IntelliGO semantic similarity measure for Gene Ontology annotations

Marie-Dominique Devignes

Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA)

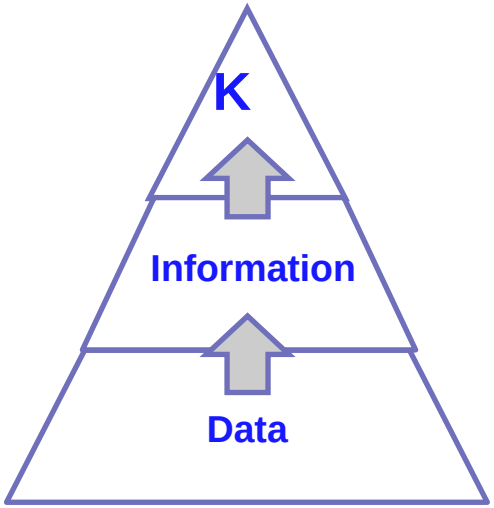
Equipe Orpailleur – INRIA Nancy Grand-Est

Outline of the talk

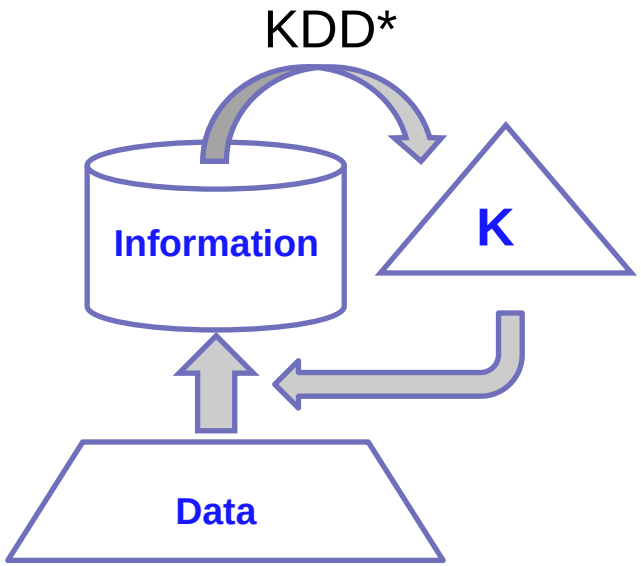
1. Introduction: Knowledge Discovery guided by Domain Knowledge
2. IntelliGO : a semantic similarity measure for GO annotations
3. IntelliGO-based clustering of genes
4. IntelliGO-based abstraction for data mining
5. Conclusion

LORIA, Team Orpailleur

- Making data talk : from data to knowledge



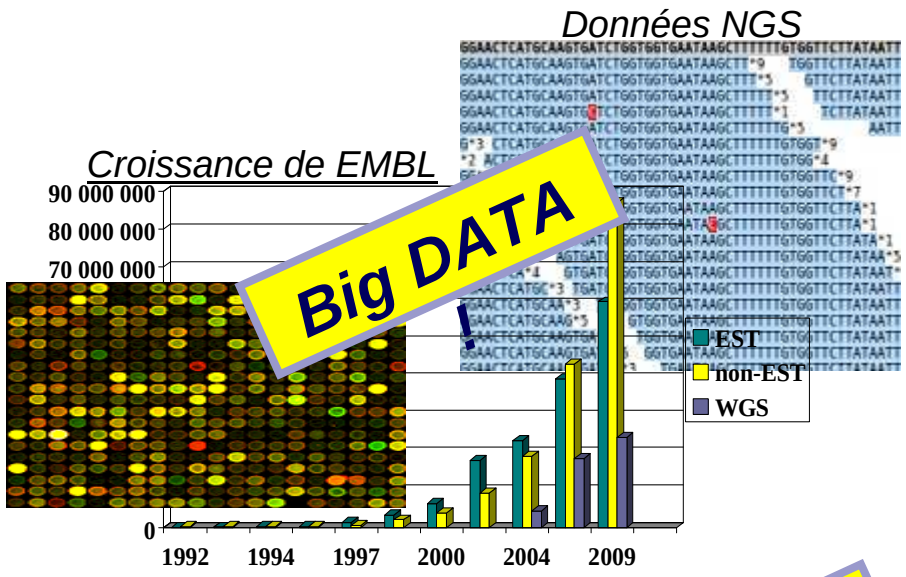
Static picture : pyramid

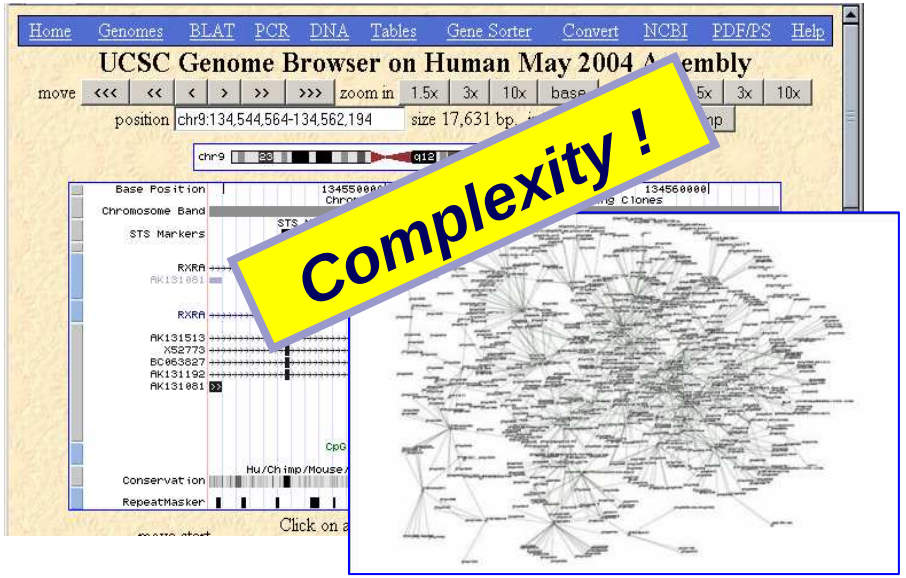


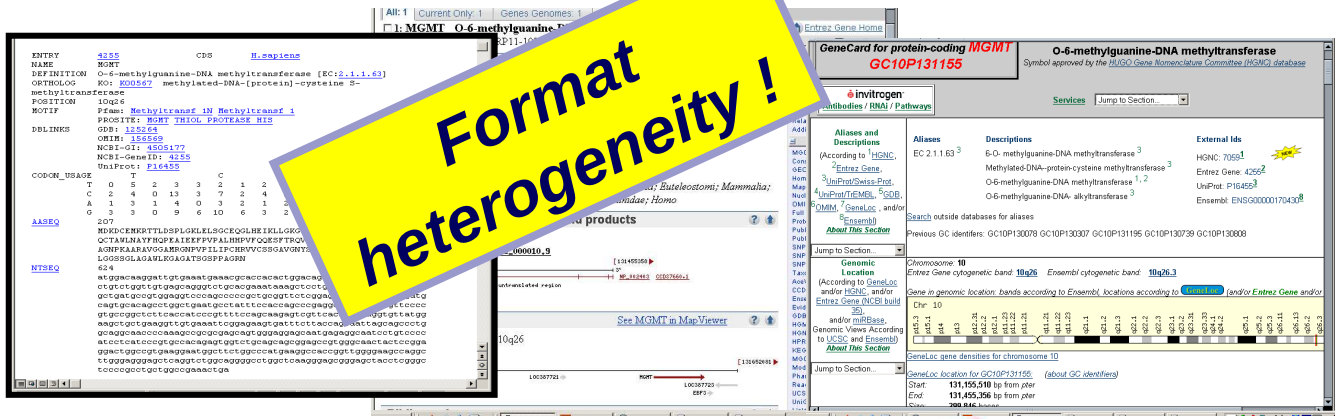
Dynamic picture : loop

** KDD : Knowledge Discovery from Databases*

The data deluge…





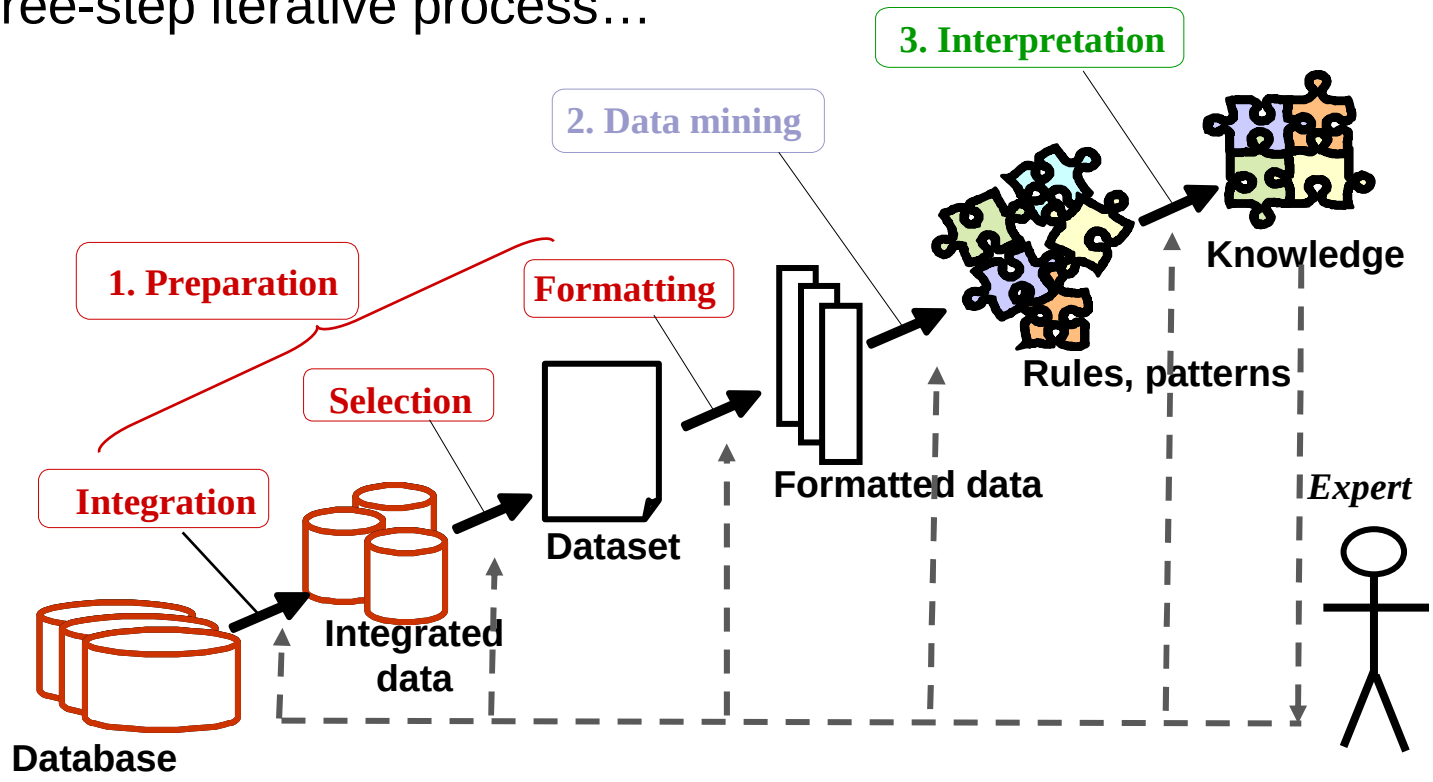


→ *No knowledge without knowledge …*
(Amedeo Napoli)

Loria Laboratoire lorrain de recherche en informatique et ses applications | cnrs | Inria | UNIVERSITÉ DE LORRAINE

Knowledge Discovery from Databases (KDD)

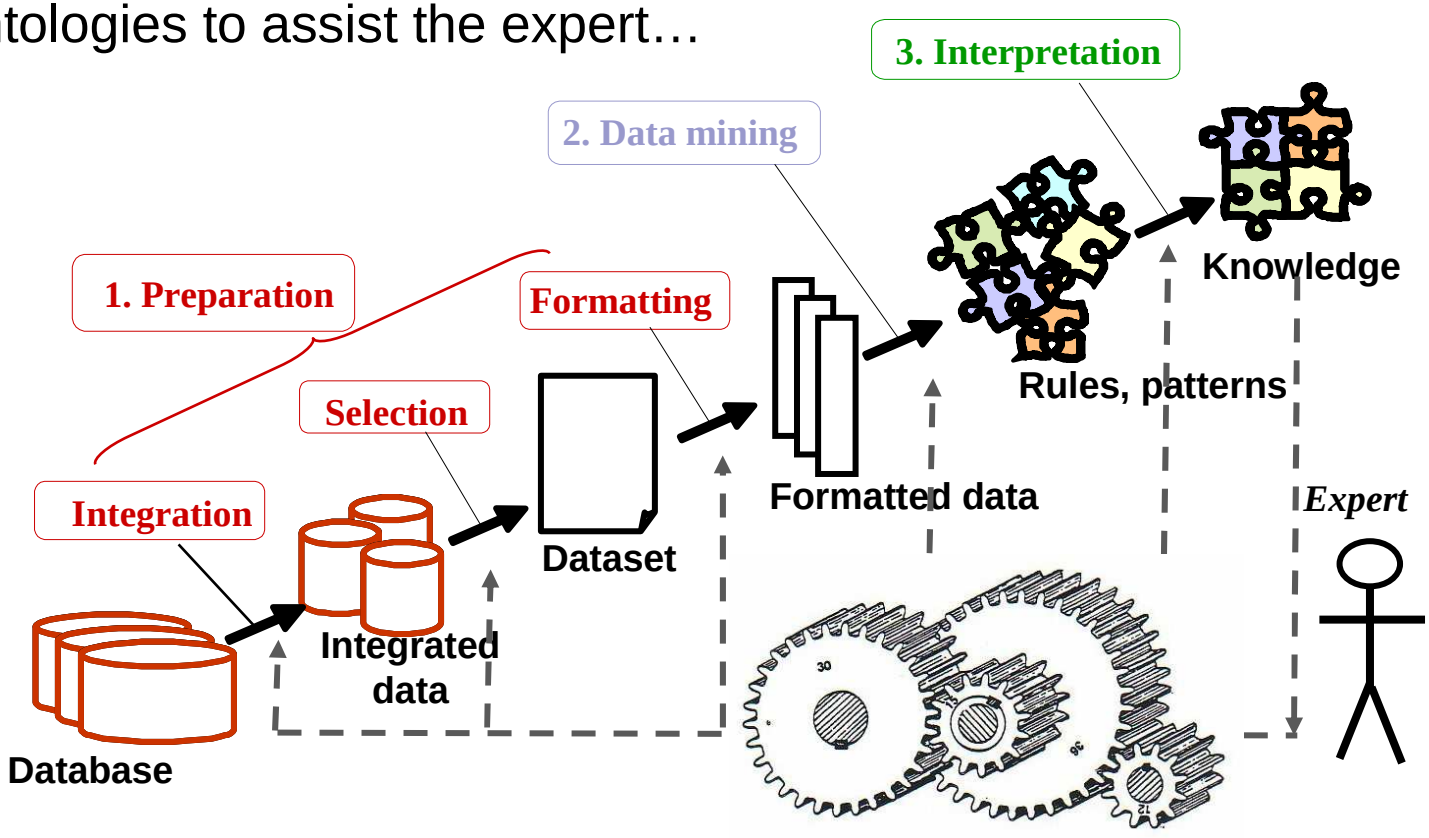
A three-step iterative process…



…interactively controlled by an expert.

Knowledge Discovery guided by Domain Knowledge : KDDK

Ontologies to assist the expert…



… at each step of the process.

Our scope today :

- How ontologies can be used to improve the data mining step :
  - Design of a semantic similarity measure on Gene Ontology
  - Use it for functional clustering of genes
    - ***Improvement of gene classification***
  - Apply it to another ontology (MedDRA) for secondary effects clustering
    - ***Data abstraction and dimension reduction***
    - ***Essential for execution of data mining programms***

Outline of the talk

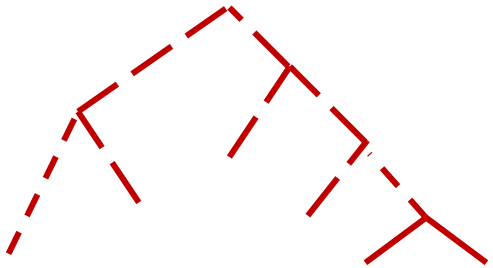
1. Introduction: Knowledge Discovery guided by Domain Knowledge
2. IntelliGO : a semantic similarity measure for GO annotations
3. IntelliGO-based clustering of genes
4. IntelliGO-based abstraction for data mining
5. Conclusion

Semantic similarity in biomedical ontologies

- Biological objects are « annotated » by terms / keywords
  - → facilitates data retrieval from database
- Annotation terms / Keywords are organized as controlled vocabularies
  - Tree : e.g. Swissprot Keywords
  - Thesaurus : e.g. MESH, UMLS
  - Ontologies : e.g. Gene Ontology
- Similarity measure comparing annotations can take advantage of semantic relationships between terms /keywords
  - → **« semantic similarity measure »**

Semantic similarity of annotations

- General intuition….



Semantic relationships between terms

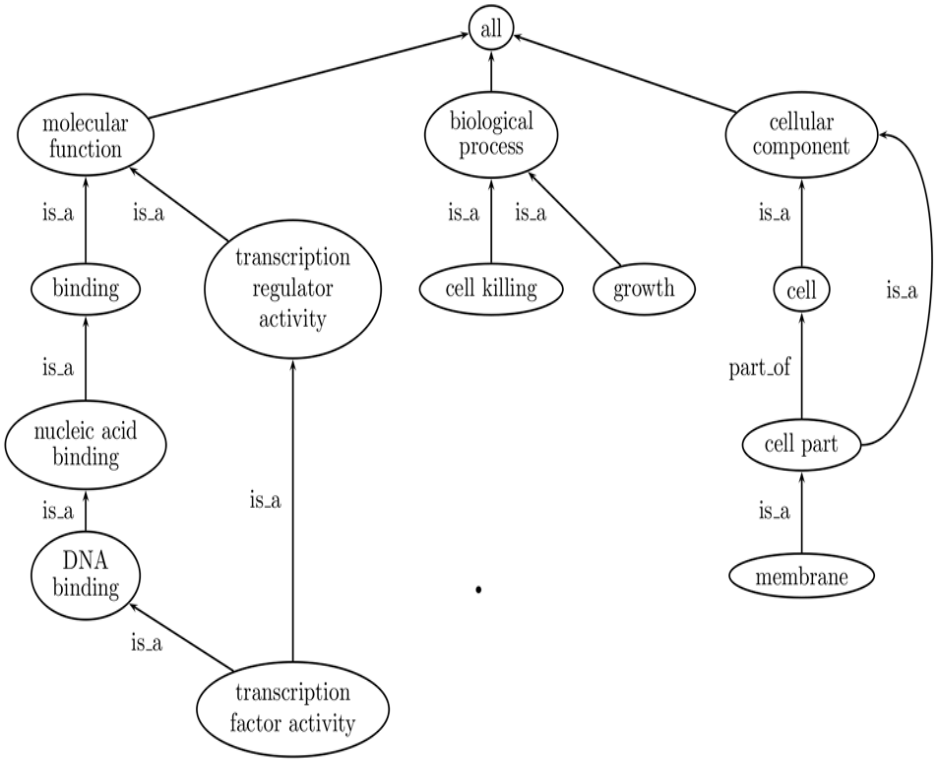
	GO-t1	GO-t2	GO-t3	…
Gene1	X	X	O	…
Gene2	X	O	X	…
…				

Objects as « feature vectors »

« Euclidean » Score	**1**	**0**	**0**	→ **1**
« Semantic » Score	**1**	**0.5**	**0.5**	→ **2**

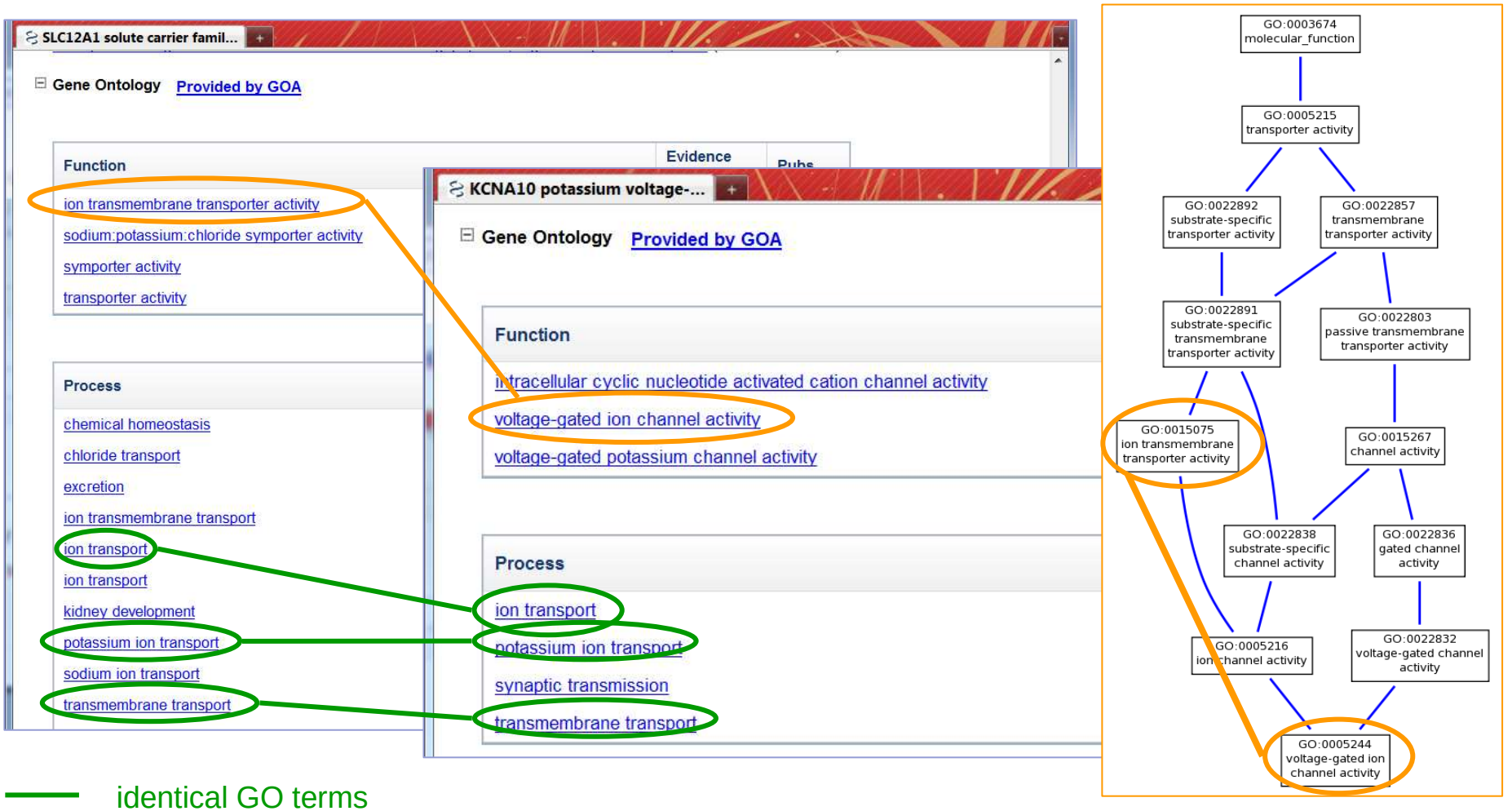
Brief outlook on GO (1)

- Annotation vocabulary (GO consortium since 2000)
  - More 30,000 terms
  - Millions of annotations (genes and proteins in various species) recorded in public databases
  - Three aspects :
    - ***Biological Process***
    - ***Cellular Component***
    - ***Molecular Function***
- Structured as a Rooted Directed Acyclic Graph
  - Nodes = terms
  - Edges = semantic relationships
    - *Is_a, part_of, related to*
  - IMPORTANT : More than one parent



from Pesquita et al., 2009

Example of similar GO annotations



identical GO terms

similar GO terms

GO evidence codes

- Traces the origin of each annotation
  - IEA (« Inferred from Electronic Annotation ») = not assigned by curator
  - All others (16) manually assigned

Experimental	Computational analysis	Author statement	Curatorial statement
EXP	ISS	TAS	IC
IDA	ISO	NAS	ND
IPI	ISA		
IMP	ISM		
IGI	IGC		
IEP	RCA		

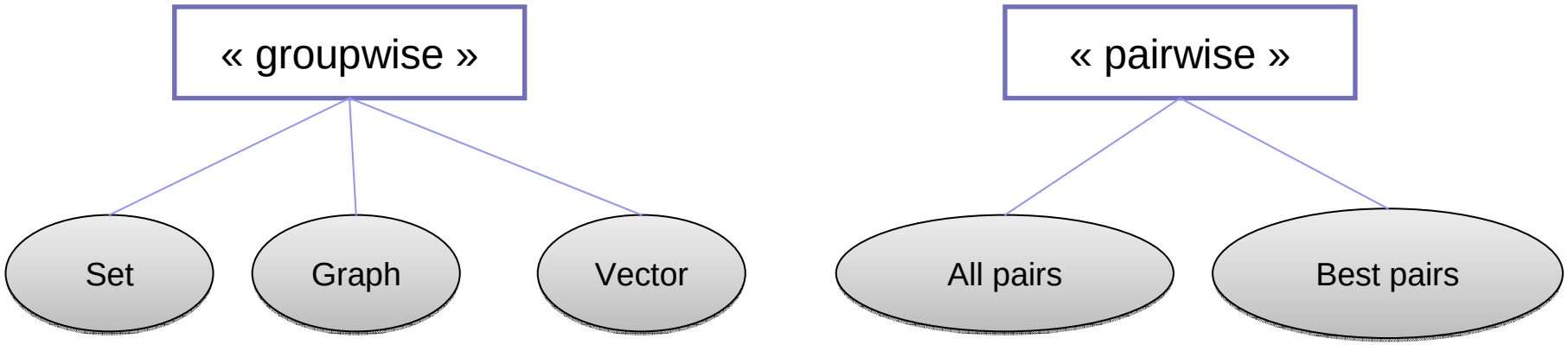
« Evidence codes cannot be used as a measure of the quality of the annotation »
http://www.geneontology.org/GO.evidence.shtml

- To be used as a filter for certain type of annotations

Semantic similarity measures : overview (1)

Pesquita et al., 2009: Semantic similarity in biomedical ontologies, PLOS Comp. Biol. July 2009

1. Gene-Gene or Protein-Protein comparison
   = comparing two lists of terms (feature vectors)

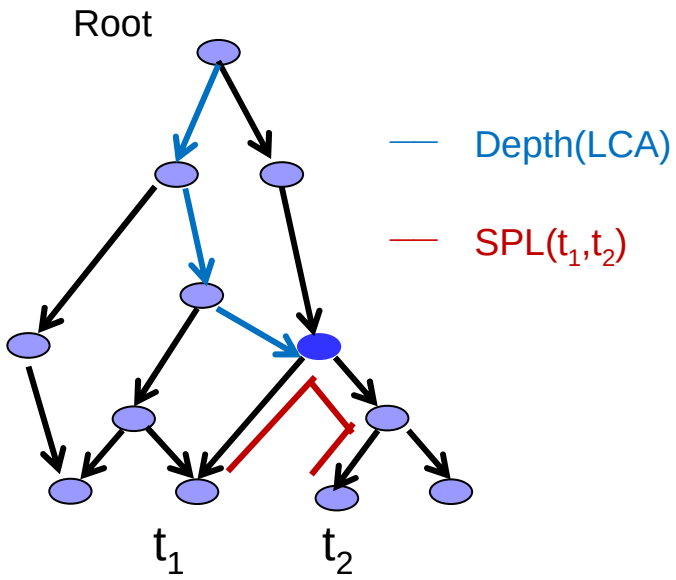


- Huang *et al.* (2007) : Vector/ Kappa-statistics (DAVID)
- Martin *et al.* (2004) : Graph/Jaccard

- Lord *et al.* (2003) : All pairs/ Average/ Resnick, Lin, Jiang measures
- Wang *et al.* (2007) : best pairs/Average/ Wang measure

Semantic similarity measures : overview (2)

2. Term-term comparison
   = exploiting the semantic relationships



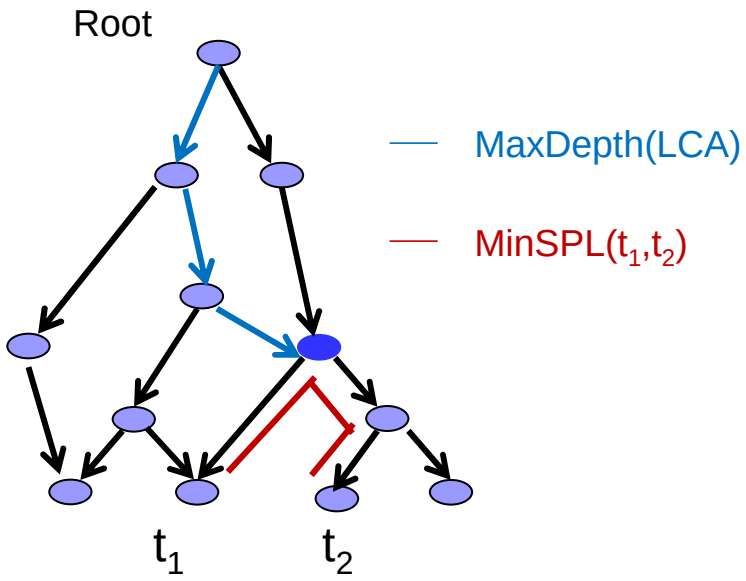
- Node-based approaches
  - Information content of lowest common ancestor (LCA)
    - *Relative to a corpus*
  - Depth(LCA)
    - *Independent of corpus*
- Edge-based approaches
  - Shortest Path Length (SPL)

IntelliGO semantic similarity measure (1)

Benabderrahmane et al., 2010 : BMC Bioinformatics

1. Term-term comparison

= combining node-based and edge-based approaches in a DAG



In a tree (MeSH, Ganesan 2003)

$$Sim_{Ganesan}(t_1,t_2) = \frac{2 \times Depth(LCA)}{Depth(t_1) + Depth(t_2)}$$

In a rooted DAG (GO, IntelliGO 2010)

$$Sim_{IntelliGO}(t_1,t_2) = \frac{2 \times MaxDepth(LCA)}{MinSPL(t_1,t_2) + 2 \times MaxDepth(LCA)}$$

IntelliGO semantic similarity measure (2)

2. Gene (or protein) representation

- Representation of genes in a vector space model

$$\vec{g} = \sum_i \alpha_i \vec{e_i}$$

$\vec{e_i}$: basis vector, one per term (t_i)

α_i : Coefficient for term t_i

- Definition of coefficients

$$\alpha_i = w(g, t_i) \times IAF(t_i)$$

$w(g, t_i)$: weight of evidence code * qualifying the assignment of feature t_i to gène g

$IAF(t_i)$: « Inverse Annotation Frequency » ~ Information Content of feature t_i in annotation corpus.

** When more than one code, take the maximal weight*

- Definition of information content

$$IAF(t_i) = \log \frac{N_{TOT}}{Nt_i}$$

N_{TOT} : Total number of genes in the corpus

Nt_i : Number of gènes with feature t_i

IntelliGO semantic similarity measure (3)

3. Gene-gene comparison

- Generalized dot-product between two gene vectors

$$\vec{g} \cdot \vec{h} = \sum_{i,j} \alpha_i \times \beta_j \times \vec{e_i} \cdot \vec{e_j}$$

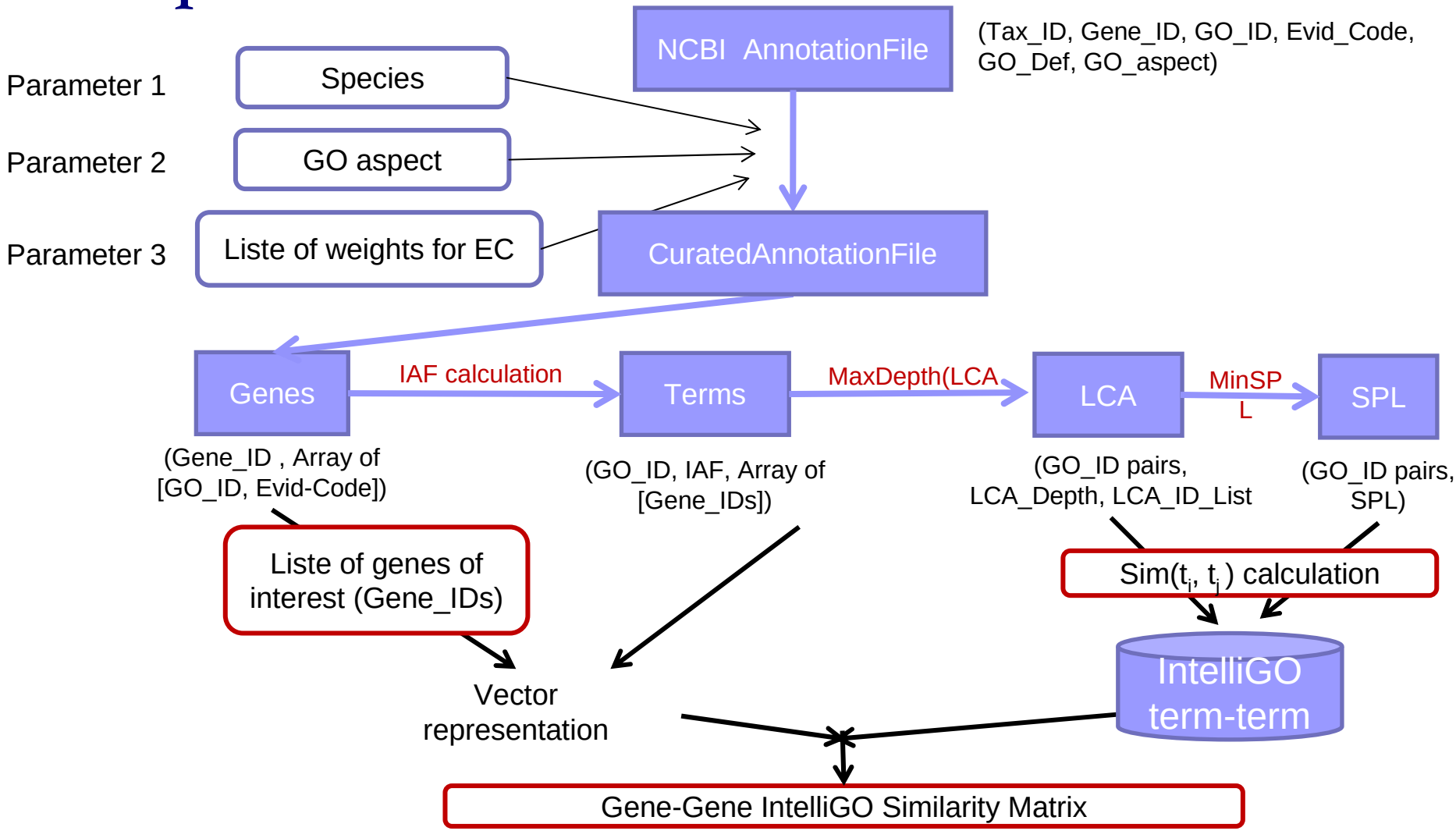
with $\vec{e_i} \cdot \vec{e_j} = \mathrm{Sim}_{IntelliGO}(t_i, t_j)$

$\begin{cases} 1 \text{ for } i = j \\ \neq 0 \text{ for } i \neq j \end{cases}$

- Generalized cosine similarity

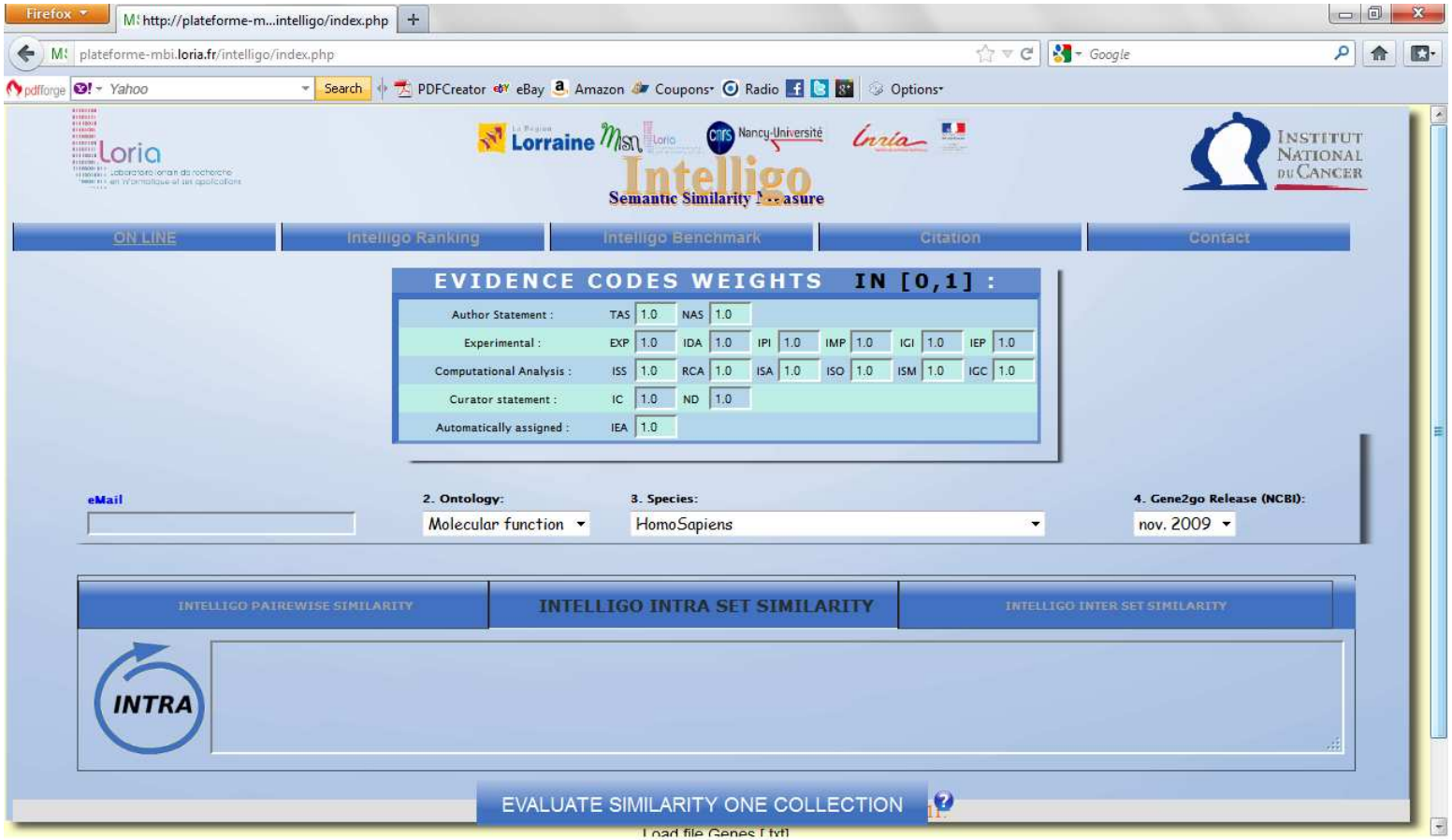
$$\mathrm{Sim}_{IntelliGO}(\vec{g}, \vec{h}) = \frac{\vec{g} \cdot \vec{h}}{\sqrt{\vec{g} \cdot \vec{g}} \times \sqrt{\vec{h} \cdot \vec{h}}}$$

Implementation



IntelliGO on-line

- http://plateforme-mbi.loria.fr/intelligo



Outline of the talk

1. Introduction: Knowledge Discovery guided by Domain Knowledge
2. IntelliGO : a semantic similarity measure for GO annotations
3. IntelliGO-based clustering of genes
4. IntelliGO-based abstraction for data mining
5. Conclusion

Gene clustering with IntelliGO : reference datasets

- Clustering means grouping together most similar objects and putting in different clusters most dissimilar objects
  - → **relies on a similarity/distance measure**
- Evaluation purpose : Reference sets of genes
  - **participating in similar biological process → pathways**
  - **sharing similar molecular functions → Pfam Clans**
  - **In two different species**

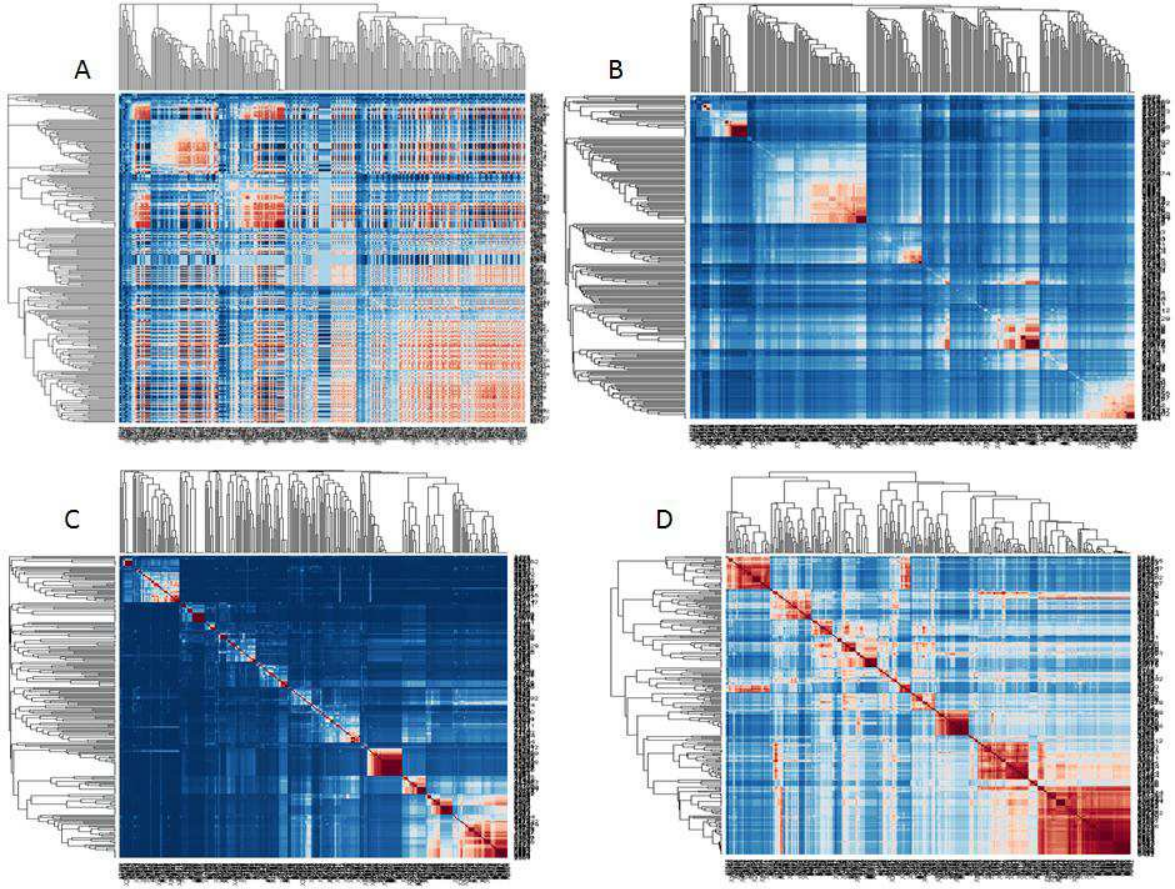
Dataset	Species	Source	Number of sets	Total genes
1	Human	KEGG pathways	13	275
2	Yeast	KEGG pathways	13	169
3	Human	Pfam Clans	10	94
4	Yeast	Pfam Clans	10	118

1. Hierarchical clustering for comparing IntelliGO with other semantic similarity measures

- For each collection of gene sets
  - List all genes from all sets
  - Pairwise similarity calculation - > distance matrix
    - *Lord's measure (normalized) : based on IC(LCA)*
    - *Al-Mubaid's measure : based on SPL*
    - *SimGIC measure : based on count of common ancestors*
    - *IntelliGO measure : based on both MaxDepth(LCA) and minSPL*
  - Hierarchical clustering and heatmap visualisation
    - *R BioConductor*

Heatmap visualisation of hierarchical clustering

Lord et al. (normalized)



Al-Mubaid

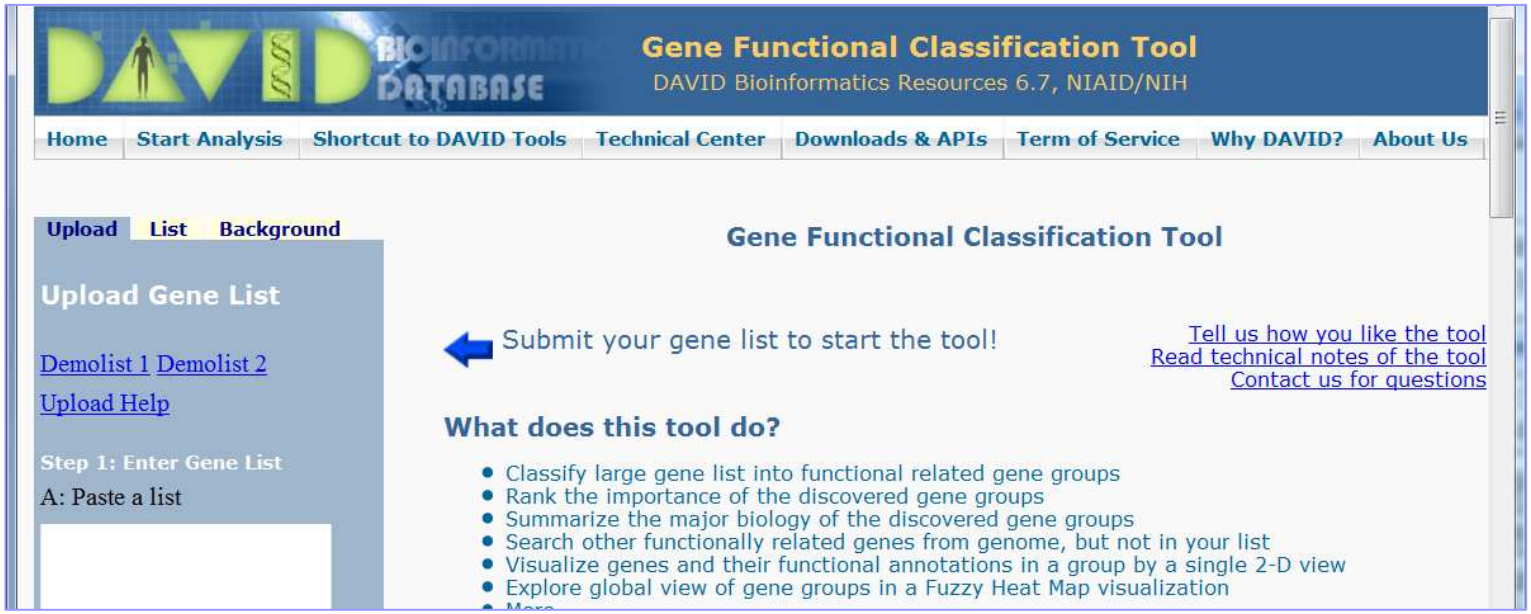
SIM_{GIC}

IntelliGO :

- well-balanced
- suggests fuzziness

2. Functional classification : the DAVID software (1)

- Functional classification = Clustering of functional annotation of genes:
- On-line tool : DAVID (Database for Annotation Visualisation and Integrated Discovery) software



DAVID functional classification

- Rich feature vectors : GO terms, domains, Enzyme Classification, Pathways, etc.
- Kappa statistics : similarity measure based on co-occurrence and co-absence of features
- Fuzzzy clustering agorithm : one gene can belong to more than one cluster.
- Optimization of K number by varying kappa threshold -> % excluded genes

	GO-t1	GO-t2	GO-t3	…	PfamD1	…
Gene1	X	X	O	…	X	…
Gene2	X	O	X	…	X	…
…						

Counting present and absent features + 'Kappa statistics'

Not a semantic similarity measure

No feature selection possible

3. IntelliGO-based fuzzy C-means

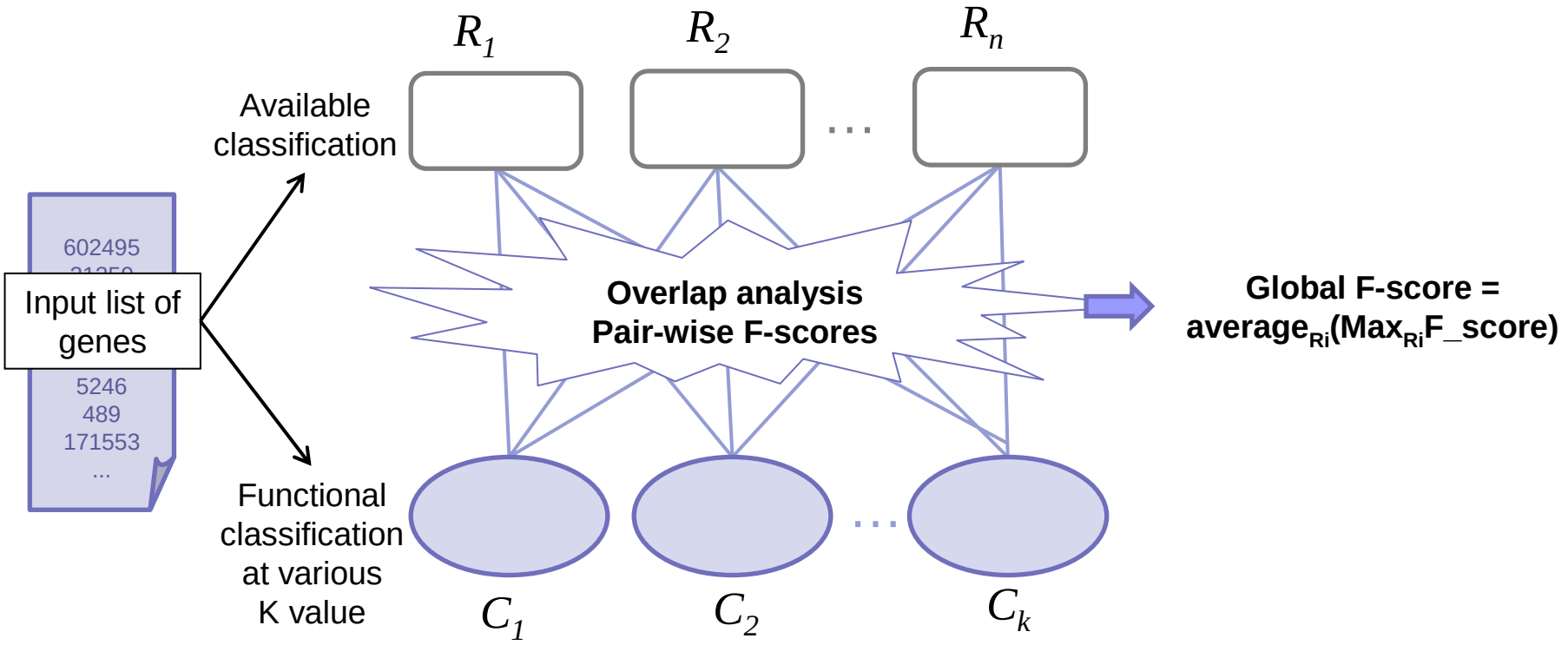
- IntelliGO distance matrix -> *fanny* implementation of fuzzy C-means in R environment, minimization of objective function $\sum_{v=1}^{k} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} u_{iv}^{r} u_{jv}^{r} d(i,j)}{2\sum_{j=1}^{n} u_{jv}^{r}}$
- Remark : not a feature vector matrix + euclidean distance → no centroid calculation, no appropriate validity index
- Membership probability matrix: one gene can belong to more than one cluster.
- -> % excluded genes

	C1	C2	C3	…
Gene1	90 %	0	8 %	…
Gene2	25 %	40 %	0 %	…
Gene3	0 %	10 %	13 %	…

Membership threshold : 20 %

	Assignt
Gene1	C1
Gene2	C1, C2
Gene 3	Excl.

4. Comparison using overlap analysis



Precision (Ci, Rj) : % genes present in Ci that belong to Rj

Recall (Ci, Rj): % genes from Rj that are found in Ci

$$F\_score\ (Ci, Rj) = \frac{2 \times \text{Precision}(Ci, Rj) \times \text{Recall}(Ci, Rj)}{\text{Precision}(Ci, Rj) + \text{Recall}(Ci, Rj)}$$

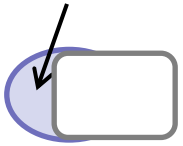
Comparison : optimal F-score and K number

	IntelliGO +Hierarchical			IntelliGO+ fanny			DAVID		
Ref sets	Opt. Fscore	Opt. K	Excl. genes	Opt. Fscore	Opt. K	Excl. genes	Opt. Fscore	Opt. K	Excl. genes
1 (13)	0.58	17	0 %	**0.70**	7	0 %	0.67	10	21%
2 (13)	0.57	17	0 %	**0.74**	9	1 %	0.68	9	18 %
3 (10)	**0.82**	18	0 %	0.77	15	5 %	0.64	11	27 %
4 (10)	0.68	13	0 %	**0.71**	12	2 %	0.70	10	41 %

Functional classification is reliable and robust with IntelliGO measure

Benabderrahmane et al., BIBM workshop IDASB 2011

5. C\R analysis : enrichment of reference sets



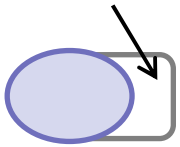
- C\R : Similar annotation but not in reference set
  - **Suggestion : include this gene in the reference set**
  - **Calculate a weight :**

$$Suggest(g, R) = \frac{\left| Enrichment\ (R) \cap Annotation\ (g) \right|}{\left| Enrichment\ (R) \right|}$$

where *Enrichment(R)* contains annotation terms specific of R (*P-value* < *Pthreshold*)

- Illustration
  - **The *STON1* gene was suggested to join human pathway hsa04130 (SNARE interactions in vesicular transport) on the basis of its BP annotation: 'intracellular transport'.**
  - **To be confirmed by biologist experts of this pathway !**

6. R\C analysis: enrichment of gene annotation



- R\C : Members of reference set not functionally classified with the rest of the set
  - **Suggestion : enrich gene annotation with terms t specific of C (present in *Enrichment*(C), given a P-value threshold)**
  - **Calculate a weight :**

$$Suggest(t) = \frac{\left| \left\{ g \,\middle|\, t \in Enrichment\ (C) \setminus Annotation\ (g) \right\} \right|}{\left| R \setminus C \right|}$$

- Illustration
  - **Adding BP term "Methionine biosynthetic process" to ARO8 gene in yeast because this term is significantly enriched in cluster C9 matching with pathway R6 (Lysine metabolism) to which ARO8 belongs to.**
  - **Indeed, the expert confirmed that ARO8 is involved in a methionine biosynthesis pathway**

Outline of the talk

1. Introduction: Knowledge Discovery guided by Domain Knowledge
2. IntelliGO : a semantic similarity measure for GO annotations
3. IntelliGO-based clustering of genes
4. IntelliGO-based abstraction for data mining
5. Conclusion

Application of IntelliGO measure to another ontology: MedDRA

- Study with MedDRA : Medical Directory of Regulatory Activities
  - **Part of UMLS, about 20,000 terms.**
  - **Used for side-effect description = subset of MedDRA previously called COSTART, about 1288 terms.**
  - **MedDRA is organized as a rDAG with five depth levels: System Organ class, High Level Group Term, High Level Term, Preferred Term, Lowest Level Term.**
  - **About 1/3 of MedDRA terms have more than one parent.**
- SIDER : Side Effect Repository at the EMBL (http://sideeffects.embl.de)
  - **About 800 drugs associated with 1288 side effect features (MedDRA terms)**
  - **Challenge : to apply symbolic data mining methods on this large dataset to discover patterns and regularities *(Emmanuel Bresso's PhD Thesis, Harmonic Pharma)***

Replacing MedDRA terms with term clusters (TC)

- Large matrices (1288 attributes) are untractable with symbolic methods
  - **Search for frequent itemsets fails because not enough features are shared !**

- Abstraction : Reduce the number of attributes without losing information
  - **Classical problem in data mining**
  - **Generalisation in a tree, not possible with a rDAG**
  - **Semantic clustering is a solution**
  - **IntelliGO similarity between MedDRA terms**

$$Sim_{IntelliGO}(t_i, t_j) = \frac{2\ Depth_{Max}[LCA\ (t_i, t_j)]}{SPL(t_i, t_j) + 2\ Depth_{Max}[LCA\ (t_i, t_j)]} \qquad DIst_{IntelliGO}(t_i, t_j) = \frac{SPL\ (t_i, t_j)}{SPL(t_i, t_j) + 2\ Depth_{Max}[LCA\ (t_i, t_j)]}$$

Hierarchical clustering of MedDRA terms

- Pairwise distances calculated for a subset of 1288 terms
- Hierarchical clustering + Kelley's optimisation of cluster number
  - ➔ 112 term clusters
  - Calculation of the most representative element
  - Validation by the expert
- Example : TermCluster T54 Erythema
  - 15 terms related to skin pathologies

Cluster terms T54	AvgDist to other cluster terms
Erythema	**0.31**
Lichen planus	0.32
Parapsoriasis	0.32
Pityriasis alba	0.32
Rash papular	0.32
Decubitus ulcer	0.35
Lupus miliaris disseminatus faciei	0.35
Pruritus	0.35
Rash	0.35
Sunburn	0.35
Vulvovaginal pruritus	0.35
Dandruff	0.37
Rash	0.37
Photosensitivity reaction	0.37
Psoriasis	0.37

Mining Cardiovascular Agents (CA) and Anti-Infective Agents (AIA)

CA_{All}

1288 attributes (terms)

	T1	T2	T3	T4	…
Drug1					
Drug2					
…					

94 drugs

CA_{TC}

112 attributes (term clusters)

	TC1	TC2	TC3	…
Drug 1				
Drug 2				
…				

94 drugs

- Idem for Anti-infective Agents : 76 drugs, 2 datasets AIA_{All} and AIA_{TC}

Frequent Closed Itemset (FCI) extraction (1)

Minimal support	50 %	60 %	70 %	80 %	90 %	100 %
CA_{All}	386	94	41	11	1	0
CA_{TC}	5,564	1,379	256	62	6	0
AIA_{All}	178	41	9	2	0	0
AIA_{TC}	654	154	30	3	0	0

- Use of Zart algorithm (Coron platform for symbolic datamining)
- FCI : maximal subsets of drugs sharing similar side effects (*All*) or similar TCs
- More FCIs are found with *TC* representation

Frequent Closed Itemset (FCI) extraction (2)

- FCIs obtained with TC representation are more informative
- Example : comparison of top- 5 FCIs obtained with AIA datasets

All representation	*TC* representation
Nausea_and_vomitting_symptoms, Vomiting (82%)	54_Erythema (88 %)
Nausea (80%)	64_Nausea_and_vomitting_symptoms (88%)
Pruritus (79%)	99_Neuromyopathy (82 %)
Nausea_and_vomitting_symptomes, Vomiting, Nausea (78%)	54_Erythema, 64_Nausea_and_ vomitting_symptoms (79%)
Headache (76%)	65_Blepharitis (78%)

Bresso et al., KDIR 2011

Conclusion

- Bio-ontologies and data mining
  - Semantic distances can lead to better clustering than euclidean distances
    - *Exploit 'omics' datasets*
    - *Overlap analysis for annotation curation*
  - Semantic abstraction make symbolic methods practicable
    - *-> extract explicit knowledge from large real-world datasets*
- Systems Biology
  - Data-driven modelling of complex systems implies KDD approaches
  - Semantic similarity may help in reducing data complexity

Acknowldegements

LORIA, Equipe Orpailleur Nancy

MD Devignes

Malika Smaïl-Tabbone

Sidahmed Benabderrahmane

Jean-François Kneib

http://plateforme-mbi.loria.fr/intelliGO

Harmonic Pharma (Start-up), Nancy

Michel Souchet

Emmanuel Bresso

Financements

INCa (bourse de thèse interdisciplinaire)

Communauté Urbaine du Grand Nancy

Contrat Plan Etat Région : MISN