On the Use of a Molecular Structure Ontology

Henry Dayringer, Matthias Nolte, Leiming Zhu, and Chris Waller
<table>
<thead>
<tr>
<th>Client Tier</th>
<th>Mid-tier Bits</th>
<th>Data and Service Tier</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Mining and Access</strong> (RGate)</td>
<td><strong>Server-side Data Integration</strong> (PLP, etc.)</td>
<td><strong>CPMart</strong></td>
</tr>
<tr>
<td><strong>Compound Design</strong> (PGVL, MoViT)</td>
<td><strong>Service Orchestration</strong> (BizTalk, etc.)</td>
<td><strong>SARMart</strong></td>
</tr>
<tr>
<td><strong>eNotebook</strong> (CeN, Symyx)</td>
<td></td>
<td><strong>Pfizer File</strong></td>
</tr>
<tr>
<td><strong>Project Team Collaboration</strong> (SharePoint)</td>
<td></td>
<td><strong>Global Monomers</strong></td>
</tr>
<tr>
<td><strong>Data Query Engine</strong></td>
<td></td>
<td><strong>Property Calculation Engine</strong></td>
</tr>
<tr>
<td><strong>Chemical Search Service</strong></td>
<td></td>
<td><strong>Model Server</strong></td>
</tr>
<tr>
<td><strong>Computational Services</strong></td>
<td></td>
<td><strong>External Data Web Services</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Literature/Patent/ Data Feeds</strong></td>
</tr>
</tbody>
</table>

User “Tier”
Outline

- What Problem are We Solving?
- What is an InChI?
- What is an Ontology?
- Description of the Molecular Structure Ontology (MSO)
- Practical Application of the MSO Technology
- Conclusion?
Structure Relationship ?s

**Structure concordance:** “What sample can I substitute in this assay?”
- Very strict business rules
- Applies only to internal compounds

**Related structure:** “What other compounds are identical (or very nearly so) to compound X?”
- Find related assay data
- Correlate internal with external compound sources
- Supports a wide variety of business processes
The IUPAC International Chemical Identifier (InChI™) is a non-proprietary identifier for chemical substances that can be used in printed and electronic data sources thus enabling easier linking of diverse data compilations. It was developed under IUPAC Project 2000-025-1-800 during the period 2000-2004. Details of the project and the history of its progress are available from http://www.iupac.org/web/ins/2000-025-1-800.

Project Goal: Develop a set of algorithms for the standard representation of chemical structures that will be readily accessible to chemists in all countries at no cost. The standard chemical representation could be used as input into existing and newly developed computer programs to generate a IUPAC name and a unique IUPAC identifier.

Source: http://www.iupac.org/inchi/
Six InChI Layers
- Main layer
- Charge layer
- Stereochemical layer
- Isotopic layer
- Fixed-H layer
- Reconnected Layer

Ambiguous information and specific information can be encoded. Search relevant “layers” for different levels of concordance.
Ontology

- The term *ontology* has its origin in *philosophy*, and has been applied in many different ways. The core meaning within *computer science* is a model for describing the world that consists of a set of types, properties, and relationship types. Exactly what is provided around these varies, but they are the essentials of an ontology. There is also generally an expectation that there be a close resemblance between the real world and the features of the model in an ontology.[3]

- What ontology has in common in both computer science and in philosophy is the representation of entities, ideas, and events, along with their properties and relations, according to a system of categories. In both fields, one finds considerable work on problems of ontological relativity (e.g., Quine and Kripke in philosophy, Sowa and Guarino in computer science)[4] and debates concerning whether a normative ontology is viable (e.g., debates over foundationalism in philosophy, debates over the Cyc project in AI). Differences between the two are largely matters of focus. Philosophers are less concerned with establishing fixed, controlled vocabularies than are researchers in computer science, while computer scientists are less involved in discussions of first principles (such as debating whether there are such things as fixed essences, or whether entities must be ontologically more primary than processes).

Source: http://en.wikipedia.org/wiki/Ontology_(information_science)
Molecular Structure Ontology

A derived set of relationships between chemical compounds based on their chemical structures.

Specifically: The result of comparisons of InChI string (or parts thereof) representations of chemical structures for the purpose of finding “related” compounds in a large set of chemical structures.
MSO Generation

- Break MOL file into bonded fragments
- For each fragment, generate Pipeline Pilot unique tautomer
- Generate InChI string for each fragment
- Truncate InChI just before the beginning of stereo information layer
- Generate an integer hashcode for the full and truncated InChI string (64 bit, collision rate < 1:20,000,000)
- Register (based on full string and its hash code) unique InChI strings, assign “InChI ID”
- Register “composition” of molecules based on InChI IDs, store repeat counts and assigned “type” (parent, salt, solvent…) keyed to “compound ID”
MSO Based Lookup

- Get “InChI ID” from “compound ID” or from structure
- Get all “truncated” hashcodes that exist with this “InChI ID”
- Get all “InChI IDs” that can produce any of the found “truncated” hashcodes (expands to all registered full InChI strings that yield the same truncated InChI string)
- Get all compounds that are known to contain any of these full “InChI IDs” (with assigned type of parent)
- Assign the relationship between the input structure and the found structures based on differences in the full InChI strings (different stereo, mirror image, different salt, etc).
<table>
<thead>
<tr>
<th>Related Molecules</th>
<th>Partial InChI string</th>
<th>Full InChI string</th>
</tr>
</thead>
<tbody>
<tr>
<td>RS-239543</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PF-04129950</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MN-014895</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP-598331</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RS-010227</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RS-005715</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RS-331086</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PNU-00070250</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UK-007553</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ST-010837</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AG-052510</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PF-01505177</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MN-002049</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0624813</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VC-99999999</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

```
InChI=1/C12H17NO2/c1-9(2)11(13)12(14)15-8-10-6-4-3-5-7-10/h3-7,9,11H,8,13H2,1-2H3
```
Conceptual: Building a Chemistry Knowledge-Base

Data Mining

Knowledge Discovery

Knowledge Presentation

Data Integration

Knowledge Access

Mapping

Abstraction

Attributes

Descriptors

D1

D135

D47

A1

... A53

... -2, -0.045, -0.2, 4.5x10e4, 74.2, 0.742

µM, nM, pM

P(efficacy)

-log10 Affinity

mM
Conceptual: Building a Chemistry Knowledge-Base
Practical Application of the MSO

Tools and opportunities for making the most of the Pfizer structure collection through the MSO technology
The Pfizer structure collection is an aggregate of both internal and external compounds. The internal compounds derive from the aggregation of the compounds of the many legacy companies from which Pfizer is derived. The external compounds derive from a variety of public sources, many of the compounds have associated biological activity data.

The Pfizer internal collection has about 11,170,000 distinct parent IDs representing 7,365,000 distinct parent compound structures.

The external collection has about 10,104,000 distinct parent IDs representing 7,800,000 distinct parent compound structures.

The total collection of 21,275,000 parent IDs representing 14,105,000 distinct compound structures.
The Molecular Structure Ontology

The service provides the following groupings of compounds related to a starting compound:

- All compounds that are identical to the starting compound, ignoring stereochemistry, double bond geometry and isotopic labeling
- All compounds with the same bonded framework of atoms (with three sub-groupings)
MSO Relatedness Layers

Framework compounds (experimental)

Three categories:
1. Same formula and bonded framework
2. Same framework, different formula
3. Same formula, different framework

Starting compound

MSO compounds
Closely related. Same atoms, same bonds. Might differ in stereochemistry, double bond geometry or isotopic labeling
MSO Related Compounds

- Begin with a compound (by ID, etc)
- Find the compounds very closely related to this compound
- For example, start with Chantix®

InChI=1/C13H13N3/c1-2-16-13-5-11-9-3-8(6-14-7-9)10(11)4-12(13)15-1/h1-2,4-5,8-9,14H,3,6-7H2/t8-,9+
MSO Related Compound

- 28 related MSO related compounds including:
  - 3 concorded IDs (business rules apply)
  - 13 identical structures (same salt and stereo, fail some business rule for concordance)
  - 9 that are a different salt
  - 2 with a different chirality
  - 1 labeled compound
MSO Related Examples

Diff salt

Diff stereo

Different isotopes

Identical parent
Current Status: MSO Service

- Current processes about 400,000 requests per day (mostly SMILES string input)
- Used to identify existing compounds in database similar to proposed compounds in new synthesis libraries
- Used to help correlate compounds in research databases (external to internal, external to external) – more powerful than canonical SMILES and much faster.
- Average processing time is about 250 ms per compound.
Experimental – Don’t Try This at Home!

New service offering, still in development
Framework Related Compounds

“Framework” derived by making all bonds single and all atoms = carbon:
Framework Generation

- Discard all frameworks that reduce to methane (single atom salts, water, etc).
- From the InChI of the framework, generate a “hashcode”.
- Register the full InChI string and the hashcode for unique frameworks.
- Register each unique molecular formula.
- Register the composition of each molecule correlated to its framework and formula.
- 20,000,000 compounds yielded about 7,000,000 unique frameworks.
Maximum frequency is 465,000 occurrences for ethane
Next most frequent is 71,000 occurrences for n-butane
About 1,100 (of about 7,000,000) frameworks occur 500 times or more
Framework Searching

- From the input ID (or structure), get the framework ID assigned and the molecular formula of the original structure component.
- Find all other compounds in the database with the same framework ID or the same formula.
- Return three sets:
  - Same framework and formula
  - Same framework, different formula
  - Same formula, different framework
Same Framework, Same Formula

- There are 13 compounds in database with the same framework and formula (after removing those related by the MSO) as the input molecule: Chantix®
- 6 have EXT IDs, 7 have internal IDs.
- Some examples (similarity scores calculated using the Accelrys SDK):

Red highlights the areas of “change” between reference and comparison compound

Similarity=.57
Same Framework, Different Formula

(23 compounds identified, some examples show below)

Similarity=.23

Similarity=.42

Similarity=.63

Similarity=.88
Same Formula, Different Framework

766 compounds identified, some examples show below. Highlighted part comes from a “lenient” maximum common substructure analysis.

Similarity=.11

Similarity=.19

Similarity=.29

Similarity=.31
Preliminary Conclusions

- The MSO service provides a unique capability to locate compounds related to a starting compound of interest.
- The service is very fast
- The capability is based on the use of the novel properties of the InChI structure representation
Data Mining and Access (RGate)

Compound Design (PGVL, MoViT)

eNotebook (CeN, Symyx)

Project Team Collaboration (SharePoint)

Server-side Data Integration (PLP, etc.)

Service Orchestration (BizTalk, etc.)

Data Query Engine

Chemical Search Service

Computational Services

Data and Service Tier

CPMart

SARMart

Pfizer File

Global Monomers

Property Calculation Engine

Model Server

HTSMart

Literature/Patent/Data Feeds

External Data Web Services
Compound ID: viagra

Options:
- As Parent
- As Salt
- As Structure
- As SMILES

Reference Structure

Compare To (MFCD10566459)

Service returned 98 compounds

Filter By:
- Exclude Geometric
- Exclude Stereo
- Exclude Mirror Image
- Exclude Isotopic
- Exclude Identical
- Exclude Internal
- Exclude External
- Exclude EXT Prefix
- Exclude Concordance
- Exclude Diff Salt
- Exclude Diff Solvent
- Exclude Diff Other
- Exclude Multiple Parents
- Match Stereo

Table:

<table>
<thead>
<tr>
<th>ParentID</th>
<th>CompoundID</th>
<th>Match Type</th>
<th>Global ID</th>
<th>Global Source</th>
<th>Global URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXT-0010179516</td>
<td>EXT-0010179516</td>
<td>Identical parent</td>
<td>93706</td>
<td>STARLITE</td>
<td>Addl Data</td>
</tr>
<tr>
<td>EXT-0010545714</td>
<td>EXT-0010545714</td>
<td>Identical parent, Diff other</td>
<td>5543276</td>
<td>IBMPAT</td>
<td>Addl Data</td>
</tr>
<tr>
<td>MFCD00867528</td>
<td>MFCD00867528</td>
<td>Identical parent, Diff salt</td>
<td>MFCD00867528</td>
<td>MLDACD</td>
<td>Addl Data</td>
</tr>
<tr>
<td>MFCD02102122</td>
<td>MFCD02102122</td>
<td>Identical parent, Diff salt</td>
<td>MFCD02102122</td>
<td>MLDACD</td>
<td>Addl Data</td>
</tr>
<tr>
<td>MFCD08063394</td>
<td>MFCD08063394</td>
<td>Diff isotopes (labeled)</td>
<td>MFCD08063394</td>
<td>MLDACD</td>
<td>Addl Data</td>
</tr>
<tr>
<td>MFCD09753069</td>
<td>MFCD09753069</td>
<td>Identical parent, Diff salt</td>
<td>MFCD09753069</td>
<td>MLDACD</td>
<td>Addl Data</td>
</tr>
<tr>
<td>MFCD0941266</td>
<td>MFCD0941266</td>
<td>Diff isotopes (labeled)</td>
<td>MFCD0941266</td>
<td>MLDACD</td>
<td>Addl Data</td>
</tr>
<tr>
<td>MFCD10566459</td>
<td>MFCD10566459</td>
<td>Identical parent, Diff salt</td>
<td>MFCD10566459</td>
<td>MLDACD</td>
<td>Addl Data</td>
</tr>
<tr>
<td>MFCD10567103</td>
<td>MFCD10567103</td>
<td>Diff isotopes (labeled), Diff salt</td>
<td>MFCD10567103</td>
<td>MLDACD</td>
<td>Addl Data</td>
</tr>
<tr>
<td>PDB-VIA</td>
<td>PDB-VIA</td>
<td>Identical parent</td>
<td>PDB-VIA</td>
<td>Pfizer</td>
<td>Addl Data</td>
</tr>
<tr>
<td>PF-00345409</td>
<td>PF-00345409</td>
<td>Identical parent</td>
<td>PF-00345409</td>
<td>Pfizer</td>
<td>Addl Data</td>
</tr>
<tr>
<td>PF-01000000</td>
<td>PF-01000000</td>
<td>Concordance</td>
<td>PF-01000000</td>
<td>Pfizer</td>
<td>Addl Data</td>
</tr>
<tr>
<td>PF-01000000</td>
<td>PF-01000000-10</td>
<td>Concordance</td>
<td>PF-01000000-10</td>
<td>Pfizer</td>
<td>Addl Data</td>
</tr>
<tr>
<td>PE-01000000</td>
<td>PE-01000000-10</td>
<td>Concordance</td>
<td>PE-01000000-10</td>
<td>Pfizer</td>
<td>Addl Data</td>
</tr>
</tbody>
</table>
Review: What Have We Done With This?

- Simplified complex structure-based queries (any atom, any bond, S-groups, etc.)
- Simplified complex business rules from an information technology perspective (no Cheshire, no structure checker scripting, no complex stereochemistry codes, etc.)
- Simplified complex framework analysis methodology…
Future: What Might We Do With This?

- Simplify complex Oracle structure cartridge technology ("grep" 🙂)
  - Discontinuous substructure (pharmacophore) search issues...
- Simplify the complex patent searching/filing/defending process
- To be determined…
Thanks

Henry Dayringer, Matthias Nolte, Leiming Zhu, and Chris Waller

chris.waller@pfizer.com