

Sofie Van Landeghem^{1,2}, Yvan Saeys^{1,2}, Chris Cornelis³, Mohammed Adberrafi Benotmane⁴ and Yves Van de Peer^{1,2}

¹ Department of Plant Systems Biology, VIB, Technologiepark 927, 9052 Gent, Belgium

² Department of Molecular Genetics, Ghent University, Ghent, Belgium

³ Computational Web Intelligence, Department of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281 (S9), B-9000 Gent, Belgium

⁴ Laboratory of Radiobiology, Studiecentrum voor Kernenergie/Centre d'etude de l'Energie Nucleaire (SCK-CEN) Boeretang 200, Mol B-2400, Belgium

E-mail : sofie.vanlandeghem@ugent.be

Introduction

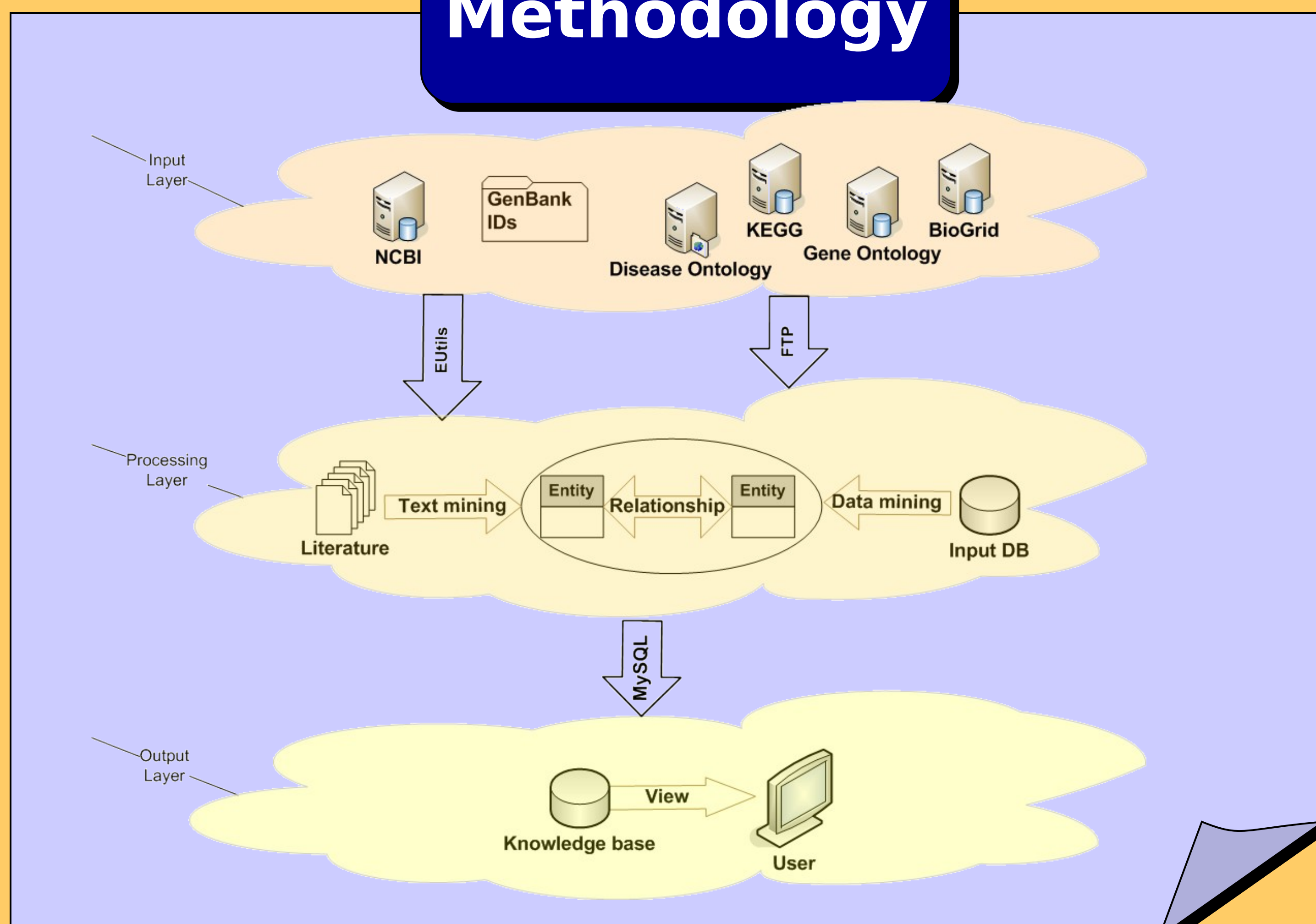
As publicly available data in the field of life sciences is increasing continuously, **data mining** techniques that automatically **extract knowledge** from online databases and scientific literature have become a time saving necessity. We present **GeneFetch**, a system that enables biologists to easily **find and browse vital information on any gene or group of genes**. A **knowledge base** is constructed automatically by retrieving facts from established sources like the National Center for Biotechnology Information and Gene Ontology, as well as by performing **text mining** on PubMed articles to find links between genes, diseases and GO terms. We offer up-to-date records by incrementally expanding the database as new data become available. GeneFetch is the first application that **fully integrates facts with text mining results**, offering a full report per gene. This summary can be used by a whole range of applications, including clustering and customized visualization. Our example web application is available at <http://bioinformatics.psb.ugent.be/webtools/genefetch>.

GeneFetch

The screenshot displays the GeneFetch web application interface, divided into four main sections:

- FILTER:** A section titled "SELECT A FILTER" where users can choose a filter. The current filter is "Mus musculus". A list of genes is shown, including NM_207707, NM_008709, J04847, NM_007956, and NM_021884. A red box highlights this list with the text "Find common properties of a group of genes".
- BIOLOGICAL PROCESSES:** A section titled "PERSONAL LIST" showing 38 biological processes related to the selected genes. Processes include cell differentiation (3), regulation of transcription, DNA-dependent (3), transcription (3), estrogen receptor signaling pathway (2), lipid binding (2), apoptosis (1), brain development (1), branching morphogenesis of a tube (1), cell cycle arrest (1), cell death (1), cell growth (1), cellular process (1), embryonic development (1), glucose homeostasis (1), gonadotropin secretion (1), imprinting (1), induction (1), interleukin-2 production (1), keratinocyte differentiation (1), learning (1), and liver development (1). A red box highlights "cell differentiation (3)" with the text "Look at the genes involved in a certain process".
- SUMMARY:** A section titled "SUMMARY NM_008709" showing related biological entities. It lists related genes, diseases, pathways, molecular functions, and biological processes. A red box highlights "cell differentiation : PubMed sentence" with the text "Show origin of extracted relationships".
- BIOLOGICAL PROCESS:** A section titled "CELL DIFFERENTIATION" showing 3 related genes in the filter for this process. Genes listed include J04847/DIPa1, NM_008709/Mycn, and NM_021884/Tsg101. A red box highlights "NM_008709/Mycn" with the text "Show origin of extracted relationships".

Methodology



Features

- Relationships between genes, diseases, pathways, molecular functions, biological processes, cellular components and articles
- Data mining of NCBI and GeneOntology databases
- Text mining on PubMed articles
- Results combined and stored in local database
- Fast access to data
- Database updated and incrementally expanded
- Profile for each gene
- Summary for each gene listing related biological entities
- Evidence stored to show origin of extracted relationships
- General properties of a group of genes
- Organism independent
- Filter functionality

Applications in systems biology

- Constructing protein-protein **interaction networks**
- **Clustering** of genes using annotations extracted from literature and through data mining
- **Prediction** of protein function and other properties through the **analysis of homologues**