Exploring the Semantics of Biomedical Ontologies

Francisco M Couto
12 April 2011
EBI External Seminar

New Entities

• Common step
  – Compare it with known entities
  – to transfer knowledge
• Common to any research area
  – Characterization of new proteins
  – Microarray Analysis
  – Information Retrieval in general

Computational Comparison

• Using digital representations
• Structural Similarity
  – Based on the syntax of the representations
  – Protein sequence (BLAST)
• Semantic Similarity
  – Based on the implicit and explicit semantics (Ontologies)
  – Protein molecular function (GO Analysis)

How to compare?

• Manually
  – Before computers
  – based on catalogues
• Problems
  – Digital Era
  – Information overload
  – UniProtKB/TreMBL (14 million sequences)
**Structural Similarity**

- Structure is “always” available
  - We normally start by identifying
    - how an entity look like
    - before finding what it does (meaning)
- Structure is less ambiguous
  - It’s easier to establish an agreement on describing
    - how an entity look like
    - than on what it does
- Performance and Scalability
  - String matching techniques (BLAST)

**Semantic Similarity**

- Solve the problems of structural similarity
  - Two entities that look similar
    - Do not imply
    - They have the same meaning
  - And two entities with similar meaning
    - Do not imply
    - that they look similar
- When we want to find similar entities
  - Based on their meaning
  - **Not on** how they look like

**Example: Similar Semantics**

![Caffeine](CHEBI:27732) ![Doxapram](CHEBI:681849)

- Different structure but
- both central nervous system stimulants (CHEBI:35337)

**Example: Similar Structure**

![Caffeine](CHEBI:27732) ![Adenosine](CHEBI:16335)

- Similar structure but
- different roles
  - adenosine is an anti-arrhythmia drug (CHEBI:38070)
  - nucleoside (CHEBI:38254)
Semantic description

• Is not always available
• Semantic Web Initiative
  – inserting machine-readable metadata
• Ontologies
  – Serve as metadata vocabularies
  – to describe entities (annotations)
• Biomedical Ontologies
  – Large effort and involvement from the community
  – 101 ontologies at http://www.obofoundry.org/
  – 82 million annotations from UniProtKB-GOA

Semantic similarity measures

• Input:
  – two ontology concepts
  – or two sets of terms annotating two entities
• Output:
  – a numerical value reflecting the closeness in meaning between them
Information Content

- measures how specific and informative a concept is

\[ IC(c) = -\log \left( \frac{freq(c)}{maxFreq} \right) \]

- Inversely proportional to frequency in a given corpus
- The frequency is also propagated to its ancestors
  - IC proportional to the depth of a concept
- Extrinsic IC
  - number of entities mapped to each concept
- Intrinsic IC
  - number of children

Example

<table>
<thead>
<tr>
<th>Concept</th>
<th>freq</th>
<th>hfreq</th>
<th>IC</th>
</tr>
</thead>
<tbody>
<tr>
<td>stimulant</td>
<td>1</td>
<td>6</td>
<td>0.0</td>
</tr>
<tr>
<td>Caffeine</td>
<td>3</td>
<td>3</td>
<td>0.3</td>
</tr>
<tr>
<td>Docapram</td>
<td>2</td>
<td>2</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Similarity based on IC

- Similarity proportional to the
  - IC of the most informative common ancestor (MICA)
    - Shared information between two concepts
    - Resnik
  - Weighted Jaccard index of two sets of concepts
    - Shared information between two entities
    - SimGIC

Example

- \( IC(\text{copper}) > IC(\text{coinage}) > IC(\text{metal}) \)
- \( Sim(\text{copper}, \text{gold}) \sim IC(\text{coinage}) \)
- \( Sim(\text{copper}, \text{palatium}) \sim IC(\text{metal}) \)
- \( Sim(\text{copper}, \text{gold}) > Sim(\text{copper}, \text{palatium}) \)
Applications

- Chemical Classification
- Chemical entity recognition and mapping
- Ontology matching and extension
- Enzyme family coherency assessment
- Epidemic and VPH information retrieval

Classification Problems

- BBB:
  - Do compounds cross the blood brain barrier?
- P-gp:
  - Are compounds substrates to the P-glycoprotein?
- estrogen:
  - Are compounds ligands to an estrogen receptor?

Structural Similarity

- Fingerprints and Machine Learning

<table>
<thead>
<tr>
<th>Testing set</th>
<th>Classification system</th>
<th>Accuracy</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>BBB</td>
<td>Artificial Neural Networks</td>
<td>75.7%</td>
<td>(Doniger et al., 2002)</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>80.9%</td>
<td>(Svrtnik et al., 2003)</td>
</tr>
<tr>
<td></td>
<td>Support Vector Machines</td>
<td>81.2%</td>
<td>(Doniger et al., 2002)</td>
</tr>
<tr>
<td>P-gp</td>
<td>Four-point Pharmacophore</td>
<td>62.7%</td>
<td>(Penucci et al., 2002)</td>
</tr>
<tr>
<td></td>
<td>Support Vector Machines</td>
<td>79.4%</td>
<td>(Xue et al., 2004)</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>80.6%</td>
<td>(Svrtnik et al., 2003)</td>
</tr>
<tr>
<td>estrogen</td>
<td>Decision Forest</td>
<td>~80%</td>
<td>(Tong et al., 2003)</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>82.8%</td>
<td>(Svrtnik et al., 2003)</td>
</tr>
</tbody>
</table>
Hybrid Approach

- Based on Semantic Similarity
  - CHEBI and KEGG pathways

\[ \text{sim}_{\text{hybrid}} = \alpha \cdot \text{sim}_{\text{structural}} + (1 - \alpha) \cdot \text{sim}_{\text{semantic}} \]

<table>
<thead>
<tr>
<th>Set</th>
<th>Claym Parameters</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{BBB}_p )</td>
<td>FP3, simGIC, all, ( \alpha = 0.28 )</td>
<td>90.9%</td>
</tr>
<tr>
<td>( \text{P-EP}_p )</td>
<td>FP4, simUI, all, ( \alpha = 0.66 )</td>
<td>87.7%</td>
</tr>
<tr>
<td>( \text{estrogen}_p )</td>
<td>FP4, simGIC, role, ( \alpha = 0.42 )</td>
<td>84.2%</td>
</tr>
</tbody>
</table>

New Predictions

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Set</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>1015</td>
<td>orthoaminic acid</td>
<td>( \text{BBB}_p )</td>
<td>0.503</td>
</tr>
<tr>
<td>2054</td>
<td>uracilamide</td>
<td>( \text{BBB}_p )</td>
<td>0.489</td>
</tr>
<tr>
<td>2060</td>
<td>L-methylsemicarbazide</td>
<td>( \text{BBB}_p )</td>
<td>0.477</td>
</tr>
<tr>
<td>3368</td>
<td>chloroquine</td>
<td>( \text{BBB}_p )</td>
<td>0.475</td>
</tr>
<tr>
<td>2430</td>
<td>acetamide</td>
<td>( \text{P-EP}_p )</td>
<td>0.474</td>
</tr>
<tr>
<td>1883</td>
<td>4-hydroxyxystereone</td>
<td>( \text{estrogen}_p )</td>
<td>0.577</td>
</tr>
<tr>
<td>5078</td>
<td>flurazol*</td>
<td>( \text{estrogen}_p )</td>
<td>0.577</td>
</tr>
<tr>
<td>5262</td>
<td>galangin</td>
<td>( \text{estrogen}_p )</td>
<td>0.577</td>
</tr>
</tbody>
</table>

* This compound is a false positive.

Structural vs. Semantic

<table>
<thead>
<tr>
<th>Alpha</th>
<th>( \text{BBB}_p )</th>
<th>( \text{P-EP}_p )</th>
<th>( \text{estrogen}_p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>81.0%</td>
<td>74.1%</td>
<td>73.3%</td>
</tr>
<tr>
<td>0.1</td>
<td>86.9%</td>
<td>74.1%</td>
<td>74.3%</td>
</tr>
<tr>
<td>0.2</td>
<td>88.9%</td>
<td>79.0%</td>
<td>74.3%</td>
</tr>
<tr>
<td>0.3</td>
<td><strong>90.2%</strong></td>
<td>76.5%</td>
<td>79.2%</td>
</tr>
<tr>
<td>0.4</td>
<td>88.2%</td>
<td>81.5%</td>
<td><strong>84.2%</strong></td>
</tr>
<tr>
<td>0.5</td>
<td>85.0%</td>
<td>81.0%</td>
<td>83.2%</td>
</tr>
<tr>
<td>0.6</td>
<td>83.0%</td>
<td>88.9%</td>
<td>78.9%</td>
</tr>
<tr>
<td>0.7</td>
<td>83.0%</td>
<td><strong>86.4%</strong></td>
<td>81.2%</td>
</tr>
<tr>
<td>0.8</td>
<td>81.0%</td>
<td>82.7%</td>
<td>78.2%</td>
</tr>
<tr>
<td>0.9</td>
<td>77.1%</td>
<td>81.0%</td>
<td>71.3%</td>
</tr>
<tr>
<td>1.0</td>
<td>71.9%</td>
<td>85.2%</td>
<td>79.2%</td>
</tr>
</tbody>
</table>

CHEMICAL ENTITY RECOGNITION AND MAPPING
A common example of a chemical substance is pure water; it has the same properties and the same ratio of hydrogen to oxygen whether it is isolated from a river or made in a laboratory. Some typical chemical substances are diamond, gold, salt (sodium chloride) and sugar (sucrose).

The scope of a scientific publication is typically narrow
- Specific protein, specific metabolic pathway, specific disease, etc
- Thus it is expected for the entities present in a document to be related
- Measure semantic similarity between the entities

Example
- “Despite a lack of data regarding their efficacy, both caffeine and doxapram have been recommended for treatment of hypercapnia in equine neonates with central nervous system damage." (PMID: 18372520)
- The fact that caffeine and doxapram are semantically similar
  - is an evidence for being correctly identified
Ontology Extension Approach

T1: Prediction of Ontology Extension
T2: Ontology Matching
T3: Ontology Learning From Text
T4: Evaluation
T5: Integration into Auxesia tool

Ontology Matching

• String Matching:

Structural level

• Similarity Propagation:

Semantic Similarity in Propagation

Dietary system Fluid secretion
Intestine 3475600

OAEI 2011 anatomy results

- Adult Mouse Anatomy (2744 classes)
- NCI Thesaurus (3304 classes) describing the human anatomy.

<table>
<thead>
<tr>
<th>System</th>
<th>precision</th>
<th>recall</th>
<th>f-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>AgreementMaker (F-measure)</td>
<td>90.3%</td>
<td>85.3%</td>
<td>87.7%</td>
</tr>
<tr>
<td>AgreementMaker (precision)</td>
<td>96.2%</td>
<td>75.1%</td>
<td>84.3%</td>
</tr>
<tr>
<td>BLOOMs-XLDB</td>
<td>96.7%</td>
<td>72.5%</td>
<td>82.9%</td>
</tr>
</tbody>
</table>

Goals

- Measure and Improve
  - functional annotation coherence of protein families.
- Enrich
  - under-annotated families with functional annotations.
- Classify
  - novel sequences into richly annotated protein families.

Approach

ENZYME FAMILY COHERENCY ASSESSMENT

CAZy: case-study

- Database specialized in catalytic modules of enzymes that degrade, modify or create glycosidic bonds and modules associated to carbohydrate adhesion

Plots

- Histograms plot frequency of protein pairs within SSM ranges
- Labels show most frequent term with highest IC for each similarity range

Semantic Similarity Analysis

EPIDEMIC AND VPH INFORMATION RETRIEVAL
Epidemiological and Virtual Physiological Human

- Epidemic
  - Diseases
  - Symptoms
  - Anatomy
  - Phenotypes
  - Chemical substances
  - Proteins/Genes
  - Clinical Records
  - Geospatial
  - Therapeutics
  - Vaccine
  - Transmission modes
  - Organisms
  - ...

- VPH
  - Diseases
  - Symptoms
  - Anatomy
  - Phenotypes
  - Chemical substances
  - Proteins/Genes
  - Clinical Records
  - Cellular Component
  - Gene Expression
  - Cell Type
  - Pathways
  - Radiology
  - ...

Information Retrieval

- Input
  - search keywords
- Output:
  - Ranked list of datasets and/or models
- Ranking Method
  - Semantic Similarity
  - Using multi-domain annotations
  - Mapping search keywords to ontology terms

Challenges

- Dataset and Model annotation
  - Ontology Selection
  - Manual or/and semi-automated
- Semantic Similarity
  - Extend to another ontologies
    - Nowadays: GO, CHEBI, HPO
    - Next: FMA
- Integrate multi-domain measures
  - Weights
OUR WEB TOOLS

cessm

- Platform to evaluate GO semantic similarity measure
- Provides a list of 13,430 protein pairs
  - 1,039 distinct proteins
  - Blast E-value $< 10^{-4}$
  - At least one Pfam and EC annotation
- Calculate their similarity with your measure
- Provides a quantitative comparison

web services

ontology/concept|instance|inputList|feature|parameter1|paramValue1/...
Thanks for your attention!

Biomedical Informatics research line
at University of Lisbon

http://xldb.fc.ul.pt/wiki/Biomedical_Informatics