SNOMED and the Art of Bioinformatics

Yves A. Lussier,
Dept of Biomedical Informatics,
Dept of Medicine
Columbia University

Integration Biomedical Databases

Medical Records
DNA Sequences
Gene Expression
Protein Structure
Lead Compounds

Comparative Genomics
Proteomics
Literature

Integration for Scientific and Technological Innovation
Problems with Linking Heterogeneous Biomedical Databases

• Burgeoning DBs
  – Lacking shared indexes
  – Heterogeneous schemas
  – Heterogeneous data
• Manual / curators
  – rate limiting
  – 95% accurate unless double independent modeling
  – expensive
• Current automated methods inadequate

Problems Linking Heterogeneous Data

• Depth
  – Granularity / Compositionality
• Breadth
  – Biological Scales
  – Context (Species, experimental conditions, tissue, cell type, cell cycle, ..., regional anatomy / Space)
• Overlapping ontologies
  – Ambiguity
  – Redundancy
  – Varying comprehensiveness of taxonomies / classification
Outline

• Introduction:
  – Computational Ontologies
  – Lattice, Graph, Networks
  – Phenotypic Networks (PN)
• From Phenotypic Databases to PN
• From Narratives to PN
• Mining PN for Knowledge Discovery

Computable Information

• Mathematics: numbers and variables
• Chemistry: periodical table of elements
• Biomedicine: terminologies for intelligent systems
Complex integrated biomedical models

Terminologies

Biology

<table>
<thead>
<tr>
<th>meters</th>
<th>ICD-10, SNOMED 1 SNOMED</th>
</tr>
</thead>
<tbody>
<tr>
<td>$10^0$</td>
<td></td>
</tr>
<tr>
<td>$10^{-1}$</td>
<td></td>
</tr>
<tr>
<td>$10^{-2}$</td>
<td></td>
</tr>
<tr>
<td>$10^{-3}$</td>
<td></td>
</tr>
<tr>
<td>$10^{-4}$</td>
<td></td>
</tr>
<tr>
<td>$10^{-5}$</td>
<td></td>
</tr>
<tr>
<td>$10^{-6}$</td>
<td></td>
</tr>
<tr>
<td>$10^{-7}$</td>
<td></td>
</tr>
<tr>
<td>$10^{-8}$</td>
<td></td>
</tr>
<tr>
<td>$10^{-9}$</td>
<td></td>
</tr>
</tbody>
</table>

Technology

- Measure Organism organs
- Tissue Cells Organelles Virus DNA bases
- Gene Ontology Quaternary code
SNOMED Information Model:
Compositional, multiaxial, multi-hierarchic

Axes

H. Pylori associated hemorrhagic Gastric Ulcer =
(4) D5-32220 Gastric (1) Ulcer (2) with haemorrhage (3) G-
C002 associated with (5) L-13551 H. pylori (6)

Lattice, Graph, Networks

- Full graphs
  - Allows cycles in the relationships
- Directed Acyclic Graphs
  - Multiple parents for one child concept
  - Simple taxonomy: one parent per child concept
- Lists (e.g. Controlled terminologies, Data Dictionaries…)
Knowledge Economy

• One of the tenets of the knowledge economy is Metcalfe’s Law.

Outline

• Introduction: Phenotypic Networks (PN)
• From Phenotypic Databases to PN
  – Knowledge Engineering / curation
  – Network of Knowledge bases
• From Narratives to PN
• Mining PN for Knowledge Discovery
The ancient art of knowledge acquisition

- Highly Dependant on:
  - Domain Experts Selection
  - Knowledge Engineer Selection
  - Double Modeling
  - Double Knowledge Representation
  - “Software Engineering Methods & Cycles”

Distributed Scientific Knowledge Across Scientific Domains

- Pathogen Phylogenetic Databases
- Host-Pathogen Interaction DBs
- Immunology, etc.
- Gene Products Databases
- Host Databases
Examples of Valuable Databases Reflecting Silos of Scientific Domain

Pathogen Phylogenetic Databases
- NIAID RCE: ICTVdb (Buchin-Osmond)
- Bacterial (Editor?)

GIDEON, etc.

Host Databases
- SNOmed
- MODs (Mouse MGI, Ratbase, etc.)
- GIDEON
- OMIM

Host-Pathogen Interactions
- NIAID RCE (Dickerman, Lussier, etc.)

Gene Products Databases
- NIAID RCE (Sobral, Gerstein, Skolnick, etc.)
- Locuslink, Suissprot, etc.

Aim: Usable, Integrated and Transdisciplinary Scientific Knowledge

Pathogen Phylogenetics Databases

Host Pathogen Interaction DBs

Gene Products Databases
Outline

• Introduction: Phenotypic Networks (PN)
• From Phenotypic Databases to PN
• From Narratives to PN
  – Natural Language Processing
• Mining PN for Knowledge Discovery

PhenoGenes – Purpose

• Facilitate cancer research by leveraging Informatics technologies
• Extract information from the EMR as well as biomedical literature
• Ultimately, build a high-throughput tool to organize information on phenotypes and genotypes
PhenoGenes - Methodology

- Mine phenotype/genotype data from EMR
- Create knowledge base of biomedical relationships in cancer from literature
- Create visualization tool for researchers
  - XML database
  - Signal transduction through GeneWays
- Link system to clinical data warehouse

Funded Phenogenes Collaboration

BioMedLEE → XML transformation → MedLEE

GENIES

XML Database

M³  GeneWays  CliniGene  Phenotype Organizer

- Clinical – biological database interoperability
- Extract and codify ST relationships, Visualize extracted pathways
- New clinical genomic relationships
- Genotypes/Phenotypes across species

Visualization Tools:
Protein, disease, pathway, etc.

Hypothesis Generation → New Discoveries → Hypothesis Validation
PhenoGenes- Components

• M³ – The Molecular Medicine Matrix
  – Developing automated methods for improved interoperatbility between clinical and biological databases

• MedLEE
  – State of the art NLP extraction and encoding system in the clinical domain

• BioMedLEE
  – MedLEE adapted to extraction and encoding of phenotypic-genotypic relations in the literature

BioMedLEE
(Friedman, Lussier 2004)
PhenoGenes- Components (3)

• Clinigenes
  – Automated methods to mine biomedical information for new clinical genomic relationships

• Phenotype Organizer System
  – Automated methods for analyzing phenotypes and underlying genotypes across species

Outline

• Introduction: Phenotypic Networks (PN)
• From Phenotypic Databases to PN
• From Narratives to PN
• Mining PN for Knowledge Discovery
Mining terminologies:
Clinigene & Genestrace
Results

<table>
<thead>
<tr>
<th></th>
<th>Exact Match</th>
<th>MMTx Strict</th>
<th>MMTx Norm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>94%</td>
<td>65%</td>
<td></td>
</tr>
<tr>
<td>Recall</td>
<td>100%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Exact Match
- MMTx Strict
- MMTx Norm

Solution to interoperability of Phenotypics Databases

• Manual Curation based interoperability
  - E.g.: UMLS

• Automated interoperability methods
  • E.g. automated indexing
Manual Indexing / Curation

Other subdomains
- Anatomy
- Biomedical literature
- Clinical repositories
- Genetic knowledge base
- Genome Annotations
- UMDA
- GO
- OMIM
- SNOMED
- UMLS

Manual Curation based interoperability

- rate-limiting for data sets using current terminologies
  - delayed and incomplete synchronization
- High throughput unattainable for uncoordinated data sets
Automated Indexing

– Rare studies in clinical genetics and molecular biology;
– Lexical matching

Records from MIM were lexically mapped to the equivalent concepts in Mesh


– Lexical and semantics

Sarkar IN, Lussier YA et al. Linking biomedical information and knowledge resources: GO and UMLS. Pac Symp Biocomputing 2003;8:427-50.

Proof-of-Concept Study:

Automated Networks of Terminologies can increase recall and precision of queries across two heterogeneous databases sharing no cross-indexes.
Methods

• Automated terminology networks
  – Databases
  – Construction of the network
  – Pathways derived from the network

• Lexico-semantic techniques
  – Lexical method
  – Semantic constrains

• Judicious Multi-Strategy Use of the Network Relationships

Target databases

• Human Disease Genes Database (HDG)
  – Manually compiled database to classify disease genes & their products according to function
  – 921 disease genes are documented in the database

• SNOMED-Clinical Term (clinical medicine)
  – Concept-based clinical terminology
  – Version used: July, 2002 ; 333,325 concepts.
Intermediating databases/terminologies

• Online Mendelian Inheritance in Man (OMIM);
  – human genes and genetic disorders
  – Total 14280 entries (Loci and diseases)

• Unified Medical Language System (UMLS);
  – 2002AB: 871,584 concepts

• SNOMED 3.5
  – SNOMED-3.5: 208,454 concepts

---

Manual Curation

Manual Curation

OMIM -> HDG

SNOMED 3.5

UMLS

OMIM

SNOMED CT

Manual curation

250

HDG

SNOMED CT

OMIM

SNOMED 3.5

UMLS

208,454

HDG

SNOMED CT

OMIM

SNOMED 3.5

UMLS

Manual curation
SNOMED CT

UMLS

OMIM

HDG

Manual curation

SNOMED 3.5

208,454

OMIM

SNOMED 3.5

514

47

37

37

HDG

UMLS

SNOMED CT
Automated Terminology Networks: ATN

Paths derived from the network

<table>
<thead>
<tr>
<th>Path Name</th>
<th>Intermediating terminologies (#)</th>
<th>Complete Path</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>3</td>
<td>HDG = OMIM = UMLS = SNOMED3.5 = SNOMED-CT</td>
</tr>
<tr>
<td>P2</td>
<td>0</td>
<td>HDG → SNOMED-CT</td>
</tr>
<tr>
<td>P3</td>
<td>1</td>
<td>HDG → UMLS → SNOMED-CT</td>
</tr>
<tr>
<td>P4</td>
<td>1</td>
<td>HDG → OMIM (Disease) → SNOMED-CT</td>
</tr>
<tr>
<td>P5</td>
<td>1</td>
<td>HDG → OMIM (Title) → SNOMED-CT</td>
</tr>
<tr>
<td>P6</td>
<td>2</td>
<td>HDG → UMLS → OMIM → SNOMED-CT</td>
</tr>
<tr>
<td>P7</td>
<td>2</td>
<td>HDG → OMIM → UMLS → SNOMED-CT</td>
</tr>
</tbody>
</table>

A = B  Manual Curation / Mapping of terms via a common index between databases A and B.
A → B  Automated Mapping / lexico-semantic mapping of terms between databases A and B.
Automated Terminology Network: ATN

OMIM -> HDG
OMIM -> UMLS
SNOMED 3.5 -> HDG
SNOMED 3.5 -> UMLS
SNOMED CT -> HDG
SNOMED CT -> UMLS

P1: Manual curation
P2: Automatic mapping

Manual curation
Automatic mapping
Automated Terminology Network: ATN

OMIM

SNOMED 3.5

HDG

SNOMED CT

UMLS

Manual curation

Automatic mapping

Result: Quantitative analysis

Manual curation  Direct automated path  ATN mapping  Multi-Strategy

Precision (%) vs. Recall (%)
Future Directions

• to investigate the accuracies of combined MC and AM terminology paths

• to support compositional mapping

• to predict the accuracy of terminological pathways large-scale networks
Use of terminologies

- Semantic Classes, terminology mappings, underused tools in biomedical research
- A focus on knowledge technologies rather than information technologies is required to reach the ultimate goal: accelerate quality research.
- After well funded databases in genomics, proteomics, and eventually clinomics will likely require databases, information-bases (terminologies) and knowledge bases.

Take Home Points:

- A distinct role for controlled terminologies, reference terminologies and ontologies,
- *Semantic classes* provide original reuse and analysis of data via,
- *Automated terminology mapping* a tool for high throughput database interoperability and accelerated biomedical discovery.
Conclusions

• Knowledge Technologies are transforming the digital economy
  – From the “Article economy” to the “excerpt economy” and the “economy of network of excerpts”.
  – Redefining a digital library
  – Redefining the roles of librarians
Acknowledgements

Laboratory
• Ying Tao
• Michael Cantor
• Hedi Hegyi
• Xiaoyan Wang
• Huinar Quek
• Neil Sarkar
• Jianrong Li
• Inderpal Kohli
• Rony Gelman

Collaborators
• Carol Friedman
• Eneida Mendonça
• James Cimino
• Edward Shortliffe