Self-Organizing Ontology of Biochemically Relevant Small Molecules

Leonid Chepelev, Michel Dumontier

Department of Biology, School of Computer Science, Institute of Biochemistry, Carleton University
Chemical Classification

• Structure dictates function.
• If represent chemical knowledge in terms of structure and function, then we can reason about the composition of biochemical networks.
  – By characterizing metabolic substrates, we will **automatically identify** potential enzyme substrates.
  – By characterizing biochemical roles, we will **automatically identify** toxic compounds, and explain their toxicity.
Believe it or not, the current approach to structural classification is manual assignment

I say ‘tis an alcohol, good sir!
• **Medical Subject Headings**

• **Hierarchically organized Controlled Vocabulary developed by the US National Library of Medicine (NLM)**
  
  – Used primarily for indexing of documents, MEDLINE abstracts from over 5400 journals

1. + Anatomy [A]
2. + Organisms [B]
3. + Diseases [C]
4. - Chemicals and Drugs [D]
   ○ Inorganic Chemicals [D01] +
   ○ Organic Chemicals [D02] +
   ○ Heterocyclic Compounds [D03] +
   ○ Polycyclic Compounds [D04] +
   ○ Macromolecular Substances [D05] +
   ○ Hormones, Hormone Substitutes, and Hormone Antagonists [D06] +
   ○ Enzymes and Coenzymes [D08] +
   ○ Carbohydrates [D09] +
   ○ Lipids [D10] +
   ○ Amino Acids, Peptides, and Proteins [D12] +
   ○ Nucleic Acids, Nucleotides, and Nucleosides [D13] +
   ○ Complex Mixtures [D20] +
   ○ Biological Factors [D23] +
   ○ Biomedical and Dental Materials [D25] +
   ○ Pharmaceutical Preparations [D26] +
   ○ Chemical Actions and Uses [D27] +
5. + Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]
6. + Psychiatry and Psychology [F]
7. + Phenomena and Processes [G]
8. + Disciplines and Occupations [H]
9. + Anthropology, Education, Sociology and Social Phenomena [I]
10. + Technology, Industry, Agriculture [J]
11. + Humanities [K]
Freely available database of curated chemicals, focused on those of interest to biology

- To replace proprietary sources
- Contains constitutionally or isotopically distinct atom, molecule, ion, ion pair, radical, radical ion, complex, conformer, etc
- Draws chemicals from IntEnz, KEGG compound, PDBeCHEM, ChEMBL
- Stores attributes: formula, mass, charge, synonyms

ChEBI Ontology
- Manually constructed ontology of chemical types
Objectives and Approach

• Objectives
  – To uncover the chemical signature of compound classes from annotated data
  – To automatically generate an integrated MeSH + ChEBI ontology

• Approach:
  – Formalize consensus fragments as OWL ontologies amenable to automated reasoning
Web Ontology Language (OWL)
Classes and class axioms

- **a class** is a set of individuals that share one or more characteristics
  - a protein

- classes can be organized in a hierarchy using **subClassOf** axioms
  - i.e. every member of C2 is a member of C1
  - subClassOf (protein molecule)

- special classes
  - **owl:Thing** is the superclass of all things
  - **owl:Nothing** is the subclass of all things, denotes an empty set

- classes can be made **disjoint** from one another
  - i.e. there is no member of C1 that is also a member of C2
  - disjointClasses (protein DNA)

- classes can be said to be **equivalent**
  - i.e. all members of C1 are members of C2 and all members of C2 are members of C1
  - EquivalentClass (Peptide Polypeptide)
Class Expressions

Class expressions are rich descriptions of classes through the logical combination of ontological primitives (classes, object properties, datatype properties, individuals)

Protein subClassOf
  molecule and ‘has direct part’ min 2 ‘amino acid residue’

Combinations specified using logical operators
  • conjunction (and), disjunction (or), negation (not)

Object or data property expressions provide a qualified cardinality over the relation
  o minimum: \( rel \) min \( \# \) \( Y \)
  o maximum: \( rel \) max \( \# \) \( Y \)
  o exact: \( rel \) exactly \( \# \) \( Y \) (minimum + maximum)
  o some: \( rel \) min 1 \( Y \)
Class Expressions

- The quantifications can be qualified by the object type
  - rel only Y – the only values allowed are of type Y

- To form complex class expressions like
  - 'molecule' and not 'dna'
  - 'has direct part' min 2 'amino acid'
  - 'is located in' only ('nucleus' or 'cytoplasm')

- and be expressed as axioms in the ontology
  - Protein subClassOf molecule and 'has direct part' min 2 'amino acid residue'
  - Transcription Factor equivalentTo 'protein'
    - and 'has disposition' some 'to bind to DNA'
    - and 'has function' some 'to regulate gene expression'
Methods

**Data:**
60 MeSH classes + 766 PubChem compounds
40 ChEBI classes + 606 3 star ChEBI compounds

**Fragmentation:**
up to 4 bonds
SMILES/SMARTS canonicalization (OpenBabel)

**Consensus:**
maximal common substructures using binary combinations of five smallest class members
Chemistry Development Kit (CDK)

**Formalization:**
URIs generated from SHA1 Hash
Semanticscience Integrated Ontology (SIO)
Axioms generated with the OWL API

**Reasoning:**
Pellet
Getting Consensus
Significantly easier with chemicals 😊

| Fragment | Contained |
|----------|-----------|
| C        | y         |
| C=O      | y         |
| CC       | y         |
| CC(=O)N  | y         |
| CC(=O)O  | y         |
| CC=O     | y         |
| CCN      | y         |
| CCO      | y         |
| CN       | y         |
| CNC      | y         |
| CNC(=O)C | n         |
| CNC=O    | y         |
| CNCC     | y         |
| CO       | y         |
| N        | y         |
| NC=O     | y         |
| NCC(=O)O | n         |
| NCC=O    | y         |
| NCCO     | y         |
| O        | y         |
| OC=O     | y         |

Principal Characteristic Sub-Graph A

Principal Characteristic Sub-Graph B
Self-Assembling Hierarchy

Feature List

| Feature   |
|-----------|
| 1 NCC(=O)O |
| 2 CC      |
| 3 CCO     |
| 4 CC=O    |
| 5 CO      |
| 6 C=O     |
| 7 CCN     |
| 8 OC=O    |
| 9 CN      |
| 10 CC(=O)O |
| 11 O      |
| 12 N      |
| 13 C      |

Organic Molecular Entity

Carboxylic Acid

Amino Acid

Biological Amino Acid
‘Flat’ Ontology
Reasoning unveils the MESH Hierarchy
Reasoning unveils the ChEBI Hierarchy
Automatic Integration: MeSH + ChEBI
Evaluation

- Test: recapitulate the original classification
  - 200 randomly selected molecules

- Results
  - 91% exactly matched assigned annotations
  - 8.5% discovered new, correct annotations
  - 92.7% correct annotations; errors found in those types requiring negation (e.g. in classical chemistry, an alcohol is not a carboxylic acid, although it is structurally correct)
Application: Enzyme Substrate Specificity

• BRENDA – database of enzyme kinetics
• 23 substrates for yeast alcohol dehydrogenase
• Obtained the consensus fragment as equivalent to that for alcohols
Structure-Based Classification

• Obtain formal definitions for controlled vocabularies used as entity annotation
• High degree of accuracy in uncovering consensus attributes, provided that the annotations are correct
• Enormous potential for application for biological analysis and understanding
• Attribute-based descriptions reduce manual labour to providing formal definitions.
• Flexible, re-usable, extensible.
Michel Dumontier
michel_dumontier@carleton.ca

Publications: http://dumontierlab.com
Presentations: http://slideshare.com/micheldumontier