Bio-ontologies for guiding integration and mining of biological data

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. Bio-ontologies and data integration
3. Bio-ontologies and data mining
   1. Functional Classification (GO)
   2. Dimension reduction (MedDRA)
4. Conclusion

Rennes, 18 octobre 2011
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.
Setting up KDDK

1. Data extraction and formatting
2. Data mining: guided by domain knowledge
3. Result interpretation and KB enrichment

Data integration: guided by domain knowledge

1. Data extraction and formatting
2. Data mining: guided by domain knowledge
3. Result interpretation and KB enrichment

What is a Knowledge Base?

- For biologists « knowledge bases » are rich, integrated databases
  - Ex: Uniprot, Kegg, OMIM, IMAGE, PharmGKB, etc.
  - Knowledge is embedded in the global schema
  - Knowledge is not used by programs for reasoning

- For computer scientists « knowledge bases » are systems where data are associated with explicit semantics (logic formulas) that can be used by programs
  - Ex: Ontologies in Description Logics (Baader et al., The DL Handbook, 2003)
  - Knowledge is used for reasoning, inferring instance classification
    - Concepts and relations are organized as consistent hierarchies from most general to most specialized
    - Reasoners are able to classify instances with respect to the concepts they instantiate.
Examples in the Life Sciences

- A « DL » knowledge base: SO-Pharm (Adrien Coulet)
  - Integrating pharmacogenomic data for mining
  - Attribute selection before mining

- Structured vocabularies
  - Gene Ontology (Sidahmed Benabderrahmane)
    - Semantic similarity measure
    - Use for functional classification of genes
  - MedDRA: Medical Dictionary for Regulatory Activities (Emmanuel Bresso)
    - Use for dimension reduction in symbolic mining methods.

Outline of the talk

1. Introduction: Knowledge Discovery guided by Domain Knowledge

2. Bio-ontologies and data integration
   Example of a DL knowledge base: SO-PHARM

3. Bio-ontologies and data mining
   1. Functional classification (GO)
   2. Dimension reduction (MedDRA)

4. Conclusion
Bio-ontologies and data integration

- A challenge for many years
  - And many more…

- Not so many examples of « true » knowledge bases

---

DL (Description Logics) Knowledge Base

- Description Logics (DL) foundations
  - Family of 1st-order logics ([Baader et al. The DL handbook, 2003](http://www.cs.man.ac.uk/~stevensr/tambis/))
  - Concepts (classes) and Roles (relationships between classes, domain and co-domain)
  - Various types of constructors (Intersection, Union, Negation, Restrictions, etc.)

- In DL, KB = T-Box + A-Box
  - T-Box = Terminology : concepts and roles
    - Atomic concepts (C) or roles (R): simple descriptions
    - Composed concepts or roles : complex descriptions (terminological axioms)
  - A-Box = Assertions on individuals/instances
    - Concept assertion : $C(a) \rightarrow$ instance $a$ belongs to concept $C$
    - Role assertion: $R(a,b) \rightarrow$ instances $a$ and $b$ are in relation through role $R$
DL (Description Logics) Knowledge Base

- Implementation thanks to semantic web technologies
  - OWL language (Ian Horrocks, Peter F. Patel-Schneider and Frank van Harmelen, 2003).
  - SPARQL: query language for OWL ontologies
  - Protégé: ontology editor (Musen, since 2000)
  - Racer, Pellet, FaCT: reasoners on DL ontologies

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
An example in pharmacogenomics (2)

Articulation of existing ontologies (15) covering various biological domains
ME0 : 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

Rennes, 18 octobre 2011

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

Rennes, 18 octobre 2011
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)
    - 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)

<table>
<thead>
<tr>
<th>SNP</th>
<th>Gene</th>
<th>Chr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>ABCCC1</td>
<td>16</td>
</tr>
<tr>
<td>7</td>
<td>ALOX5</td>
<td>10</td>
</tr>
<tr>
<td>1</td>
<td>CYSLTR1</td>
<td>X</td>
</tr>
<tr>
<td>8</td>
<td>LTA4H</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>LTC4S</td>
<td>12</td>
</tr>
</tbody>
</table>

Drug treatment: Montelukast 6 months

Phenotype items
- Exa: Asthma exacerbation
- Per: % Change in Forced Expiratory Volume

Knowledge discovery in SO-Pharm KB

1. Exploration of the Graph of Role Assertions
2. Formal Concept Analysis
3. Interpretation, insertion into the ontology

Association Rules

New concepts, roles, and role assertions

T-Box
A-Box

Instances of a given concept

RAA : Role Assertion Analysis

Coulet et al., Advances in Experimental Medicine and Biology, 2011
Conclusion (for Part 2)

- Proof of concept but...
  - Still difficult to handle large datasets with semantic web technologies

- Data representation for mining: still under study
  - The Bio2RDF project is a tool to convert bioinformatics data and knowledge bases to RDF format. It is a kind of generalized rdfizer for bioinformatics applications, and it is a place for the semantic web life science community to develop and grow. (Michel Dumontier, Ottawa Carleton University)


Functional classification (GO)

Plan de l’exposé

1. Introduction: Knowledge Discovery guided by Domain Knowledge

2. Bio-ontologies and data integration

   Example of a DL knowledge base: SO-PHARM

3. Bio-ontologies and data mining
   1. Functional classification (GO)
   2. Dimension reduction (MedDRA)

4. Conclusion
Bio-ontologies for guiding the mining step

1. Data extraction and formatting
2. Data mining
3. Result interpretation and KB enrichment

Data mining: guided by domain knowledge

Semantic Clustering (1) on GO
Example 1 : Semantic clustering

- Clustering means grouping together most similar objects and putting in different clusters most dissimilar objects
  - relies on a similarity/distance measure

- Clustering functional annotation of genes: functional classification
  - Example of DAVID tool (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

Rennes, 18 octobre 2011

Example 1 : Semantic clustering (2)

- Clustering is semantic when it relies on a semantic similarity measure

- A semantic similarity measure:
  - Is defined between biological objects annotated by feature lists from a structured vocabulary (bio-ontology)
  - Takes into account the ontological relationships between features

⇒ Various strategies exist for
(i): capturing feature similarity in a bio-ontology
(ii): aggregating these similarities for comparing biological objects

Rennes, 18 octobre 2011
Term-term semantic similarity (1)

To compare two features, members of an ontology

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

Term-term semantic similarity (2)

To compare two features, members of an ontology

- Wu et al. (2006) : Depth of LCA ; Lowest Common Ancestor, Shortest Path Length
- Poze et al. (2008) : Depth of LCA
- Othman et al. (2007) : IC/Depth/number of children; Distance
Gene-gene semantic similarity

- To compare two feature lists annotating genes or gene products

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

« Groupwise » approaches
- Martin et al. (2004) : Graph/Jaccard on lists enriched with term ancestors
- Chabalier et al. (2007) : Vectors compared with the cosine measure

IntelliGO: a hybrid vector-based semantic similarity measure

- Representation of genes in a vector space model
  \[ \overrightarrow{g} = \sum_{i} \alpha_i \overrightarrow{e}_i \]
  \( \overrightarrow{e}_i \) : basis vector, one per feature \( t_i \)
  \( \alpha_i \) : Coefficient for feature \( t_i \)

- Definition of coefficients
  \[ \alpha_i = w(g, t) \times IAF(t) \]
  \( w(g, t) \) : weight of evidence code * qualifying the assignment of feature \( t_i \) to gene \( g \)
  \( IAF(t) \) : « 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
  \[ IAF(t) = \log \frac{N_{TOT}}{N_{t_i}} \]
  \( N_{TOT} \) : Total number of genes in the corpus
  \( N_{t_i} \) : Number of genes with feature \( t_i \)
Adapting the Generalized Cosine Similarity Measure


- 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:

\[ \overrightarrow{e_i} \cdot \overrightarrow{e_j} = \begin{cases} 1 & \text{And } i, j \neq t \neq j, e_i \cdot e_j \neq 0 \\ \frac{2 \times \text{Depth}[\text{LCA}(t, t)]}{\text{Depth}(t) + \text{Depth}(t')} & \end{cases} \]

IntelliGO term-term similarity


- 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) is not unique

\[ \overrightarrow{e_i} \cdot \overrightarrow{e_j} = \text{Sim}_{\text{IntelliGO}}(t_i, t_j) = \frac{2 \times \text{MaxDepth}[\text{LCA}(t_i, t_j)]}{\text{MinSPL}(t_i, t_j) + 2 \times \text{MaxDepth}[\text{LCA}(t_i, t_j)]} \]
IntelliGO gene-gene similarity

- Generalized dot-product between two gene vectors
  \[ \vec{g} \cdot \vec{h} = \sum_{i,j} \alpha_i x \beta_j x \vec{e}_i \cdot \vec{e}_j \]
  avec \( \vec{e}_i \cdot \vec{e}_j \neq 0, \ \forall i, i \neq j \)

- Generalized cosine similarity
  \[ \text{Sim}_{\text{IntelliGO}}(\vec{g}, \vec{h}) = \frac{\vec{g} \cdot \vec{h}}{\sqrt{\vec{g} \cdot \vec{g}} \times \sqrt{\vec{h} \cdot \vec{h}}} \]

Les étapes de l’implantation

1. Espèce
2. Aspect de GO
3. Liste des poids des codes d’évidence

Fichier NCBI: AnnotationFile (Tax_ID, Gene_ID, GO_ID, Evid_Code, GO_Def, GO_aspect)
Fichier spécifique: CuratedAnnotationFile (GO_ID, GO_aspect)

Genes
(Gene_ID, Array of [GO_ID, Evid-Code])

Terres
(GO_ID, IAF, Array of [Gene_IDs])

LCA (requêtes sur GO database)

SPL
(GO_ID, GO_ID, LCA_Depth, LCA_ID_List)

Calcul du SCL

Liste de termes d’intérêt (Gene_IDs)

Liste des mesures de similarités entre gènes 2 à 2
Validation on benchmark datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Species</th>
<th>Source</th>
<th>Number of sets</th>
<th>Total genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Human</td>
<td>KEGG pathways</td>
<td>13</td>
<td>275</td>
</tr>
<tr>
<td>2</td>
<td>Yeast</td>
<td>KEGG pathways</td>
<td>13</td>
<td>169</td>
</tr>
<tr>
<td>3</td>
<td>Human</td>
<td>Pfam Clans</td>
<td>10</td>
<td>94</td>
</tr>
<tr>
<td>4</td>
<td>Yeast</td>
<td>Pfam Clans</td>
<td>10</td>
<td>118</td>
</tr>
</tbody>
</table>

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

Heatmap visualisation of hierarchical clustering

Lord et al. (normalized)

Al-Mubaid

SIM

IntelliGO

Rennes, 18 octobre 2011
Fuzzy C-means clustering: optimal F-score and K number comparison with DAVID tool

<table>
<thead>
<tr>
<th>Dataset (Sets)</th>
<th>IntelliGO</th>
<th></th>
<th>DAVID</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Optimal F-score</td>
<td>Optimal K number</td>
<td>Optimal F-score</td>
<td>Optimal K number</td>
</tr>
<tr>
<td>1 (13)</td>
<td>0.62</td>
<td>14</td>
<td>0.67</td>
<td>10</td>
</tr>
<tr>
<td>2 (13)</td>
<td>0.67</td>
<td>14</td>
<td>0.68</td>
<td>9</td>
</tr>
<tr>
<td>3 (10)</td>
<td>0.75</td>
<td>11</td>
<td>0.64</td>
<td>11</td>
</tr>
<tr>
<td>4 (10)</td>
<td>0.82</td>
<td>11</td>
<td>0.70</td>
<td>10</td>
</tr>
</tbody>
</table>

Functional classification is reliable and robust with IntelliGO measure

Benabderrahmane et al., BIBM workshop IDASB 2011

Rennes, 18 octobre 2011

Bio-ontologies for guiding the mining step

3. Result interpretation and KB enrichment

Semantic Clustering (2) on MedDRA

1. Data extraction and formatting

Domain Knowledge

Annotation vocabulary

2. Data Mining

Data mining: guided by domain knowledge

Rennes, 18 octobre 2011
Semantic clustering for data reduction

- 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 Bresco's PhD Thesis, Harmonic Pharma)

Replacing terms with term clusters (TC)

- Such large matrices are untractable with symbolic methods
  - Search for frequent itemsets fails because not enough features are shared!

- 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
Hierarchical clustering of MedDRA terms

- Pairwise distances calculated for a subset of 1288 terms
- Hierarchical clustering (Wards) + 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

<table>
<thead>
<tr>
<th>Cluster terms</th>
<th>AvgDist to other cluster terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Erythema</td>
<td>0.31</td>
</tr>
<tr>
<td>Lichen planus</td>
<td>0.32</td>
</tr>
<tr>
<td>Parapsoriasis</td>
<td>0.32</td>
</tr>
<tr>
<td>Phyllitis alba</td>
<td>0.32</td>
</tr>
<tr>
<td>Rash papular</td>
<td>0.32</td>
</tr>
<tr>
<td>Decubitus ulcer</td>
<td>0.35</td>
</tr>
<tr>
<td>Lupus miliaris</td>
<td>0.35</td>
</tr>
<tr>
<td>disseminatus faciei</td>
<td>0.35</td>
</tr>
<tr>
<td>Pruritus</td>
<td>0.35</td>
</tr>
<tr>
<td>Rash</td>
<td>0.35</td>
</tr>
<tr>
<td>Sunburn</td>
<td>0.35</td>
</tr>
<tr>
<td>Vulvovaginal pruritus</td>
<td>0.35</td>
</tr>
<tr>
<td>Dandruff</td>
<td>0.37</td>
</tr>
<tr>
<td>Rash</td>
<td>0.37</td>
</tr>
<tr>
<td>Photosensitivity reaction</td>
<td>0.37</td>
</tr>
<tr>
<td>Psoriasis</td>
<td>0.37</td>
</tr>
</tbody>
</table>

Rennes, 18 octobre 2011

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

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

Rennes, 18 octobre 2011
Frequent Closed Itemset (FCI) extraction (1)

<table>
<thead>
<tr>
<th>Minimal support</th>
<th>50 %</th>
<th>60 %</th>
<th>70 %</th>
<th>80 %</th>
<th>90 %</th>
<th>100 %</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA_{All}</td>
<td>386</td>
<td>94</td>
<td>41</td>
<td>11</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>CA_{TC}</td>
<td>5,564</td>
<td>1,379</td>
<td>256</td>
<td>62</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>AIA_{All}</td>
<td>178</td>
<td>41</td>
<td>9</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AIA_{TC}</td>
<td>654</td>
<td>154</td>
<td>30</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

- 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

- Nausea and vomiting symptoms, Nausea (80%)
- Pruritus (79%)
- Nausea and vomiting symptoms, Nausea (78%)
- Headache (76%)

**TC representation**

- 54_Erythema (88 %)
- 64_Nausea_and_vomitting_symptoms (88%)
- 99_Neurologyopath (82 %)
- 54_Erythema, 64_Nausea_and_vomiting_symptoms (79%)
- 65_Blepharitis (78%)

Bresso et al., KDIR 2011
Conclusion

- Bio-ontologies and integration: still difficult
  - Data are still difficult to handle as ontology instances
  - Use bio-ontologies to standardize metadata
  - Bio2RDF initiative by Michel Dumontier

- Bio-ontologies and mining
  - Bio-ontologies as structured sources of features
  - Semantic clustering: complementary to GO-term enrichment
  - Symbolic methods become practicable -> extract explicit knowledge from large real-world datasets

Participants

LORIA, Equipe Orpailleur Nancy
MD Devignes
Malika Smail-Tabbone
Adrien Coulet
Sidahmed Benabderrahmane
Jean-François Kneib

Financements
Projet Eureka GenNet
Communauté Urbaine du Grand Nancy
Contrat Plan Etat Région : MISN
INCa (bourse de thèse interdisciplinaire)

KIKA medical
Hôpital Saint Antoine Paris
Pascale Benlian (MD)
PhenoSystems
David Atlan

Harmonic Pharma
Michel Souchet
Emmanuel Bresso

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