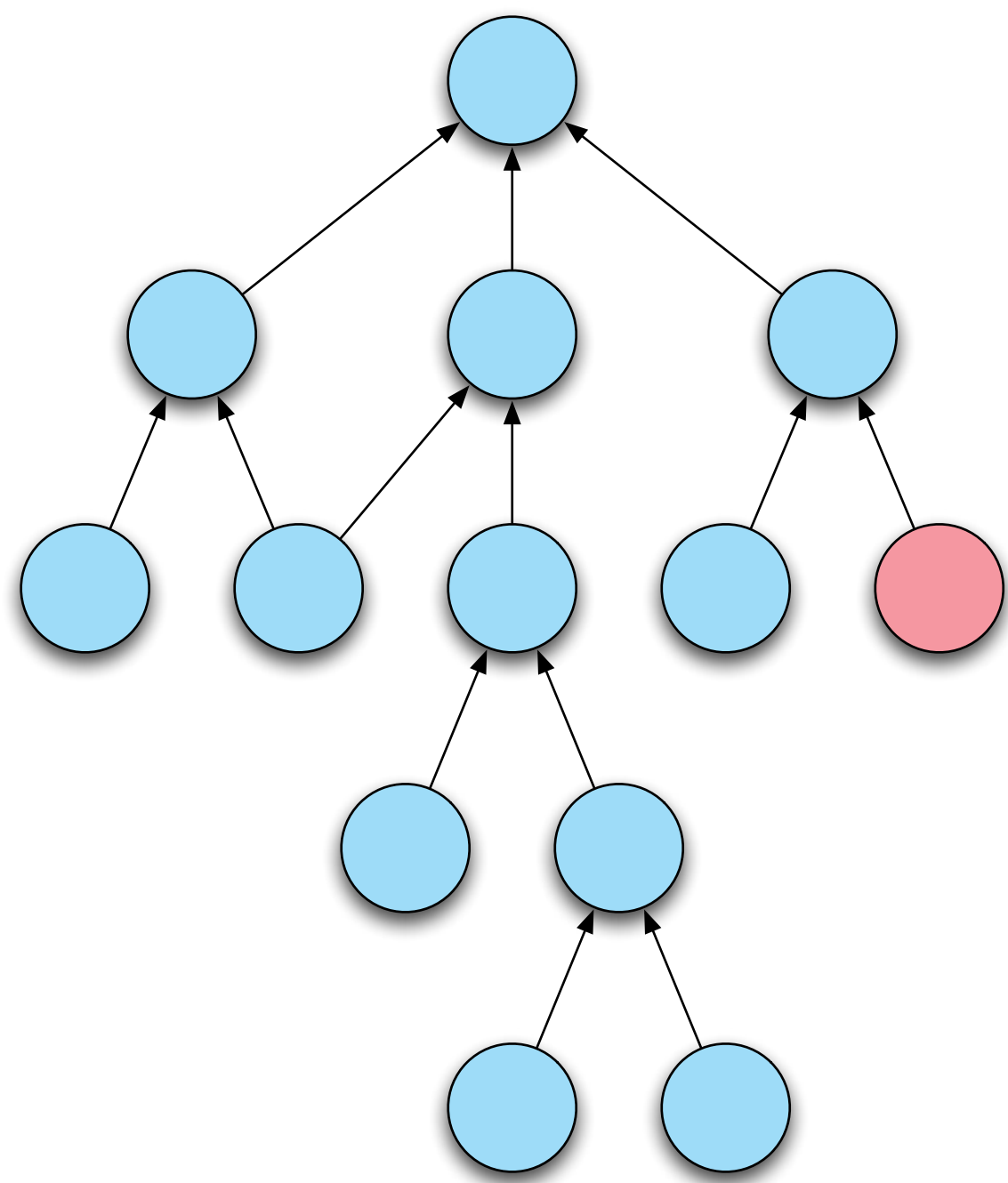
How?

1. Identify GO areas that would benefit from enrichment (hotspots).
2. Use external resources to retrieve new and relevant terms and relationships.
3. Integrate new terms and relationships in GO's hierarchy.

Task 1: Find hotspots

GO hotspots are areas where the current level of detail is not answering the community's needs:

- > Terms with high number of direct annotations in comparison with their siblings.
- > Terms whose usage has increased significantly in a given period of time.
- > Focus on mid-level terms, since changes to upper-level terms would disrupt GO's basic structure and low-level terms are already detailed.



Task 2: Retrieve new terms

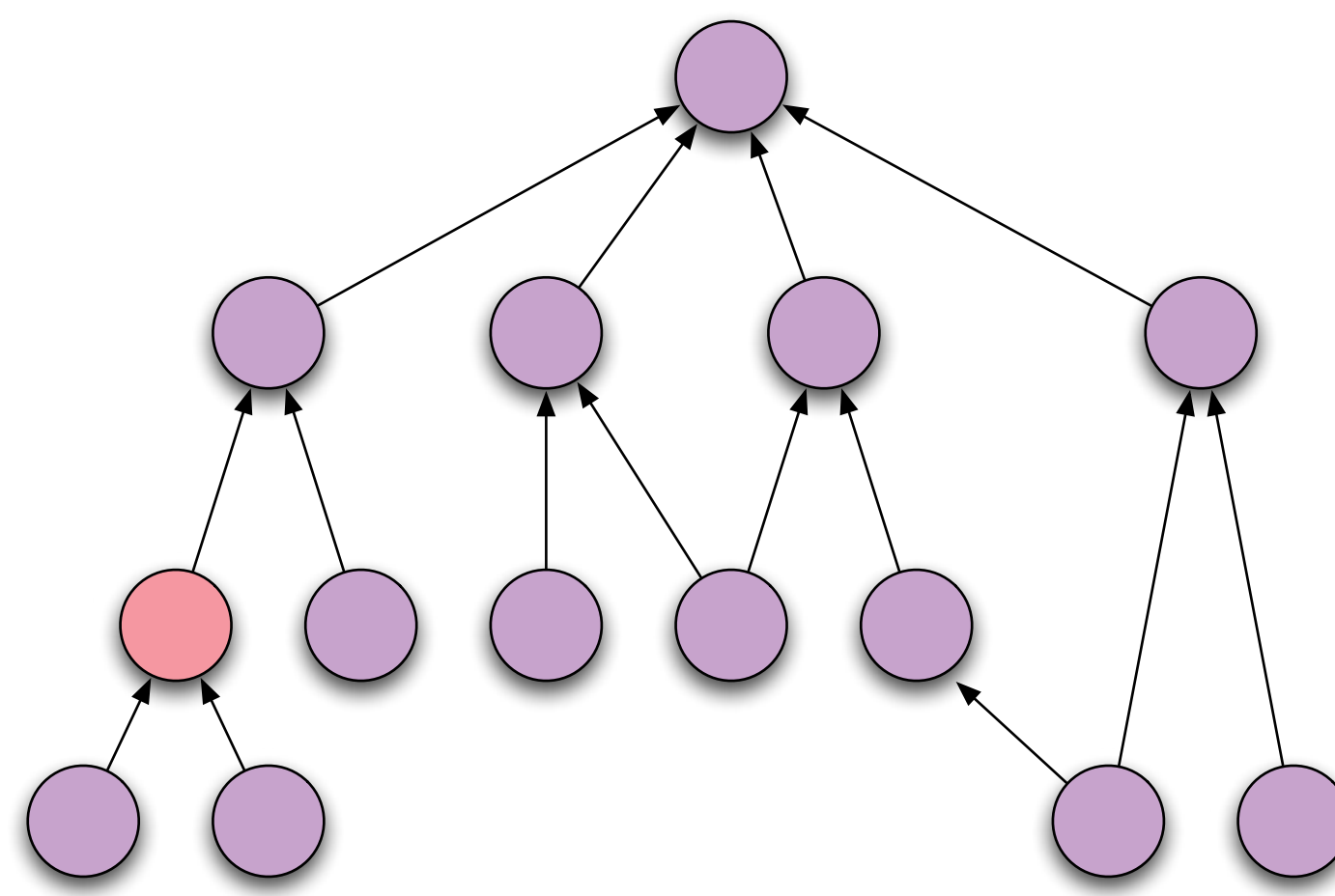
Most biological knowledge is stored in scientific publications, and there is a considerable number of publicly available abstracts online (18M in Pubmed).

Text mining techniques need to be used to retrieve the relevant documents, and extract from them the relevant terms and relationships.



Another important source of biological knowledge is already in structured form in other BioOntologies.

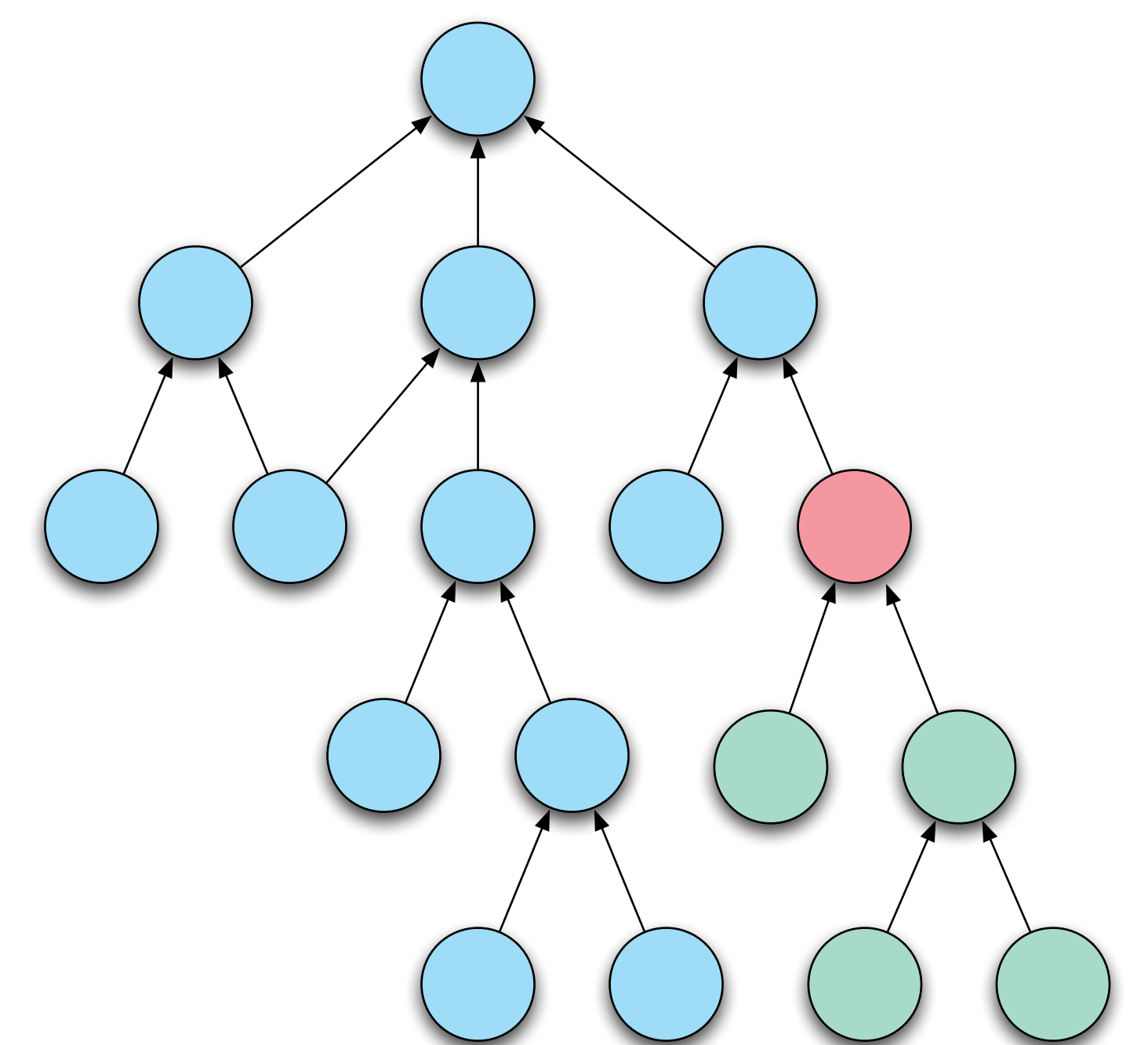
To benefit from this knowledge, the two ontologies need to be aligned, by finding the relations between the two ontologies' terms.



Task 3: Integrate into GO

Since not all new terms found may be direct descendants of the hotspot term, they need to be hierarchized, in order to reflect their degree of specificity.

Both clustering techniques and natural language processing can be employed to this end, since we will need to find the relative specificity of each term in relation to the others.



The Gene Ontology

- > describes gene products' characteristics in 3 areas:
 - >> molecular function
 - >> biological process
 - >> cellular component
- > organized as a directed acyclic graph, where terms are nodes and the relationships between them are edges.
- > used by GOA and others databases/projects to annotate gene products.

Ontology: knowledge representation scheme for a given domain and with a given scope, whereby the concepts of that domain are formally and objectively defined, and whose structure reflects the relationships between those concepts.

GO statistics

