Semantic Infrastructure for Automated Small Molecule Classification and Data Mining for Lipidomics

Christopher J.O. Baker
University of New Brunswick, Canada

CSHALS
Feb 24th, 2011, Boston, USA

bakerc@unb.ca
Activity Statement

• A Lipid Chemist has discovered a new compound and wants to identify which class of lipid it belongs to, what proteins it is related to, and what is known about this class of lipid.
## Semantic Infrastructure

### BACKGROUND

- Lipid Ontology
- LipidMaps Database
  - Smiles
- SADI framework
- DL-Reasoner

### FOREGROUND

- Lipid Eicosanoid Ontology
- SADI Semantic Web Services
  - Functional Group Annotation
  - Classification
  - Evaluation on Eicosanoids
  - Text Mining for Lipids
  - Lipid Protein Databases
Lipid Ontology: A History

2008

FIRST-ISWC 2007 - Early Text Mining / Simple Ontology / modeling lipid nomenclature

BMC Bioinf. 2008 - Baseline Ontology / Text Mining / Visual Query

ISMB 2008 - Text Mining / Apoptosis / Data Mining for Transitive Relations

OBO 2009 - Online Listing

AMIA 2009 - Text Mining / Ovarian Cancer

HS Low MSc 2009 - Formal Lipid Ontology

ICBO 2009 - OWL-DL Ontology for classification of Lipids

IntOnt 2010 - Lipid Ontologies

ACS 2010 - Semantic Chemistry - Corpus Annotation with Lipid Ontology Axioms

2009

CSHALS 2011 - Semantic Infrastructure for Automated Small Molecule Classification (Live DEMO)
Objectives

• **Formalize and represent:**
  – Lipid nomenclature and classification hierarchy in (OWL-DL)

• **Provide definitions:**
  – Independent of graphical descriptions
  – Amenable to inference and classification-based reasoning
  – Semantically explicit

• **Evaluate:**
  – a systematic and formalized OWL-DL definitions of lipids for testing appropriateness of existing nomenclature to lipid structures.
Lipids

• Lipids can be defined by their functional groups and other structural descriptors.
  
  Mycolic Acid

• Analysis, reuse and extension of existing chemical nomenclatures and ontologies.

• Definition of lipid classes using DL Axioms
  – Necessary (and sufficient) conditions
  – Cardinality
Reuse of existing ontologies.

- Limit Compound to Lipid
Functional Groups in Lipids

**Organic Group**

Total no. of simple organic group = 95 (2009), = 150 (2011)

Extensions required to support lipid characterization

Organic group From Chemical Ontology
Definition of lipid classes: DL Axioms

• As we traverse down the hierarchy
  • Specify more specific organic group classes
  • Specify cardinality

• hasPart some Carboxylic_Acid_derivative_Group
• hasPart some Aldehyde
• hasAldehyde exactly 3
Figure 2. DL-definitions of \textit{Alpha\_mycolic\_acid}, \textit{Fatty\_alcohol} and \textit{Docosanoid}
Docosanoid Example

LC_Fatty_Acyl
(hasPart some Carboxylic_Acid) and (hasPart some Alcohol) and (hasPart some Alkenyl_Group) and (hasPart some Cyclopentenone) and (hasAcyl_Chain exactly 1)

hasPart only (Carboxylic_Acid or Alcohol or Alkenyl_Group or Cyclopentenone or Acyl_Chain)
LiPro Ontology

Annotations: LC_Hydroxyeicosatrienoic_acid

Description: LC_Hydroxyeicosatrienoic_acid

Equivalent classes +

Superclasses +

LC_Eicosatrienoic_acid_derivative

Inferred anonymous superclasses

- Lipid
  - hasProperPart some Carboxylic_Acid_derivative_Group
  - hasProperPart some Primary_Acyl_Chain
  - hasProperPart some Alkyl_Chain

- LC_Fatty_Acyl
  - hasProperPart some Alkenyl_Group
  - hasProperPart some Carboxylic_Acid
  - hasProperPart exactly 1 Primary_Acyl_Chain

- LC_Eicosanoid
  - hasProperPart some (Alcohol
    - or Hydroperoxide
    - or Peroxide)
  - hasPart only (Alcohol
    - or Alkenyl_Group
    - or Carboxylic_Acid
    - or Hydroperoxide
    - or Peroxide
    - or Primary_Acyl_Chain)
  - hasProperPart exactly 3 Alkenyl_Group

Members +

Disjoint classes +
Axioms in LiPro and LEO

```
LC_Eicosanoid
  and hasProperPart some (Alcohol
    or Hydroperoxide
    or Peroxide)
  and hasPart only (Alcohol
    or Alkenyl_Group
    or Carboxylic_Acid
    or Hydroperoxide
    or Peroxide
    or Primary_Acyl_Chain)
  and hasProperPart exactly 3 Alkenyl_Group
```

```
LC_Eicosanoid
  and hasProperPart some Alkenyl_Group
  and hasProperPart some Carboxylic_Acid
  and hasProperPart some Primary_Acyl_Chain
  and hasProperPart some (Alcohol
    or Hydroperoxide
    or Peroxide)
  and hasProperPart exactly 3 Alkenyl_Group
```
Ontology-based Classification

Wolstencroft, Stevens, Haarslev (2007)

Chepelev, Riazanov, Kouznetsov, Low, Dumontier, Baker (2011)
Generating N3 Instances from LIPID - SMILES

- CheckMol (2005)
- Open Babel (2009)

Chepelev and Dumontier (2010)
Performance of Lipid Classification

- 150 lipids, correctly assigned to a class consistent to the LIPID MAPS asserted classes in 96% of lipid entities considered
- 8% of all lipids that received classification to a more general class in the classification hierarchy than the target class
- 4% of lipids received a classification that was entirely incorrect
Multiple classifications

- 7% of all lipids were classified into a LIPID MAPS-(in)consistent classes simultaneously e.g.

- **LMFA03010001:**
  - correctly classified as a *lipro:LC_Prostaglandin* and the LIPID MAPS class FA0301,
  - wrongly classified as a *lipro:LC_Isoprostane* (LIPID MAPS class FA0311).

- **Non-structural** information is required to differentiate the two classes - synthetic route.

- Insufficient class definitions / additional axioms required in Lipro (LEO)
Errors in manually curated Lipidmaps DB

- Acids LMFA03050020 and LMFA03050021 are classified as (FA0305) hydroxy/hydroperoxyeicosatrienoic acids in the LIPID MAPS database.

- In the Lipid Eicosanoid Ontology, hydroxyl/hydroperoxyeicosanoic class = LIPID MAPS FA0305 and has equivalent class expression.
  
  hasProperPart exactly 1 Primary_Acyl_Chain
  and hasProperPart some Carboxylic_Acid
  and hasProperPart exactly 3 Alkenyl_Group
  and hasProperPart some (Alcohol or Hydroperoxide or Peroxide)

- LMFA03050020 and LMFA03050021 possess: 3 alkenyl functional groups, a carboxylic acid functional group, and one epoxide on the main chain, but NO alcohol, hydroperoxide, or peroxide functional groups necessary to classify them into FA0305.
Resolved!

- LMFA03050020 and LMFA03050021 possess: three alkenyl functional groups, a carboxylic acid functional group, and one epoxide on the main chain, but NO alcohol, hydroperoxide, or peroxide functional groups necessary to classify them into FA0305.

- Our classification identifies them as Epoxideicosatrienoic acids where: Epoxideicosatrienoic acids / LIPID MAPS class (FA0308)

  hasProperPart exactly 1 Primary_Acyl_Chain
  and hasProperPart some Epoxy
  and hasProperPart some Carboxylic_Acid
  and hasProperPart exactly 3 Alkenyl_Group

- Reassignment required to class epoxyeicosatrienoic acid class, FA0308 was identified by explicit class definitions in our lipid ontology
Evaluation Framework

Input
- Lipid Molecule
- functional groups

LMSD ID

LMSD Database

Classification workflow

LIPRO/Ontology

LIPRO/LMSD Alignment

Evaluation module

LMSD Classification

Evaluation

Predicted Classification
Evaluation Sample

Agreement Level: 3-0=3
LiPro Prediction specificity Level: 5-3=2
LIPID MAPS specificity level: 4-3=1
SADI facilitates novel data discovery, interoperability, and integrative behaviours that closely mirror the needs and expectations of our end-user community simply by indexing services based on this predicate.

• Semantic Web data vs data derived from Web Service.
• SADI simply comprises a set of standards-compliant conventions and suggested best-practices for data representation and exchange between Web Services that fully utilizes Semantic Web technologies.

• SADI mandates the inclusion of a single required annotation in the Web Service metadata that describes the biological relationship ("predicate") that is created between the input and output data of that Service.
SADI’s Collaborative Architecture

SADI’s Semantic Open Collaborative Registry

SADI-Compliant Web Services
SADI’s Collaborative Architecture

Publish ontologies to, or select ontologies from The Web, that describe the service interface and function.

SADI’s Semantic Open Collaborative Registry

SADI-Compliant Web Services
SADI’s Collaborative Architecture

Notify the SADI registry that the Service exists

SADI’s Semantic Open Collaborative Registry

SADI-Compliant Web Services
SADI’s Collaborative Architecture

SADI analyses and indexes the “semantics” of each interface; (discover potential “collaborations” between independent services)

SADI’s Semantic Open Collaborative Registry

SADI-Compliant Web Services
SADI’s Collaborative Architecture

User comes along with data, and a desired outcome, and asks SADI to help find a solution.

SADI’s Semantic Open Collaborative Registry

SADI-Compliant Web Services
SADI’s Collaborative Architecture

SADI uses logical reasoning to match-make between the data and the semantics of each service, based on the desires of the User.
SADI’s Collaborative Architecture

SADI Registry provides a menu of choices

SADI’s Semantic Open Collaborative Registry

SADI-Compliant Web Services
SADI’s Collaborative Architecture

User interacts with their chosen service(s)

SADI-Compliant Web Services

SADI’s Semantic Open Collaborative Registry
FG Annotate Service

- Created by Leonid Chepelev and Michel Dumontier, Carleton University
- Takes a SMILES description of a molecule as input:
  \[ \text{CCC}[\text{C@H]}(O)/\text{C}=\text{C}/[\text{C@H}]1[\text{C@H]}(O)\text{C}[\text{C@H]}(O)[\text{C@H}]1\text{CC}(=\text{O})\text{CCCC}(=\text{O})\text{O} \]
- Identifies substructures that are functional groups:
  4 x Alkyl_Chain, 1 x Primary_Acyl_Chain, 1 x Carboxylic_Acid, ...
- Packaged as a SADI service: SMILES description of a molecule $\rightarrow$ lipid description with $\text{hasProperPart}$ and instances of functional groups
- Ready for programming-free integration with the lipid classification service: the use of LiPrO ensures semantic interoperability
FG Annotate Service I/O
Ontology-based Classification

- Express a description of a molecule in terms of identified participating functional groups as an OWL class:
  (hasProperPart exactly 1 Primary_Acyl_Chain) and
  (hasProperPart exactly 1 Carboxylic_Acid) and
  (hasProperPart exactly 4 Alkyl_Chain) and

- Use an OWL reasoner to test if this class is a subclass of a LiPrO class.

- Packaged as a SADI service: lipid description with hasProperPart and instances of functional groups \( \rightarrow \) predicted lipid class.

- Potentially, can be used (1) to classify lipids, possibly new, identified in high-throughput experiments or (2) to validate existing manual classifications.

- Can be much more useful in conjunction with some software that can detect functional groups in low-level chemical structure descriptions.
Classifier Service I/O

Input

Output

service:input2321

lipid has LiPro class

lipro:Isoprostane

has proper part

has proper part

has proper part

has proper part

isa

isa

isa

isa

x 4

lipro:Primry_Acyl_Chain

lipro:Alkyl_Chain

lipro:Carboxylic_Acid
CardioSHARE is a unique framework for querying distributed data and performing data analysis using Semantic Web standards. CardioSHARE's two main innovations are an enhancement to a standard SPARQL query engine, which enables the required data to be retrieved dynamically from Web Services, and the ability to use OWL class restrictions to drive the discovery and execution of Web Services capable of generating that class' defining properties, thus allowing naive data to be "lifted" into more complex OWL classifications. Both of these behaviours are accomplished by mapping predicates onto Web Services capable of producing RDF data that satisfy those predicates. Our initial focus has been on integration with the bioNoby project: a set of 15000+ interoperable biomarkers web services. CardioSHARE effectively brings this established pool of resources into conformance with Semantic Web standards. Given that much of the data from CardioSHARE is generated dynamically based on analysis of incoming query data, the effective size of the "virtual" triplestore is unmeasurable, limited only by the number of conceivable inputs.

To learn more about the system, please see our article CardioSHARE: Web Services for the Semantic Web.
Integration of FG Annotate an Classifier

SELECT ?liProClass ?LIPIDMAPSClass
FROM <http://unbsj.biordf.net/lipids/service-data/LMFA03010001.rdf>
WHERE
{
    # annotator service + classifier service
    # molecule structure -> functional groups -> LiPr0 lipid class

    # LiPr0 lipid class -> LIPID MAPS lipid class
}
Connecting Lipids to Proteins

- A biologist wants to look at proteins related to the identified lipid classes via metabolic or signalling pathways.
- LIPID MAPS provides the mapping. We wrap it as a SADI service: top-level lipid class → UniProt ID.
- Had to write another service to map a LIPID MAPS class to its top-level superclass (providing rdfs:subClassOf).

Five services are automatically combined to do the job: annotator → classifier → LiPrO to LM → LM top level → lipid-protein map. The user only deals with a declarative query.
FROM <http://unbsj.biordf.net/lipids/service-data/LMFA03010001.rdf>
WHERE
{
    # Same as in the previous queries:

    # LIPID MAPS lipid class -> top level LIPID MAPS lipid class
    ?LIPIDMAPSClass rdfs:subClassOf ?LMCategory .

    # top level LIPID MAPS lipid class -> protein URI
    ?LMCategory sio:isRelatedTo ?protein .
}
Future Work

1. Revisit Ontology Axioms
2. Revisit Functional Group annotation
3. Engage Lipid Community to assess classifications / improve evaluation script
4. Iterate of ontology improvements in 1 and 2
5. Extend beyond Eicosanoids
6. Deploy classification in real-time Lipidomics
Acknowledgments

National University of Singapore
• Markus R. Wenk
• Hong Sang Low

Institute of Infocomm Research
Singapore
• Rajaraman Kanagasabai
• Wee Tiong Ang

National University Hospital,
Singapore
• Mahesh Choolani
• Narishiman Kothandra

Carleton University, Canada
• Michel Dumontier
• Leonid Chepelev

University of New Brunswick, Canada
• Alexandre Riazanov
• Alexandre Kouznetsov
• Artjom Klein
• Jonas B Laurila

Concordia University, Canada
• Rene Witte

University of Bremen, Germany
• Alexander Garcia
Funding Sources

• A*STAR, Singapore
• NUS Office of Life Science (R-183-000-607-712),
• ARF (R-183-000-160-112),
• BMRC A*STAR (R-183-000-134-305),
• Singapore NRF under CRP award No. 2007-04,
• NBIF, New Brunswick, Canada.
• NSERC, Discovery grant. Canada
• Quebec -New Brunswick University Co-operation in Advanced Education - Research Program, Government of New Brunswick, Canada
• NSERC Discovery Grant, Canada 2009
• CANARIE, Canada 2009-2011