COMP 150 CSB –
Computational Systems Biology

Computing Molecular Similarities

Soha Hassoun

Department of Computer Science (primary)
Department of Chemical and biological Engineering
Department of Electrical and Computer Engineering

Tufts UNIVERSITY

Materials for slides and reading for this week

- (reading) Hattori, M., Okuno, Y., Goto, S., & Kanehisa, M. (2003). Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *Journal of the American Chemical Society*, *125*(39), 11853-11865.

- (reading) Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, *28*(1), 31-36.

- Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D., & Pletnev, I. (2013). InChI-the worldwide chemical structure identifier standard. *Journal of cheminformatics*, *5*(1), 7.

- Cereto-Massagué, A., Ojeda, M. J., Valls, C., Mulero, M., Garcia-Vallvé, S., & Pujadas, G. (2015). Molecular fingerprint similarity search in virtual screening. Methods, 71, 58-63.

- Bajusz, D., Rácz, A., & Héberger, K. (2015). Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?. Journal of cheminformatics, 7(1), 20.

Why care about molecular similarity?

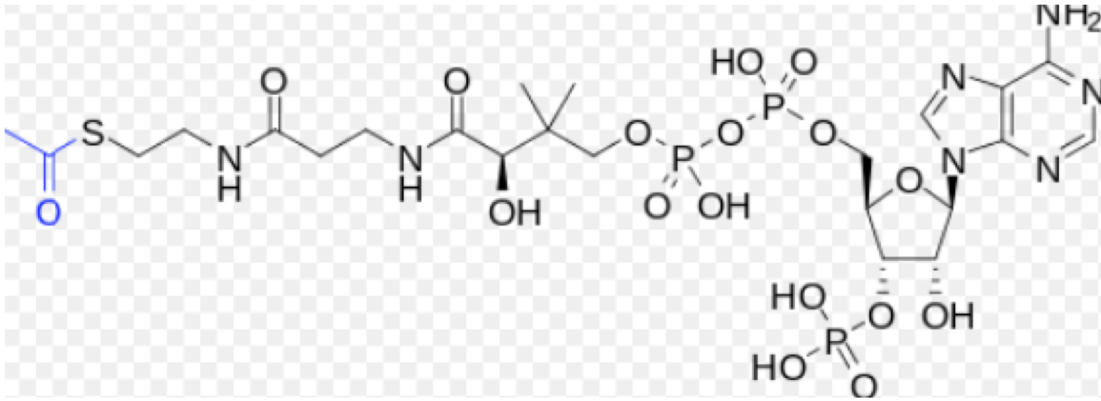
- Find similar structures to facilitate constructing biosynthesis pathways
- Further correlation of chemical and genetic information
- Better understand metabolic networks

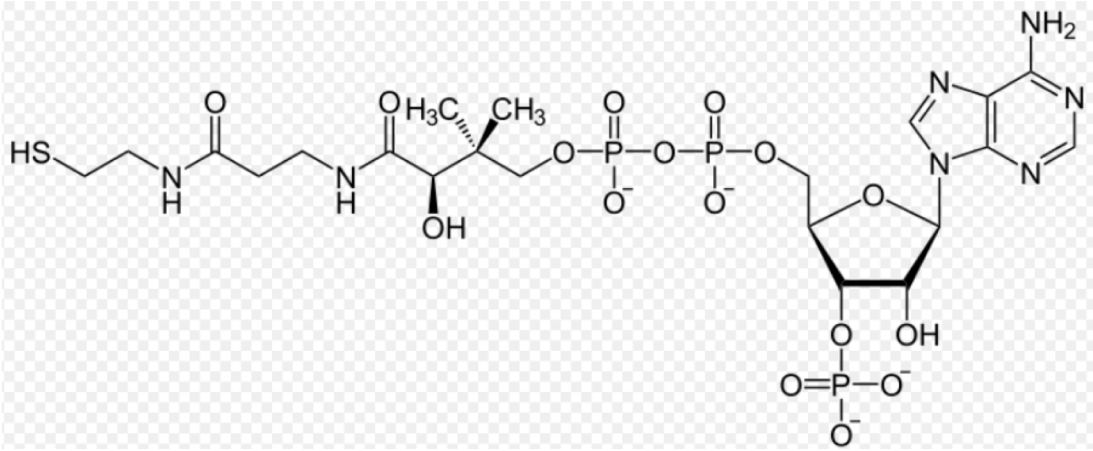
How do we compute molecular similarity? Outline

- Graph-based methods
  - Subgraph isomorphism
  - Subgraph isomorphism is NP-hard
  - Example: Using Bron–Kerbosch Algorithm as in SIMCOMP
- Text-based methods – The Full Molecules
  - Key idea: Use strings to represent molecules
    - Example Strings:
      - SMILES
      - SMARTS
      - inChI and InChIKey
- Fingerprints – vectors capturing molecular features
  - Substructure fingerprints
- Computing similarity

Key idea for finding molecular similarity using graph

- Atoms are mapped to graphs:
  - Atoms → nodes
  - Bonds → edges
- Given two molecule M1 and M2, with graphs G1 and G2, identify the maximum common subgraph





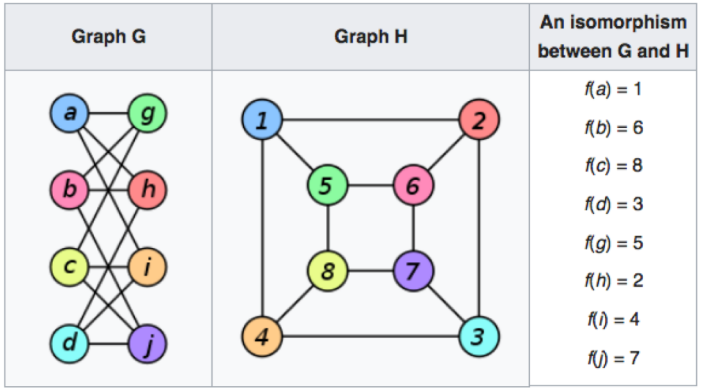
Graph & Subgraph isomorphism

Graph Isomorphism: An isomorphism of graphs G and H is a bijection (one-to-one correspondence) between the vertex sets of G ar $f: V(G) \to V(H)$

such that any two vertices u and v of G are adjacent in G if and only if f(u) and f(v) are adjacent in H.

Subgraph isomorphism:

Does G contain a subgraph that is isomorphic to H



Maximum Common Subgraph

- A common subgraph of G1 and G2, CS(G1,G2), is a graph which is isomorphic to a subgraph of both G1 and G2. The maximal common subgraph of G1 and G2, MCS(G1,G2), is the CS(G1,G2) whose cardinality is not smaller than that of any other CS(G1,G2).

Maximal common subgraph is NP-hard

- Characteristics of NP-hard class of problems (that are not NP complete):
  - We currently do not know if there are polynomial time algorithms that can solve these class of problems
  - We also do not know if they are verifiable in polynomial time

Notes:

NP-complete problems:

we don't know if we can solve in polynomial time

we can verify in polynomial time

P problems:

we can solve in polynomial time

we can verify in polynomial time

Clique Finding in the Association Graph

The association graph AG possesses all possibilities of vertex matches between two initial graphs G1 and G2

A clique in AG corresponds to a common subgraph between G1 and G2.

The largest clique based on the number of matching vertexes becomes the largest match of our interest.

The initial problem of finding the MCS(G1,G2) can be reduced to the problem of finding the MCL(AG).

Bron-Kerbosch Algorithm To Final Maximal Cluster

https://www.youtube.com/watch?v=132XR-RLNoY

Bron–Kerbosch Algorithm(simplified)

1. Take any single vertex in Graph as a subset *(uses set R)*
2. Add as many vertices to this subset (which are adjacent to every other vertex in it), as possible. *(uses sets R,P)*
3. If subset is not already reported, report as maximal clique *(uses set X)*
4. Remove the vertice used in step 1 from further considerations
5. Repeat step 1 to 4 for all vertices in Graph

Tailoring the algorithm to the application

- Each vertex is an atom type
  - What is considered a match?
- All-or-nothing matching:
  - Based on exact match
  - Each vertex of the association graph is weighted as only one or zero, called all-or-none weighting here, depending on whether two vertexes from the original graphs do or do not match
- Loose weighting:
  - allows partial matches for the same atom species with different environments
  - a(V) returns the atom species of vertex V, and c is the constant value between 0 and 1.

SIMCOMP algorithm

- first obtain all cliques with the maximum number of vertexes by the clique-finding algorithm
- calculate the sum of weights for each clique to select the largest weighted one

Improvements in the Bron-Kerbosch Algorithm

- Stop the recursion in the algorithm once we obtain a candidate set of MCL
- Start searching for a better common subgraph, using quasi-MCS from the candidate set:
  - Eliminate small simply connected common subgraphs whose cardinality is less than a threshold
  - Extend non-small SCCSs one by one greedily until no more atom pairs can be included
- Threshold:
  - R_{max} for termination of the usual clique finding algorithm
  - S_{min} for consideration of the greedy search around each of the SCCSs found

Normalized Score for Compound Similarity

- Jaccard Coefficient, or Tanimoto coefficient:
  - Ratio of size of the common substructure divided by the size of the non redundant set of substructures
  - 0 no common substructures; 1 identical structures

$$\frac{|\mathrm{MCS}(G_1,G_2)|}{|G_1| + |G_2| - |\mathrm{MCS}(G_1,G_2)|}$$

See more on jaccard in later part of lecture

Comparison of all compound pairs in KEGG

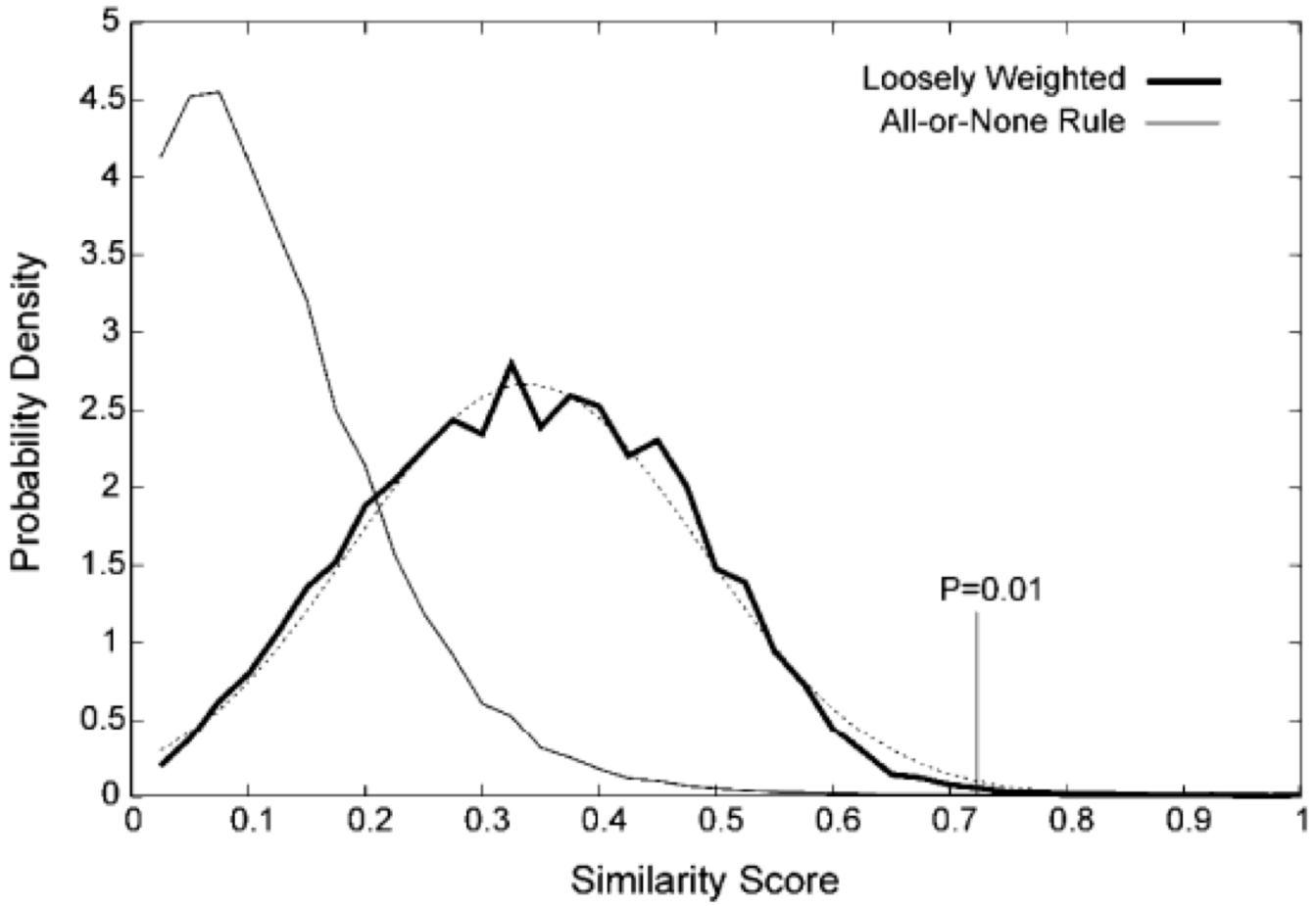


Figure 4. Distribution of normalized similarity scores for all possible pairs of chemical compounds in KEGG. The thick line is the probability density distribution with the loose weighting condition, and the thin line is that for the all-or-none weighting condition. Here the thick line can be fitted with a normal distribution, drawn in a dashed line, whose statistical parameters are $\mu = 0.338$ and $\sigma = 0.150$. According to this normal distribution P-value $= 0.01$ for the right tail corresponds to score $= 0.723$, as indicated in the figure.

Clustering analysis of similar compounds

- Cluster all pairs with similarity score > .723
- Total number of clusters found was 3970:
  - consisting of 1871 singletons and 2099 non-singletons,
  - maximum size cluster contained 64 compounds.

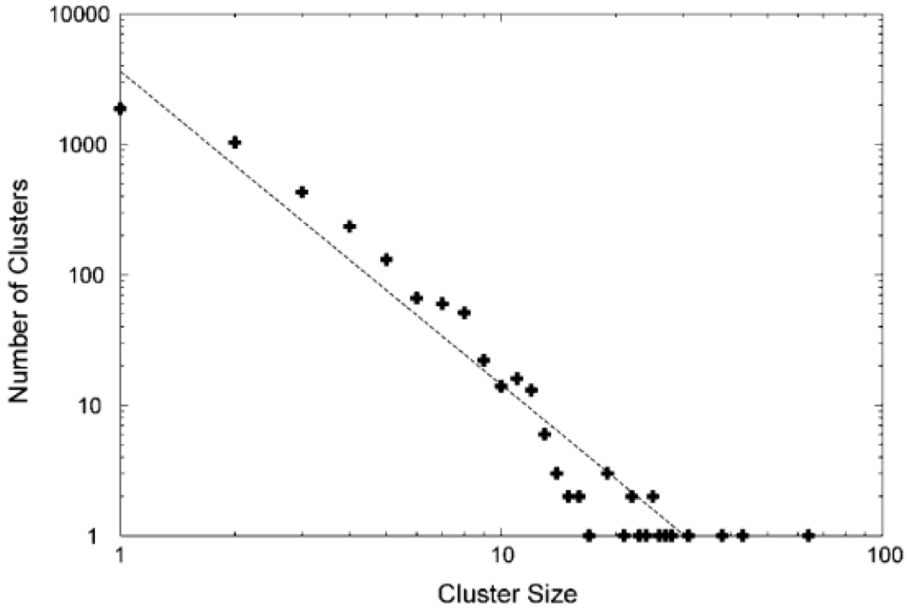


Figure 5. Size distribution of similar compound clusters that are identified by the complete linkage analysis with the threshold similarity score of 0.723 (the degree of confidence is 99%). In this log−log plot, the horizontal axis is the cluster size or the number of compounds belonging to the cluster, and the vertical axis is the number of clusters with a given size. The dashed line is the regression line, indicating that the size distribution of clusters approximately follows the power-law, $P(k) \propto k^{-\gamma}$, with $\gamma = 2.41$.

Most of the top 10 largest clusters are highly correlated with specific metabolic pathways

Table 1. Top Ten Largest Clusters of Similar Chemical Compounds

no.	size	common formula	description of members	KEGG pathways map numbers[a]						
				C	L	N	AA	CC	second	AtR
1	64	C_6O_6	hexose, its uronic acid, glycoside	10, 30, 52				**500**		
2	43	C_6O_5	ketohexose, aldohexose, aldarate	30, 40, 51, 52, 53						
3	38	C_5O_5P	ribose and phosphoric acid group of nucleic acids							**970**
4	31	C_6O_8P	phosphorylated hexose	**51**, 52		520				
5	28	C_5O_5	ketopentose, hexose lactone	40, **53**						
6	27	C_9O	containing a cinnamate skeleton				350, 360		940	
7	26	C_5O_4	aldopentose, pentoside	**40**		520				
8	25	C_{10}	containing a menthol skeleton						900	
9	25	$C_{27}O$	containing a cholesterol skeleton		**100**					
10	24	C_8O_6N	*N*-acetylated hexosamine					**530**		

[a] The pathway map numbers are simplified; for example, 40 stands for map00040 in KEGG. The most frequently observed pathways are shown in bold. Abbreviations for the pathway categories are: C, carbohydrate metabolism; L, lipid metabolism; N, nucleotide metabolism; AA, amino acid metabolism; CC, metabolism of complex carbohydrates; second, biosynthesis of secondary metabolites; and AtR, aminoacyl-tRNA synthesis.

Pathway-oriented Clustering

- Perform cluster analysis on 2294 compounds known to be in KEGG pathways
- Most of the KEGG metabolic pathway maps could be divided into several parts of chemical compound clusters
- See next slide

Figu

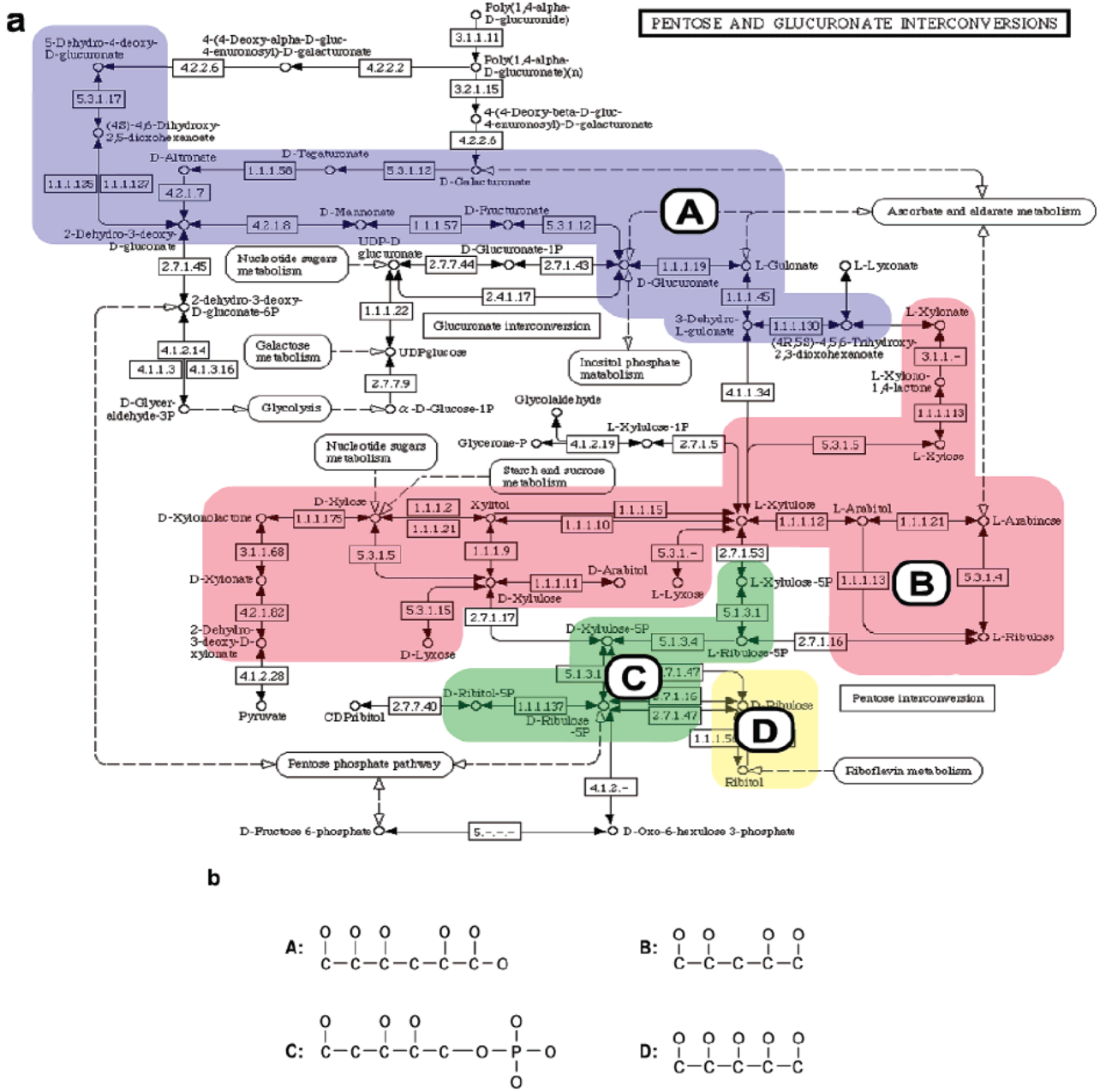


Figure 7. An example of similar compound clusters mapped onto a specific pathway. a is the result of the pathway-oriented clustering for the metabolic pathway of pentose and glucuronate interconversions, whose accession number is map00040 in the KEGG/PATHWAY database. After clustering 2294 metabolites that appear on any of the KEGG pathway maps, non-singleton clusters were superimposed on each of the pathway maps. Here, chemical compounds included in the same shaded region exhibit high structural similarities and high connectivities along the pathway in map00040. There are four major clusters of such chemical compounds in this pathway map: **A**, **B**, **C**, and **D** whose schematic representations of common components are drawn in b.

Correlation Between Chemical Information And Genomic Information (Fig 8 Of 2003 Hattori Paper)

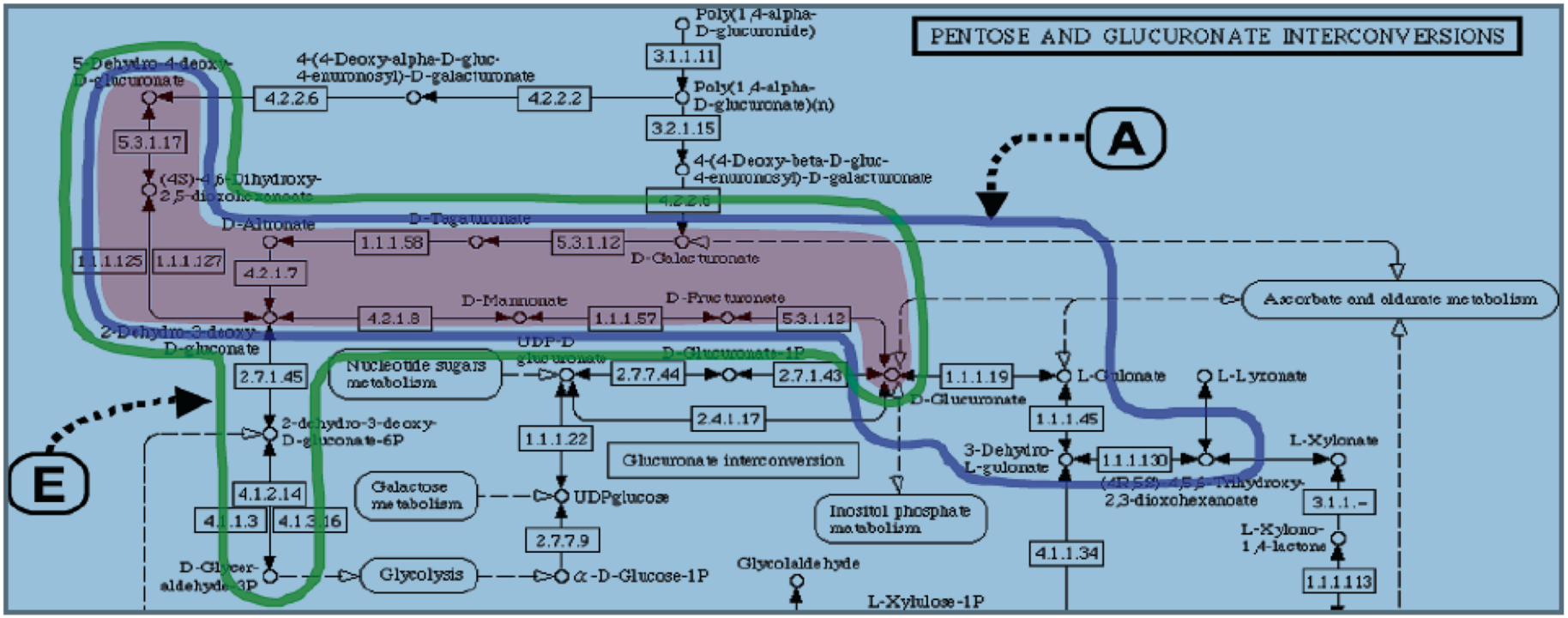


Figure 8. Example of the correlation between chemical information and genomic information. The area designated by A corresponds to the cluster of similar compounds shown in Figure 7. The area designated by E corresponds to the cluster of genomic associations where genes coding for the enzymes are closely located on selected genomes according to the KEGG ortholog group table. Thus, the shaded area is the overlap of chemical and genomic clusters.

SIMCOMP, Now

See Hattori, M., Tanaka, N., Kanehisa, M., & Goto, S. (2010). SIMCOMP/SUBCOMP: chemical structure search servers for network analyses. Nucleic acids research, 38(suppl_2), W652-W656.

- SIMCOMP - find similar molecular structures
  - http://www.genome.jp/tools/simcomp/
  - Try Pyruvate C00022
- SIMCOMP2 – compute similarities between 2 molecules
  - http://www.genome.jp/tools/simcomp2/
- SUBCOPM -- Find superstructures
  - http://www.genome.jp/tools/subcomp/
- Added many options (post processing, chiral check, atom-based vs KEGG Atom Types, etc…)

Advances in subgraph isomorphism for molecules

- Rahman, S. A., Bashton, M., Holliday, G. L., Schrader, R., & Thornton, J. M. (2009). Small molecule subgraph detector (SMSD) toolkit. Journal of cheminformatics, 1(1), 12.
- Key idea: use multiple algorithms depending on the situation
- Publically available toolkit (SIMCOMP is not)

Summary of graph based molecular similarity

- Similarity threshold – consider the specific application
- Algorithms are approximations, and not exact.
  - Consider faster algorithms, and fusion of algorithms
- Reported numbers are approximations (lower bounds on similarity)
- Maybe consider function (+…) , along with graph-based similarity
- Is the analysis complete?
  - Power law figure, they were not because they left out singletons

Non-Graph molecular representations

- Motivation – substructure searches are computationally expensive. Come up with representation that is:
  - **Standardized**
  - **Compact**
  - **Ideally human readable**
  - **Can apply a reaction transformation to it**

- SMILES (Simplified Molecular Input Line System)
  - development was initiated by David Weininger in the 1987 based on graphs.
  - Parentheses are used to indicate branching points and numeric labels designate ring connection points.

SMILES Specification Rules - Atoms

- Atoms are represented by their atomic symbols
- Attached hydrogens and charges are specified inside brackets
- Non organic elements must be in brackets

C	methane	(CH4)
P	phosphine	(PH3)
N	ammonia	(NH3)
S	hydrogen sulfide	(H2S)
O	water	(H2O)
Cl	hydrochloric acid	(HCl)

[H+]	proton
[Fe+2]	iron (II) cation
[OH-]	hydroxyl anion
[Fe++]	iron (II) cation
[OH3+]	hydronium cation
[NH4+]	ammonium cation

[S]	elemental sulfur
[Au]	elemental gold

SMILES Specification Rules - Bonds

Single, double, triple, and aromatic bonds are represented by the symbols -, =, #, and :, respectively.

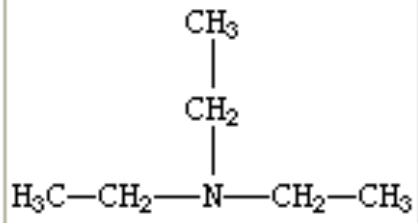
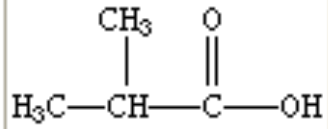
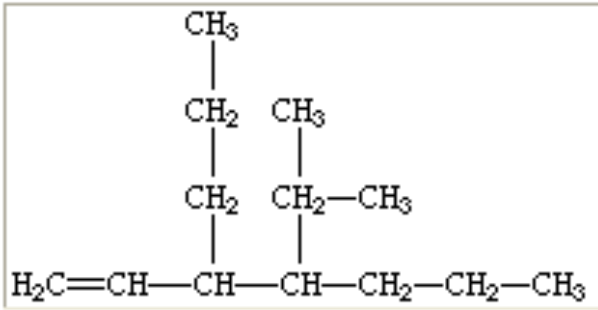
Adjacent atoms are assumed to be connected to each other by a single or aromatic bond (single and aromatic bonds may Be, an usually, omitted).

Structure	Valid SMILES
	C=CCC=CCO
CH2=CH-CH2-CH=CH-CH2-OH	C=C-C-C=C-C-O
	OCC=CCC=C

CC	ethane	(CH3CH3)
C=O	formaldehyde	(CH2O)
C=C	ethene	(CH2=CH2)
O=C=O	carbon dioxide	(CO2)
COC	dimethyl ether	(CH3OCH3)
C#N	hydrogen cyanide	(HCN)
CCO	ethanol	(CH3CH2OH)
[H][H]	molecular hydrogen	(H2)

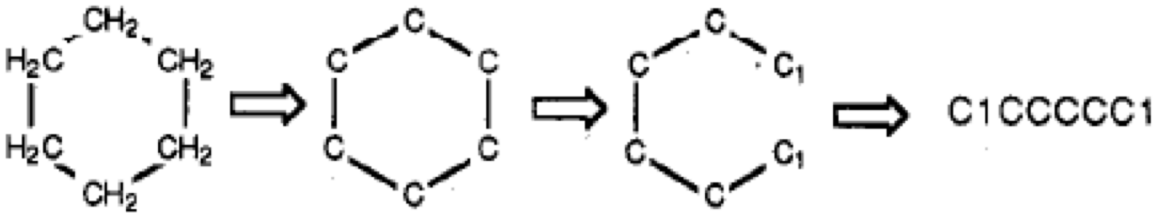
SMILES Specification Rules - Branches

Branches are specified by enclosing them in parentheses, and can be nested

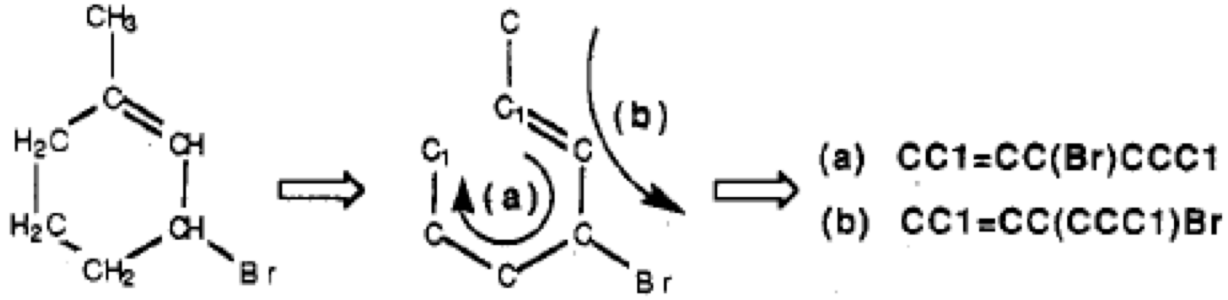
		
CCN(CC)CC	CC(C)C(=O)O	C=CC(CCC)C(C(C)C)CCC
Triethylamine	Isobutyric acid	3-propyl-4-isopropyl-1-heptene

SMILES Specification Rules – Cyclic structures

- Cyclic structures are represented by breaking one bond in each ring

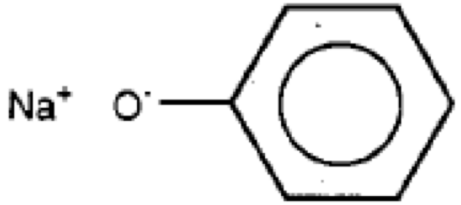


- There are many different but equally valid descriptions of the same structure



SMILES Specification Rules - Disconnected Structures

Disconnected compounds are written as individual structures separated by a "." (period).



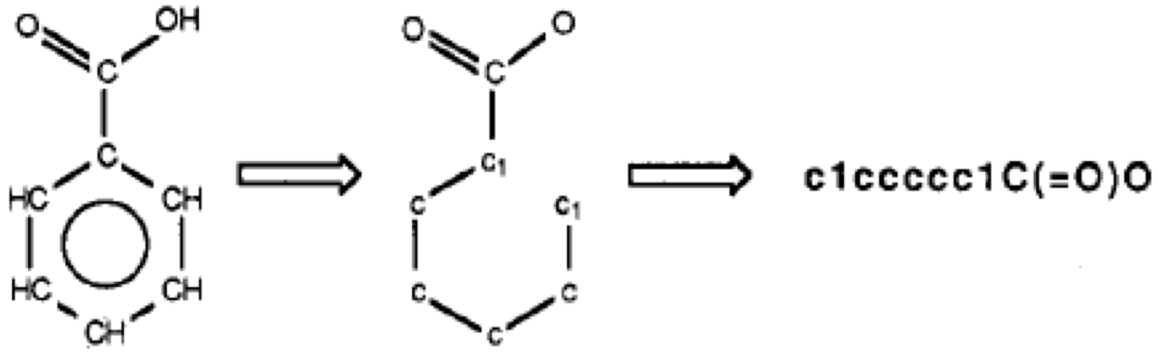
[Na+].[O-]c1ccccc1

or

c1cc([O-].[Na+])ccc1

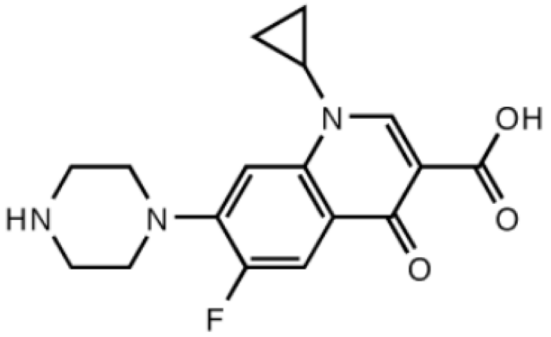
SMILES Specification Rules - Aromaticity

- Aromatic structures may be distinguished by writing the atoms in the aromatic ring in lower case letters,

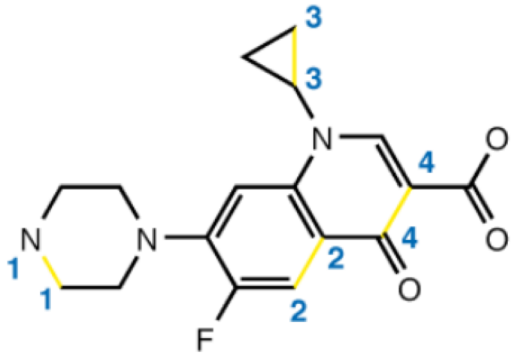


Putting it Together
Ciprofloxacin

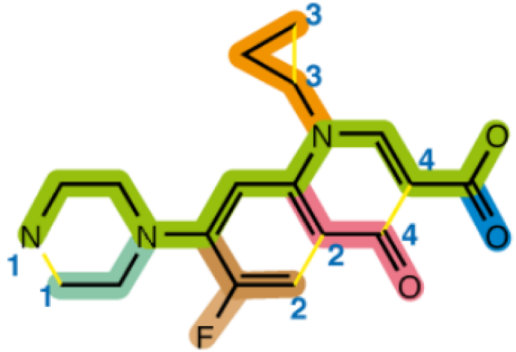
A



B



C



D

N1CCN(CC1)C(C(F)=C2)=CC(=C2C4=O)N(C3CC3)C=C4C(=O)O

https://en.wikipedia.org/wiki/Simplified_molecular-input_line-entry_system#/media/File:SMILES.png

Are SMILES representation canonical?

- There is a follow-up SMILES paper:
  - Weininger, D., Weininger, A., & Weininger, J. L. (1989). SMILES. 2. Algorithm for generation of unique SMILES notation. Journal of Chemical Information and Computer Sciences, 29(2), 97-101.

- An open source implementation:
  - http://openbabel.org/dev-api/canonical_code_algorithm.shtml

Addendum: SMART to search SMILES

- Based on some form of regular expression search
  - Examples:
    - [C,N] is an atom that can be aliphatic C or aliphatic N
    - SMARTS bond symbol ~ (tilde) matches any bond
    - [H] means hydrogen
    - [*H2] means any atom with exactly 2 hydrogens attached
    - Allow for logical operators:
      - ! → "NOT"
      - AND→ "&" (high precedence) or ";" (low precedence)
      - OR→ ","
      - Exampes:
        - [!C;R] (NOT aliphatic carbon) AND in rin
        - [c,n&H1] any arom carbon OR H-pyrrole nitrogen

See for details: http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html

Pros/Cons with SMILES

Pros:

Cons:

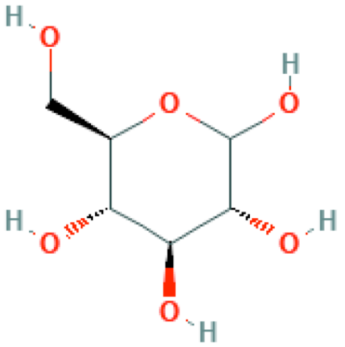
- Non-canonical
  - Different answers if algorithm is based on the ordering in the string
  - Searching for similarities or # of unique compounds is more difficult
  - As hard as maximal common subgraph
- Cant annotate 3D structures

InChI: International Chemical Identifier

- Developed by members of the of IUPAC, the International Union of Pure and Applied Chemistry
- Motivation:
  - Old style "registry-based" CAS system generated a new number that was a function of prior assigned number
  - Develop a standard for "structured-assigned" system
  - Get buy-ins from various stake holders
  - Internet search creates new opportunities for search ➔ compact form
  - Unique labels
  - Organic Chem and also non-organic
  - Hierarchical approach encoding molecules with different levels of granularity
  - Open Source, non-proprietary, but develop mechanism to certify correctness
  - (low importance): Ability to be human read/parsed and manually edited

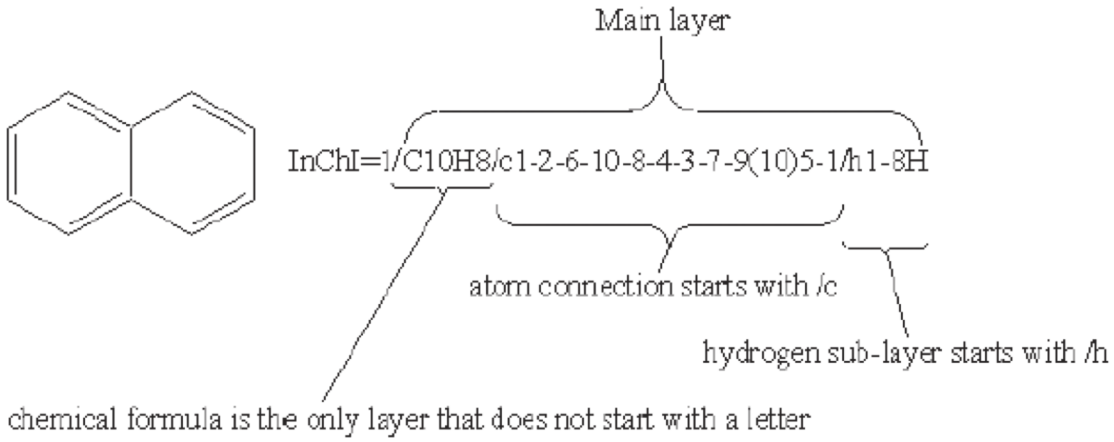
InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6?/m1/s1

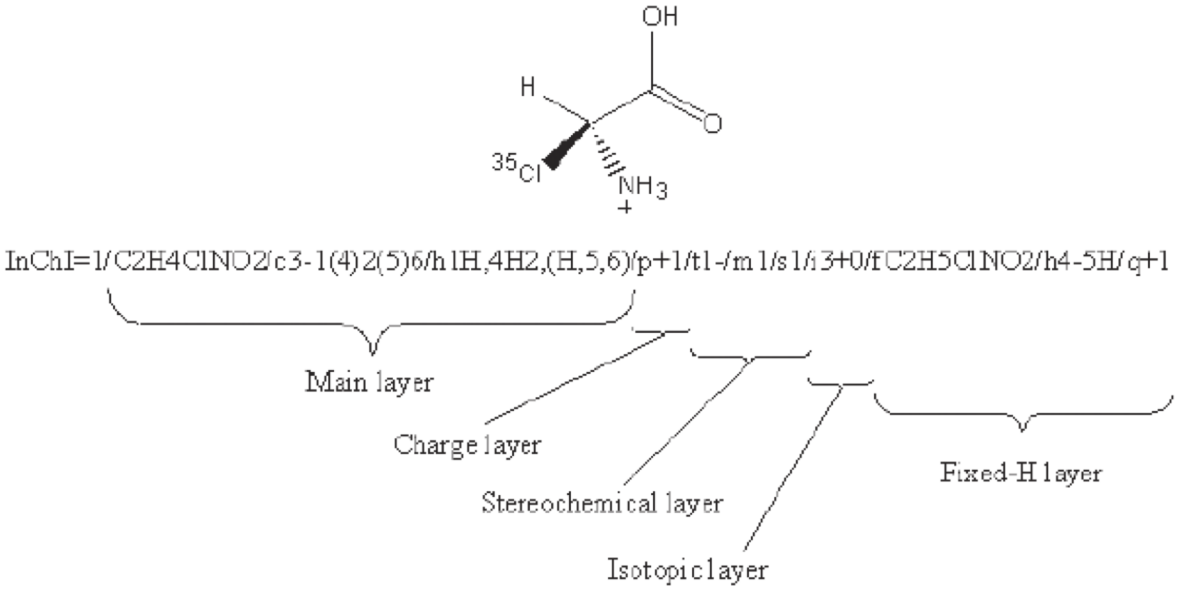
Layered Approach



- Main Layer:
  - Formula - No prefix
  - Skeletal connections layer - prefixed with '/c'
    - represents connections between skeletal atoms by listing the canonical numbers in the chain of connected atoms (branches are given in parentheses).
  - Hydrogens layer - prefixed with '/h'
    - Lists bonds between the atoms in the structure,
    - Partitioned into as many as three sublayers.
      - Sublayer 1: all bonds other than those to non-bridging H-atoms
      - Sublayer 2: bonds of all immobile H-atoms
      - Sublayer 3: possible multiple locations of any mobile H-due to tautomerism
- Non-Main Layers:
  - Charge layer –
    - Sublayer prefixed with /q stating net charge of structure
    - Sublayer for protonation/deprotonation sublayer '/p', indicates the net number of protons removed
  - Stereochemistry layer
  - Isotopic layer
  - …

InChI Structure





http://www.academia.edu/2603426/How_can_the_international_chemical_identifier_InChI_be_extended_to_non_trivial_chemicals

Internet search problematic with InChI

- Internet search would only work with a shortened string
  - Some InChI representations are ridiculously long

- Solution: develop InChI Key
  - A hashed version of InChI
    - Maps an arbitrary long InChI entry into a fixed size one

InChIKey

- Condensed, 27 character standard InChIKey is a hashed version of the full standard InChI
  - using the SHA-256 algorithm
- InChIKeys consist of
  - 14 characters hashing connectivity information of the InChI
  - followed by 9 characters hashing remaining layers of the InChI
  - followed by a single character indication the version of InChI used
  - followed by single checksum character
- Example:
  - InChI=1S/C17H19NO3/c1-18-7-6-17-10-3-5-13(20)16(17)21-15-12(19)4-2-9(14(15)17)8-11(10)18/h2-5,10-11,13,16,19-20H,6-8H2,1H3/t10-,11+,13-,16-,17-/m0/s1
  - BQJCRHHNABKAKU-KBQPJGBKSA-N

Generating InChI and InChI Key

- Many available packages – from drawings, mol files, SMILES, file etc.
- Look for certified versions

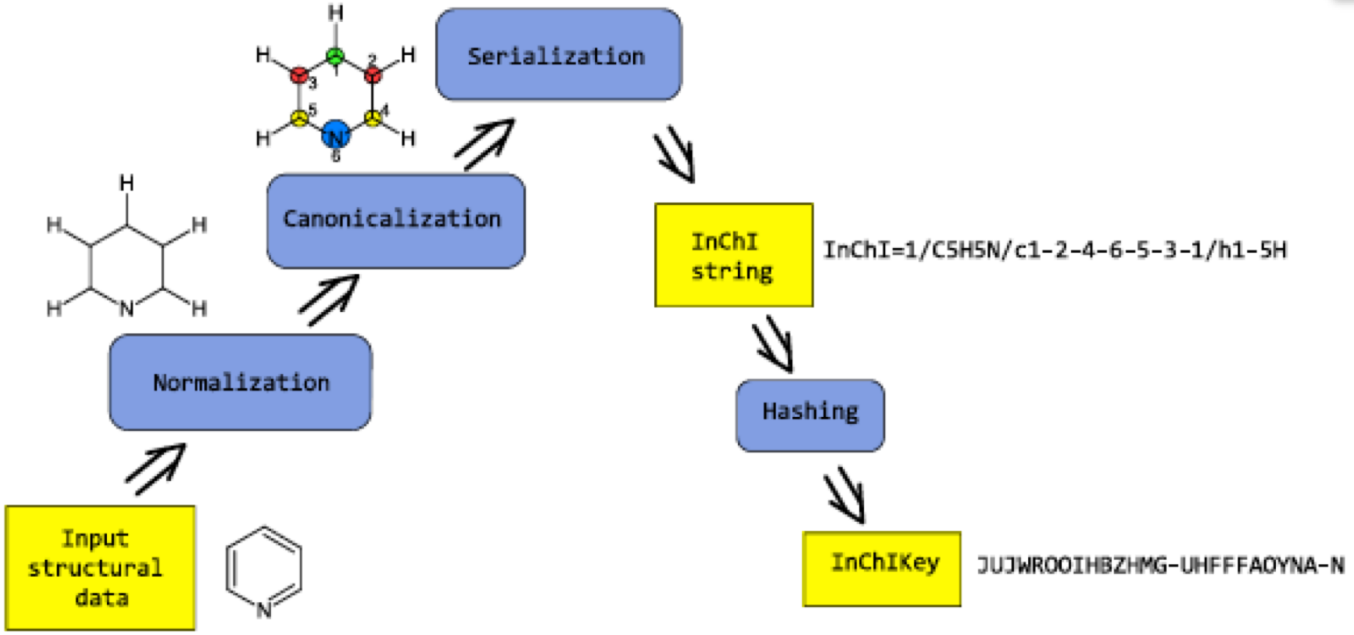


Figure 9 General workflow of InChI/Key generation.

Heller, S. R., McNaught, A., Pletnev, I., Stein, S., & Tchekhovskoi, D. (2015). InChI, the IUPAC international chemical identifier. Journal of cheminformatics, 7(1), 23.

Pros/Cons of InChI

Fingerprints

- Originates in virtual screening (VS) for drug discovery
  - There is a drug target that needs to be activated or inhibited: a protein receptor or enzyme
  - VS is a computational technique to search libraries of small molecules to identify structures that are most likely to bind to a drug target
- Driven by the "Similar Property Principle"
  - Molecules that are structurally similar are likely to have similar properties
  - If a molecule M that has not been tested for biological activity but is structurally similar to a reference molecule (known to have some activity), then M is likely to be active, and more likely to be active than any other dissimilar molecules
- Strategy
  - Compute similarity of Ms to reference structure and only biologically test the top ranked molecules
- Fingerprint format
  - Arrays of "features", where each array entry is 0 or 1 representing presence or absence of a feature

Types of Molecular Fingerprints: 1. Substructure keys

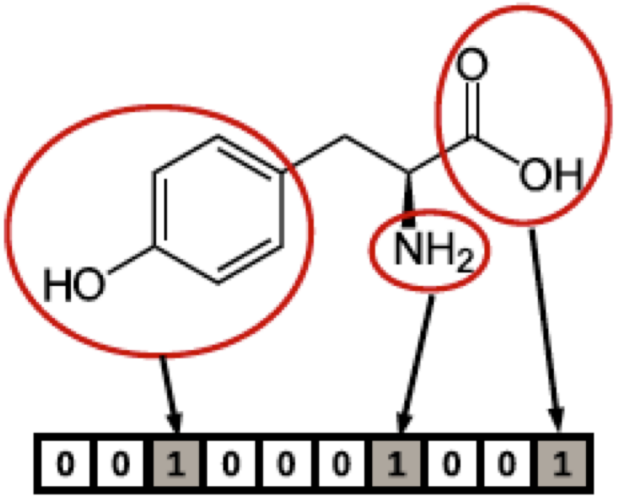


Fig. 1. A representation of a hypothetical 10-bit substructure fingerprint, with three bits set because the substructures they represent are present in the molecule (circled).

- Useful when molecules have substructure
- Easy to generate

Substructure keys – example keys

- MACCS fingerprint
  - Two variants, 960 and 166 keys, based on SMARTS patterns
- PubChem fingerprint
  - 881 substructure keys
- BCI fingerprints
  - 1052 keys
  - Allow for different # of bits, customized by user
- TGD and TGT fingerprints
  - 735 and 13,824 bits
  - TGD encodes atom-pair descriptors using seven-atom features and distances up to 15 bonds
  - TGT encodes triplets of four-atom features using three graph distances divided into 6 distance rages
  - Commercially available through MOE

Types of Molecular Fingerprints: 2. Topological or Path-Based Fingerprints

- Analyze fragment of molecule following a linear path up to a certain # of bonds
- + Any molecule can produce a meaningful fingerprint
- Bit Collision= bit set by multiple paths

- Example:
- Daylight fingerprint
- OpenEye Tree fingerprint

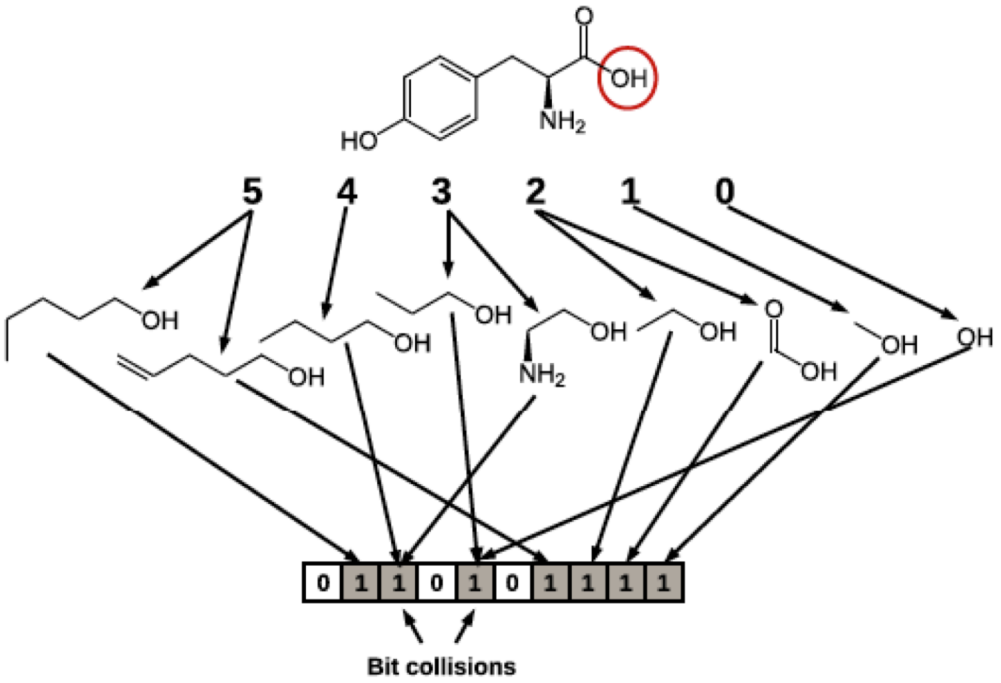
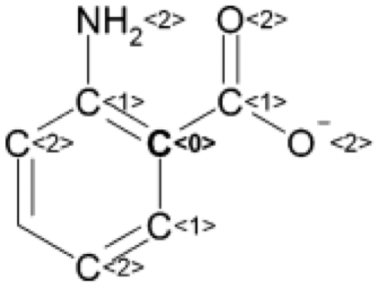


Fig. 2. A representation of a hypothetical 10-bit topological fingerprint, in this case a linear path-based fingerprint with fragments up to a length of 5. All fragments found from the starting atom (circled) are shown, and the fragment length and corresponding bit in the fingerprint are indicated. There are two bit collisions, which are bits that are set by more than one fragment; these are likely in fingerprints with a reduced number of bits. Only fragments and bits for a single starting atom are shown; for the full fingerprint, this process would be carried out for every atom in the molecule. Circular fingerprints use a similar approach, but building fragments within a radius of the starting atom instead of linear fragments.

Types of Molecular Fingerprints: 3. Circular Fingerprints

- Look at environment of each atom up to a determined radius
- Example: Molprint2D



Layer	0	1	2
	C.ar	C.ar	C.ar
		C.ar	C.ar
		C.2	N.2
			O.co2
			O.co2

Figure 1. Illustration of the descriptor generation step, applied to an aromatic carbon atom. The distances ("layers") from the central atom are given in angular brackets. In the first step, Sybyl mol2 atom types are assigned to all atoms in the molecule. In the second step, count vectors from the central atom (here C⟨0⟩) up to a given distance (two bonds from the central atom apart) are constructed. Molecular atom environment fingerprints are then binary presence/ absence indicators of count vectors of atom types.

Bender, A., Mussa, H.Y., Glen, R. C., & Reiling, S. (2004). Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. Journal of chemical information and computer sciences, 44(5), 1708-1718.

Types of Molecular Fingerprints: 3. Circular Fingerprints - continued

- Example --
  - ECFP – de facto standard circular fingerprint
    - ECFP4 (diameter of 4)
    - ECFP - Variation: keep count of frequency of features not just 0/1 entries
    - FCFP (functional-class finger print):
      - Index atom's role and not exact match
      - Used for pharamcophoric fingerprints
  - Keeps *all* identifiers as it builds out in diameter; realized fingerprints represents both very small local substructures (each atom collecting identifiers only from its immediate neighbors) as well as large substructures

Rogers, D., & Hahn, M. (2010). Extended-connectivity fingerprints. Journal of chemical information and modeling, 50(5), 742-754.

Types of Molecular Fingerprints: 4. Hybrid fingerprints

- Combine two or more fingerprints
  - Example: Unity 2D combines structural keys and connectivity path fragments
  - MP-MFP: 171 bits: 110 bits are structural keys, and 61 are property descriptors

Pros/Cons on Fingerprints

General observations about similarity

- Fingerprints are great, but how are they useful?
  - Often, ranking molecules in similarity against a reference

- Need a metric(s) to measure similarity (or dissimilarity) between molecules
  - The metric must be selected based on the application
  - No single metrics is "best"
  - Let's look at few

Similarity Metrics

Table 1
Some similarity coefficients and distances used with fingerprints.

Measure	Expression	Range
Tanimoto/Jaccard coefficient	$\frac{c}{a+b-c}$	0 to 1
Euclidean distance	$\sqrt{a+b-2c}$	0 to N
City-block/Manhattan/Hamming distance	$a+b-2c$	0 to N
Dice coefficient	$\frac{2c}{a+b}$	0 to 1
Cosine similarity	$\frac{c}{\sqrt{ab}}$	0 to 1
Russell–RAO coefficient	$\frac{c}{m}$	0 to 1
Forbes coefficient	$\frac{cm}{ab}$	0 to 1
Soergel distance	$\frac{a+b-2c}{a+b-c}$	0 to 1

Where, given the fingerprints of two compounds, A and B, m equals the total amount of bits present in the fingerprints, a equals the amount of bit set to 1 in A, b equals the amount of bits set to 1 in B and c equals the amount of bits set to 1 in both A and B.

Cereto-Massagué, A., Ojeda, M. J., Valls, C., Mulero, M., Garcia-Vallvé, S., & Pujadas, G. (2015). Molecular fingerprint similarity search in virtual screening. Methods, 71, 58-63.

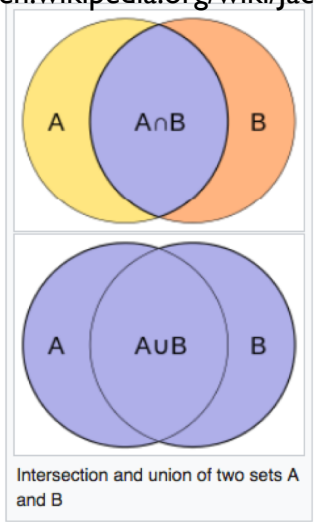
Jaccard Index, Jaccard Similarity Coefficient

https://en.wikipedia.org/wiki/Jaccard_index

- Circa 1901, by Paul Jaccard
- Measures similarity of two samples
- Intersection over Union

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

  - if A and B are empty, then J(A,B) = 1



Intersection and union of two sets A and B

- Jaccard distance (Soergel distance) measures "dissimilarity" between samples

$$d_J(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}.$$

- In 1960s, based on an IBM report:
- Tanimoto similarity == Jaccard Index
- Tanimoto distance != Jaccard Distance
  - Tanimoto_distance = -log_2 (Tanimoto similarity)

Hamming Distance

- Hamming distance between two strings of equal length:

  the number of positions at which the corresponding symbols are different.

- Euclidean distance is the sqrt of the **Hamming distance**

Euclidean distance	$\sqrt{a+b-2c}$	0 to N
City-block/Manhattan/Hamming distance	$a+b-2c$	0 to N

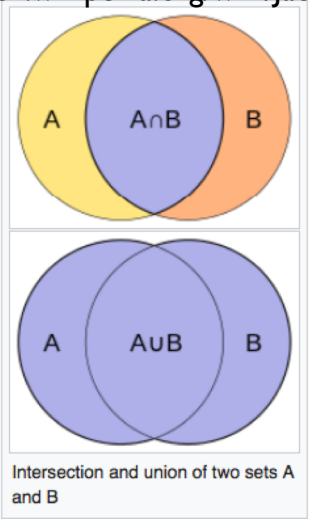
- Used extensively in coding theory

Dice coefficient

- Known also as Sørensen index
- For two samples X and Y,

$$QS = \frac{2|X \cap Y|}{|X| + |Y|}$$

https://en.wikipedia.org/wiki/Jaccard_index



Intersection and union of two sets A and B

- Difference from Jaccard Index:
  - Jaccard counts the overlap only ONCE in both numerator and denominator

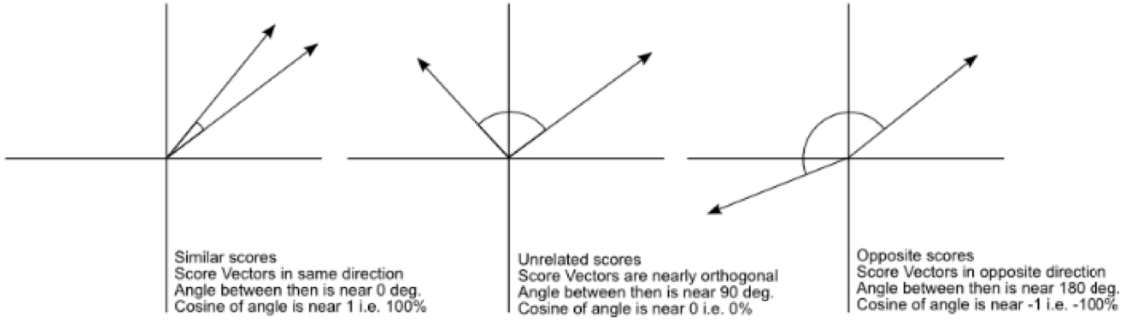
Cosine Similarity

- Measures the cosine of angle between two vectors
- Measure of orientation and not magnitude on a normalized space

http://blog.christianperone.com/2013/09/machine-learning-cosine-similarity-for-vector-space-models-part-iii/

$$\vec{a} \cdot \vec{b} = \|\vec{a}\| \|\vec{b}\| \cos\theta$$

$$\cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \|\vec{b}\|}$$



The Cosine Similarity values for different documents, 1 (same direction), 0 (90 deg.), -1 (opposite directions).

- Dot product $\vec{a} \cdot \vec{b} = \sum_{i=1}^{n} a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$
- Distance $|\mathbf{x}| = \sqrt{\sum_{k=1}^{n} |x_k|^2}$, becomes a sqrt of count of "1"s in x
- For fingerprints: $\frac{c}{\sqrt{ab}}$

Metric Fusion

- Use Metric Fusion to avoid biasing similarity analysis by any one metric
  - *n* similarity metrics produce a "fused" metric
  - Must define a "fusion rule"
    - Max, Min, median, averages based on scores
    - Rank-based, where molecules are ranked

Which Metric is Best for Molecular Similarity?

- "Best" in comparison to what?
  - There are a series of studies to determine this
  - Bajusz, 2015: metric that on its own produces the most similar rankings to those produced using the fusion of the other metrics

Bajusz, D., Rácz, A., & Héberger, K. (2015). Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?. Journal of cheminformatics, 7(1), 20.

Which Metric is Best for Molecular Similarity?

- Most recent, Bajusz, 2015, used sum of ranking differences, and analysis of variance methods and analyzed several metrics. Conclusions:
  - Each similarity metric produced more reliable rankings than random numbers.
  - Cosine, Dice, Tanimoto and Soergel similarities were identified as the best (equivalent) similarity metrics
  - Euclidean and Manhattan distances are far from being optimal.
    - Dependent on molecule size
    - But deviation from others makes them orthogonal and useful in some circumstances

Bajusz, D., Rácz, A., & Héberger, K. (2015). Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?. Journal of cheminformatics, 7(1), 20.

Summary: Molecular representations and similarity metrics

- Multiple ways of representing molecules
- Multiple similarity metrics
- Determine your application first before selecting either