

Entrepôts de données personnalisés pour l'identification de gènes candidats - Construction de jeux de données intégrés, guidée par les connaissances du domaine

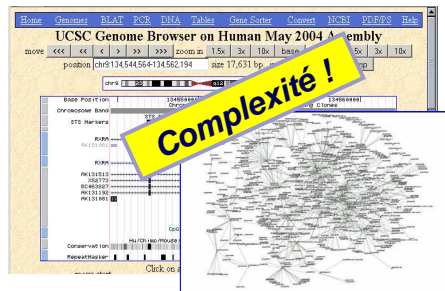
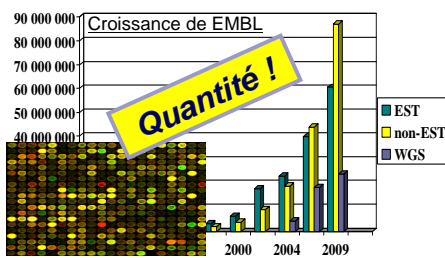
Marie-Dominique Devignes

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

Equipe Orpailleur

Nancy

Intégration de données hétérogènes en biologie



Paradoxe : Trop d'info tue l'info

Lille, 4 mai 2010

Multiplicité des bases de données biologiques

- Bref historique
 - 1965 : Atlas of Protein Sequences and Structures : Margaret Dayhoff
 - > → *PIR en 1986 (Université Georgetown)*
 - 1980's : EMBL/GenBank/DDBJ ; SwissProt
 - 2K's : Compilations Nucleic Acids Research (Janvier)
 - > *Janv 2003 : > 300 BD*
 - > *Janv 2010 : > 1200 BD*
- Problématique de l'intégration de ces bases de données
 - ❖ Dès 1994-1995 le problème est soulevé
 - > *P Karp, D Markowitz, S Davidson*
 - > *L. Stein, Nature 2002 - > 2006 « A bioinformatic nation » « Cyberinfrastructure »*
 - > *R Altman 2005 Concept de « Ressourcome »*
 - ❖ Workshops: DILS, International Symposium on Integrative Bioinformatics, etc.

Lille, 4 mai 2010



3/55

Motivation: la recherche « in silico »

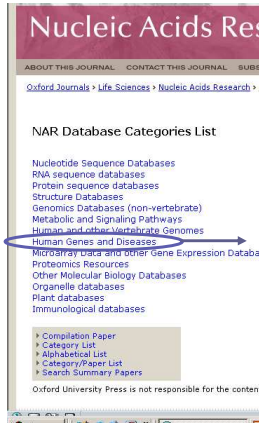
- Double problème :
 - ❖ **Pb1** : Choix des bases de données (BD) à interroger/explorer
 - ❖ **Pb2** : Intégration des données collectées
- Exemple :
 - « Parmi les 50 gènes de ce locus remanié sur le chromosome X, lesquels sont exprimés dans le cerveau et annotés par des termes GO en rapport avec la migration neuronale »
 - ❖ **Pb1** : Recherche de BDs
 - > *Sur l'expression des gènes*
 - > *Sur l'annotation des gènes*
 - > *Sur les gènes du chromosome X ou leurs orthologues dans d'autres espèces*
 - ❖ **Pb2** : Intégration des données collectées
 - > *Mise en relation des données provenant de BD distinctes*
 - > *Détection possible d'incohérences*

Lille, 4 mai 2010



4/55

Pb1 : Choix des bases de données



Human Genes and Diseases

- Protein Mutant Database
- General human genetics databases
 - BodyFams3D
 - Comparative Toxicogenomics Database
 - DC-CST
 - GenAtlas
 - GeneCards
 - Genetics Home Reference
 - HAGR - Human Ageing Genomics
 - HCAD - Human Chromosome A
 - HERVd - Human Endogenous R
 - HGNC Database
 - Human PAML Browser
 - MST Breakpoint Mapper
 - MuDB
 - OMIM - Online Mendelian Inher
 - SNP2MMD
- General polymorphism databases
 - ALFRED
 - CTGA
 - Cyriopt national mutation datab
 - D-HaploDB
 - Database of Genomic Variants
 - dbSNP
 - dBRIP
 - F-SNP
 - FINDBase
 - HapMap Project
 - HGVbase
 - HGVs Databases
 - HuRef
 - Hvrbase++
 - IPD - Immuno Polymorphism D
 - OSNP
 - OMIM - Online Mendelian Inher
 - PharmGKB
 - PhenomicDB
 - PolyDoms
 - Polymorphix
 - rSNP Guide
 - SNAP
 - SNP2MMD
 - SNP@Ethnos
 - SNPeffect
 - SNPlogit
 - TopoSNP
 - TPMD - Taiwan polymorphic mi
 - VarySysDB
 - YH database

Cancer gene databases

- Atlas of Genetics and Cytogenetics in Oncology and Haematology
- Cancer Chromosomes
- CancerGenes
- CanGEM
- CGED - Cancer Gene Expression Database
- ChimerDB
- COSMIC - Catalogue Of Somatic Mutations In Cancer
- CTDdatabase
- Database of Germline p53 Mutations
- DOOC
- EHC0
- HPTA
- Human p53, human hprt, rodent lacI and rodent lacZ databases
- IARC TP53 Database
- ITTACA
- MethyCancer
- MarkCa
- Mouse Tumor Biology Database
- OncoDB_HCC
- PubMeth
- RTCGD - Retroviral Tagged Cancer Gene Database
- SNP500Cancer
- SV40 Large T-Antigen Mutant Database
- Tumor Associated Gene database
- Tumor Gene Family Databases (TGDs)

Gene-, system- or disease-specific databases

- ALPbase
- AlzGene
- Androgen Receptor Gene Mutations Database
- AutDB
- Atlas of Genetics and Cytogenetics in Oncology and Haematology
- BGED - Brain Gene Expression Database
- BTKBase
- CarpeDB
- CASRD
- Collagen Mutation Database
- Cytokine Gene Polymorphisms
- dBER
- DNARepliation.net
- ELCO DB - Expression-based Imprint Candidate Organiser
- EndoNet
- EPCoNDB
- EpiDB - Erythropoiesis Database
- ERGD - Estrogen Responsive Genes Database
- ERGR
- EyeSite
- FUNPFP
- GCDB
- GOLD.db - Genomics Of Lipid-associated Disorders
- HaemB
- HbVar
- HDBase

Lille, 4 mai 2010

5/55

Exemple d'annuaire interrogeable en ligne

Les BD du catalogue NAR sont indexées automatiquement par des métadonnées « Subject » issues du vocabulaire MeSH

The screenshot shows the BioRegistry website. The main page has a navigation bar with 'HOME', 'HELP', 'STATISTICS', and 'ADMINISTRATION'. Below the navigation bar, there is a section titled 'Please choose one querying mode:' with five options: 1. Query by Subject, 2. Query by Keyword, 3. Query by Text, 4. Query by Category, and 5. Query by Name or Identifier. A blue arrow points to '1. Query by Subject'. Below this, there is a 'Choosing Subject' section with a list of subjects and a 'Query Subjects' section with a search form. The search form has a 'Query options' section with 'Boolean search' and 'AND' selected, and a 'Full display' checkbox checked. A 'Search' button is at the bottom right.

Devignes MD, Franiatte P, Messai N, Napoli A et Smail-Tabbone M (2010) *BioRegistry : automatic extraction of metadata for biological database retrieval and discovery. International Journal on Metadata, Semantics and Ontologies (sous presse).*

Lille, 4 mai 2010



6/55

Pb2 : Intégration de données hétérogènes

Entrez-Gene NCBI
Le gène MAGED1

MétaDonnées

UniProt
La protéine MAGD1_HUMAN

MétaDonnées

Correspondances / Mappings

Lille, 4 mai 2010

Nancy-Université

7/55

Pb2 : Intégration de données

- Différents niveaux
 - ❖ Données
 - ❖ MétaDonnées
 - ❖ Correspondances / Mapping
 - ❖ Références croisées

- Rôle des connaissances
 - ❖ Comparaison des données
 - Exemple des annotations fonctionnelles : rôle des vocabulaires, mesures de similarité sémantiques
 - ❖ Signification des MétaDonnées
 - Schéma, Indexation
 - ❖ Correspondances / Mappings
 - Connaissances de l'expert

PLAN DE L'EXPOSÉ

- I. Les différentes stratégies pour l'intégration de données
- II. ACGR : Un entrepôt de données personnalisé pour l'identification de gènes candidats
- III. MODIM : Généralisation à la construction de jeux de données pour la fouille de données

I. Stratégies

II. ACGR

III. MODIM

Lille, 4 mai 2010



9/55

PLAN DE L'EXPOSÉ

- I. Les différentes stratégies pour l'intégration de données
 1. **Essai de classification**
 2. **Avantages-Inconvénients de quelques systèmes fonctionnels**
 - *Systeme d'interface unifiée*
 - *Base de données intégrée*
 - *Workflows et web services*
 3. **Un système « hybride : l'entrepôt personnalisé**
- II. ACGR : Un entrepôt de données personnalisé pour l'identification de gènes candidats
- III. MODIM : Généralisation à la construction de jeux de données pour la fouille de données

I. Stratégies

II. ACGR

III. MODIM

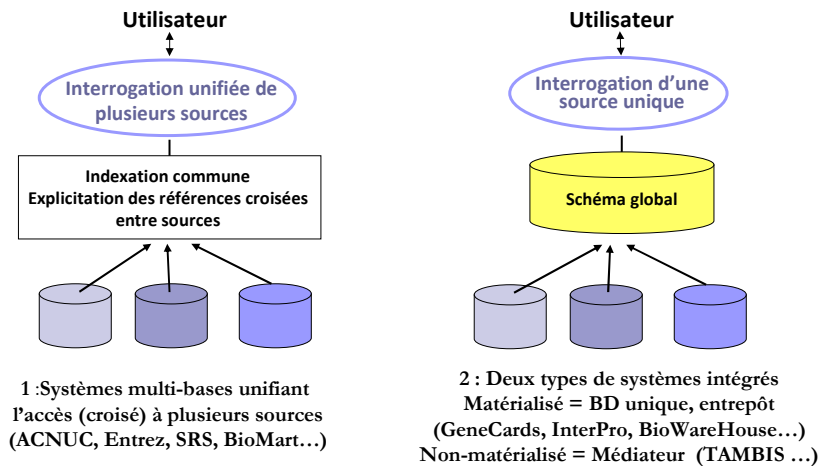
Lille, 4 mai 2010



10/56

1. Essai de classification (1)

Avec ou sans schéma intégrateur



Lille, 4 mai 2010

11/55

1. Essai de classification (2)

Avec flots de traitements et/ou visualisation

- Enchaînement d'outils et d'accès aux bases de données
 - ❖ Accès standardisés par la technologie des services web
 - ❖ Notion de workflow
 - ❖ Exemple Taverna (UK MyGRID)
- Plateforme d'intégration et de visualisation
 - ❖ Exemple ONDEX (représentation à base de graphes)

Lille, 4 mai 2010

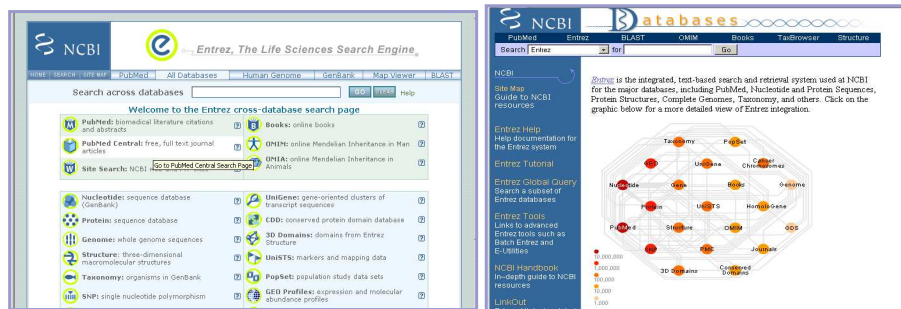
12/55

2. Avantages-Inconvénients

2.1 Interface d'interrogation unifiée

■ Exemple Entrez-NCBI

- ❖ Entrez: ensemble des BD de biologie moléculaire du NCBI + un système d'interrogation particulier associé à ces banques



Lille, 4 mai 2010

13/55

2. Avantages-Inconvénients

2.1. Interface d'interrogation unifiée (suite)

■ Avantages

- ❖ Préserve l'autonomie des BD
- ❖ Utilisation facile, intuitive
- ❖ Références croisées entre les BD
- ❖ Exportation dans divers format
 - Permet la constitution d'un jeu de données personnelles
- ❖ Existence de systèmes ouverts, personnalisables en terme de nombre et nature des BD interfacées (SRS, BioMart)

■ Limites

- ❖ NCBI-Entrez : système fermé, seulement les BD du NCBI
- ❖ SRS : Maintenance, nécessité de ré-indexer régulièrement pour tenir compte des mises à jour
- ❖ Pas d'exploitation des relations entre les données elles-mêmes
- ❖ Connaissances requises a priori sur les BD et leur qualité

Lille, 4 mai 2010

14/55

2. Avantages-Inconvénients

2.2 Base de données intégrée / Entrepôts

■ Exemple BioWarehouse

- ❖ Système d'entrepôt configurable permettant d'importer des BD entières et de gérer les données localement selon un schéma global simple et extensible

BMC Bioinformatics

Software
BioWarehouse: a bioinformatics data warehouse
Thomas J Lee¹, Yannick Pouliot¹, Valerie David WJ Stringer-Calvert², Jessica D Tenenbaum

Address: ¹Bioinformatics Research Group, SRI International, Menlo Park, US and ²Stanford Medical Informatics, Stanford University, Stanford, USA
Email: Thomas J Lee - tomlee@sri.com; Yannick Pouliot - ypouliot@rcn.com; Priyanka Gupta - priyanka0902@gmail.com; David WJ Stringer-Calvert - ds@stanford.edu; Jessica D Tenenbaum - jessiet@stanford.edu; Peter D Karp* - pkarp@ai.sri.com
* Corresponding author

<http://biowarehouse.ai.sri.com/PublicHouseOverview.html>

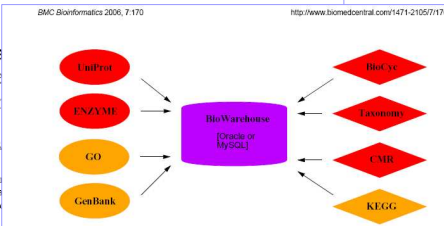


Figure 1

Lille, 4 mai 2010

15/55

2. Avantages-Inconvénients

2.2 Base de données intégrée / Entrepôts (suite)

■ Avantages

- ❖ Robustesse
- ❖ Rapidité d'accès
- ❖ Volumes traités
- ❖ Nettoyage des données, résolution des conflits entre les données
- ❖ L'utilisateur n'a pas besoin de connaître les détails des modèles de données de chacune des BD sources

■ Inconvénients

- ❖ Problème des mises à jour : repeupler l'entrepôt
- ❖ Complexité du système
- ❖ Coût de l'ajout d'une nouvelle BD (développements nécessaire, adaptation du schéma global)

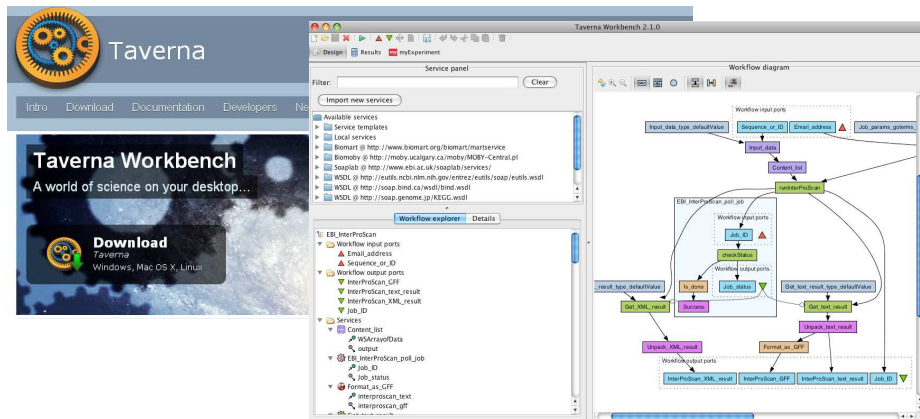
Lille, 4 mai 2010

16/55

2. Avantages et Inconvénients

2.3 Plateforme de « workflows »

- Exemple Taverna MyGrid <http://www.taverna.org.uk/introduction/what-is-taverna/>



Lille, 4 mai 2010

17/55

2. Avantages et Inconvénients

2.3 Plateforme de « workflows » (suite)

- Avantages
 - ❖ Exécutions répétitives d'enchaînements d'étapes
 - ❖ Gain de temps
 - ❖ Possibilité de partage des workflows
 - ❖ Flexibilité (plus de 1600 services web accessibles)
- Inconvénients
 - ❖ Prise en main relativement lourde
 - ❖ Pas d'intégration des résultats des traitements
 - Travail d'analyse postérieur à l'exécution des workflows
- Repris par d'autres exposés de la journée

Lille, 4 mai 2010

18/55

2. Avantages-Inconvénients

2.4 Plateformes d'intégration et de visualisation

- Exemple ONDEX
 - ❖ Combiner intégration de BD et fouille de texte grâce à des méthodes d'analyse à base de graphes

The image shows the Ondex website interface on the left and a network graph visualization on the right. The website header includes the Ondex logo and the text 'Data integration and visualisation'. Below the header, there are navigation links: Home, About, Application projects, Screenshots, and Download. The main content area contains a paragraph describing the platform's capabilities in integrating and visualizing biological data. The network graph on the right is a complex, multi-colored graph with nodes labeled 'enzymes', 'genes', 'complexes', 'proteins', 'reactions', 'transcription factors', 'enzyme classes', 'compounds', 'pathways', and 'treatments'. The graph is connected to a data table on the right side of the interface.

Köhler et al., *Bioinformatics* 2006/

Lille, 4 mai 2010



19/55

Downloaded from <http://bioinformatics.oxfordjournals.org/>

2. Avantages-Inconvénients

2.4 Plateformes d'intégration et de visualisation (suite)

- Avantages
 - ❖ Prise en compte de la visualisation
 - ❖ Puissance de calcul
 - ❖ Lien avec les technologies du web sémantique (graphes RDF)
- Inconvénients
 - ❖ Prise en main relativement lourde
 - ❖ Coût d'adaptation à un problème d'intérêt difficilement évaluable

Lille, 4 mai 2010



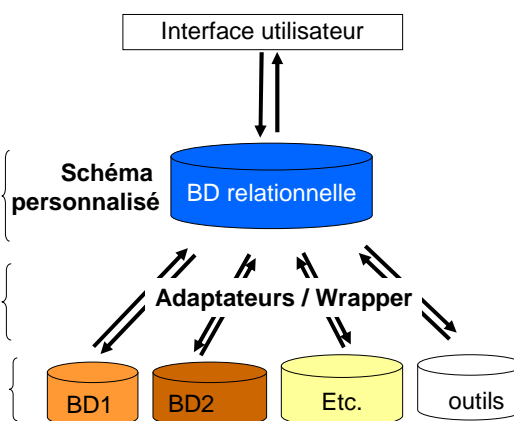
20/55

3. L'approche « entrepôt personnalisé »

Une solution hybride et flexible à l'échelle du laboratoire

Avantages et motivation

- Modélisation des données en fonction d'un problème posé
- Robustesse d'une BD relationnelle
- Wrappers génériques
- Notion de scénario
- Choix des BD ouvert
- Respect de l'autonomie des sources (mises à jour)



Lille, 4 mai 2010

21/55

PLAN DE L'EXPOSÉ

I. Les différentes stratégies pour l'intégration de données

II. ACGR : Un entrepôt de données personnalisé pour l'identification de gènes candidats

1. La question posée et sa modélisation
2. L'approche ACGR - Réalisation
3. Résultats pour 2 syndromes rares

III. MODIM : Généralisation à la construction de jeux de données pour la fouille de données

I. Stratégies

II. ACGR

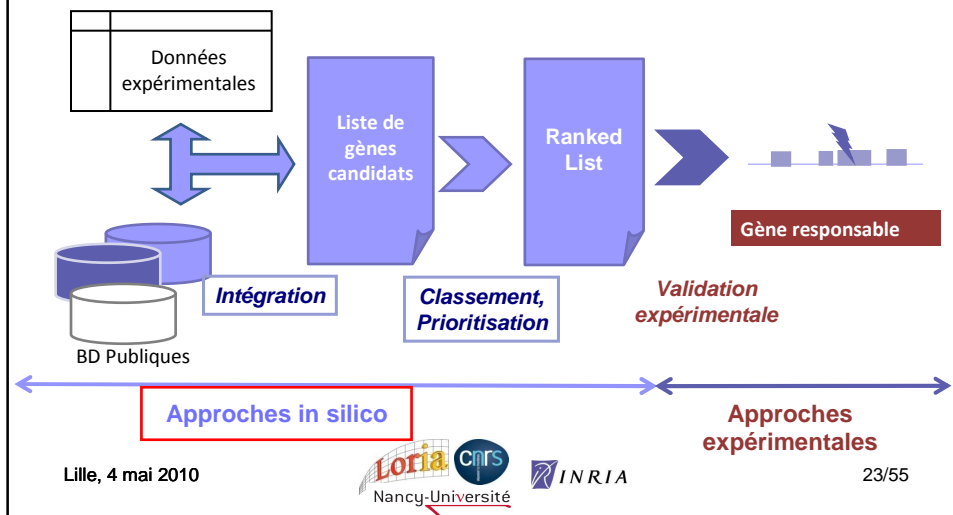
III. MODIM

Lille, 4 mai 2010

22/55

1. La question posée :

Identifier des gènes candidats pour une maladie donnée



La notion de gène candidat (1)

- Réalité intuitive pour le biologiste
 - ❖ Définition du Webster Medical Dictionary
 - *Any gene thought likely to cause a disease. The gene may be a candidate because it is located in a particular chromosome region suspected of being involved in the disease or its protein product may suggest that it could be the disease gene in question.*
- Difficile à appréhender pour l'informaticien
 - *Likely to cause a disease*
 - *Region suspected of being involved,*
 - *may suggest,*
 - *it could be the disease gene...*

La notion de gène candidat (2)

Essai de formalisation

- Un gène candidat pour une maladie est un gène qui est en relation avec une maladie
 - *Est localisé dans la région chromosomique liée à la maladie*
 - *Est exprimé dans le tissu affecté par la maladie*
 - *Est induit ou réprimé chez les patients atteints de la maladie*
 - *Est annoté de façon similaire à la maladie*
- Le gène candidat peut aussi être un gène qui est en relation avec un autre gène qui lui est en relation avec la maladie
 - *Notion de gène intermédiaire*
 - *Gène orthologue dans un organisme modèle*
 - *Gène en interaction*

Lille, 4 mai 2010

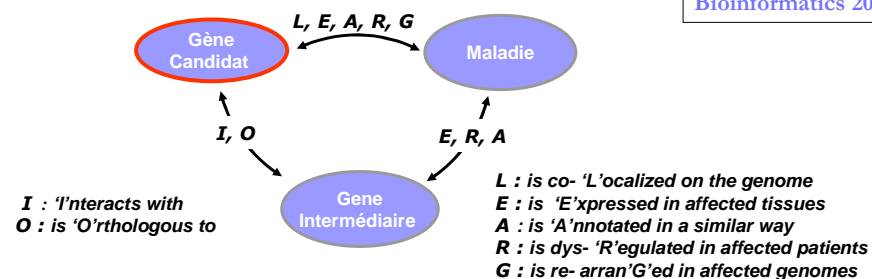


25/55

Modélisation ACGR* (1)

Définition des gènes candidats

→ *Combinatoire de relations gène-maladie et gène-gène*



***ACGR:**
Approach for
Candidate Gene
Retrieval
Yilmaz S et al.,
Bioinformatics 2009

→ *Mise en œuvre des connaissances du domaine, rôle de l'expert du domaine*

Lille, 4 mai 2010

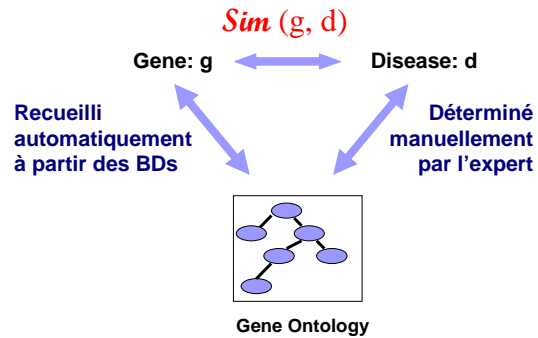


26/55

Modélisation ACGR (2)

Mesure de similarité

→ Similarité « sémantique » entre un gène et une maladie à partir des annotations fonctionnelles issues d'un même vocabulaire



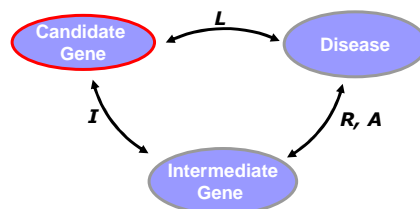
Lille, 4 mai 2010

27/55

Modélisation ACGR (3)

Résultats attendus

- Tableaux présentant les gènes selon les définitions possibles de gènes candidats, classés selon la similitude avec la maladie
 - ❖ Ex : Liste des gènes co-localisés avec la maladie et interagissant avec un gène intermédiaire dérégulé chez les patients malades / ou annotés de façon similaire à la maladie.



Lille, 4 mai 2010

28/55

Modélisation ACGR (4)

Inventaire des données à collecter et des sources à explorer

Données à collecter	Choix des sources
Annotations GO de la maladie	OMIM + travail de l'expert
Gènes (homme, souris, droso)	NCBI GENE, MGD, Flybase
Localisation chromosomique	NCBI GENE
Annotations GO (BP)	NCBI GENE
Interactants	BIND, HPRD, via NCBI GENE
Outil de classement	GO-Family (GO ToolBox)
Données expérimentales (Analyses transcriptomiques, CGH, etc.)	Fichiers Excel « maison »

→ Mise en œuvre des connaissances du domaine, rôle de l'expert du domaine

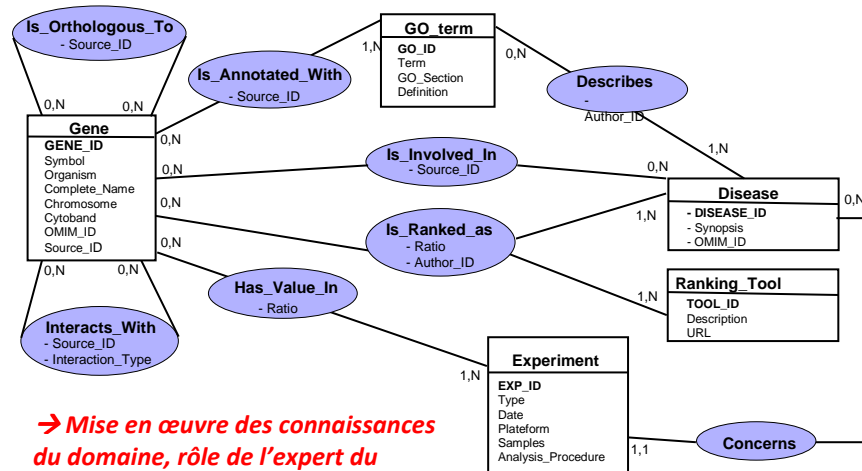
Lille, 4 mai 2010



29/55

Modélisation ACGR (5)

Modèle conceptuel de données



→ Mise en œuvre des connaissances du domaine, rôle de l'expert du domaine

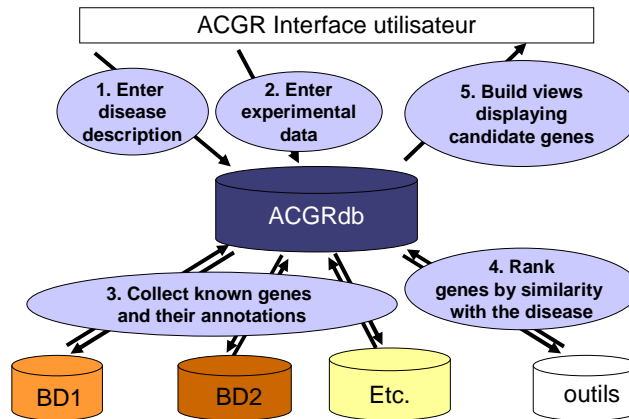
Lille, 4 mai 2010



30/55

2. L'approche ACGR - Réalisation

Fonctionnalités mises en oeuvre



Lille, 4 mai 2010

31/55

Création de la BD ACGR

Modèle relationnel de données

Gene (Gene_ID , Symbol, Organism, Complete_name, Chromosome, Cytoband, OMIM_ID, Source_ID)
GO_Term (GO_ID , Term, GO_section, Definition)
Gene_GO_Term (Gene_ID , GO_ID , Source_ID)
Orthology (Gene_ID1 , Gene_ID2 , Source_ID)
Interaction (Gene_ID1 , Gene_ID2 , Source_ID, Interaction_Type)
Disease (Disease_ID , Synopsis, OMIM_ID)
Disease_GO_Term (Disease_ID , GO_ID , Author_ID)
Involvement (Gene_ID , Disease_ID , Source_ID)
Ranking_Tool (Tool_ID , Description, URL)
Gene_Disease_Rank (Gene_ID , Disease_ID , Author_ID , Tool_ID , Value)
Experiment (Exp_ID , Type, Date, Platform, Analysis_procedure, Disease_ID)
Gene_Experiment (Gene_ID , Exp_ID , Ratio)

→ **Création de la base de données ACGR (MySQL)**

Lille, 4 mai 2010

32/55

Peuplement de la BD ACGR Wrappers

- Collecte de données dans les BD publiques
 - ❖ Système Xcollect
 - <http://www.loria.fr/~devignes/>
 - ❖ Définition de scénarios
 - *Pour chaque étape : URL de la ressource, requête, données à collecter délimitées dans le document HTML par des expressions régulières,*
 - ❖ Mapping des données collectées sur le modèle relationnel
- Outils de classement des gènes
 - ❖ Pour remplir la table Gene_Disease_Rank
 - ❖ Un wrapper par outil
 - ❖ Exemple du programme GO-Family de GO-ToolBox
 - *Pourcentage de termes communs entre les deux listes en incluant tous les ancêtres de chaque terme*

Lille, 4 mai 2010



33/55

Création des vues ACGR (1) Définitions

- Exemples de vues définies sur ACGRdb
 - ❖ Vue 1: Gènes classés selon leur similarité d'annotation GO avec la maladie
 - *Remarque : interclassement des gènes d'espèces différentes*
 - ❖ Vue 2: Orthologues humains des gènes d'organismes modèles classés selon leur similarité d'annotation GO avec la maladie
 - *Un gène candidat peut avoir un orthologue extrêmement bien annoté chez la souris ou la drosophile*
 - ❖ Vue 3: Gènes interactants des gènes de la Vue 1
 - ❖ Vue 4: Orthologues humains de gènes interactants des gènes des organismes modèles de la Vue 1
- Variantes avec données d'expression
 - ❖ Quand celles-ci sont disponibles

Lille, 4 mai 2010



34/55

Création des vues ACGR (2)

Requêtes SQL

- **Vue 1** : Gènes classés selon leur similarité d'annotation GO avec la maladie
 - ❖ SQL expression (a est l'identifiant de la maladie concernée):

```
CREATE VIEW View1 AS
(SELECT g.Symbol, g.Organism, r.Value, g.Cytoband
FROM Gene g, Gene-Disease-Rank r, Disease d
WHERE g.Gene_ID = r.Gene_ID AND r.Disease_ID=d.disease_ID AND
d.Disease_ID = a
ORDER BY r.Value DESC)
```
- La vue est dynamique : exécutable après chaque mise à jour de la BD
- Autres vues disponibles sur <http://bioinfo.loria.fr/projets/acgr>

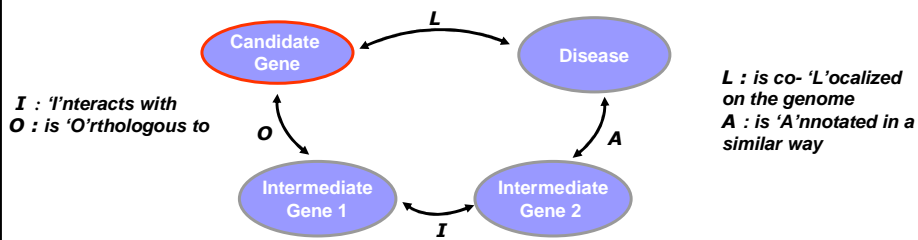
3. Résultats pour 2 syndromes rares

- Syndrome de Goltz - FDH (OMIM #305600)
 - *Hypoplasie dermique focale : atrophie et pigmentation anormale de la peau, hernies graisseuses à travers les défauts du derme, papillomes des muqueuses ou de la peau, anomalies des doigts.*
 - *CGH-array : CNV en Xp11.23*
 - *Gène responsable identifié en 2007 : PORCN (Porcupine)*
- Syndrome d'Aicardi (OMIM %304050)
 - *Triade caractéristique : agénésie du corps calleux, anomalies chorio-rétiniennes, spasmes en flexion. Retard mental*
 - *Génétique : chromosome X*
 - *Gène responsable encore inconnu*
 - *Etudié au Laboratoire de Génétique Humaine de la Faculté de Médecine de Nancy (Prof Jonveaux)*

Exemple de résultat ACGR (1)

Le syndrome de Goltz

- Description de la maladie
 - *Skin defects* → GO term « *Skin development* »
 - *Digit anomalies* → GO term « *Embryonic digit morphogenesis* »
 - *Skeletal defects* → GO term « *Embryonic skeletal morphogenesis* »
- Modèle de gène candidat



Lille, 4 mai 2010

37/55

Exemple de résultat ACGR (1)

Le syndrome de Goltz (suite)

- Tableau des premiers gènes de la BD ACGR-Goltz

Gène en interaction avec le gène orthologue au gène candidat				Gene Orthologue au gène candidat				Gène candidat		
Symbol	Organism	Cytoband	Rank (%)	Interactant_Symbol	Source	Interactant_Cytoband	Interactant_Rank (%)	Human_Ortholog	Ortholog_Cytoband	Ortholog_Rank (%)
Wnt7a	mouse	6 39.5 cM	31	Porcn	BIND	X 2.15 cM	5	PORCN	Xp11.23	7
Ngfr	mouse	11 55.6 cM	22	Ndn	BIND	7 28.0 cM	11	NDN	15q11.2-q12	11
Ngfr	mouse	11 55.6 cM	22	Ndnl2	BIND	7 C	13	NDNL2	15q13.1	13

Lille, 4 mai 2010

38/55

Exemple de résultat ACGR (2)

Le syndrome d'Aicardi

■ Input 1: disease description

- *Corpus callosum agenesis* → 5 GO terms « Forebrain development », « Corpus callosum development », « Corpus callosum morphogenesis », « neuron migration », « neural plate development ».
- *Chorio-retinal lacunae* → 1 GO term « Camera-type eye morphogenesis »
- *Infantile spasms* → no appropriate GO term

■ Input 2 : Dysregulated genes

- ANOVA list of 300 genes (Saliha Yilmaz PhD Thesis)

■ Input 3 : Copy Number Variation (CGH array)

Lille, 4 mai 2010



39/55

Exemple de résultat ACGR (2)

Le syndrome d'Aicardi (suite)

- Vue 1 : Gènes classés par similarité avec la maladie et co-localisés ou dérégulés



L : is co- 'L'ocalized on the genome
R : is dys- 'R'egulated in affected patients
A : is 'A'nnnotated in a similar way

Deux conclusions sur les listes de gènes obtenues dans la BD ACGR:

1. Les gènes sur le chromosome X qui sont bien classés ne sont pas dérégulés chez les patientes
2. Les gènes sur le chromosome X qui sont ré-régulés chez les patientes ne sont pas bien classés du point de vue de la similarité des annotations fonctionnelles

Lille, 4 mai 2010

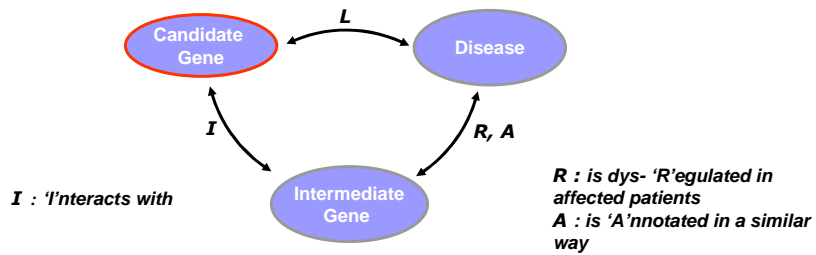


40/55

Exemple de résultat ACGR (2)

Le syndrome d'Aicardi (suite)

■ Vue 2 : Gènes qui interagissent avec un gène similaire ou dérégulé



Lille, 4 mai 2010

41/55

■ Tableau des résultats de la Vue 2

Gène en interaction avec le gène candidat					Gène candidat				
Symbol	Species	CytoLoc	Rank (%)	Exp_Value	Interactant	Source	Interactant_Cyto loc	Interactant_Rank (%)	Interactant_Exp_Value
DLX5	human	7q22	50	1	MAGED1	HPRD	Xp11.23	3	0
UBE3A	human	15q11-q13	22	1	UBQLN2	HPRD	Xp11.23-p11.1	8	0
CXCL10	human	4q21	21	1	CXCR3	HPRD	Xq13	10	0
IGF1	human	12q22-q23	21	1	IGSF1	BIND	Xq25	6	0

Lille, 4 mai 2010

42/55

Conclusion sur l'approche ACGR

- ACGR prototype
 - ❖ Dans la BD ACGR, tout gène qui s'y trouve a une « bonne raison » d'y être.
 - ❖ ACGR permet d'intégrer des données expérimentales personnelles avec les données publiques
 - ❖ ACGR permet d'exploiter de nombreuses définitions de gènes candidats
- Perspectives
 - ❖ Séquencer les gènes candidats du syndrome d'Aicardi pour détecter des mutations (en cours)
 - ❖ Introduire d'autres méthodes de classement
 - ❖ Appliquer cette approche à d'autres types de gènes candidats : chéomogénomique, médecine chinoise

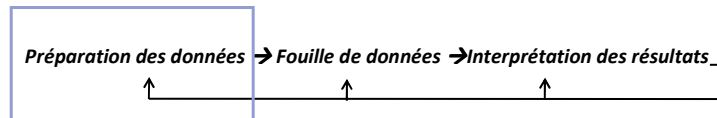
PLAN DE L'EXPOSÉ

- I. Les différentes stratégies pour l'intégration de données
- II. ACGR : Un entrepôt de données personnalisé pour l'identification de gènes candidats
- III. MODIM : Généralisation à la construction de jeux de données pour la fouille de données
 - I. Principe: une aide à la construction d'entrepôts personnalisés
 - II. Prototypage: description
 - III. Tests

MODIM

« Model-Driven Data Integration for Mining »

- Contexte de « Data Mining »: Fouille de données en vue de l'extraction de connaissances
 - ❖ KDD (Knowledge Discovery in Databases) : 3 étapes itératives



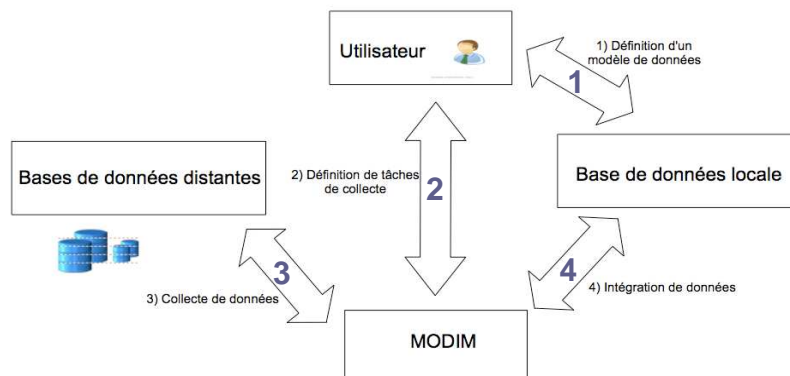
Objectifs de MODIM:

1. Intégrer les données pertinentes pour le problème posé
2. Construire des jeux de données pour la fouille

Lille, 4 mai 2010

45/55

1. Aide à la construction d'entrepôts personnalisés (1)



Lille, 4 mai 2010

46/55

1. Aide à la construction d'entrepôts personnalisés (2)

- Spécification du problème
 - ❖ Par l'utilisateur, expert du domaine
 - ❖ Choix des données à collecter, des sources de données à interroger
 - ❖ Création d'un modèle conceptuel (~ représentation des connaissances)
- Création de la BD MODIM
 - ❖ Sauvegarde du modèle
 - ❖ S'appuie sur un système de gestion de BD
- Configuration des tâches MODIM pour la collecte et la sauvegarde données
 - ❖ Selon un modèle spécifié en XML
- Exécution de ces tâches
 - ❖ Peuplement de la BD

Lille, 4 mai 2010



47/55

2. MODIM : Application web

Home **Models** **Task creation** **Task edition** **Task enactment** **Task library** **Tutorial**

Model Driven Data Integration for Mining

Welcome to Modim web interface

This web interface is developed under the modim project. It is dedicated to biological applications of data mining.
It allows users to create and execute scenarios for mining and integration data in databases hosted on the local server gbmserv.
It contains three features :

- task creation which creates a task
- task edition that lets you modify a task
- task enactment which allows the execution of a task to collect data from remote database and to integrate them into a local database.

**Conçu et développé
par Birama Ndiaye,
IJD, Equipe Orpailleur
INRIA Nancy Grand-
Est**

Lille, 4 mai 2010



48/55

2.1 Création d'une BD

Home Models Task creation Task edition Task enactment Task library Tutorial

A

Create new database

Enter the database name Path2GO Create

B

Enter the table name KEGGpathway

Attributes

Name KPname

Type varchar

Primary key No

Constraints

remove

add attribute

Choose the database Path2GO Create table

Lille, 4 mai 2010

Etc.

49/55

2.2 Définition d'une tâche de collecte Pour une BD et un type de donnée d'entrée

Home Models Task creation Task edition Task enactment Task library Tutorial

Definition

Enter the task name KEGG_pathway

Task description Query KEGG pathway database with a KEGG pathway ID to collect the name, the class and the organism this pathway comes from. *: Avoid accented characters

Choose the database path2go

Input

Input name KP_ID

Import from database Choose the database

Validate Subtask design

Spécifie la BD que l'on veut peupler

Spécifie le type d'entrée pour les requêtes associées à cette tâche

Lille, 4 mai 2010

50/55

2.3 Définition d'une sous-tâche Pour une BD distante à interroger

Un élément output pour chaque attribut de la table à remplir

Xpath ou expression régulière selon le format retourné par la BD interrogée

Lille, 4 mai 2010

51/55

2.4 Mise en œuvre de la collecte

Moteur d'exécution des tâches et des sous-tâches programmé en PHP
Interrogation des BD spécifiées dans les sous-tâches
Extraction des données spécifiées
Ecriture dans la base de données concernée

Lille, 4 mai 2010

52/55

MODIM : Bilan

- Aide à la réalisation de BD personnalisées à double titre
 - ❖ Selon un modèle personnalisé
 - ❖ Avec possibilité d'importer des données personnelles
- Accès à la BD via un système de gestion de BD relationnelle classique
 - ❖ PostgreSQL
 - ❖ Interrogation en langage SQL
 - ❖ Possibilité de créer des « Vues » dynamiques comme avec ACGR
 - ❖ Possibilité d'enrichir l'entrepôt par des fonctionnalités complémentaires si nécessaire
- Prototype en cours de test
 - ❖ Stage de M1 : prise en main par une biologiste en 2 mois
 - ❖ Ouvert à collaboration si vous avez des données à collecter et à intégrer !

Lille, 4 mai 2010



53/55

CONCLUSION

- Proposition d'entrepôt personnalisé
 - ❖ Approche générique et non généraliste
 - ❖ Garder l'utilisateur biologiste au cœur du système
- Technologies du web sémantique
 - ❖ Mise à disposition d'« entrepôts RDF » (ORACLE 10g & 11g)
 - ❖ *ex d'utilisation: AlzPharm*
 - ❖ Données représentées par des graphes de triplets
 - ❖ *ex: (Gene ABC, HasSnp, rs23456)*
 - ❖ Langage de requête SPARQL
- Rôle des ontologies
 - ❖ Maîtriser leur utilisation à des points-clé du processus d'intégration
 - ❖ Systèmes à base de connaissance et de raisonnement : à venir...
 - ❖ *ne passent pas encore l'échelle en terme de quantité de données à traiter en biologie*

Lille, 4 mai 2010



54/55

Participants

Laboratoire de Génétique
Faculté de Médecine Nancy

Saliha Yilmaz

Prof Philippe Jonveaux

Prof Bruno LeHeup



LORIA, Equipe Orpailleur
Nancy

MD Devignes

Malika Smaïl-Tabbone

Birama Ndiaye (ingénieur CDD INRIA)

Cédric Bicep, Lu Zhang, Anaïs Gigant
(stagiaires)



Thrombosis Research Institute
& King's College

John Louis McGregor and his team



Financement

Communauté Urbaine du Grand Nancy

Contrat Plan Etat Région : Intelligence
Logicielle

AAL association A.A.L. - Syndrome d'Aicardi

Lille, 4 mai 2010



55/55