Cheminformatics in R for Analyzing Chemical Genomics Screens

Tyler Backman and Thomas Girke

December 10, 2012

- Introduction
- CMP Structure Formats
- Similarity Searching
  - Background
  - Fragment Similarity Search Methods
  - Structural Descriptors
  - Similarity Coefficients
  - Alternatives
- CMP Properties
- CMP Libraries
- References

Outline

Introduction

CMP Structure Formats

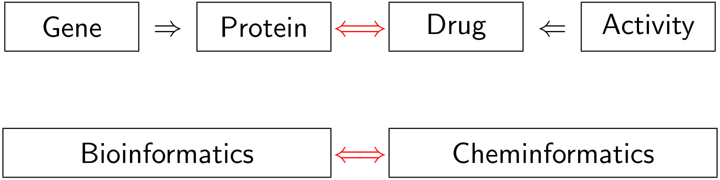
Similarity Searching
- Background
- Fragment Similarity Search Methods
- Structural Descriptors
- Similarity Coefficients
- Alternatives

CMP Properties

CMP Libraries

References

Informatics in Chemical Genomics



Gene ⇒ Protein ⟺ Drug ⇐ Activity

Bioinformatics ⟺ Cheminformatics

Why Compound Analysis in R

- Open source approach
- Numeric nature of all compound analyses
- Efficient data structures
- Access to unlimited number of clustering tools
- Expandability: programming environment

Outline

Introduction

CMP Structure Formats

Similarity Searching

- Background
- Fragment Similarity Search Methods
- Structural Descriptors
- Similarity Coefficients
- Alternatives

CMP Properties

CMP Libraries

References

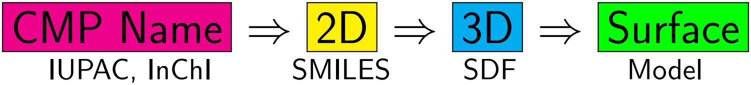
Structure Formats

Computer Readable Representations of Chemical Compounds

Differences in Computing Biosequences and CMPs

- DNA/proteins
  - Linear strings, one connection type, usually no branch points or ring closures
- Compounds
  - Several connection types, many branch points and/or ring closures

Utility of Stucture Formats



CMP Name ⇒ 2D ⇒ 3D ⇒ Surface

IUPAC, InChI — SMILES — SDF — Model

- Nomenclature to uniquely represent chemicals
- Computer representation and manipulation
- Format interconversions
- Representation of stereochemistry and 3D formats

Different Names for Different Purposes

Trivial and Brand Names
: Short, easy to pronounce names that lack chemical information. Often ambiguous and not very precise.

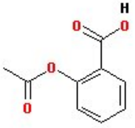
IUPAC
: Unambiguous naming conventions defined by the International Union of Pure and Applied Chemistry (IUPAC). Not very useful for computational approaches.

InChI
: InChI (International Chemical Identifier) is the latest and most modern line notation. It resolves many of the chemical ambiguities not addressed by SMILES, particularly with respect to stereo centers, tautomers, etc.

Most Commonly Used Structure Formats

- Chemical nomenclature
  - Trivial names: aspirin, acetylsalicylic acid
  - IUPAC: 2-acetoxybenzoic acid
  - InChI: 1.12Beta/C9H8O4/c1-6(10)13-8-5-3-2-7(8)9(11)12/h1H3,2-5H,(H,11,12)
- Line notations
  - SMILES: CC(=O)Oc1ccccc1C(=O)O
  - Other: WLN, ROSDAL, SLN, etc.
- Connection tables hold 3D & annotation information
  - SDF (structure definition file)
  - MDL Molfile
  - Other: PDB, CML, etc.

Aspirin



SMILES

SMILES: Simplified Molecular Input Line Entry System

- Tutorial: http://www.daylight.com/smiles/smiles-intro.html
- Online rendering: http://www.daylight.com/daycgi/depict
- Non-canonical SMILES for manual entry
- Canonical SMILES needs to be computer generated
- Canonicalization: single ('correct') representation of several posibilities
  - OCC - ethanol
  - CCO - ethanol
- Canonical format important for databases

SSMILES

SSMILES is an extremely simplified subset of SMILES that consists only of four rules:

1. Atoms are represented by atomic symbols
2. Double bonds are '=', triple bonds are '#'
3. Branching is indicated by parentheses
4. Ring closures are indicated by pairs of matching digits.

SMILES Rules 1

C

Methane: CH_4. Hydrogens are added according to valence rules.

N-C=O

Formamide. Single '-', double '=', triple '#' and aromatic bond ':'.

NC=O

Formamide. Bonds do not need to be specified in unambiguous cases.

NC(CO)=O

2-hydroxyacetamide. Side-chains of branch points in parentheses. The leftmost atom inside parentheses is attached to the atom to the left of the parentheses.

C1CCNCC1

Piperidine. If there is a ring, a matching pair of digits means that the two atoms to the left of the digits are bonded.

SMILES Rules 2

c1ccccc1O

Phenol. Aromatic atoms are represented as lowercase letters. Note also that the bonds default to aromatic and single, as appropriate.

[Pb]

Lead. The typical organic atoms, B, C, N, O, P, S, F, Cl, Br, are drawn without brackets. All other elements must have square brackets, and all their bonds including hydrogens must be specified.

[OH-]

Unusual valence and charge are represented in square brackets '[]'.

c1ccccc1[N+](=O)[O-]

Nitrobenzene. Another example using square brackets to be specific about charge location.

SMILES Rules 3

[Na+].[O-]c1ccccc1

Sodium phenoxide. The '.' (period or "dot") is used to represent disconnections.

[13CH4]

Isotopes are specified in brackets by prefixing the desired integral atomic mass. Connected hydrogens must be specified in brackets.

F/C=C/F

Trans-difluoroethene. Cis/trans configurations around double bonds are specified by slashes: 'C/C=C\C' (cis) and 'C/C=C/C' (trans).

N[C@@H](C)C(=O)O

L-alanine (from N, H-methyl-carboxy appear clockwise). Chirality is specified with '@' and '@@'. @ means anti-clockwise and @@ means clockwise.

N[C@H](C)C(=O)O

D-alanine (from N, H-methyl-carboxy appear anti-clockwise).

SMARTS Is a Query Expression System for SMILES

SMARTS: SMiles ARbitrary Target Specification

- Motivation: superset of SMILES to expresses molecular patterns
- Regular expression system for molecules represented in SMILES format

Connection Table Formats: SDF and Mol

Molfile: header block and connection table (a, b)
SDfile: extension of Molfile (a, b, c)

(a) Header block

- (a1) CMP name or blank line
- (a2) software, date, 2/3D, ...
- (a3) blank line

(b) Connection table (CT)

- (b1) counts line: n atoms, n bonds, chiral, ...
- (b2) atom block: x,y,z coordinates, atoms, mass diff., charge, ...
  2D representation when z coordinates all zero
- (b3) bond block: atom 1, atom 2, bond type, stereo specs, ...
- (b4) CT delimiter

(c) Annotation data

- (c1) <data header>
- (c2) data
- (c3) blank line
- (c4) continues like c1-3
- (c5) SDF delimiter ($$$$)

Example: SDF Format

```
a1      NSC85228 ethanol 1
a2      APtclserve02230600142D 0 0.00000 0.00000NCI NS
a3
b1      9 8 0 0 0 0 0 0 0 0999 V2000
b2      2.8660 -0.250 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0
b2      3.7321 0.2500 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
b2      4.5981 -0.250 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
b2      2.3291 0.0600 0.0000 H 0 0 0 0 0 0 0 0 0 0 0 0
b2      4.1306 0.7249 0.0000 H 0 0 0 0 0 0 0 0 0 0 0 0
b2      3.3335 0.7249 0.0000 H 0 0 0 0 0 0 0 0 0 0 0 0
b2      4.2881 -0.786 0.0000 H 0 0 0 0 0 0 0 0 0 0 0 0
b2      5.1350 -0.560 0.0000 H 0 0 0 0 0 0 0 0 0 0 0 0
b2      4.9081 0.2869 0.0000 H 0 0 0 0 0 0 0 0 0 0 0 0
b3      1 2 1 0 0 0 0
b3      2 3 1 0 0 0 0
b3      1 4 1 0 0 0 0
b3      2 5 1 0 0 0 0
b3      2 6 1 0 0 0 0
b3      3 7 1 0 0 0 0
b3      3 8 1 0 0 0 0
b3      3 9 1 0 0 0 0
b4      M END
c1      >< NSC >
c2      85228
c4      >< CAS >
c4      64-17-5
c4      >< SMILES >
c4      CCO
c5      $$$$
```

Outline

Introduction

CMP Structure Formats

Similarity Searching

- Background
- Fragment Similarity Search Methods
- Structural Descriptors
- Similarity Coefficients
- Alternatives

CMP Properties

CMP Libraries

References

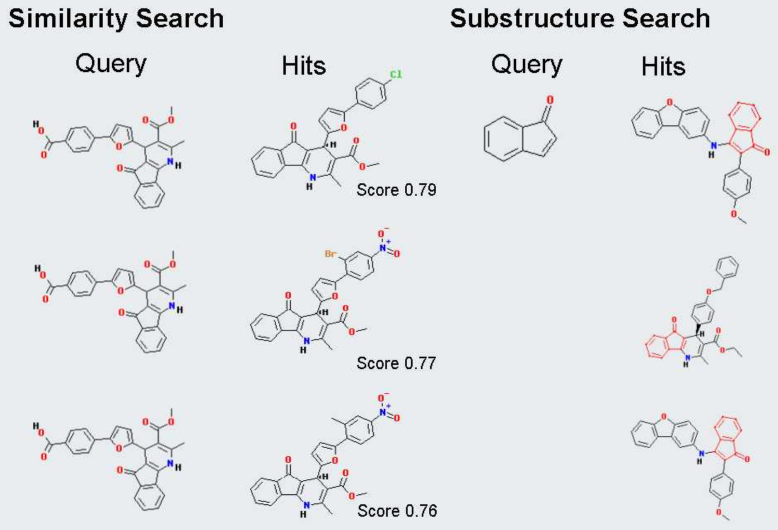
Similarity Searching

How to define similarities between compounds?

Knowledge-Based Approaches

1. Identical Structure Search
2. Superstructure Search
3. Substructure Search

CMP Similarity Searching

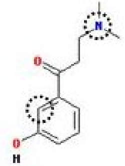


Important Compound Search Methods

1. Identical Structure Search
2. Substructure and Superstructure Searches
   - Knowledge-based approaches
3. 3D Similarity Searches (*e.g.* pharmacophore searching)
   - Slow and inaccurate
4. 2D Fragment Similarity Searching
   - Fast and accurate

   1. Involves 2 major steps
      - Structural descriptors
      - Similarity measure
   2. Structural descriptors in similarity searching
      - Atom pairs: C12N03_06
      - Atom sequences: C12C13C13C02C02N03
      - Fingerprints: rules to enumerate all fragments in common structures

Example



2D Fragment Similarity Search Methods

Involve two major steps

- Structural descriptors
- Similarity measure

Major types of structural descriptors

- Structural keys
- Fingerprints
- Atom pairs and atom sequences

Structural Keys

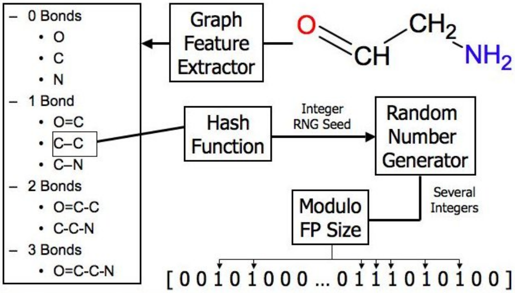
- Structural descriptors are based on lookup library of known "functional" substructures.
- Pre-compute presence of relevant substructures up front and encode them in bit-vector.
- Example of structural keys:
    - Presence of atoms (C, N, O, S, Cl, Br, etc.)
    - Ring systems
    - Aromatic, Phenol, Alcohol, Amine, Acid, Ester, ...
- Disadvantages:
    - Lookup library tends to be incomplete.
    - Sparsely populated vectors.

Fingerprints

- Fingerprints are generated directly from the molecule itself and not from a reference set of substructures.
- The algorithm examines each molecule and generates the following patterns:
  - One for each atom.
  - One representing each atom and its nearest neighbors (plus the bonds that join them).
  - One representing each group of atoms and bonds connected by paths up to 2, 3, 4, ... bonds long.
  - For example, the molecule OC=CN would generate the following patterns:
    - 0-bond paths: C, O, N
    - 1-bond paths: OC, C=C, CN
    - 2-bond paths: OC=C, C=CN
    - 3-bond paths: OC=CN,

Fingerprints

- No pre-defined patterns.
- Record counts presence or absence of structural fragments.
- Patterns are often encoded into fixed length (binary) vectors for fast similarity searching.
- Fast algorithms.
- Abstract, hard to traceback meaning of individual bits.



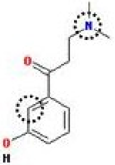
⇓

Database of binary fingerprints

Atom Pair and Atom Sequence Similarity Searching

- Like fingerprints atom pairs are generated directly from the molecule itself and not from a reference set of substructures [Chen & Reynolds 2002].
- Atom pairs are defined by:
    - the length of the shortest bond path between two atoms,
    - while the terminal atoms in this path are described by:
        - their element type
        - their number of pi electrons
        - their number of non-hydrogen neighbors
    - Example: C12N03_06
- Atom sequences:
    - similar to atom pairs, but all atoms in bond path are described.
    - Example: C12C13C13C02C02N03

Example



- Conversion of atom pairs/sequences to binary vectors of constant length is usually not performed, but would be possible.

Similarity Coefficients

1. Euclidean

$$\sqrt{\frac{c+d}{a+b+c+d}} \tag{1}$$

2. Tanimoto coefficient [Tanimoto 1957]

$$\frac{c}{a+b+c} \tag{2}$$

3. Tversky index [Tversky 1977]

$$\frac{c}{\alpha * a + \beta * b + c} \tag{3}$$

4. Many more similarity coefficients, see: [Holliday 2003]

Legend for variables:

a: count of features in CMP A but not in CMP B

b: count of features in CMP B but not in CMP A

c: count of features in both CMP A and CMP B

d: count of features absent in CMP A and CMP B

α and β: weighting variables

Global versus Local Similarity Searches

Global Search

⇓

- 2D fragment-based
- Misses local similarities

+ Fast
− Utility

Local Search

⇓

- Substructure Search
- Superstructure Search
- Local Similarity Search (MCS)

− Slow
+ Utility

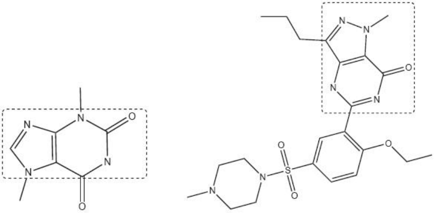
Utility for Clustering

Global Similarity

Common Fragments

Why Local Similarity Searches?

Solves two major issues of current similarity search approaches



1. Less restrictive than substructure searches
2. Allows scoring of local similarities

Possible solution: most common substructure (MCS) searches

Alternatives: 3D Searches & Docking

Conformer Predictions

Prediction of the most stable conformers in 3D space.

3D Searches

Uses shape and topological indices to query a 3D conformer database.

3D Substructure searches

Related to pharmacophore searches

Docking

Computational modeling of the possible binding modes of a ligand to a target site.

Important Compound Databases

Compound Databases

- PubChem
- DrugBank
- NCI
- ChemBank
- ChemNavigator
- SciFinder
- ChemMine

Outline

- Introduction
- CMP Structure Formats
- Similarity Searching
  - Background
  - Fragment Similarity Search Methods
  - Structural Descriptors
  - Similarity Coefficients
  - Alternatives
- CMP Properties
- CMP Libraries
- References

CMP Property Predictions

Property Descriptors

Compound Descriptors

Structural descriptors

- Atom pairs, fingerprints
- many others

Property descriptors

- Formula
- Molecular weight
- Octanol/Water partition coefficient (logP)
- Hydrogen Bond Acceptors
- Hydrogen Bond Donors
- Acidic groups
- Rotatable bonds
- over 300-3000 additional ones

Drug-likeness Filters

Lipinski Rules

In a selection of 2245 compounds from the World Drug Index Lipinski identified four property cutoffs that were common in 90% of these drugs (Lipinski et al, 1997, Adv Drug Deliv Rev: 23, 3-25). These property filters are known as the "Rule of Five" (all multiple of 5):

- MW $<$ 500g/mol
- lipophilicity: logP $<$ 5
- n H-bond donors $<$ 5 (e.g. OH and NH)
- n H-bond acceptors $<$ 10 (e.g. N and O)

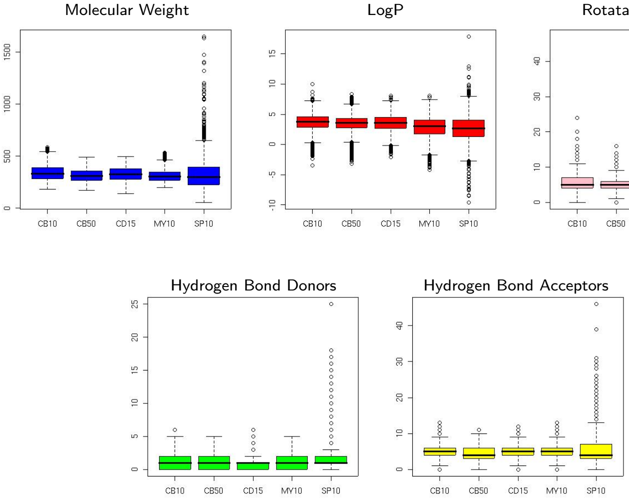
Extended Lipinski Rules

- n rotable bonds $<$ 10

ADMET Rules

- Criteria for predicting adsorption, distribution, metabolism, excretion and toxicity (ADMET) more improtant for pharmaceutical industry than chemical genomics.

Extended Lipinski Descriptors


Molecular Weight

LogP

Rotatable Bonds

Hydrogen Bond Donors

Hydrogen Bond Acceptors

Outline

Introduction

CMP Structure Formats

Similarity Searching

- Background
- Fragment Similarity Search Methods
- Structural Descriptors
- Similarity Coefficients
- Alternatives

CMP Properties

CMP Libraries

References

Compound Libraries

Which chemicals are of interest?

Library Synthesis and Assembly

Topic of a combinatorial chemistry course.

- Combinatorial synthesis
- Diversity oriented synthesis (DOS)
- Diversity collections from many sources

⇔ Virtual libraries: rationally or randomly designed

Chemical Space

Space comparisons[1]

- Chemical space of small CMPs: 10^{60} structues (theoretical number of small CMPs with MW $\leq$500)
- Feasible CMP volume for HTS approaches: 10^{6}
- Number of small CMPs in an organism much smaller: 10^{3}-10^{4}
- Protein space: 10^{390} structures (theoretical number of proteins with 300 AA)
- Number of proteins in an organism much smaller: 10^{3}-10^{5}

Critical questions

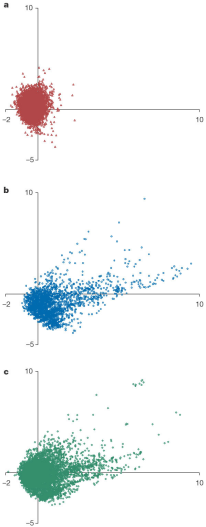
How big is the biological relevant chemical space and how can we design screening libraries that cover this space?

[1] Ref: Dobson C (2005) Nature 432, 824-828

Compound Libraries

- Combinatorial libraries[a]
  $\sim 10^7$ avail. CMPs
- Bioactive compound libraries[b]
  $\sim 10^3$ avail. CMPs
- Natural compound libraries[c]
  $\sim 10^3$ avail. CMPs
- Metabolic compound libraries
  $\sim 10^3$ avail. CMPs
- Compound collections
  Any combination of the above
- Virtual libraries
  $\sim$ limited by computer power

PCA Plots[1]



[1] Dobson C (2005) Nature 432, 824-828

Library Assembly

Drug-likeness, property bias and structural diversity

Library Assembly

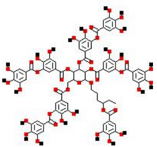
Selection Criteria

- Diverse set compounds, bioactives and natural products
- Minimum overlap within and between libraries
- Structural diversity (J Chem Inf Comput Sci 44: 643-651)
- Majority drug-like: 'plant Lipinski rules' (Pest Manag Sci 58: 219-233)
- Elimination of undesirable side groups (filters)
- Resupply and price

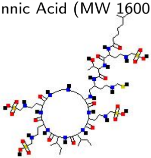
Clustering Methods

- Principal component analysis (PCA)
  - Reduction technique of multivariate data to principal compoments to identify hidden variances
- Multidimensional scaling
  - Displays distance matrix of objects in spacial plot
- Hierarchical Clustering
  - Iterative joining of items by decreasing similarity
- Binning Clustering
  - Uses provided similarity cutoff for grouping of items

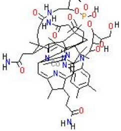
Property PCA



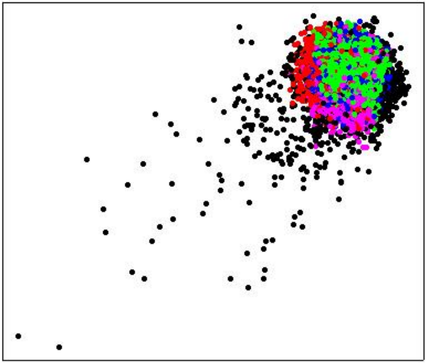
Tannic Acid (MW 1600)



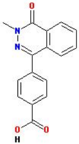
Colistimethate (MW 1400)



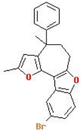
Cobalamine (MW 1100)



Comb1 Comb2 Comb3 Comb4 Bioact



56421 (MW 270)



19737 (MW 390)

Outline

Introduction

CMP Structure Formats

Similarity Searching

- Background
- Fragment Similarity Search Methods
- Structural Descriptors
- Similarity Coefficients
- Alternatives

CMP Properties

CMP Libraries

References

References

Bohacek R et al (1996) The art and practice of structure-based drug design: a molecular modeling perspective. Med Res Rev: 16, 3-50.

Butcher R & Schreiber (2003) A small molecule suppressor of FK506 that targets the mitochondria and modulates ionic balance in Saccharomyces cerevisiae. Chem Biol: 10, 521-531.

Butcher R & Schreiber (2005) Identification of Ald6p as the target of a class of small-molecule suppressors of FK506 and their use in network dissection. Proc Natl Acad Sci U S A: 101, 7868-7873.

Chen, X, Reynolds, C H (2002) Performance of similarity measures in 2D fragment-based similarity searching: comparison of structural descriptors and similarity coefficients J Chem Inf Comput Sci, 42: 1407-1414.

Dobson C (2005) Chemical space and biology. Nature: 432, 824-828.

Haggarty S (2005) The principle of complementarity: chemical versus biological space. Curr Opin Chem Biol: 9, 296-303.

Holliday, J D, Salim, N, Whittle, M, Willett, P (2003) Analysis and display of the size dependence of chemical similarity coefficients. J Chem Inf Comput Sci, 43: 819-828.

Tan D (2005) Diversity-oriented synthesis: exploring the intersections between chemistry and biology. Nat Chem Biol: 1, 74-84.

Tanimoto, TT (1957) IBM Internal Report, 17th Nov.

Tversky, A (1977) Psychological Reviews, 84, 327-352.