FROM MOLECULAR FORMULAS TO MARKUSH STRUCTURES

DIFFERENT LEVELS OF KNOWLEDGE REPRESENTATION IN CHEMISTRY

Michael Braden, PhD

ACS / San Diego/ 2016
Overview

• ChemAxon – Who are we?
• Examples/use cases:
  – Create and Report: Marvin Live
  – Store and Search: Biomolecule Toolkit
  – Analyze and Report: Markush and ChemCurator
• The Big Picture
Pharmaceutical and Biotech Industry – Drug Discovery

- Small molecules
- Biological entities
- Create, search, store, analyze, and share structural and related data

General Chemistry

Polymer Chemistry

Petrochemistry

Agrochemistry

Flavors and Fragrances

Analytical Chemistry
Vision

- Integrated
- Entity and data agnostic
- Collaborative
- Fast
- Accurate
- Ease of use
- Contextual
Levels of Chemical Information Representation

- Constitutional – 0/1D bulk properties
- Topological – 2D structures/graphs and fingerprints
- Geometric – 3D structures
- Energetic – response to probing 3D structures
CREATE AND REPORT

Marvin Live – Molecule Design and Real Time Collaboration
Use Case 1 - Industry

- Company A - bringing in consultants every week
- Company B - multiple US East Coast sites, frequent travelling between
- Company C - global enterprise, facilities on 3 continents
- Lack of co-location
  - Knowledge silos
  - Duplicate effort
Data access plugins
Mining available data
Explore property space

Calculated Properties
- Mass: 227.26
- cLogP: 2.96
- TPSA: 86.15 Å²
- pKa (str. acidic): 4.5
- pKa (str. basic): 3.3
- FSP3: 0.07
- Solubility: -4.29 logS
- H-bond acceptors: 3
- H-bond donors: 2

Compound registry
Similarity matches: 235
Use Case 2 - Academia

- University professor’s “office hours”
- Tutoring, working on publications
Reporting

Mark 20m
hi guys

Wilhelm 18m
hi Mark

Mark 14m
i think #5 should be the top item for next project

Kathie 2m
agreed

CJ 2m
agreed

Send a message... Send

Bob: check IP

check IP

Kathie: synthesize

save for later

do tox study.. but why

M. Braden / ACS / San Diego / 2016
Reporting - Plugin

- automatic capture (who, what, when)
- save important structures
  - for presentations
  - to desktop tools (SD file, SMILES)
  - to ELN, compound registration
STORE AND SEARCH

Biomolecule Toolkit – Draw, Store, and Search Biomolecules
What is the informatics problem?

Cheminformatics

Sequence based tools

Chemical structure-based tools

Bioinformatics

Representation?

Tools?

M. Braden / ACS / San Diego / 2016
Unmet needs

• Store the description of a diverse set of entities relevant for R&D in a single data management environment
• access stored entities with tools common in cheminformatics and bioinformatics
• interconvert between domain-specific file format standards
• pass data between collaborators in different domains
Solution: Biomolecule Toolkit

- Complete, consistent, unambiguous, machine-readable descriptions for any molecular and non-molecular asset in a research and development setting
  - Biomolecules described as sets of attribute value pairs
  - Molecular entities described as combination of chemical building blocks
  - Rules for connecting the building blocks are defined at the atomic level for providing full chemical detail on the resulting Biologic
  - Ambiguity is supported at various levels
  - Ability to complement description with metadata
Use case: Search

- Sequence-based search
- Chemical substructure search
- Keyword search

Query
Substructure: \(\text{H}_2\text{N}^-\text{CH}_3\)
AND
ProjectID = "PID2"
Search example: lysine derivatives, 6-Aha, SMCC linker
<table>
<thead>
<tr>
<th>CID</th>
<th>Name</th>
<th>Helm</th>
</tr>
</thead>
<tbody>
<tr>
<td>CXN1</td>
<td>tagged peptide</td>
<td>RNA1[R(G)R(A)R(P(T)R(P(T)R(P(A)R(P(C)RPA)]PEPTIDE1[S.I.G.N.A.L]CHEM1:SMCC$RNA1,CHEM1,1:R1-1:R1</td>
</tr>
<tr>
<td>CXN5</td>
<td>Circulin C</td>
<td>PEPTIDE1[A.G.C.S.K.K.K.K.Y.R.N.G.I.P.C.G.E.S.C.V.F.I.P.C.I.T.S.V]$PEPTIDE1,PEPTIDE1,17:R3-3:R3</td>
</tr>
<tr>
<td>CXN27</td>
<td>unnatural replacement</td>
<td>PEPTIDE1[S.Y.N.T.H.E.T.I.O[C(Aha)]$</td>
</tr>
</tbody>
</table>

Search example: lysine derivatives, 6-Aha, SMCC linker
<table>
<thead>
<tr>
<th>CID</th>
<th>Name</th>
<th>Helm</th>
</tr>
</thead>
<tbody>
<tr>
<td>CXN1</td>
<td>tagged peptide</td>
<td>RNA1[R(G)R(P(R(T)R(T)R(P)R(A)R(R(C)R(A))]PEPTIDE1[S.G.N.A.L]CHEM1:SMCC:SRNA1,CHEM1,1:R1-1::R1</td>
</tr>
</tbody>
</table>

Search example: lysine derivatives, 6-Aha, SMCC linker + ProjectID=“PID2“
Architecture

- Registration Client
- Plexus Connect
- JChem for Office
- Instant JChem

Web Service Layer
JChem Base / Biomolecule Toolkit API
Data Access Layer
Registration DB
ANALYZE AND REPORT

Markush and ChemCurator – Analyze Chemistry in Papers and Patents
Markush Representation

- R-groups
- Atom lists
- Bond lists
- Position variations
- Link nodes
- Repeating units
- Homology groups
Markush technologies

- Search
- Enumeration
- Hit visualization
- Non-hit visualization
- Overlap
- Composer
Non-hit visualization

Differentiating structure parts visualization in target Markush and Query structure.

Markush Editor

ChemCurator

UGM / San Diego / 2015
Markush Overlap

Overlapping chemical space calculation

- Percentage of overlap
- Overlapping Markush

UGM / San Diego / 2015
Markush Composer
Computer-assisted chemical data extraction

- English, Chinese and Japanese N2S
- Structure Checker
- Markush Editor
- Search and representation
General document curation

- Files (XML, PDF, HTML)
- Google Patents
- IFI CLAIMS
- Images (CLiDE & OSRA)
Compounds extraction view

Project explorer
Compound list
Annotated document
Selected structures

UGM / San Diego / 2015
Markush extraction view

Annotated document

Markush editor

Structure checker

Selected structures

Example structures
THE BIG PICTURE
The Big Picture

• What should we be doing now?
• How should we be doing it?
• Where are things going in the future?