Bio-ontologies

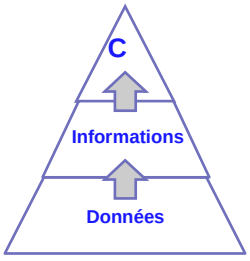
Marie-Dominique Devignes

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

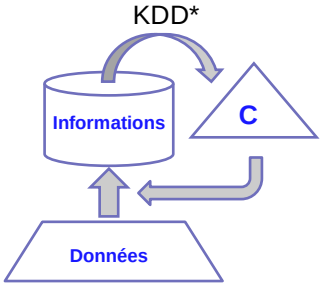
Equipe Orpailleur – INRIA Nancy Grand-Est

LORIA, Equipe Orpailleur

- Faire parler les données : passer des données aux connaissances



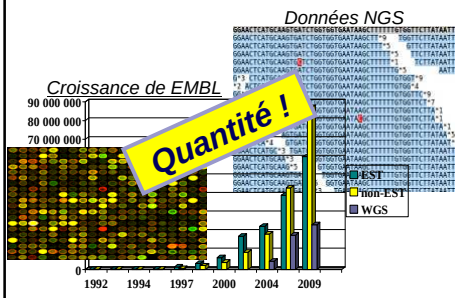
Vision statique, pyramidale

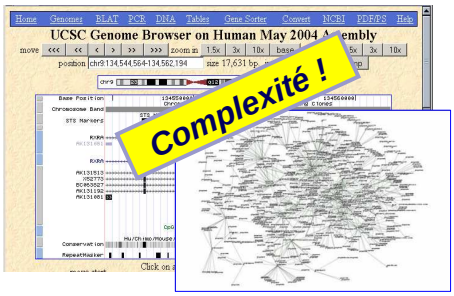


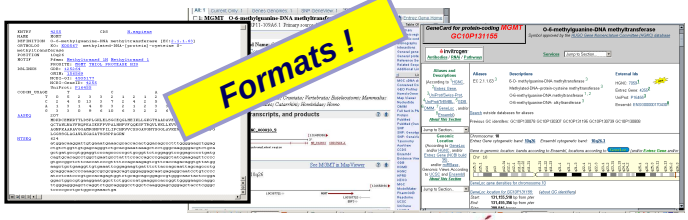
Vision dynamique, en boucle

** KDD : Knowledge Discovery from Databases*

Exploitation des bases de données biologiques







Paradoxe : Trop d'info tue l'info !

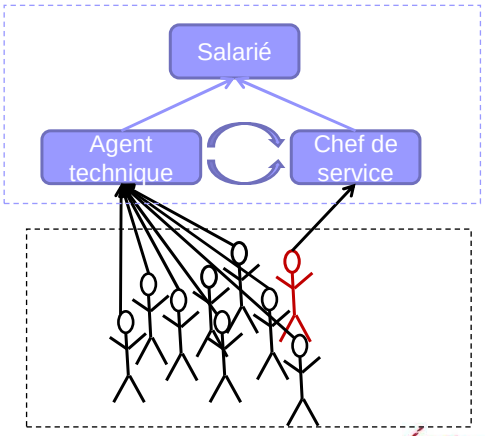
Les Bio-Ontologies

1. Introduction: définitions et enjeux des bio-ontologies
2. Bio-ontologies et annotation des contenus : recherche d'information
   - → ***Exemple BioPortal et Resource Index***
3. Bio-ontologies et intégration de données
   - → ***Exemple SO-Pharm, RDF stores***
4. Bio-ontologies et fouille de données
   - → ***Exemple Gene Ontology: similarité sémantique***
5. Conclusion: orientations de recherche actuelles

Introduction : A. Quelques définitions

- Qu'est-ce qu'une ontologie ?
  - **Vos réponses ?...**
  - **Des sens différents selon les communautés**
    - ***Philosophie : sens métaphysique défini par Aristote***
      - L'Ontologie est « La **science** de l'**être** en tant qu'être »
    - ***Sciences de l'Information et Informatique : sens informatique ou calculatoire (« computational »)***
      - Une ontologie est un artefact informatique particulier servant à modéliser la structure d'un système en utilisant des concepts et des relations (Guarino et al. Handbook on Ontologies, 2009) (→ exemple minimaliste diapo suivante)
      - « Formal, explicit specification of a shared conceptualisation » Studer 1998 (d'après Gruber 1993 et Borst 1997)
    - ***Bioinformatique : sens pragmatique***
      - Notion floue pouvant être réduite à un vocabulaire contrôlé structuré en hiérarchie de termes - chef de file GO : « Gene ontology »

Exemple minimaliste



Ontologie

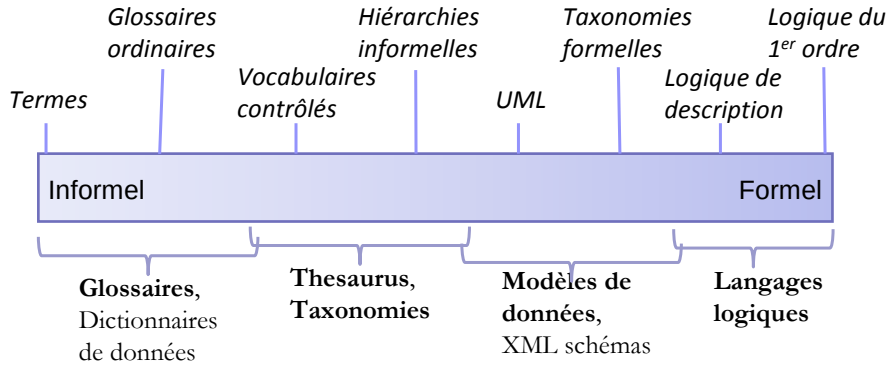
Concepts

Relations

Portion de réalité, système

Ex : service d'une entreprise

Gradualité dans la spécification formelle d'une conceptualisation



D'après Uschold M, SIGMOD Record, 2004

(1) Vocabulaire contrôlé : exemple de Glossaire

(2) Synonymes : exemple de Thesaurus MeSH: Medical Subject Headings

MeSH Descriptor Data

Return to Entry Page

Standard View. Go to Concept View; Go to Expanded Concept View

MeSH Heading	DNA Damage
Tree Number	G05.355.180
Annotation	do not use / drug eff & / rad eff for DNA damage induced by drugs or radiation; do not confuse with DNA FRAGMENTATION; see note there
Scope Note	Injuries to DNA that introduce deviations from its normal, intact structure and which may, if left unrepaired, result in a MUTATION or a block of DNA REPLICATION. These deviations may be caused by physical or chemical agents and occur by natural or unnatural, introduced circumstances. They include the introduction of illegitimate bases during replication or by deamination or other modification of bases; the loss of a base from the DNA backbone leaving an abasic site; single-strand breaks; double strand breaks; and intrastrand (PYRIMIDINE DIMERS) or interstrand crosslinking. Damage can often be repaired (DNA REPAIR). If the damage is extensive, it can induce APOPTOSIS.
Entry Term	DNA Injury
Entry Term	Genotoxic Stress
Entry Term	Injury, DNA
Entry Term	Stress, Genotoxic
See Also	Mutation
See Also	Pyrimidine Dimers

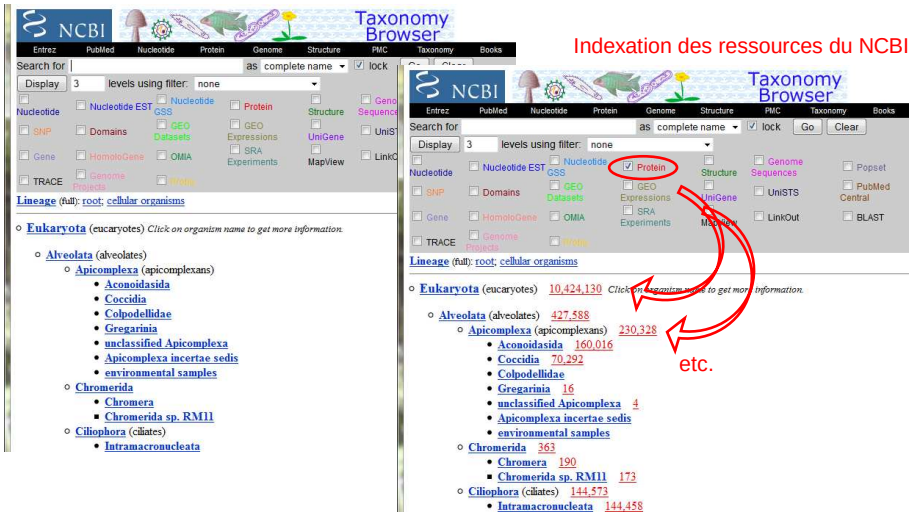
= Vocabulaire d'indexation pour MEDLINE

Hiérarchie

Synonymes

Relations

(3) Classes et hiérarchie : exemple de Taxonomie



(4) Classes et héritage: exemple de Modèle UML

UML : Unified Modelling Language

SBML : Systems Biology Markup Language

spécialisation versus généralisation

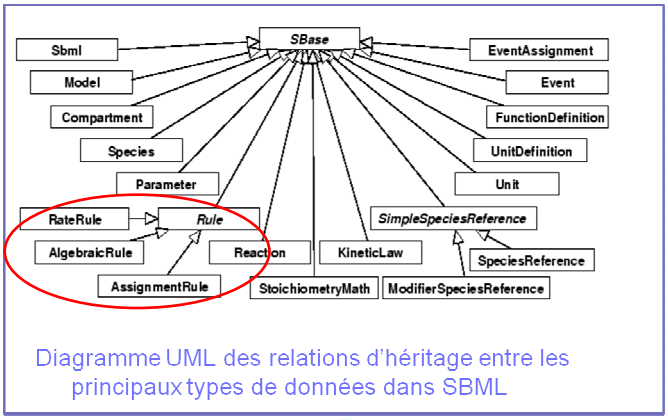
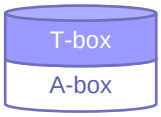


Diagramme UML des relations d'héritage entre les principaux types de données dans SBML

(4) Classes (concepts), héritage, relations et logique : les ontologies formelles



Base de connaissance = ontologie

Raisonnement sur les concepts (T-box, T comme Terminologie) :

Satisfaisabilité : un concept est satisfaisable si on peut démontrer qu'il en existe des instances

Subsomption : C $\sqsubseteq$ D si toutes les instances de C sont aussi instances de D.

Equivalence : C $\equiv$ D si C subsume D et D subsume C (C et D ont les mêmes instances

Exclusion mutuelle :(A $\sqcap$ $\neg$ A) est une proposition non satisfaisable

Inférence : prouver que C $\sqsubseteq$ D en prouvant que C $\sqcap$ $\neg$ D est insatisfaisable.

Raisonnement sur les instances (A-box, A comme Assertion) :

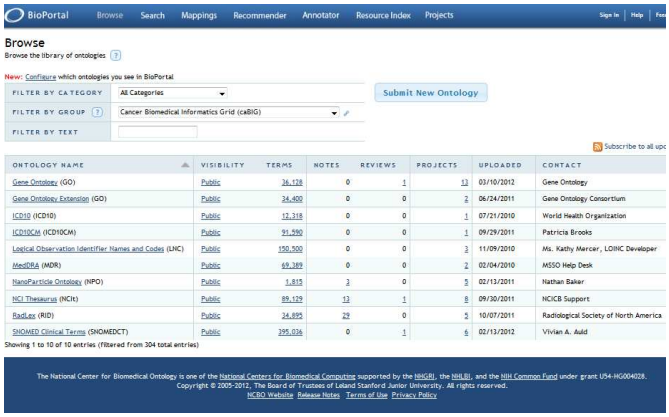
Cohérence (consistency) -> intégration de données

Validation d'instance (instance checking) -> classification … C(a), C(b), D(e), R(a, e) etc.

Technologies du web sémantique (2000, OWL 2004)

BioPortal at NCBO

http://bioportal.bioontology.org/ontologies



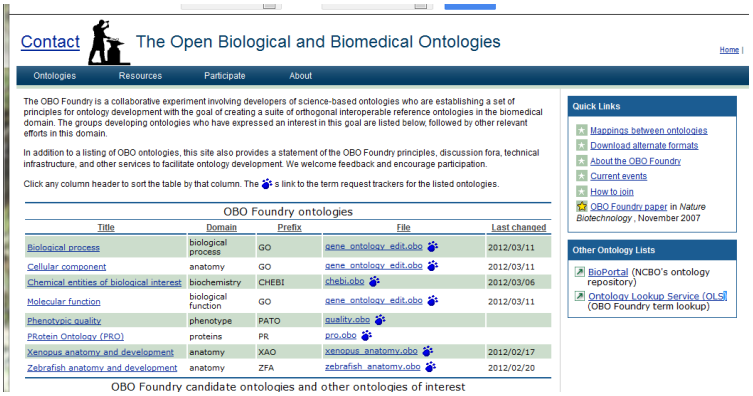
National Center for Biomedical Ontologies (Stanford)

➢300 formal bio-ontologies

➢Editeur pour les Bio-ontologies au format OWL : Protégé

OBO foundry

http://obofoundry.org/



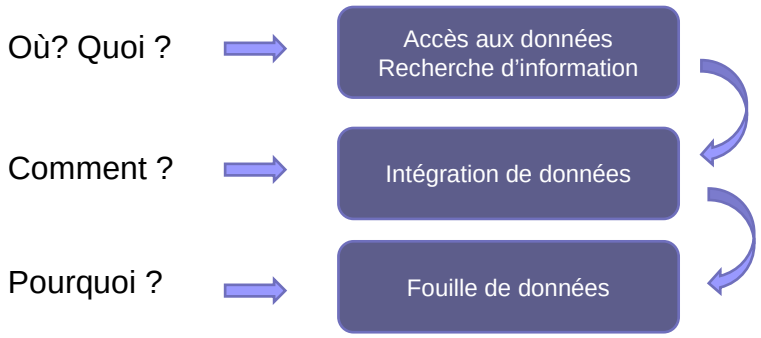
Open and Biomedical Ontologies Smith, Ashburner et al., 2007 (Berkeley)

OBO format
Editeur : OBO-edit

~ 82 ontologies

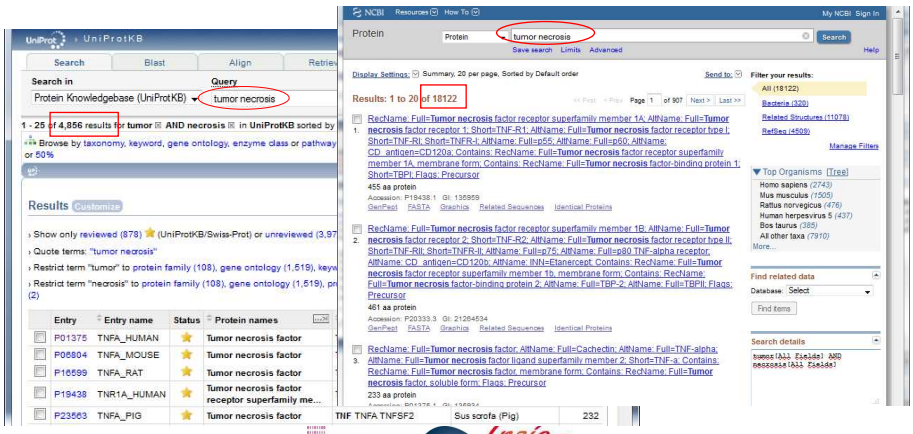
Introduction : B. Quels enjeux pour les Bio-ontologies ?

- Biologie du XXIème siècle : un déluge de données !!!



Bio-ontologies et annotation des contenus : recherche d'information

- Aller au-delà des systèmes propres à chaque ressource



L'apport des bio-ontologies pour la recherche d'information (1/3)

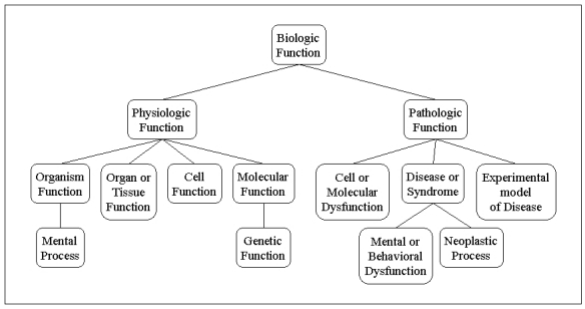
- 1. Interrogation « intelligente » des ressources
  - **Concepts - > Vocabulaire contrôlé**
    - ***Tumor necrosis factor alpha = tumor necrosis superfamily 2, etc.***
    - ***Genetic variant = genetic variation = ? Polymorphism***
    - ***Exploitation des synonymes par le MeSH pour interroger MedLine (transparent)***
  - **Relations -> Organisation hiérarchique des concepts**
    - ***Exemple MeSH***
      - Tumor necrosis factor alpha is_a_child_of tumor necrosis factors
      - Tumor necrosis factors is_a_child_of monokine, etc.
  - **➔ Utiliser les bio-ontologies pour capitaliser des connaissances et construire une interrogation intelligente des ressources**
    - ***Portail d'interrogation commun à plusieurs ontologies***
      - UMLS (1986 – aujourd'hui) : les pionniers
      - Biogateway (2009-2010) non maintenu ?
      - BioPortal (2011-2012) en évolution permanente !

UMLS : Unified Medical Language System

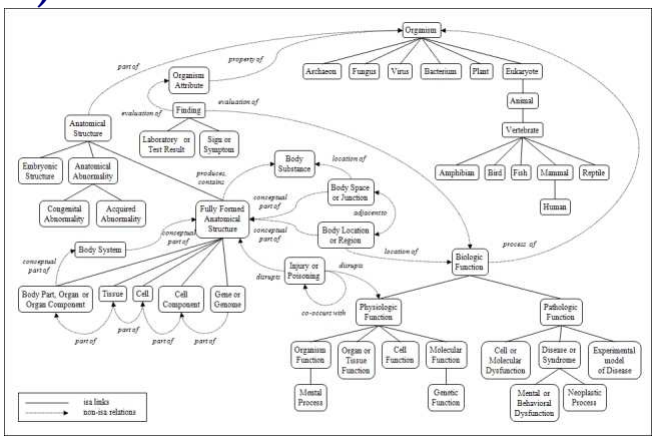
Depuis 1986 ! La référence pour les vocabulaires contrôlés biomedicaux…

- Trois outils (« knowledge sources »)
  - **MetaThesaurus : plus de 130 vocabulaires (MeSH, ICD10, SnoMed, etc.)**
  - **UMLS semantic network : types sémantiques (133) et leurs relations (54)**
    - ***Depuis 2003 : upper-level ontology***
  - **SPECIALIST Lexicon and Lexical Tools: Outils de Traitement du Langage Naturel**

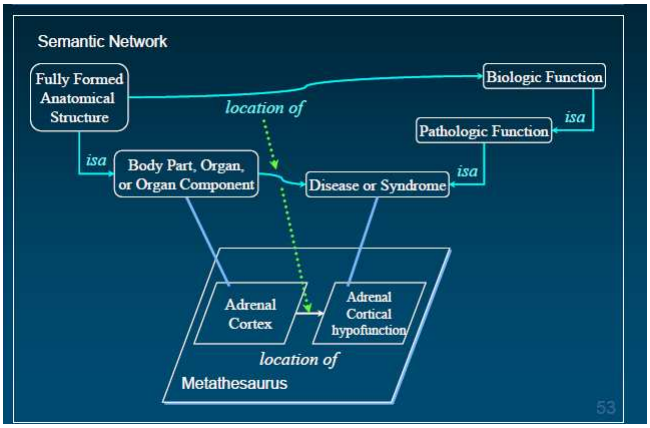
UMLS semantic network : les types sémantiques (extrait)



UMLS semantic network : les relations (extrait)



UMLS semantic network : exemple de mapping



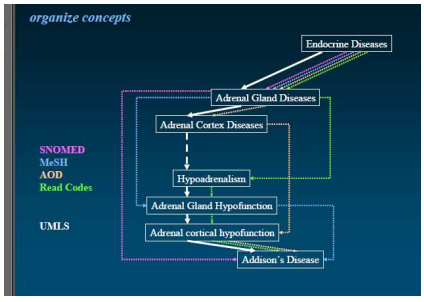
Les missions d'UMLS

- Utiliser les outils de traitement automatique des langues pour unifier les langages : trouver les synonymes, les regrouper en concepts
- Catégoriser ces concepts par type sémantique partir du "réseau sémantique"
- Incorporer les relations

et les attributs fournis par les vocabulaires

- Donner accès aux données

dans un format commun



Exemple avec la Maladie d'Addison

L'apport des bio-ontologies pour la recherche d'information (2/3)

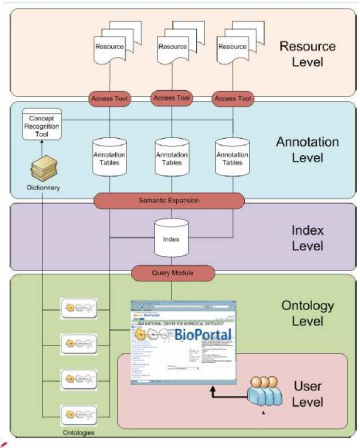
- 2. Annotation sémantique des ressources
  - **Principe : associer le (les) terme(s) les plus appropriés d'une ontologie aux différents contenus d'une ressource**
    - ***Prototype : associer des termes GO aux gènes***
    - ***Généralisation à tout type de ressource et à toutes les ontologies possibles !!***
    - ***Problème de l'automatisation***
  - **Expansion sémantique**
    - ***Fermeture transitive (« transitive closure ») à travers les relations is_a***
      - Ex : melanoma is_a melanocytic neoplasm (in NCI thesaurus)
    - ***Utilisation des « mappings » entre ontologie***
      - Ex : treatment (in MeSH) <-> therapeutic procedure (in SNOMED-CT)
    - ***Aggrégation et score***
      - Regrouper les annotations identiques (même termes dans plusieurs ontologies)
      - Tracer l'origine de l'annotation : directe versus expansion sémantique

L'apport des bio-ontologies pour la recherche d'information (3/3)

- 3. Interrogation des ressources
  - **Langage d'interrogation particulier (web sémantique)**
    - ***Forrmalisme des ontologies: OWL (« OntologyWeb Language »)***
    - ***Descriptions des ressources : RDF (« Resource Description Framework »)***
    - ***Langage d'interrogation : SPARQL (« Simple Protocol and Resource Query Language »)***
  - **Interface web « user-friendly »**
    - ***BioGateway***
    - ***BioPortal***

NCBO resource index: ontology-based search and mining of biomedical resources

- By: Clement Jonquet, Paea LePendu, Sean Falconer, Adrien Coulet, Nalatya F Noy, Mark A Musen and Nigam H Shah, 2011, Web semantics : Science, Services,and Agents on the World Wide Web 9, 316-324
  - **Stanford Center for Biomedical Informatics Research, LIRMM and LORIA**
  - **NCBO : National Center for Biomedical Ontologies**

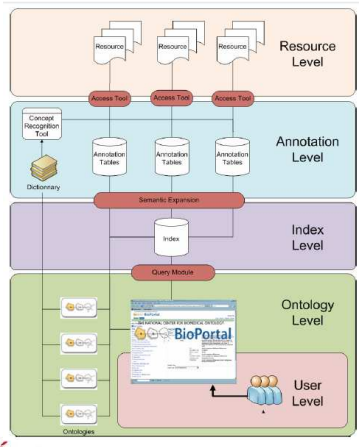


NCBO resource index: ontology-based search and mining of biomedical resources

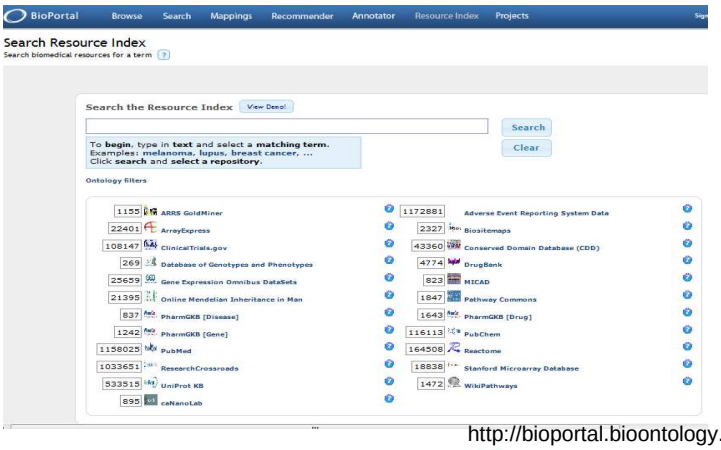
23 ressources annotées : UniProt, GO, ArrayExpress, GEO, PharmGKB, etc. soit environ 4,4 millions d'entrées

>14,6 milliards d'annotations après expansion sémantique (environ 2 millions d'annotations directes)

> 300 ontologies du BioPortail NCBO : GO, NCI thesaurus, ICD10, etc. , soit environ 5,8 millions de concepts d'ontologie



NCBO resource index: ontology-based search and mining of biomedical resources



➔Demo

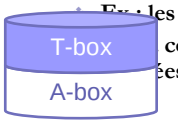
http://bioportal.bioontology.org/resources

Les Bio-Ontologies

1. Introduction: définitions et enjeux des bio-ontologies
2. Bio-ontologies et annotation des contenus : recherche d'information
   - ➔ *Exemple BioPortal et Resource Index*
3. Bio-ontologies et intégration de données
   - ➔ *Exemple SO-Pharm, RDF store*
4. Bio-ontologies et fouille de données
   - ➔ *Exemple Gene Ontology: similarité sémantique*
5. Conclusion: orientations de recherche actuelles

Des bases de données intégrées aux bases de connaissances

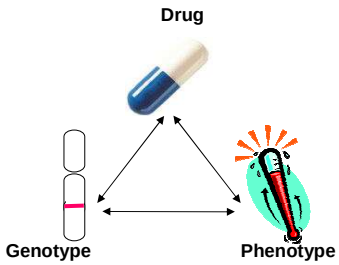
- Pour les biologistes les bases de connaissance sont en fait des bases de données intégrées
  - **Ex: Uniprot KB, Kegg, OMIM, IMAGE, PharmGKB, etc.**
  - **Dans une BD, la connaissance est présente au niveau du modèle de données**
  - **Pas d'utilisation par des programmes pour raisonner**
- Pour les informaticiens, les bases de connaissances sont des systèmes dans lesquels les données sont associées à des connaissances explicites et formelles qui peuvent être utilisées par des programmes
  - **Ex : les Ontologies en Logique de Description ou OWL (cf introduction)**
  - **connaissance peut être utilisée pour raisonner (cohérence des ées, validation de nouvelles instances etc.)**



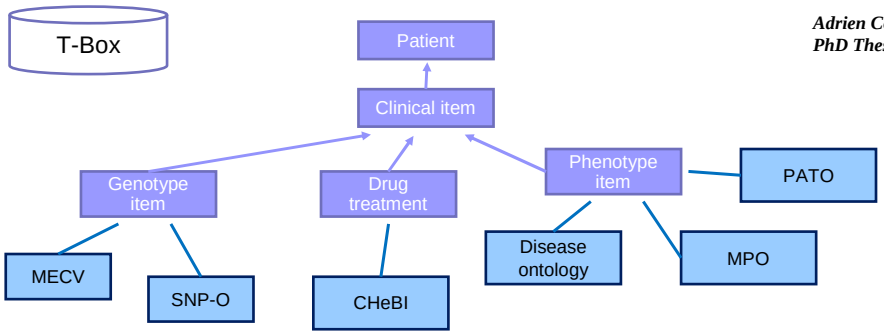
An example in pharmacogenomics (1)

- Goal of pharmacogenomics
  - **Identify individual genome variations (Genotype)**
  - **… that influence adverse reaction (Phenotype)**
  - **… to drug treatment (Drug)**
- GenNet Project
  - **KIKA medical + Phenosystems + LORIA / Orpailleur**
- Example: SNP variants in geneCYP2D6 (Desmeules et al., 1991)
  - ***More or less active forms of a given enzyme***
  - ***Fast or slow transformation of codein into morphin***
  - ***Intoxication or absence of reaction to a given treatment***

Adrien Coulet PhD Thesis



An example in pharmacogenomics (2)

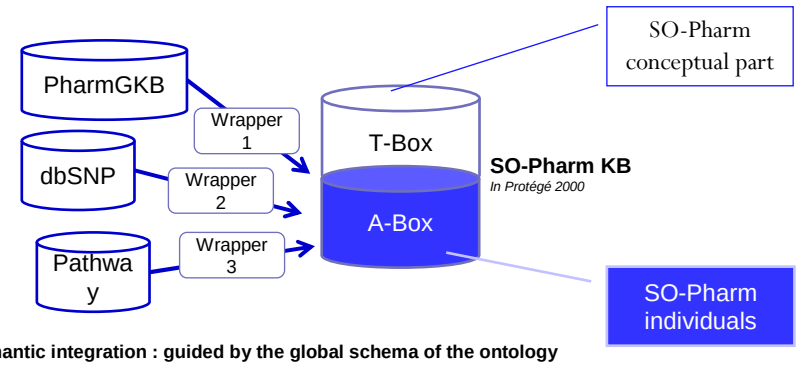


Adrien Coulet PhD Thesis

Articulation of existing ontologies (15) covering various biological domains

MEO : Mutation Event Controlled Vocabulary ; SNP-O : Single Nucleotide Polymorphism Ontol. ; CHeBI : Chemical Entities of Biological Interest ; MPO :Mammalian Phenotype Ontol., PATO : Phenotype and Trait Ontology

An example in pharmacogenomics (3)



Semantic integration : guided by the global schema of the ontology

Set of mappings between each data source and the ontology (*Poggi et al., 2008 ; Coulet PhD Thesis, 2008)*

Advantages : Consistency, lack of redundancy, new properties inferred by reasoners

Integration of a PharmGKB clinical trial in SO-Pharm KB

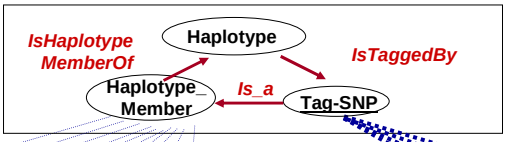
- Diversity of responses to Montelukast (Singulair)
  - Lima et al., 2006 published a study about maintenance treatment of asthma
  - Set of 61 patients, genotyped on 26 SNPs localized on 5 different genes (Leukotriene pathway)

A-Box

- Definition of mapping relations = populating the A-box
  - ➔ 61 assertions of the concept Patient e.g. Patient(pa01)
  - ➔ 162 assertions of the concept Clinical item and subconcepts e.g. ClinicalItem(exa:yes)
  - ➔ many assertions of various roles between the concepts e.g. HasClinicalItem(pa01, exa:yes)
  - + Integration of data from external databases (dbSNP, KEGG pathways)

Example of use: Attribute selection guided by an ontology

SNP-Ontology



Pa	HCF	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10	…
01	+	AA	AC	GG	GC	TA	GG	CA	AA	AC	TA	…
02	+	AA	AC	GG	GT	CA	GG	CA	AT	AC	TA	…
03	-	AT	AT	CG	GC	TA	GG	TA	AT	AC	AA	…
…												

Pa	HCF	SNP2	SNP4	SNP8	SNP9	…
01	+	AC	GC	AA	AC	…
02	+	AC	GT	AT	AC	…
03	-	AT	GC	AT	AC	…
…						

Complete dataset:
125 patients, 289 SNPs
➔ > 6900 frequent itemsets

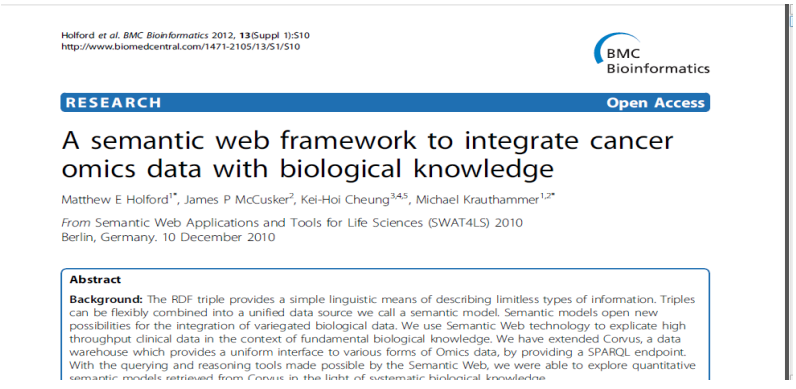
Reduced dataset :
125 patients, 198 SNPs
➔ ~ 300 frequent itemsets

(Coulet et al., BMC Bioinformatics 2008)

Généralisation

- Encore peu d'exemples d'utilisation des ontologies comme bases de connaissances
  - **Lourdeur des technologies du web sémantique (par rapport aux SGBDR)**
  - **Difficultés à gérer de grands volumes de données**
- Dévelopement d'entrepôts de tripletsRDF
  - **Projet Bio2RDF : convertir toutes les données au format RDF (resource description framework)**
    - ***Application aux sciences du vivant du projet Linked Data pour le web***
    - ***Michel Dumontier, Ottawa Carleton University***

Applications ciblées



Holford *et al. BMC Bioinformatics* 2012, **13**(Suppl 1):S10
http://www.biomedcentral.com/1471-2105/13/S1/S10

BMC Bioinformatics

RESEARCH Open Access

A semantic web framework to integrate cancer omics data with biological knowledge

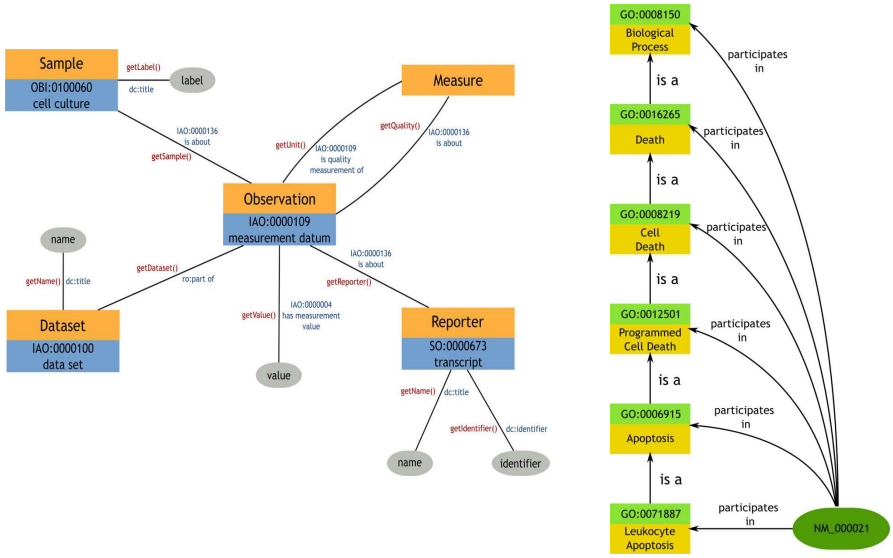
Matthew E Holford[1*], James P McCusker[2], Kei-Hoi Cheung[3,4,5], Michael Krauthammer[1,2*]

From Semantic Web Applications and Tools for Life Sciences (SWAT4LS) 2010
Berlin, Germany. 10 December 2010

Abstract

Background: The RDF triple provides a simple linguistic means of describing limitless types of information. Triples can be flexibly combined into a unified data source we call a semantic model. Semantic models open new possibilities for the integration of variegated biological data. We use Semantic Web technology to explicate high throughput clinical data in the context of fundamental biological knowledge. We have extended Corvus, a data warehouse which provides a uniform interface to various forms of Omics data, by providing a SPARQL endpoint. With the querying and reasoning tools made possible by the Semantic Web, we were able to explore quantitative

Intégration de données d'expression et de methylation pour 7 lignées de mélanome avec les annotations GO pour tout le génome humain, les réseaux de gènes et les gènes cibles des facteurs de transcription



RDF : Resource Description Framework

- RDF triple : (Subject, Property, object)

	Subject	Property	Object
Ex:	Sequence NM_00021	Participates_in	Leucocyte apoptosis GO:0071887
	(URN Refseq)	(URN property)	(URN GO)

- URN : Universal Resource Name, LSID : LifeScience Identifier

urn:lsid :adresseWebResponsableBD:nomBD:identifiant_dans_BD

- Représentation sous forme de graphes interrogrables par SPARQL

Proof of concept (Holford et al., 2012)

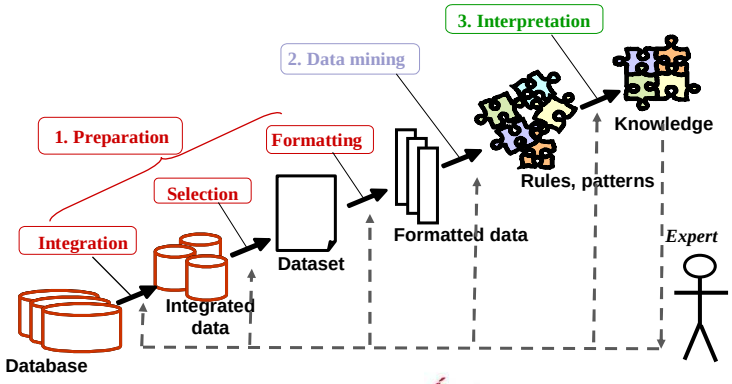
- We were able to generate a testable hypothesis to explain how Decitabine fights cancer – namely that it targets apoptosis-related ene promoters predominantly in Decitabine-sensitive cell lines, thus conveying its cytotoxic effect by activating the apoptosis pathway.
- Our research provides a framework whereby similar hypotheses can be developed easily

Les Bio-Ontologies

1. Introduction: définitions et enjeux des bio-ontologies
2. Bio-ontologies et annotation des contenus : recherche d'information
   → *Exemple BioPortal et Resource Index*
3. Bio-ontologies et intégration de données
   → *Exemple SO-Pharm, RDF stores*
4. Bio-ontologies et fouille de données
   → *Exemple Gene Ontology: similarité sémantique*
5. Conclusion: orientations de recherche actuelles

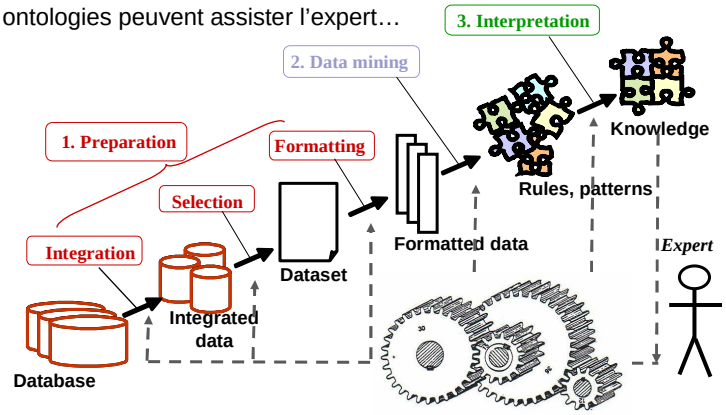
Fouille de données et connaissances

- Fouille de données : recherche de régularités dans les données
  - Etape au cœur du processus d'extraction de connaissances



Knowledge Discovery guided by Domain Knowledge : KDDK

Des ontologies peuvent assister l'expert…



… à chaque étape du processus.

Classification sémantique (1)

- Classer est une façon de fouiller les données
  - **Classification supervisée : recherche les règles qui conduisent les objets à appartenir à telle ou telle classe, à la base des systèmes de prédiction.**
  - **Classification non supervisée : recherche à identifier des sous-goupes d'objets similaires dans un ensemble d'objets (« clustering »), puis à les interpréter.**
- Nombreuses méthodes de classification non supervisée
  - **Classification hiérarchique ascendante (heatmaps d'Eisen pour les données d'expression)**
  - **Méthode des K-means avec K, nombre de cluster, à optimiser**
  - **Partitions exactes ou floues**

Classification sémantique (2)

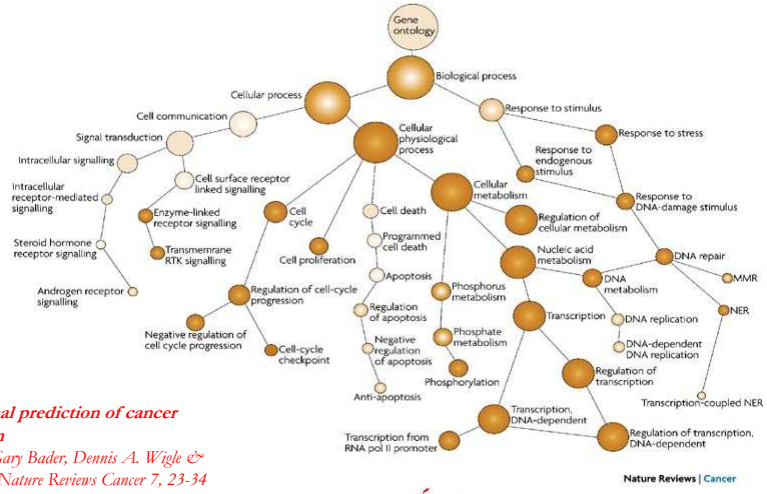
- Nombreuses mesures de similarité ou de distance
  - **Objets décrits par les valeurs prises par des descripteurs : dimensions du jeu de données**
  - **Distances entre ces objets (exemple distance euclidienne)**
  - **Considère les dimensions indépendantes**
- Mesure de similarité sémantique :
  - **Pour tenir compte des relations qui peuvent exister entre les descripteurs**
  - **Notamment lorsque ces descripteurs sont les termes d'une ontologie**
  - **Le cas des annotations GO**

Pesquita et al., 2009: Semantic similarity in biomedical ontologies, PLOS Comp. Biol. July 2009, Volume 5 | Issue 7 | e1000443

Gene Ontology (1)

- Historique rapide
  - **98 : Consortium pour l'annotation des génomes modèles (souris, drosophile, levure)**
  - **Vocabulaire contrôlé et relations (is_a, part_of , regulates)**
  - **Graphe Acyclique Dirigé (DAG) : plus d'1 parent par terme**
  - **Trois aspects : biological process, molecular function, cellular component**
- Aujourd'hui, statistiques :
  - **> 20 laboratoires participants (GOA : ouvert à tous les génomes)**
  - **> 30 000 termes**
  - **Relation is_a complète**
    - ***chaque terme a un chemin is_a complet jusqu'à la racine***
  - **>160 millions d'annotations**
  - **Notion de code d'évidence (Exp, Comp,…)**

Gene Ontology (2) : extrait (BP, cancer)

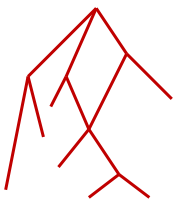


Computational prediction of cancer gene function
Pingzhao Hu, Gary Bader, Dennis A. Wigle & Andrew Emili. Nature Reviews Cancer 7, 23-34 (January 2007)

Gene Ontology (3)

- Disponibilité
  - Termes et hiérarchies AmiGO, myGO database
  - Annotations GOA, gene2GO (NCBI)
- Versions bonnes pratiques
  - GONG (GO next generation) -> version OWL cohérente
  - Traduction OWL (BioPortal), OBO (OBO Foundry), RDF

Mesures de similarité fonctionnelle entre gènes



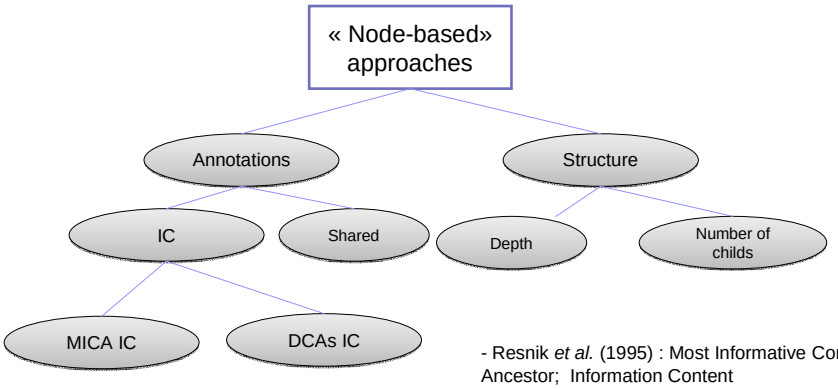
	GO-t1	GO-t2	GO-t3	…
Gene1	X	X	O	…
Gene2	X	O	X	…
…				

Deux niveaux de calcul

(i): Similarité des termes dans le graphe GO = similarité sémantique

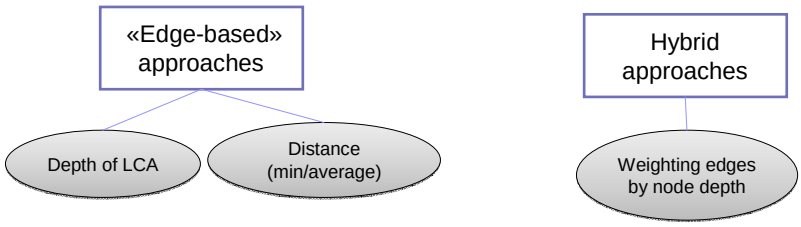
(ii): Similarité des objets (gènes) annotés par les termes = similarité fonctionnelle

Similarité sémantique « terme-terme »(1)



- Resnik *et al.* (1995) : Most Informative Common Ancestor; Information Content
- Bodenreider *et al.* (2005) : Shared annotations

Similarité sémantique « terme-terme » (2)

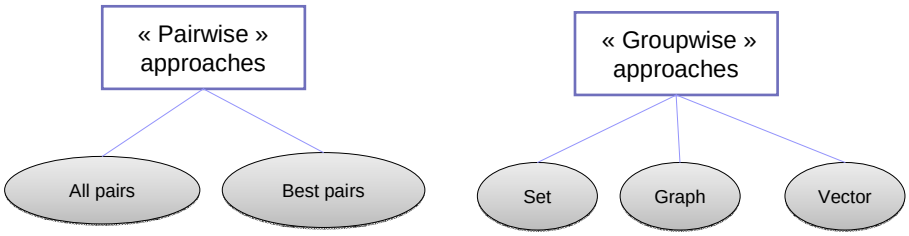


-Wu *et al.* (2006) : Depth of LCA : Lowest Common Ancestor, Shortest Path Length
- Pozo *et al.* (2008) : Depth of LCA

- Othman *et al.* (2007) : IC/Depth/number of children; Distance

Similarité fonctionnelle « gène-gène »

- Diverses façon d'agréger les similarités terme-terme



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

-Martin *et al.* (2004) : Graph/Jaccard on term lists enriched with term ancestors

-Chabalier et al. (2007) : Vectors compared with the cosine measure

>>>>http://web.cbio.uct.ac.za/ITGOM

IntelliGO: modèle vectoriel et cosinus généralisé

Benabderrahmane S. et al.(2010) BMC Bioinformatics 11:588.

- Representation of genes in a vector space model

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

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

α_i : Coefficient for feature 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

Le principe du cosinus généralisé

Ganesan P, Garcia-Molina H, Widom J (2003) Exploiting hierarchical domain structure to compute similarity. Transactions on Information Systems, 21 : 64 - 93

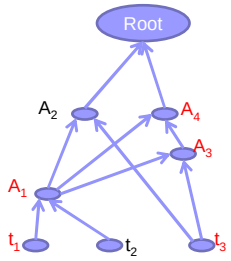
- Method proposed for « tree »-hierarchies of terms (MeSH) in document retrieval
- Principle: consider that the dimensions of the vector space are not orthogonal to each other
- Consequence in dot product:

$$\vec{e_i} \cdot \vec{e_i} = 1 \quad \text{And} \quad \forall i, i \neq j, \quad \vec{e_i} \cdot \vec{e_j} \neq 0$$

$$\vec{e_i} \cdot \vec{e_j} = \frac{2 \times \text{Depth}[\text{LCA}(t_i, t_j)]}{\text{Depth}(t_i) + \text{Depth}(t_j)}$$

Adaptation de la similarité terme-terme au DAG

- GO is a rDAG (rooted Directed Acyclic Graph)
- In a rDAG, each term can have several parents and therefore several paths to the Root
- Consequence: LCA is not unique, Depth (t_i) is not unique



$$\vec{e_i} \cdot \vec{e_j} = \text{Sim}_{IntelliGO}(t_i, t_j) = \frac{2 \times \text{MaxDepth}[\text{LCA}(t_i, t_j)]}{\text{SPL}(t_i, t_j) + 2 \times \text{MaxDepth}[\text{LCA}(t_i, t_j)]}$$

Similarity fonctionnelle IntelliGO

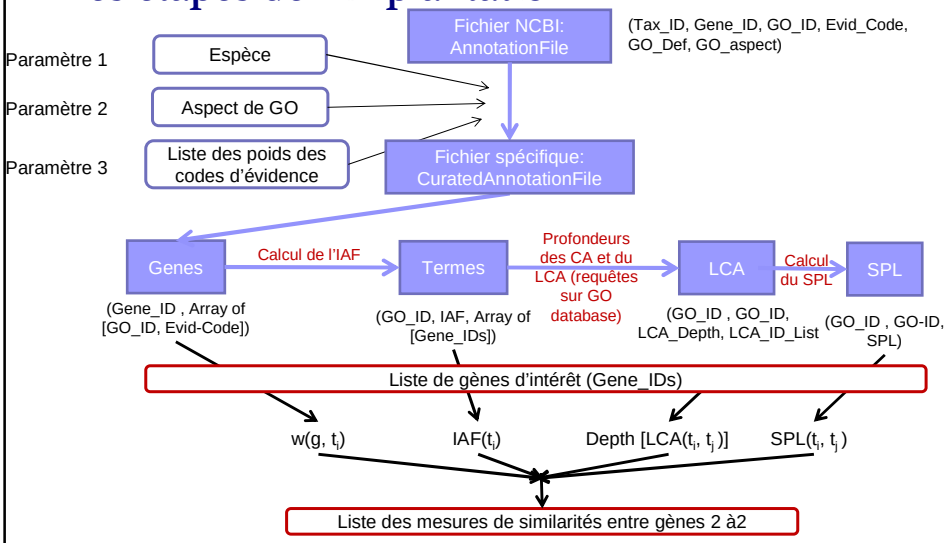
- Generalized dot-product between two gene vectors

$$\vec{g} \cdot \vec{h} = \sum_{i,j} \alpha_i \times \beta_j \times \vec{e_i} \cdot \vec{e_j} \qquad \text{avec } \vec{e_i} \cdot \vec{e_j} \neq 0, \ \forall i, i \neq j$$

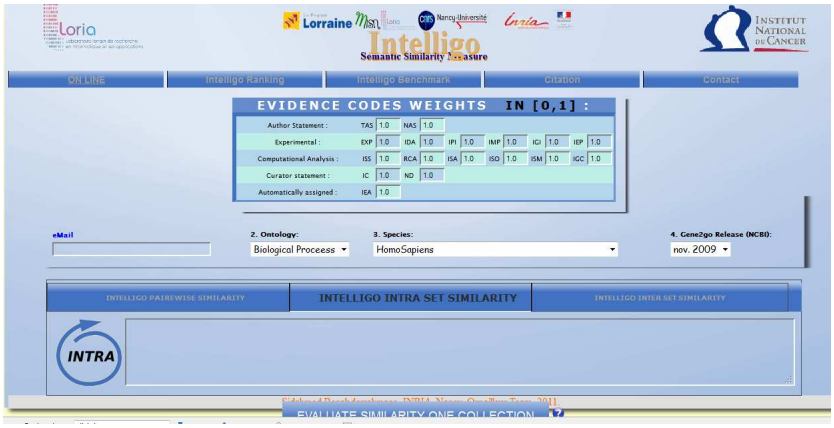
- Generalized cosine similarity

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

Les étapes de l'implantation



Mise à disposition sur la plateforme MBI



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

Validation sur des jeux de données témoins

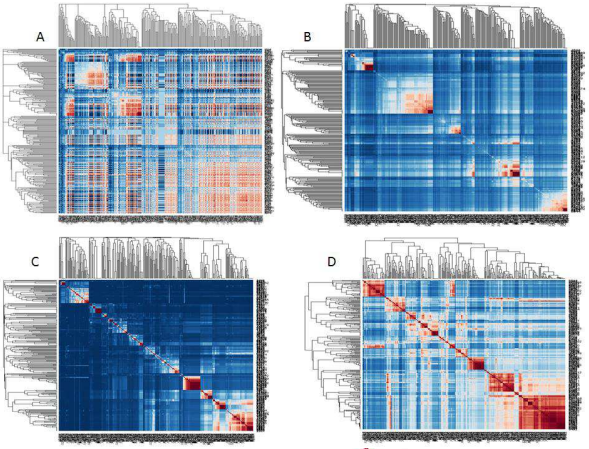
Dataset	Species	Source	Number of sets	Total genes
1	Human	KEGG pathways	13	275
2	Yeast	KEGG pathways	13	169
3	Human	Pfam Clans	10	94
4	Yeast	Pfam Clans	10	118

- For each dataset
  - **Calculate pair-wise gene-gene similarities**
  - **Apply hierarchical clustering : heatmap**
  - **Or Fuzzy C-means clustering : F-score**

Comparaison avec 3 autres mesures

- Visualisation heatmap d'un clustering hiérarchique

Lord et al. (normalized)



Al-Mubaid

SIM_{GIC}

IntelliGO

Comparaison avec l'outil de classification DAVID (1)

- Outil en ligne de classification fonctionnelle des gènes
  - **DAVID : Database for Annotation Visualisation and Integrated Discovery**

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

Similarity measure based on counting present and absent features:
measured by ***Kappa statistics***
=>No Semantics

Comparison avec l'outil de classification DAVID (2)

- Fuzzy C-means clustering : optimal F-score and K number

	IntelliGO		DAVID		
Dataset (Nber of sets)	Optimal global F-score	Optimal K number	Optimal global F-score	Optimal K number	Excluded genes
1 (13)	0.62	14	0.67	10	21%
2 (13)	0.67	14	0.68	9	18 %
3 (10)	0.75	11	0.64	11	27 %
4 (10)	0.82	11	0.70	10	41 %

>>> Functional classification is reliable and robust with IntelliGO measure

Benabderrahmane et al., BIBM workshop IDASB 2011

Conclusion : recherches actuelles (1)

- Interrogation intelligente et transversale grâce à l'annotation sémantique de ressources et documents
  - **Pour la construction de nouvelles ontologies**
- Intégration de données guidée par les connaissances du domaine
  - **Problème du volume des données : développements technologiques nécessaires**
- Fouille de données et extraction de connaissances
  - **Classification fonctionnelle plus performante, sélection d'attributs, réduction de dimensions, etc.**

Conclusion : recherches actuelles (2)

- Sciences du vivant : champ d'application privilégié des technologies du web sémantique
  - **Nombreuses ontologies formelles OBO Foundry, BioPortal**
  - **Enjeu majeur de l'exploitation des masses de données biologiques**

Quelques ouvrages et articles

- Staab S and Studer R (eds) **Handbook on Ontologies**. *International Handbooks on Information Systems*, DOI 10.1007/978-3-540-92673-3. Springer Verlag, Berlin Heidelberg, 2009.
- Clement Jonquet, Paea LePendu, Sean Falconer, Adrien Coulet, Nalatya F Noy, Mark A Musen and Nigam H Shah (2011) **NCBO Resource index:ontology-based search and mining of biomedical resources**. *Web semantics : Science, Services,and Agents on the World Wide Web* 9, 316-324.
- Antezana E, Blondé W, Egana M, Rutherford A, Stevens R, DeBaets B, Mironov V and Kuiper M (2009) **BioGateway: a semantic systems biology tool for the life sciences**. *BMC Bioinformatics* 10 : S11.
- Coulet A, Smail-Tabbone M, Napoli A, and Devignes MD (2010) **Ontology-Based Knowledge Discovery in Pharmacogenomics**. *Advances in Computational Biology*, book series Advances in Experimental Medicine and Biology, AEMB, Springer
- Holford ME, McCusker JP, Cheung KH and Krauthammer M (2012) **A semantic web framework to integrate cancer omics data with biological knowledge**. *BMC Bioinformatics* 13, S10
- Mironov V, Seethappan N, Blondé W, Antezana E, Splendiani A and Kuiper M (2012) **Gauging triple stores with actual biological data**. *BMC Bioinformatics* 13, S3.
- Benabderrahmane S., Smail-Tabbone M, Poch O., Napoli A. and Devignes MD (2010) **IntelliGO: a new vector-based semantic similarity measure including annotation origin.** *BMC Bioinformatics* 11:588.
- Bresso E, Benabderrahmane S., Smail-Tabbone M, Marchetti G, Karaboga AS, Souchet M, Napoli A. and Devignes MD . **Use of domain knowledge for dimension reduction. Application to mining of drug side effects.** 4th *International Conference on Knowledge Discovery and Information Retrieval* (KDIR'2011), Paris 24-28 oct 2011.

Participants

LORIA, Equipe Orpailleur Nancy

MD Devignes

Malika Smaïl-Tabbone

Adrien Coulet

Sidahmed Benabderrahmane

Jean-François Kneib

Amedeo Napoli

KIKA medical

Hôpital Saint Antoine Paris

Pascale Benlian (MD)

PhenoSystems

David Atlan

Harmonic Pharma

Michel Souchet

Emmanuel Bresso

Financements

Projet Eureka GenNet

Communauté Urbaine du Grand Nancy

Contrat Plan Etat Région : MISN

INCa (bourse de thèse interdisciplinaire)

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