

INTRODUCCIÓN A LA QUIMIOINFORMÁTICA

Fernán Agüero
Instituto de Investigaciones Biotecnológicas, UNSAM



INTRODUCTION TO CHEMINFORMATICS

Cheminformatics is a relatively new field of information technology that focuses on the collection, storage, analysis, and manipulation of chemical data. The chemical data of interest typically includes information on small molecule formulas, structures, properties, spectra, and activities (biological or industrial). Cheminformatics originally emerged as a vehicle to help the drug discovery and development process, however cheminformatics now plays an increasingly important role in many areas of biology, chemistry, and biochemistry. The intent of this unit is to give readers some introduction into the field of cheminformatics and to show how cheminformatics not only shares many similarities with the field of bioinformatics, but that it can also enhance much of what is currently done in bioinformatics.

-- David Wishart

CHEMINFORMATICS – QUÉ ES?

“The application of computational techniques to the discovery, management, interpretation and manipulation of chemical information and data extracted therefrom”.

Chemistry plans a structural overhaul.

Nature 419:4-7 (2002)

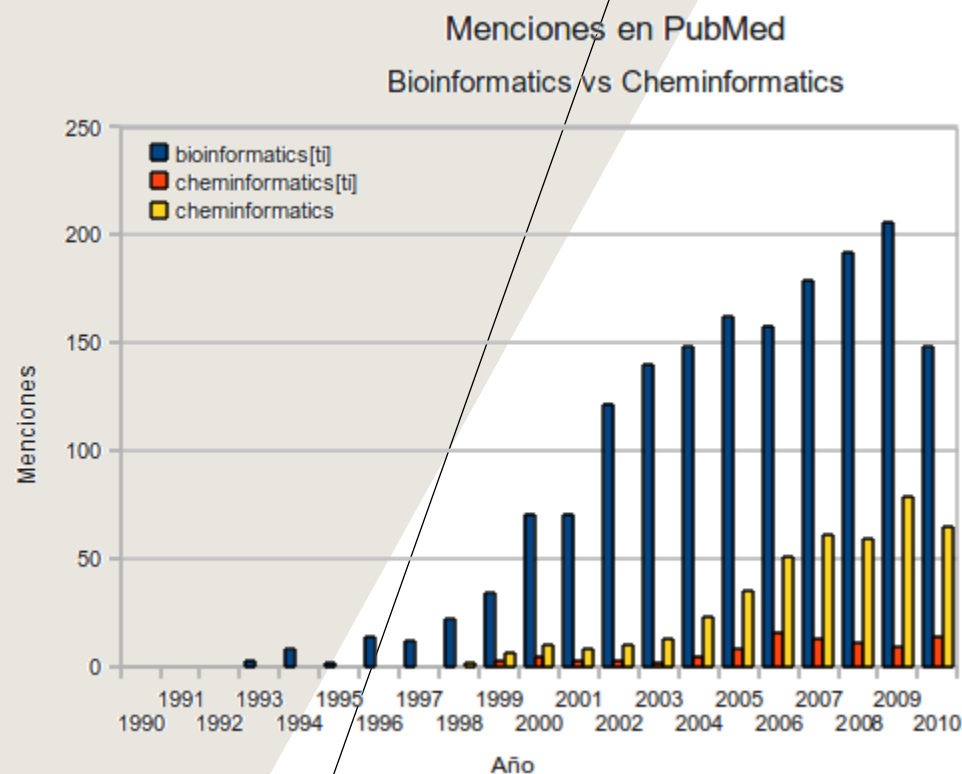
Se la conoce como:

- Computational chemistry
- Theoretical chemistry
- Molecular modeling

Nace con el desarrollo de la mecánica cuántica a principios del siglo XX

Parece haber pasado desapercibida en la revolución “ómica”

En activo desarrollo y expansión a partir de la introducción de las computadoras



CHEMINFORMATICS EN LA LITERATURA

Term	<i>Google</i>	<i>Google Scholar</i>	<i>Web of Knowledge</i>	<i>Scopus</i>
Chemical documentation	695,000	66	1	34
Chemical informatics	50,400	129	20	39
Chemical information management	978	42	4	28
Chemical information science	779	17	2	5
Chemiinformatics	2,230	2	2	2
Cheminformatics	320,000	447	83	250
Chemoinformatics	191,000	5636	99	473

Table 1. Occurrences of search terms in *Google*, *Google Scholar*, the *Web of Knowledge* and *Scopus*

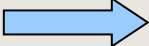
Google:

- “Bioinformatics” (2023): ~ 240 millones de páginas
- “Cheminformatics” (2023): ~ 1.6 millones de páginas

Willett P (2007). *A bibliometric analysis of the literature of chemoinformatics*. **Aslib Proceedings**, 60: 4-17

CUESTIONES QUÍMICAS

La química se ocupa de esto

estructura  propiedades

Compuestos

- Propiedades Físicas (→ Energía)
- Propiedades Químicas (Estructura, Reactividad)
- Propiedades Biológicas (→ Actividad)
- Separaciones de mezclas de compuestos

• **Aspectos estáticos**

Transformaciones

- Reacciones químicas
- **Aspectos dinámicos**

Y tiene estos desafíos

propiedades  estructura

Inferencia

- Qué compuestos (estructuras) van a mostrar una determinada propiedad?
 - Inhibición de una actividad enzimática X (ej. drogas)
 - Propiedades mecánicas y elásticas definidas (ej. polímeros)
- Definir caminos óptimos para la síntesis de compuestos
 - Reacciones
 - Materiales iniciales
- Predecir estructuras
 - A partir de datos experimentales (ej NMR)
 - Compuestos desconocidos

Síntesis, Abstracciones,
Predicciones

Insight, Wisdom

Utilización de información para
aplicaciones (memorización de
datos)

Knowledge

Datos ordenados, refinados
y puestos en contexto

Information

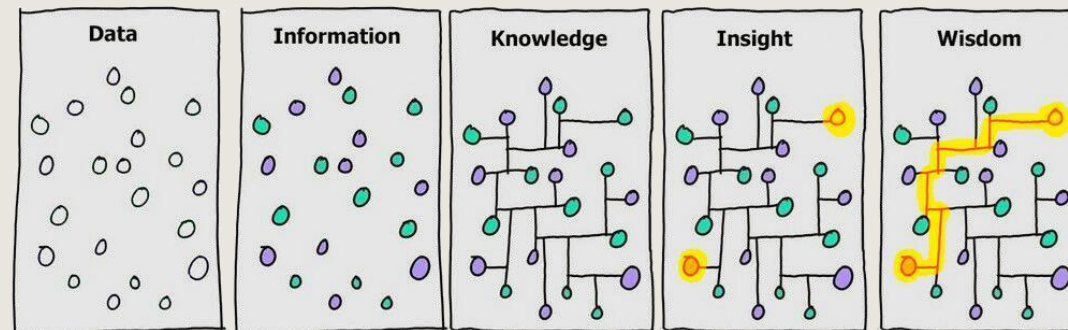
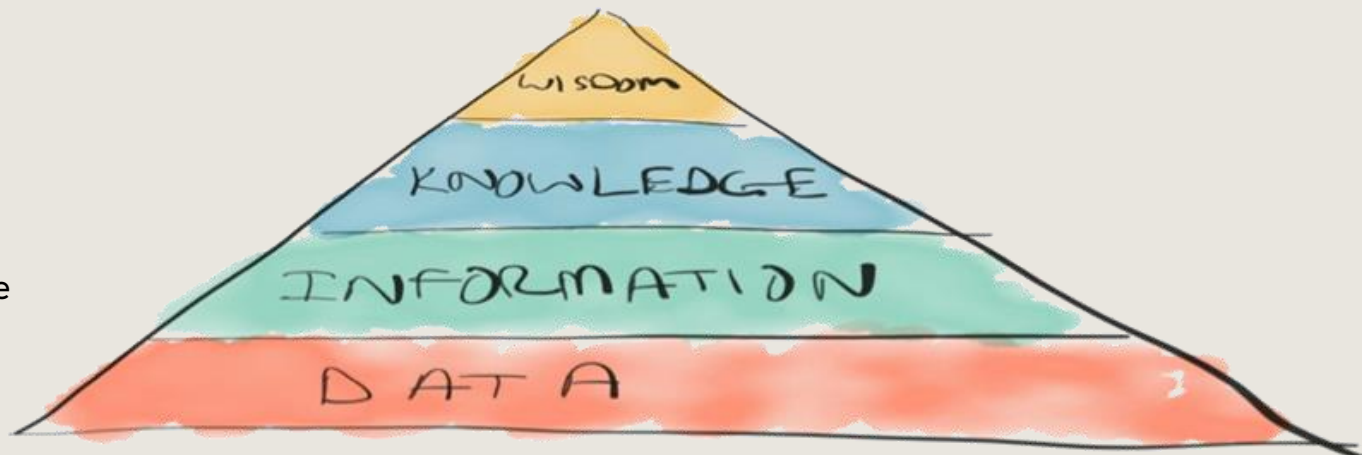
Experimentos, Mediciones

Data

Predecir

EL DESAFÍO DE LA QUIMIOINFORMÁTICA

Transformar datos
en conocimiento



- El curso de una reacción química en un solvente determinado, a una temperatura dada y usando un catalizador definido
- La actividad biológica de un compuesto X contra una proteína target Y

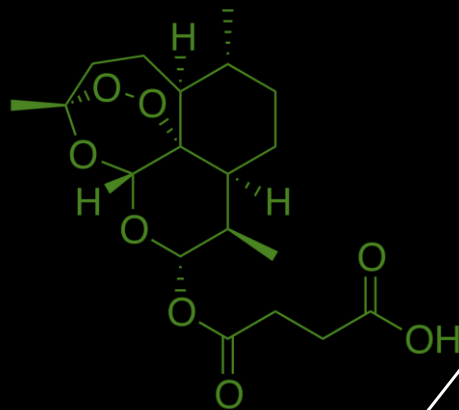
MEDITACIONES FILOSÓFICAS

“This thing, what is it in itself, in its own constitution? What is its substance and material? And what its causal nature?”

– *Marcus Aurelius*

“The history of chemistry is an elaboration of these three questions as applied to molecules: What is the essence of a molecule? What is it made of? What will it do?”

-- *Anthony Nicolls et al. Molecular shape and medicinal chemistry: a perspective. 2010. J Med Chem 53: 3862*



TEORÍA VS MODELOS



En esencia son lo mismo, pero

REPRESENTACIÓN DE COMPUESTOS QUÍMICOS

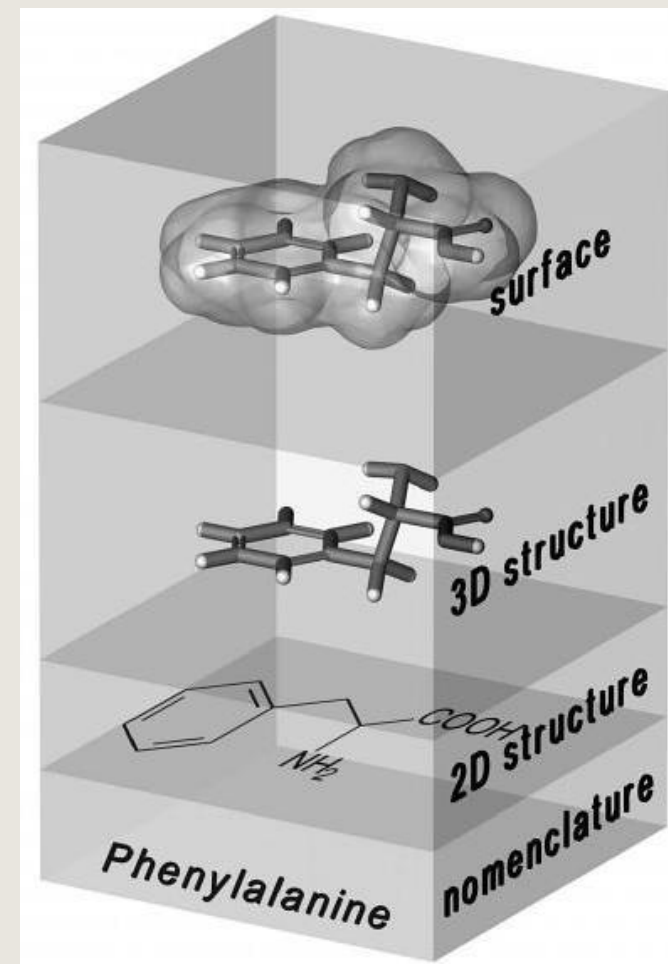
2D Structure vs 3D Structure

2D: Lenguaje natural “universal” entre químicos

- Explica la topología de una molécula
- Qué átomos están conectados mediante qué enlaces
- No explica el arreglo tridimensional de los átomos

3D: Requiere datos adicionales

- Posición de los átomos en el espacio
- Ángulos y distancias de los enlaces



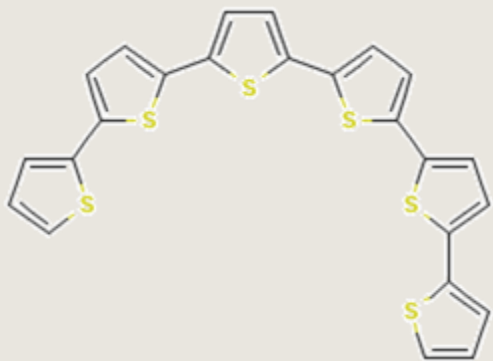
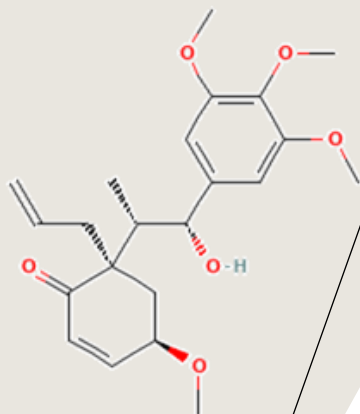
Hierarchical scheme for representations of a molecule with different content of structural information.

Tomado de J Gasteiger & T Engel (2003).

Moléculas con
nombres
populares raros:



Traumatic acid
Erotic acid
Commic acid
Diabolic acid
Megaphone
Sexitiophene



NOMENCLATURA QUÍMICA

Histórica

aqua fortis (nitric acid)
oil of vitriol (sulfuric acid)
sweet oil of vitriol (diethyl ether)

Trivial

Fenilalanina
Ibuprofeno

Popular, pero difícil de sistematizar

IUPAC

2-amino-3-phenylpropanoic acid
2-[4-(2-methylpropyl)phenyl]propanoic acid

Sistemático, pero los nombres pueden ser largos!

Fórmula empírica

C₉H₁₁NO₂
C₁₃H₁₈O₂

Ambiguo: varios compuestos pueden tener la misma fórmula

REPRESENTACIÓN DE COMPUESTOS QUÍMICOS: SMILES

SMILES (Simplified Molecular Input Line Entry System)

Introducido en 1986 por David Weininger

Representa moléculas en forma lexicográfica

Usa conceptos de grafos | Nodos conectados a través de *aristas* o *arcos*

Reglas:

Los átomos se representan con sus respectivos símbolos:

C, N, Br, Na, Cl, O, F

MAYUSCULAS → alifáticos; minúsculas → aromáticos

Los hidrógenos son implícitos

Los átomos vecinos aparecen juntos

Se usan paréntesis cuando hay más de un vecino: ramificaciones

Enlaces dobles se representan usando '='

Enlaces triples se representan usando '#'

Quiralidad: '@' (contrario a las agujas del reloj)

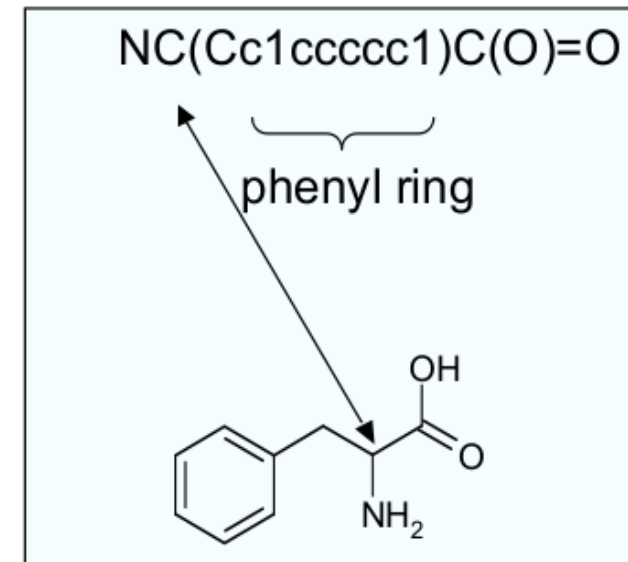
'@@' (en el sentido de las agujas del reloj)

Anillos: números a continuación de los átomos que abren/cierran el ciclo

Más información y reglas en:

<https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html>

https://en.wikipedia.org/wiki/Simplified_molecular-input_line-entry_system



Otros ejemplos:

Ciclohexano: C1CCCCC1

Benceno: C1=CC=CC=C1 (Kekulé)

Benceno: c1ccccc1

Etanol: CCO

Piridina: C1=CC=NC=C1 (Kekulé)

Piridina: c1ccncc1

Acido acético: CC(=O)O

Acido cianhídrico: C#N

L-alanina: N[C@@H](C)C(=O)O

L-alanina (sin especificar quiralidad): N[CH](C)C(=O)O

Cloruro de Sodio: [Na+].[Cl-]

ANILLOS EN SMILES

Linealizar y Etiquetar

Linealizar el anillo en cualquier parte

Benceno: ccccc(C=CC=CC=C)

Dioxano: occocc, ccocco, coccoc

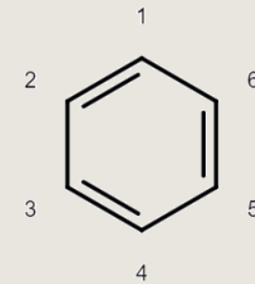
Agregar etiquetas numéricas para indicar el inicio y cierre del anillo

Benceno: c1ccccc1

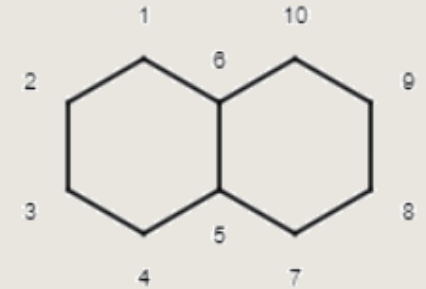
Dioxano: O1CCOCC1, C1COCCO1, C1OCCOC1

*Las etiquetas numericas pueden empezar en cero (0) **pero** rara vez se usa*

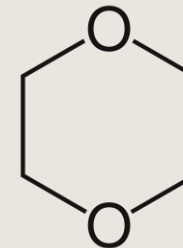
Decalin: C1CCCC2C1CCCC2, C1CCCC2CCCCC12



benceno



decalin



dioxano

REPRESENTACIÓN DE PATRONES EN MOLECULAS

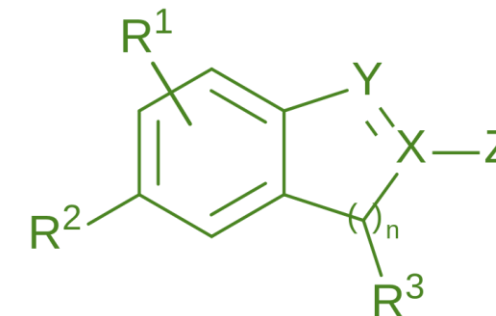
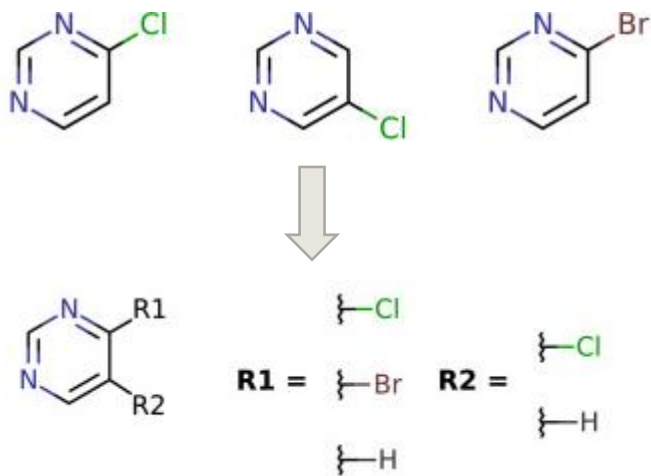
Al principio hubo **Markush structures**:

Representan varias estructuras posibles

Grupos R variables

Descripción general de una molécula con **ambigüedad** en algunas posiciones

Son comunes en patentes, y en libros de texto.



https://es.wikipedia.org/wiki/Estructura_de_Markush



Eugene A. Markush

REPRESENTACIÓN DE PATRONES: SMARTS

SMARTS - A Language for Describing Molecular Patterns

Representación lexicográfica de partes de una molécula

Es una extensión de **SMILES**

Concepto similar al de **expresiones regulares** (regex) en texto.

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

Reglas (las mismas que SMILES), y además:

Representación de patrones para átomos:

- * cualquier átomo

- a aromático

- A alifático ... hay más reglas para átomos

Representación de patrones para enlaces:

- ~ cualquier enlace

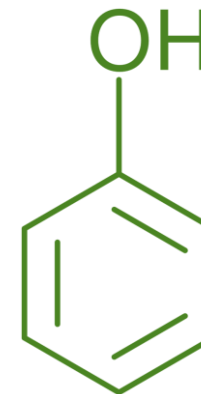
- @ cualquier enlace en un anillo

- / enlace dirigido "arriba"

- \ enlace dirigido "abajo"

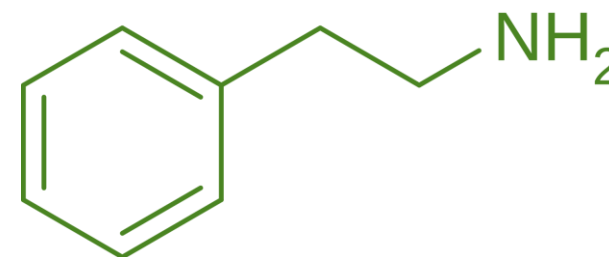
Más información y reglas en:

<https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html>



SMARTS: [OH]c1ccccc1

hydroxyl-group attached to 6 aromatic carbons in a ring

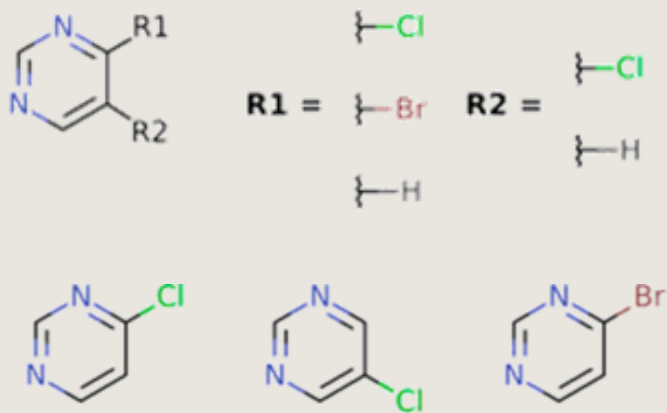


SMARTS: NCCc1ccccc1

Aliphatic nitrogen attached to 2 aliphatic carbons attached to 6 aromatic carbons in a ring

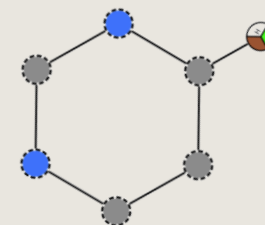
FROM MARKUSH TO SMARTS

Original molecules

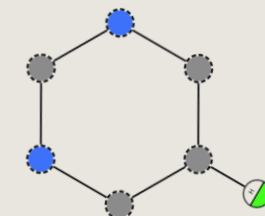


SMARTS PATTERNS

n1cnc([Cl,Br,H])cc1



n1cncc([Cl,H])c1



SMARTS.PLUS SmartView: <https://smarts.plus/>

SMARTS – A LANGUAGE FOR DESCRIBING MOLECULAR PATTERNS

Una representación SMILES es un patrón SMARTS válido

[OH]c1ccccc1 (phenol)

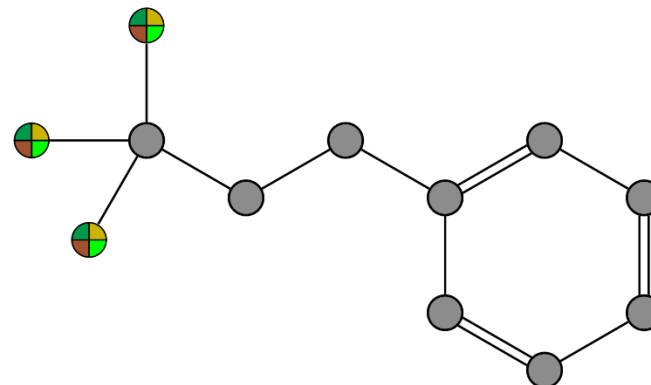
Patrones SMARTS simples

[C,N]1CCCCC1

[Cl,Br,F,I]C([Cl,Br,F,I])([Cl,Br,F,I])CCC1=CC=CC=C1

C-C=C-C=C~*~[++]

[Cl,Br,F,I]C([Cl,Br,F,I])([Cl,Br,F,I])CCC1=CC=CC=C1



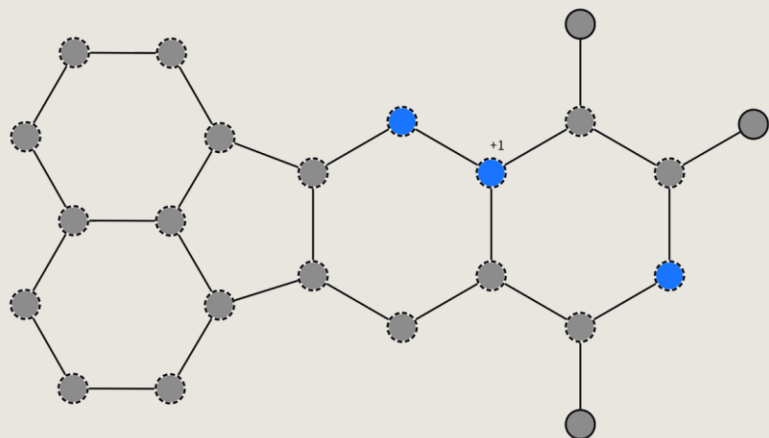
Picture created by the SMARTSviewer [<https://smarts.plus/>].
Copyright: ZBH - Center for Bioinformatics Hamburg.

LEGEND			
	default bond		Cl or Br or F or I
			aliphatic C
	aliphatic		

Referencias

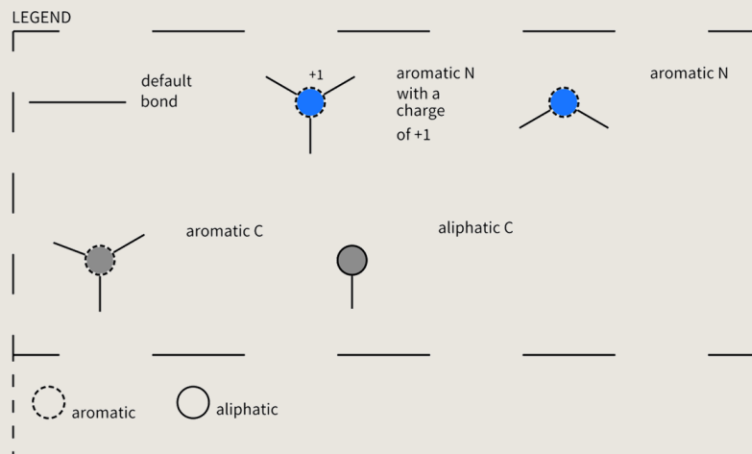
<https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html>
<https://smarts.plus/>

PATRONES SMARTS PARA BÚSQUEDAS



Referencias
<https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html>
<https://smarts.plus/>

Picture created by the SMARTSviewer (<https://smarts.plus/>).
 Copyright: ZBH - Center for Bioinformatics Hamburg.



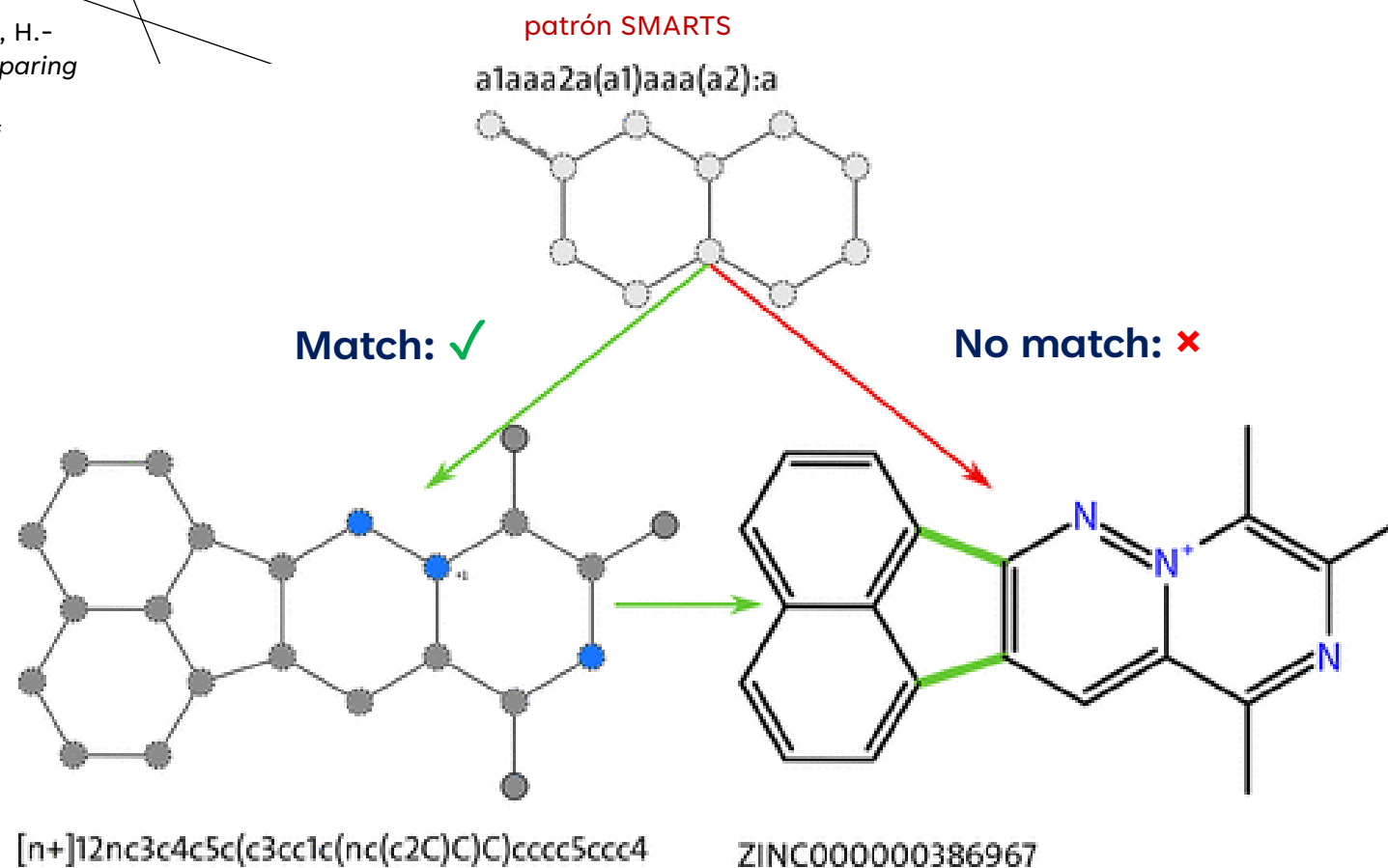
[n+]12nc3c4c5c(c3cc1c(nc(c2C)C)C)cccc5ccc4

C = carbono alifático

c = carbono aromático

MATCHING SMARTS PATTERNS

Schmidt, R., Ehmki, E. S. R., Ohm, F., Ehrlich, H.-C., Mashychev, A., & Rarey, M. (2019). *Comparing Molecular Patterns Using the Example of SMARTS: Theory and Algorithms*. *Journal of Chemical Information and Modeling*. doi:10.1021/acs.jcim.9b00250



TESTING AT SMARTS.PLUS

<https://smarts.plus/>



View Compare Search Create

Compare two SMARTS expression with respect to subset relation (Does expression A match whenever B matches?) or similarity and receive a visualization of the node mapping.

SMARTS pattern:

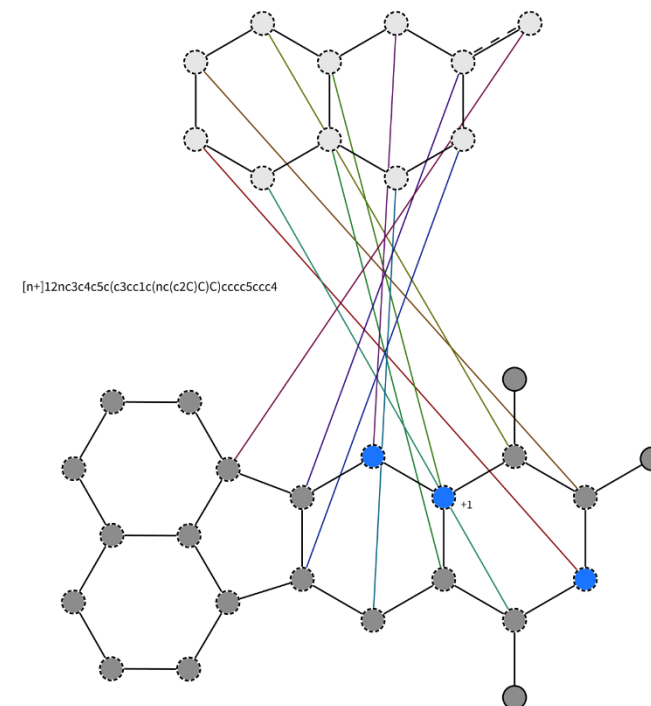
CC1=NC(C)=C(C)[N]2=C1C=C1C3=CC=CC4=C3C(=CC=C4)C1=N2

SMARTS to compare:

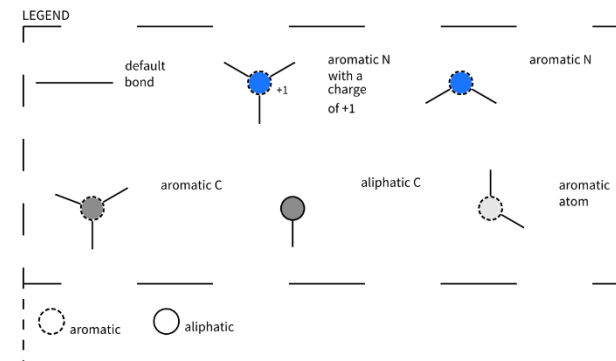
[n+]12nc3c4c5c(c3cc1c(nc(c2C)C)C)cccc5ccc4

More Options

Go!



Picture created by the SMARTSviewer (<https://smarts.plus/>).
Copyright: ZBH - Center for Bioinformatics Hamburg.



[NX3,NX4+][CX4H]([*])[CX3](=[OX1])[O,N]
Generic amino acid: low specificity
[NX3,NX4+][CX4H]([*])[CX3](=[OX1])[O,N]

SMARTS EXAMPLES

Amino Acids

Generic amino acid: low specificity:

[NX3,NX4+][CX4H]([*])[CX3](=[OX1])[O,N]

Other interesting examples

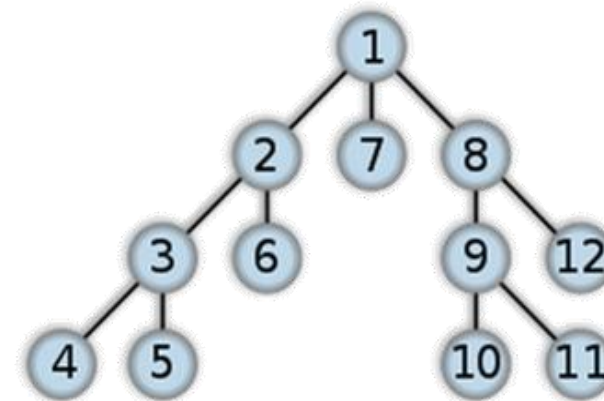
https://www.daylight.com/dayhtml_tutorials/languages/smarts/smarts_examples.html

REPRESENTACIÓN DE COMPUESTOS QUÍMICOS: SMILES

SMILES, relación con Teoría de Grafos

SMILES es una cadena de texto (ASCII)

Es el producto de escribir los símbolos (átomos) a medida que se recorre el grafo químico (la molécula) de modo *depth-first*



Order in which the nodes are expanded

Class	Search algorithm
Data structure	Graph
Worst case performance	$O(V + E)$ for explicit graphs traversed without repetition, $O(b^d)$ for implicit graphs with branching factor b searched to depth d
Worst case space complexity	$O(V)$ if entire graph is traversed without repetition, $O(\text{longest path length searched})$ for implicit graphs without elimination of duplicate nodes

Depth-first Tree/Graph Traversal:

http://en.wikipedia.org/wiki/Depth-first_search

CANONIZACIÓN DE MOLÉCULAS: ALGORITMO DE MORGAN

Canonización: Representar la conectividad de una molécula de manera uniforme

Una estructura con n átomos puede ser descripta de $n!$ maneras diferentes

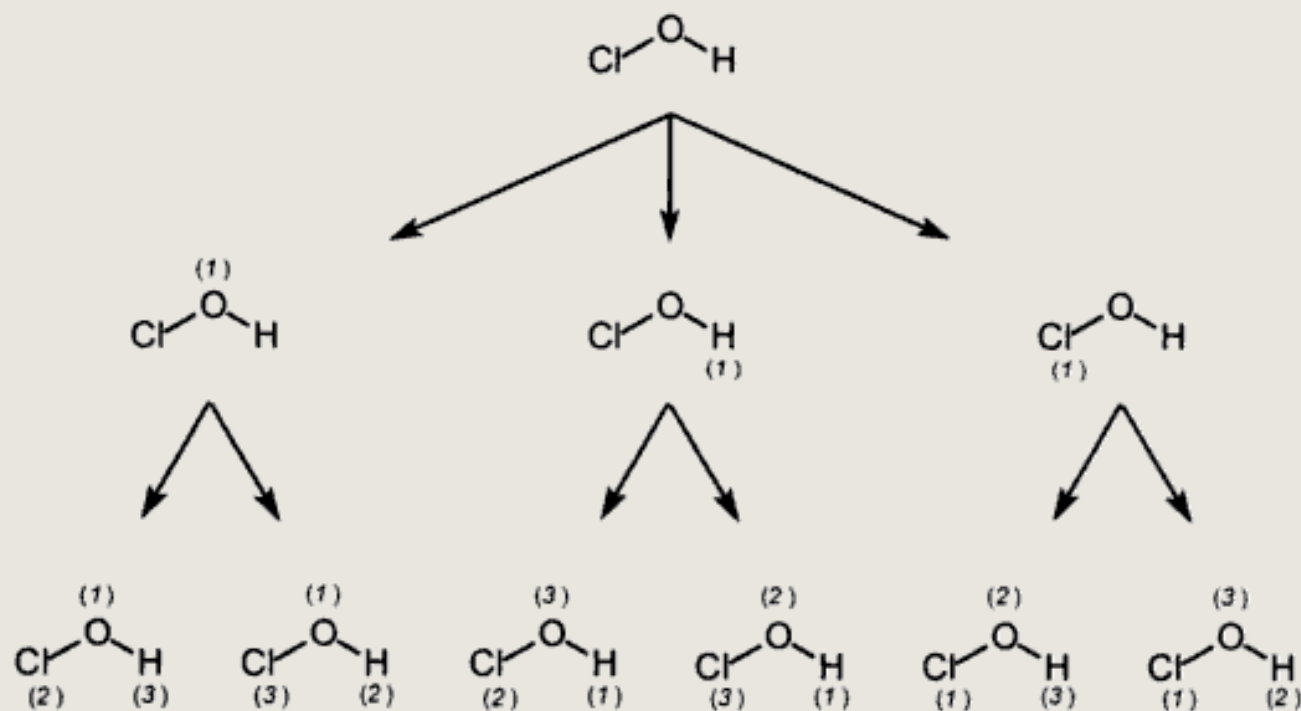


Figure 2-41. Six different possibilities for numbering the atoms in a hypochlorous acid molecule.

El algoritmo de Morgan es viejo pero lo vamos a usar para aprender el concepto de **canonización**!

Hay variantes nuevas!

Schneider N, Sayle RA, Landrum GA. Get Your Atoms in Order--An Open-Source Implementation of a Novel and Robust Molecular Canonicalization Algorithm. J Chem Inf Model. 2015 Oct 26;55(10):2111-20. doi: 10.1021/acs.jcim.5b00543. Epub 2015 Oct 15. PMID: 26441310.

CANONIZACIÓN: ALGORITMO DE MORGAN

Paso 1: clasificar átomos de acuerdo a conectividad (vecindad)

Estructuras conteniendo C, N, O, H y halógenos se clasifican en cuatro categorías dependiendo del número de enlaces (no H)

Paso 2: Iteraciones

En una segunda iteración los valores de conectividad de cada átomo se incrementan de acuerdo al de los vecinos siguiendo una serie de reglas:

- Sumas (átomos internos) o transferencia de valores (átomos terminales)
- *Extended connectivity*

Las iteraciones siguen hasta que los valores de EC son iguales o menores a los de la iteración anterior

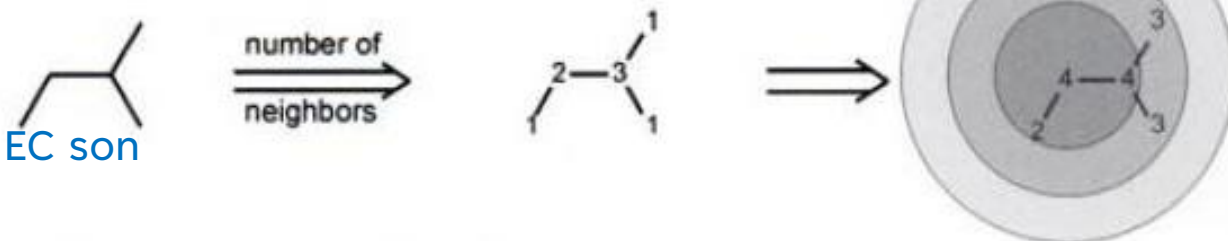


Figure 2-43. The EC value or the atom classification of each atom, respectively, is calculated by summing the EC values of the directly connected neighboring atoms of the former sphere (relaxation process).

CANONIZACIÓN: ALGORITMO DE MORGAN

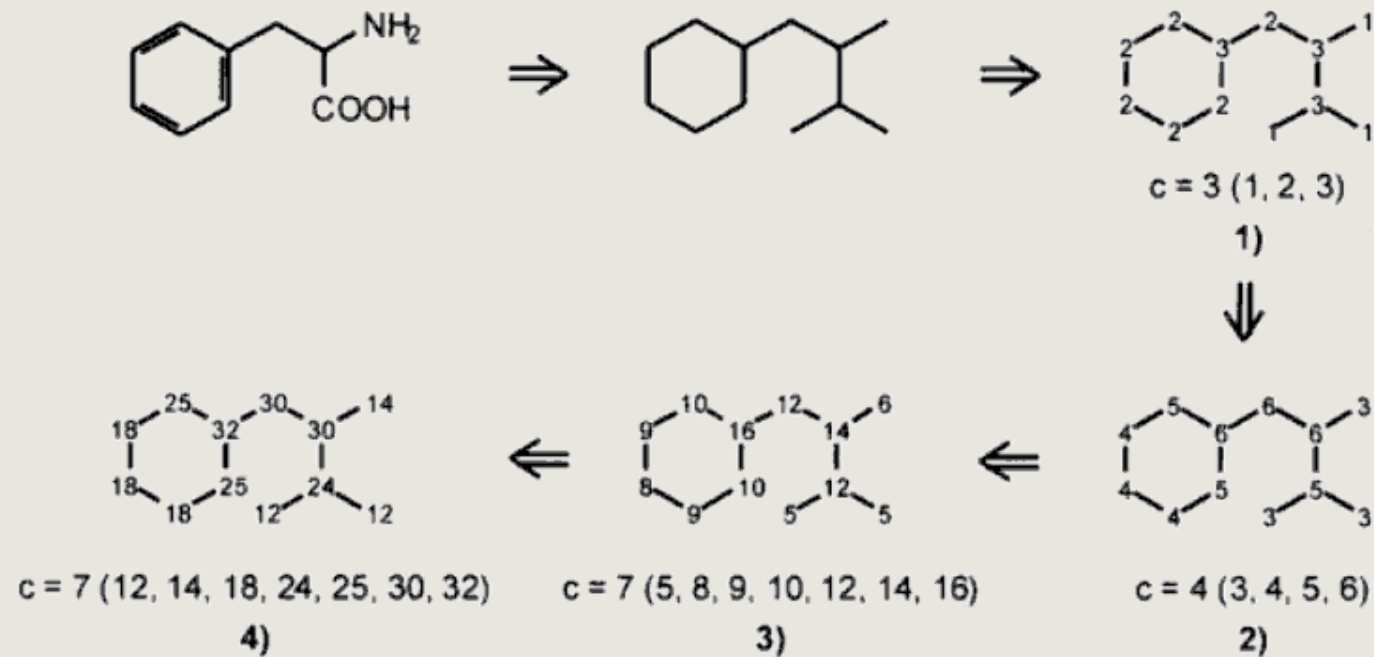


Figure 2-44. The EC values of the atoms of phenylalanine (without hydrogens) are calculated by considering the class values of the neighboring atoms. After each relaxation process, c , the number of equivalent classes (different EC values), is determined.

The process is repeated until the number of different EC values is lower than or equal to the number of different EC values in the previous iteration.

CANONIZACIÓN: ALGORITMO DE MORGAN

Paso 3: Asignación de números de átomos únicos

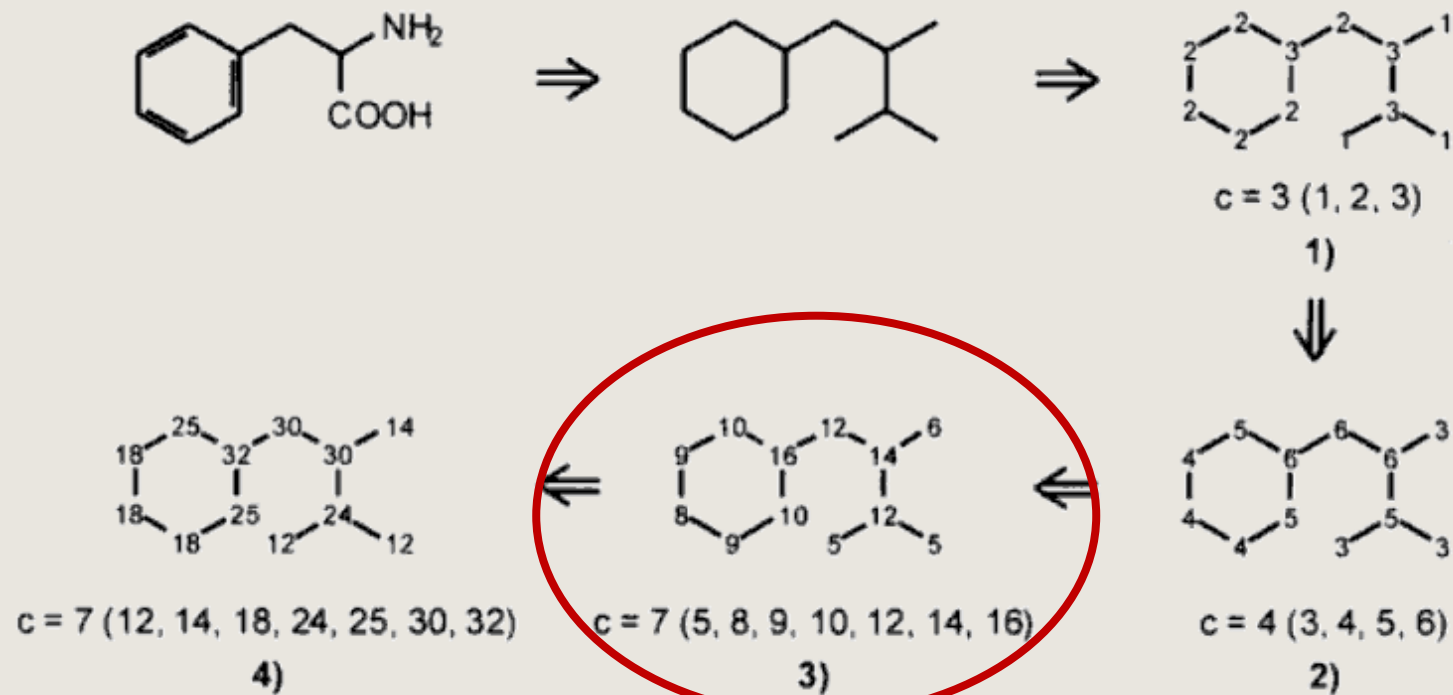


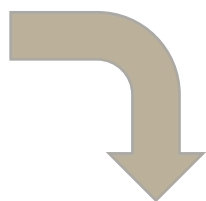
Figure 2-44. The EC values of the atoms of phenylalanine (without hydrogens) are calculated by considering the class values of the neighboring atoms. After each relaxation process, c , the number of equivalent classes (different EC values), is determined.

Se comienza por el paso en el que se obtiene el mayor EC por primera vez.

El átomo número 1 es el que tiene el mayor valor de EC en este paso.

El átomo 2 es el que sigue en la secuencia de valores EC.

El que implementa
RDKit (Python)



Schneider N, Sayle RA, Landrum GA. Get Your Atoms in Order--An Open-Source Implementation of a Novel and Robust Molecular Canonicalization Algorithm. J Chem Inf Model. 2015 Oct 26;55(10):2111-20. doi: 10.1021/acs.jcim.5b00543. Epub 2015 Oct 15. PMID: 26441310.

Krotko DG. Atomic ring invariant and Modified CANON extended connectivity algorithm for symmetry perception in molecular graphs and rigorous canonicalization of SMILES. J Cheminform. 2020 Aug 20;12(1):48. doi: 10.1186/s13321-020-00453-4. PMID: 33431026; PMCID: PMC7439248.

CANONIZACIÓN DE MOLECULAS

El algoritmo es de 1965! Es viejo!

Hay moléculas problemáticas que no son fáciles de canonizar.

El problema general que intenta resolver es el de
Canonización de Grafos

- Es un problema computacional complejo
- Relacionado con problemas de isomorfismo de grafos
- Hay muchas otras maneras (algoritmos) de resolverlos:
http://en.wikipedia.org/wiki/Graph_canonization

En resumen:

Después de aplicar un método de canonización

INTERVALO

15 minutos

May your morning coffee
give you the strength
to make it to your
mid-morning
coffee.



som_{ee}cards

REPRESENTACIÓN DE COMPUESTOS QUÍMICOS: INCHI

InChI –*I*nternational *C*hemical *I*dentifier

Introducido recientemente (2005) por IUPAC
(International Union of Pure and Applied Chemistry)

Objetivos

Establecer un identificador (nomenclatura, etiqueta) *único* y *no propietario* para cada molécula

Que pueda ser utilizado tanto en medios impresos como electrónicos y que facilite la búsqueda de compuestos

Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D. InChI, the IUPAC International Chemical Identifier. J Cheminform. 2015 May 30;7:23. doi: 10.1186/s13321-015-0068-4. PMID: 26136848; PMCID: PMC4486400.

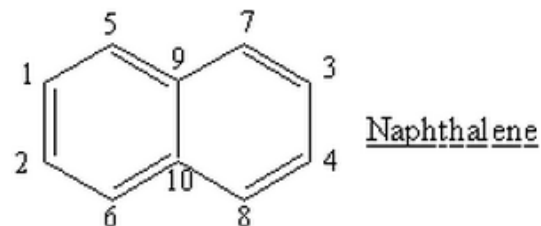
REPRESENTACIÓN DE COMPUESTOS QUÍMICOS: INCHI

Formato de un identificador InChI

Es una cadena de texto (ASCII) compuesta por *segmentos* (layers) separada por *delimitadores* (/)

Cada capa contiene distintos tipos de información estructural

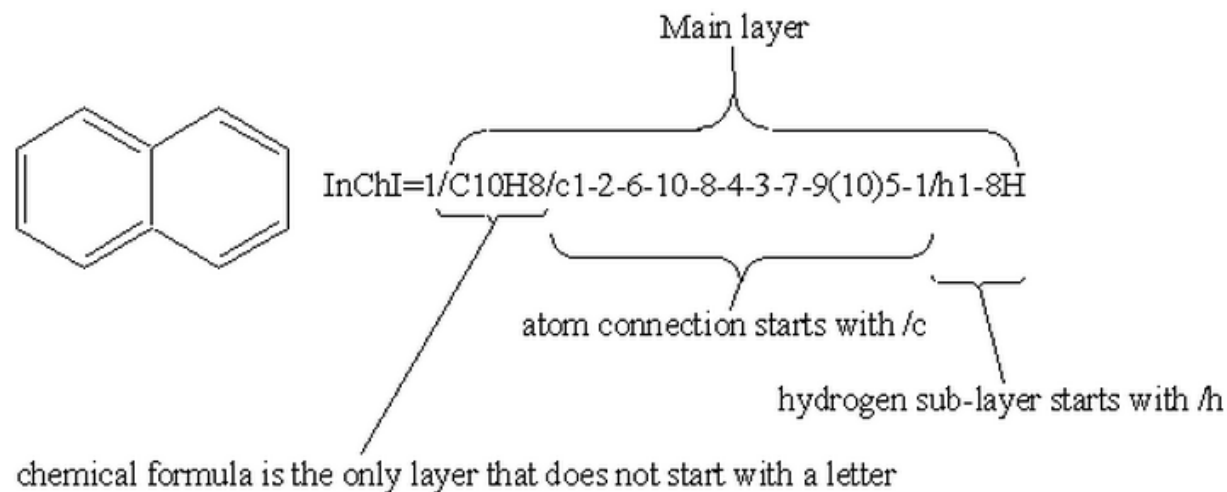
Los números dentro de una capa representan la numeración canónica de los átomos de la primera capa (fórmula) excepto los hidrógenos.



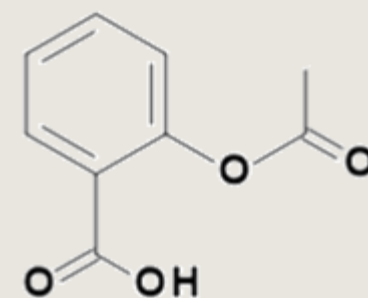
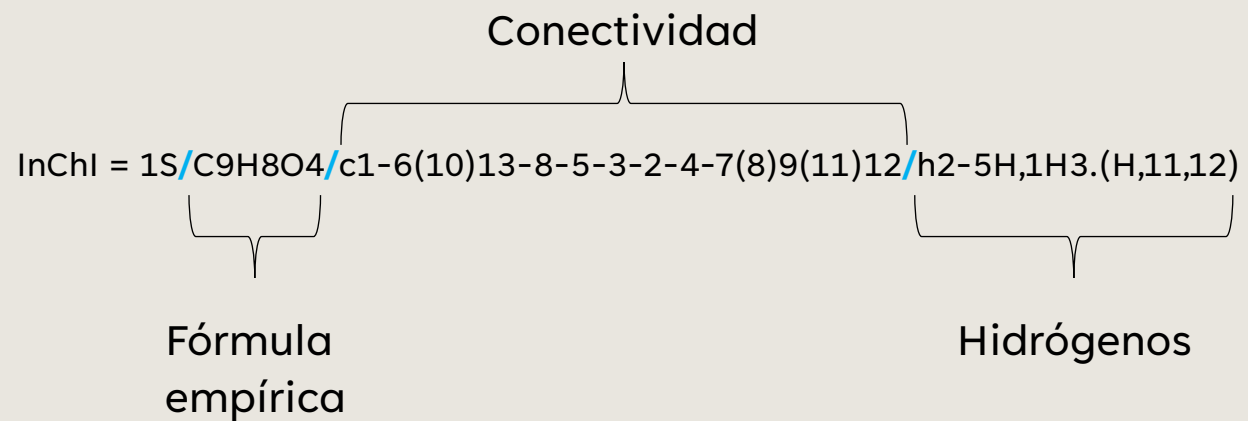
Ejemplos:

Agua: InChI = 1/H2O/h1H2

Benceno: InChI = 1/C6H6/c1-2-4-6-5-3-1/h1-6H



INCHI IDENTIFIER: MAIN LAYER



Aspirin

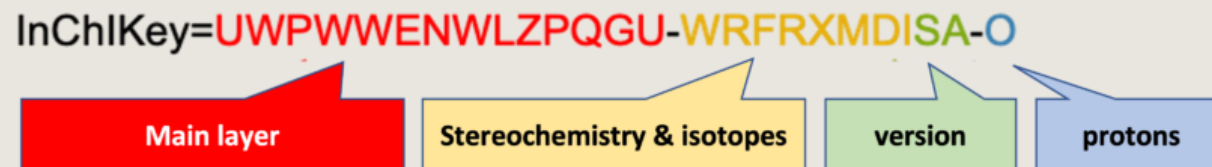
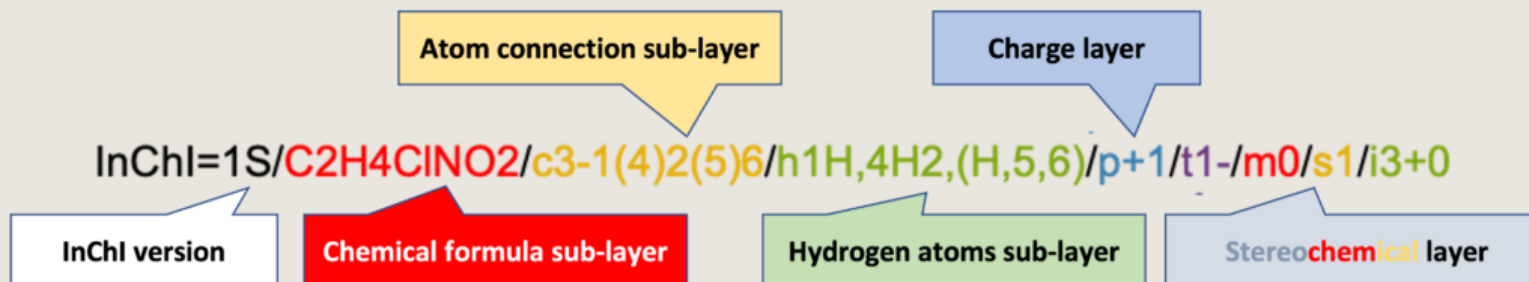
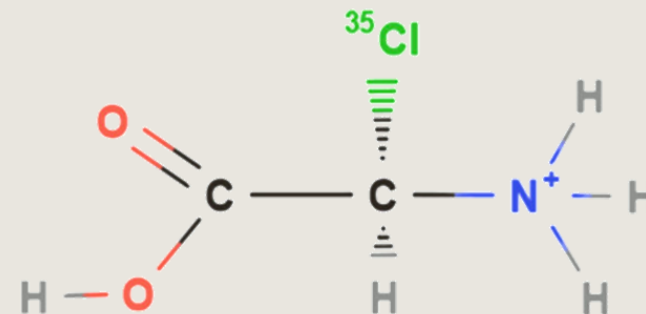
InChI = 1S/H2O/h1H2

REPRESENTACIÓN DE COMPUESTOS: INCHI

Representa la información en capas (layers)

Permiten elegir el nivel de detalle que uno quiere incluir

Sólo la capa principal es mandatoria

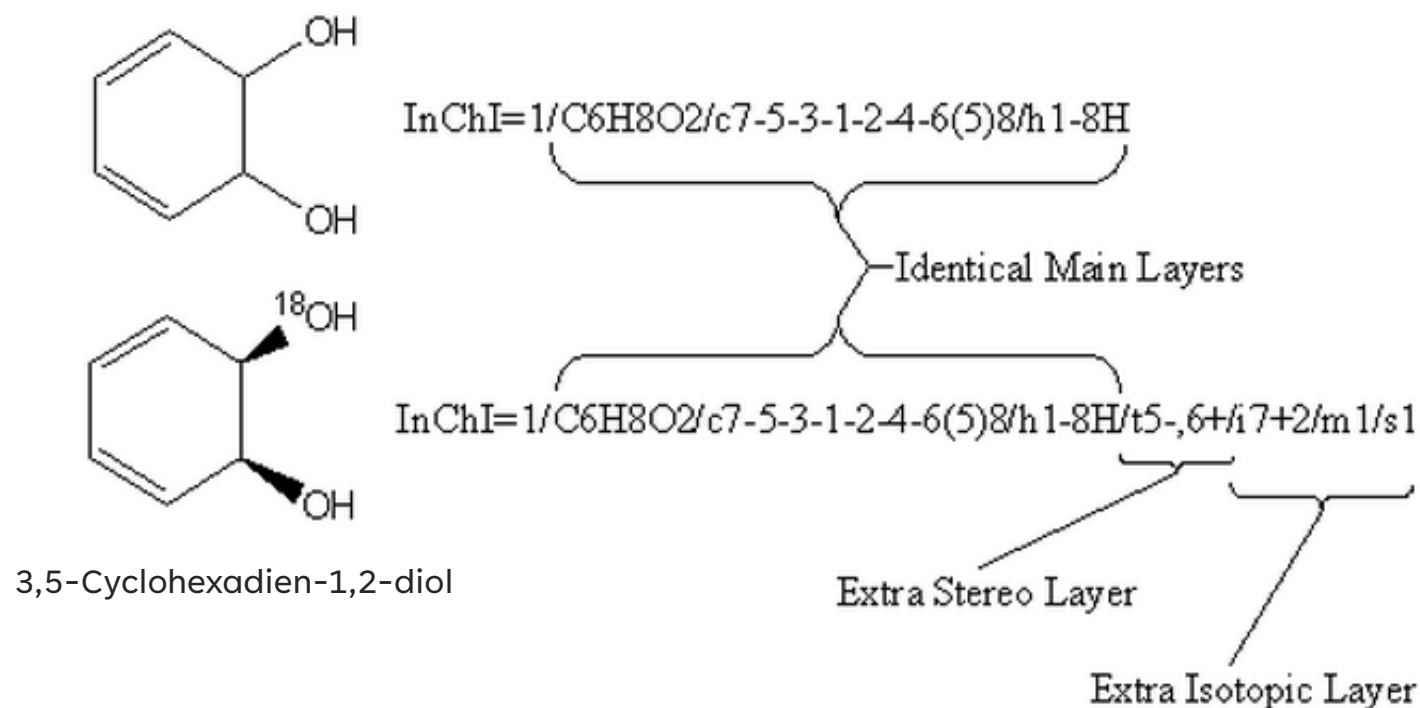


Tomado de: InChI Trust,
<https://www.inchi-trust.org/>

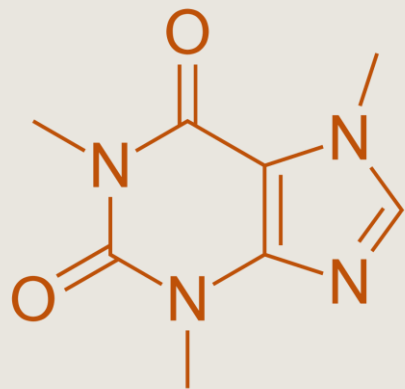
REPRESENTACIÓN DE COMPUESTOS QUÍMICOS: INCHI

Si dos InChIs son iguales, los compuestos también lo son.

Pero los compuestos pueden estar representados con diferente nivel de detalle



INCHI VS SMILES



Caffeine

Tomado de: InChI Technical FAQ
<https://www.inchi-trust.org/technical-faq-2>

Valid SMILES for Caffeine (not complete)

```
[c]1([n+])([CH3])[c]([c]2([c]([n+]1[CH3])[n][cH][n+]2[CH3]))[O-])[O-]  
CN1C(=O)N(C)C(=O)C(N(C)C=N2)=C12  
Cn1cnc2n(C)c(=O)n(C)c(=O)c12  
Cn1cnc2c1c(=O)n(C)c(=O)n2C  
O=C1C2=C(N=CN2C)N(C(=O)N1C)C  
CN1C=NC2=C1C(=O)N(C)C(=O)N2C
```

InChI: 1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3

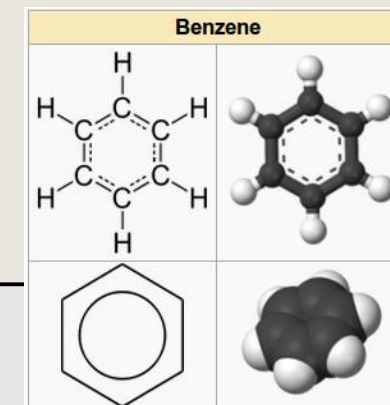
InChI Key: RYYVLZVUVIJVGH-UHFFFAOYSA-N

REPRESENTACIÓN DE COMPUESTOS: MOLFILES

MDL, Molfile | Formato creado por MDL (ahora Symyx)

Contiene información sobre: Átomos, enlaces, conectividad y *coordenadas espaciales*

Permite representar moléculas tanto en **2D** como en **3D**



Descripción del formato

1. Header

```
benzene
ACD/Labs0812062058
```

2. Comment

```
6 6 0 0 0 0 0 0 0 0 0 1 V2000
```

3. General information (counts)

6 atoms, 6 bonds, ..., V2000 standard

```
1.9050 -0.7932 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
1.9050 -2.1232 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
0.7531 -0.1282 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
0.7531 -2.7882 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
-0.3987 -0.7932 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
-0.3987 -2.1232 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
```

4. Spatial coordinates

X, Y, Z, element, extra information

5. Bonding information

1st atom, 2nd atom, bond type, extra information

```
2 1 1 0 0 0 0
3 1 2 0 0 0 0
4 2 2 0 0 0 0
5 3 1 0 0 0 0
6 4 1 0 0 0 0
6 5 2 0 0 0 0
```

8. Final del registro.

```
M END
$$$$
```

6. Atom type

7. Non-standard values

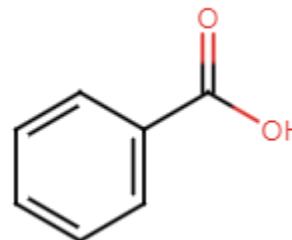
Isotopes, valence, charge

MOLFILES: BOND BLOCK

Anatomy of a MOL file
ChemInformatics 2017 (LibreTexts Chemistry)



OC(=O)C1=CC=CC=C1



chemdraw-Dec-2016.cdx
ChemDraw12011615112D

First atom
row number

Second atom
row number

First atom row number	Second atom row number	Bond type	Bond stereochemistry
1	2	1	0
2	3	2	0
3	4	1	0
4	5	2	0
5	6	1	0
6	1	2	0
5	7	1	0
7	8	1	0
7	9	2	0

M END

1	2	1	0
2	3	2	0
3	4	1	0
4	5	2	0
5	6	1	0
6	1	2	0
5	7	1	0
7	8	1	0
7	9	2	0

0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0

Bond
type

Bond
stereochemistry

REPRESENTACIÓN DE COMPUESTOS: SDF FILES

NGC00015959-03.sdf

MOLFILE

Anotaciones

```
NGC00015959-03
Marvin 07111412562D

25 30 0 0 0 0          999 V2000
 3.4098 -1.3130 0.0000 N 0 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 4.8329 -1.3130 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 3.4098 -2.1380 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 4.1248 -2.5436 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 2.6948 -2.5436 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 4.8329 -2.1380 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 4.1248 -0.8937 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 5.5547 -0.8937 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

 1 3 1 0 0 0 0
 1 7 2 0 0 0 0
 1 25 1 0 0 0 0
 2 7 1 0 0 0 0
 2 6 2 0 0 0 0
 2 8 1 0 0 0 0
 3 4 2 0 0 0 0
 3 5 1 0 0 0 0
 4 13 1 0 0 0 0
 4 6 1 0 0 0 0
 5 9 1 0 0 0 0

M CHG 1 1 1
M END
> <Formula>
C20H14NO4
> <FW>
332.3289
> <DSSTox_CID>
25204
> <Active>
1
```

CHEMICAL DATABASES

PubChem, NCBI | <https://pubchem.ncbi.nlm.nih.gov/>

repositorio abierto de information sobre moléculas y sus actividades biológicas

ChEMBL, EBI | <https://www.ebi.ac.uk/chembl/>

Repositorio abierto de bioactividades de moléculas, extraídas de la literatura

ChemSpider, Royal Society of Chemistry | <http://www.chemspider.com/>

NIST Chemistry Web Book | <https://webbook.nist.gov/>

DrugBank | <http://www.drugbank.ca/>

Zinc Databases | <https://zinc.docking.org/>

commercially-available compounds for virtual screening

PubChem

ChEMBL

ChemSpider
The free chemical database

NIST NATIONAL INSTITUTE OF
STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

ZINC20

REPRESENTACIÓN DE COMPUESTOS QUÍMICOS: GRAFOS

Un grafo es una estructura *abstracta* que contiene *nodos* conectados con *aristas* (o *arcos*)

“Los grafos son redes (networks) de puntos y líneas”

En inglés: *nodes*, *edges*

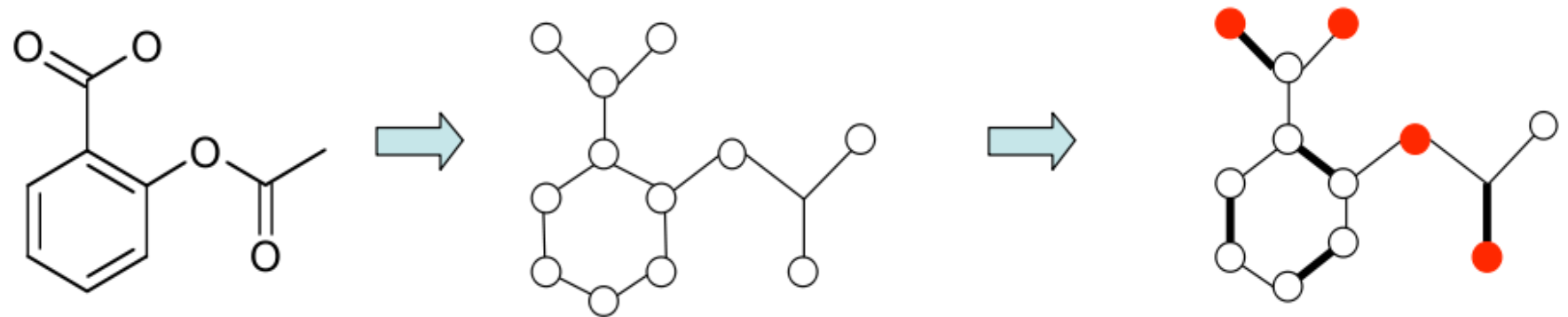
Moléculas químicas pueden representarse como *grafos*:

Los átomos como nodos

Los enlaces como aristas

Se pueden asociar propiedades a cada nodo (ej número atómico), y a cada arista (ej número y/o tipo de enlace)

En el grafo final pueden entonces distinguirse distintos tipos de nodos y aristas



UN DESVÍO: HISTORIA DE LOS GRAFOS

El problema de **los 7 puentes de Königsberg**.

La ciudad de Königsberg se encuentra dividida por el río Pregel

Incluye 2 islas que se conectan con tierra mediante 7 puentes

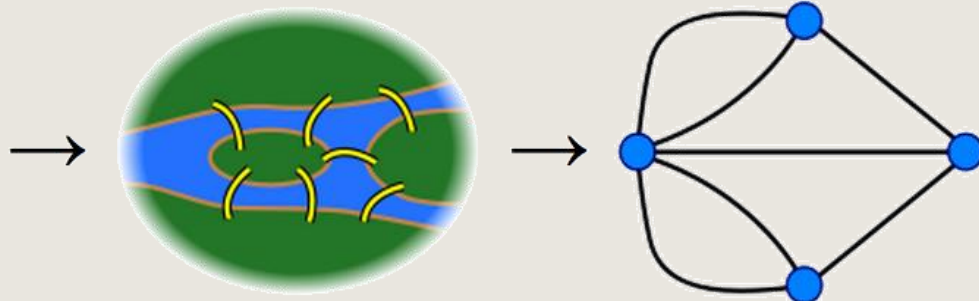
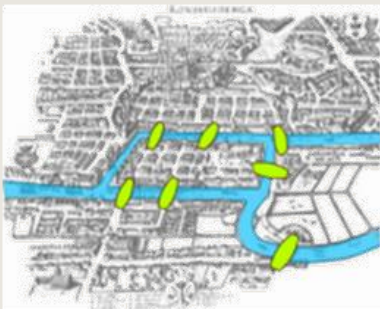
El problema: Encontrar un camino a través de la ciudad que cruce cada puente una sola vez. Hay que cruzar todos los puentes. Sólo se puede acceder a las islas cruzando un puente.

En 1735 Leonard Euler demostró que el problema no tiene solución.

El razonamiento:

La elección del camino dentro de cada porción de tierra era **irrelevante**

La única característica de la ruta elegida importante era la secuencia de puentes cruzados



Leonard Euler (1707-1783)

Abstracción del problema:

En una lista de porciones de tierra (**nodos**)

Y una lista de puentes (**aristas**)

Sólo la información de **conectividad** era relevante!

GRAFOS: PROPIEDADES Y OPERACIONES

Propiedades de los grafos:

Grado de conectividad de los nodos (degree)

Direccionalidad de las aristas

Intensidad (sentido vectorial) de cada arista

Las aristas pueden tener asociado un valor numérico (peso, largo, costo)

Posibilidad de identificar los nodos

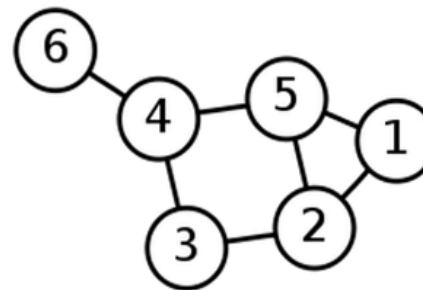
Son elementos de un conjunto

Grafos etiquetados (labeled)

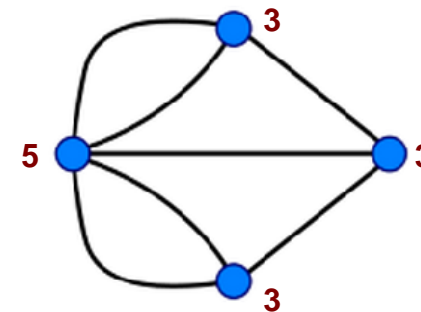
vs no-etiquetados (unlabeled)

Operaciones con grafos (algunos ejemplos):

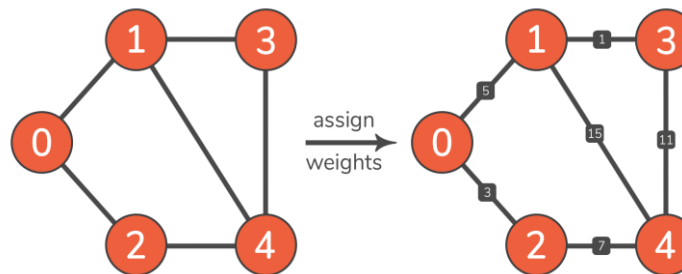
Complementación, Unión, Suma, Intersección, Diferencia, ...



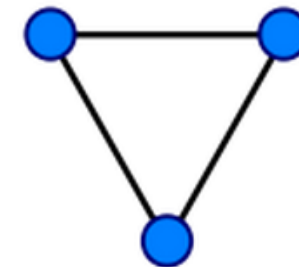
Un grafo etiquetado



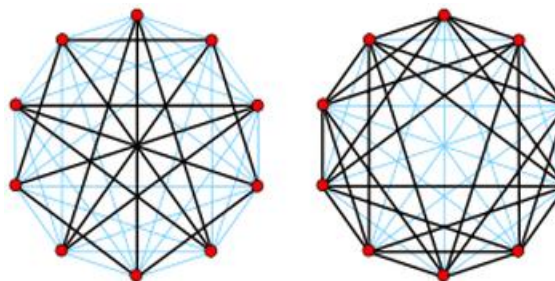
Grados de los nodos



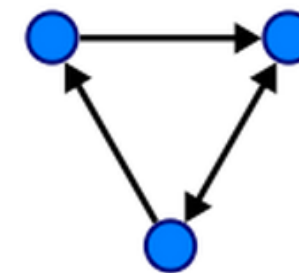
Un grafo pesado



Grafo simple o regular



Un grafo y su complemento



Grafo dirigido (red)

PROBLEMA: ENCONTRAR MOLÉCULAS IGUALES

Problema
frecuente en
química

Si representamos moléculas como **grafos**:

- dos moléculas son la misma si es posible redibujar una de ellas de manera que se vea idéntica a la otra: **Isomorphic graphs**

Problema visualmente interesante, pero la solución es obvia: **solo la conectividad es relevante!**

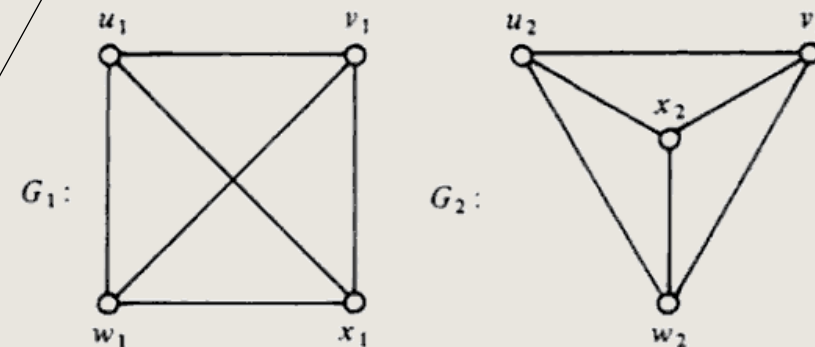
G_1 : **nodos** = {u1, v1, w1, x1}

aristas = { {u1,v1}, {u1,w1}, {u1,x1}, {v1,x1}, {v1,w1}, {x1,w1} }

G_2 : **nodos** = {u2, v2, w2, x2}

aristas = { {u2,v2}, {u2,w2}, {u2,x2}, {v2,x2}, {v2,w2}, {x2,w2} }

Problema computacionalmente sencillo (usualmente)



PROBLEMA MÁS DÍFICIL: ENCONTRAR MOLÉCULAS CON GRUPOS SIMILARES

Foye's Principles of Medicinal Chemistry (2008).
T Lemke, DA Williams. Wolters Kluwer

Otro problema común

Identificar compuestos que comparten grupos químicos similares

Farmacóforos – grupos químicos responsables de actividad farmacológica

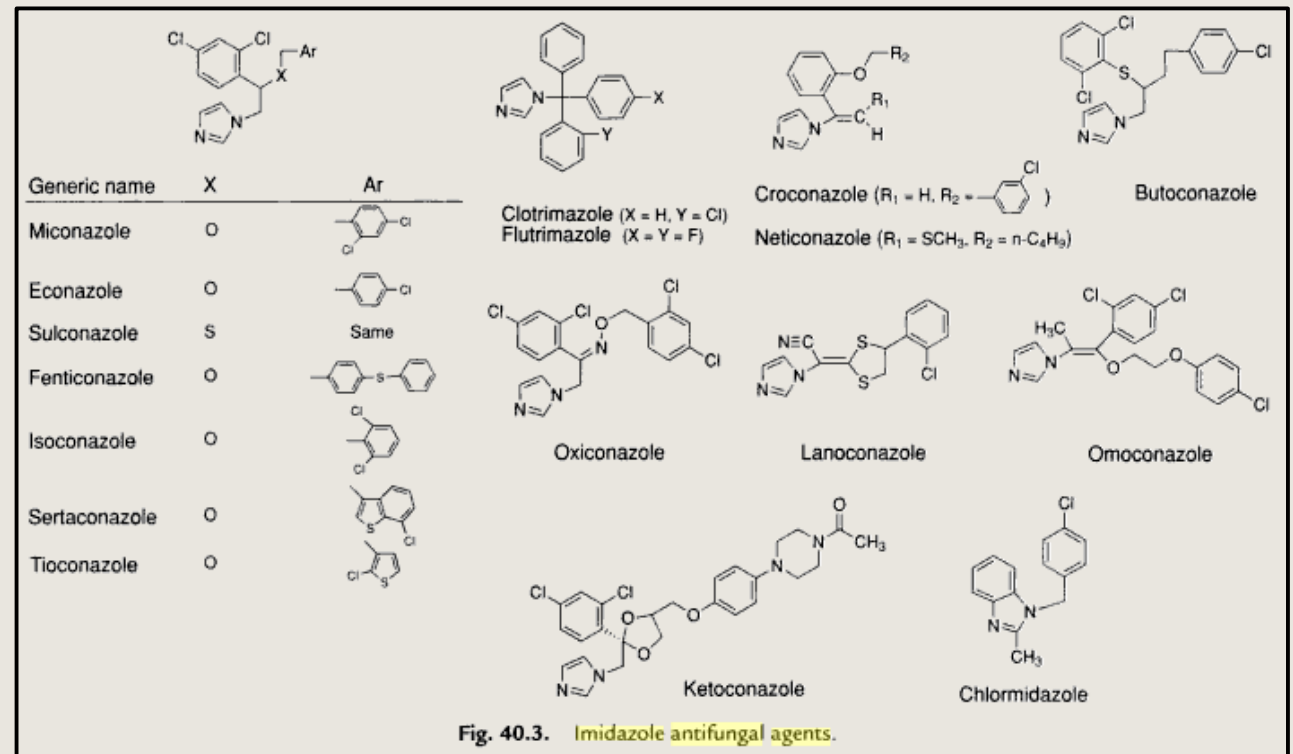
Grupos reactivos – carbonilos, aldehidos, cetonas,

Aplicaciones

Agrupar compuestos químicos en familias

Desarrollo de nuevas Drogas

Inferencia



PROBLEMA MÁS DÍFICIL: ENCONTRAR MOLÉCULAS CON GRUPOS SIMILARES

Computacionalmente: *subgraph isomorphism problem*

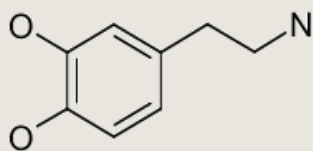
Encontrar un grafo determinado (fijo) dentro de otro grafo

Encontrar el **máximo subgrafo compartido** entre dos grafos

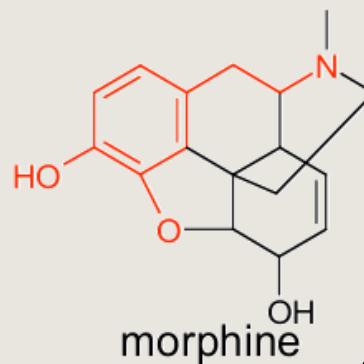
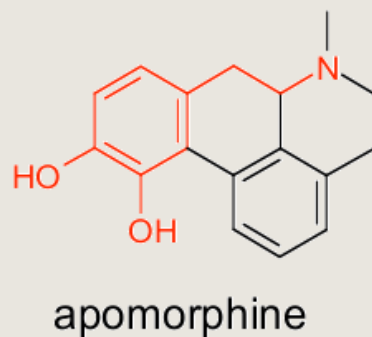
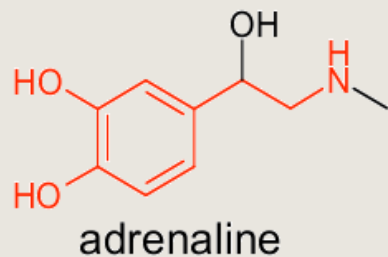
Es un problema computacionalmente difícil!

El tiempo se incrementa exponencialmente con el tamaño del problema (en este caso el número de nodos del grafo)

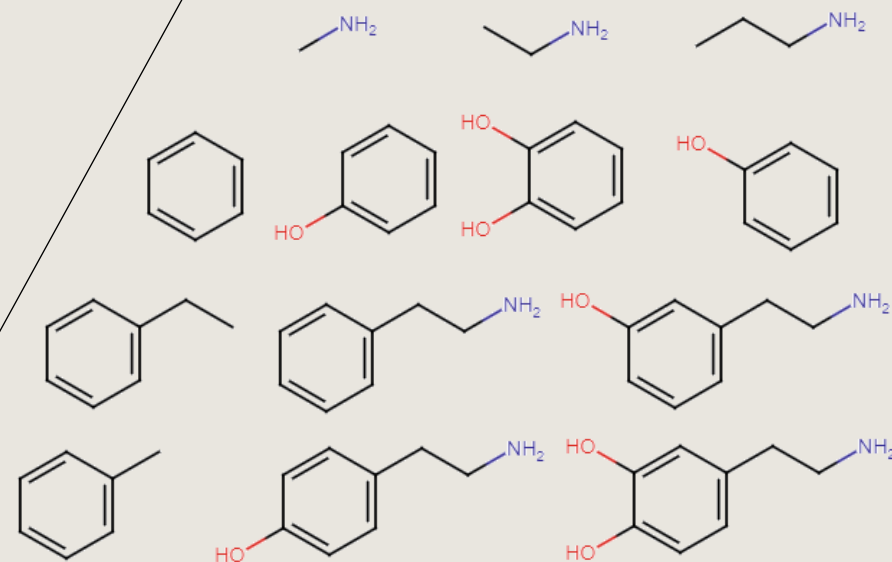
Query:



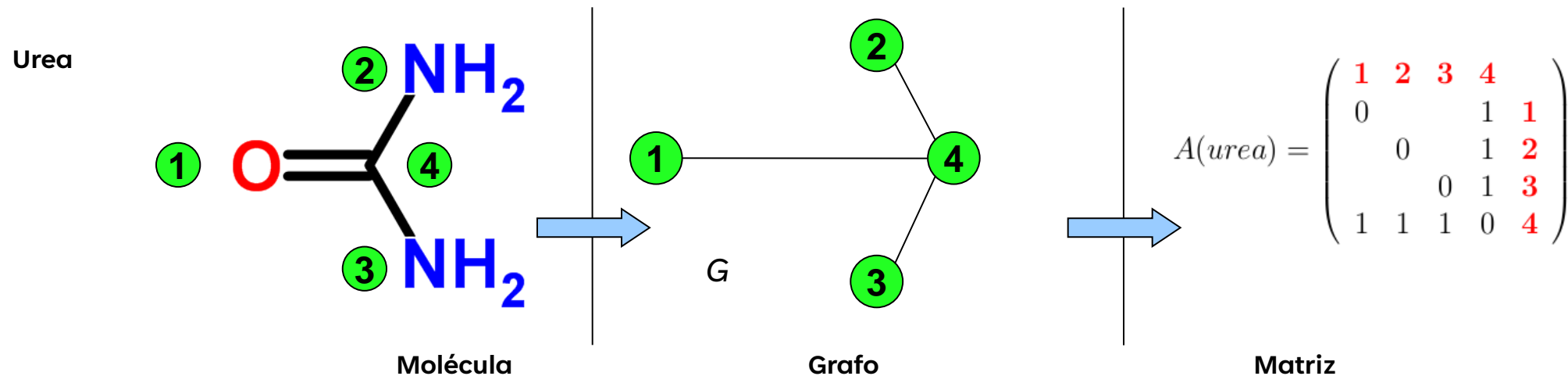
Hits:



Subgrafos compartidos



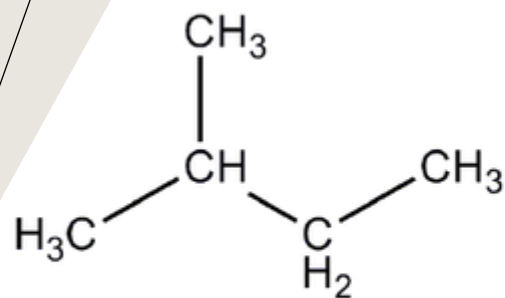
BÚSQUEDA DE SUBESTRUCTURAS: MATRICES DE ADYACENCIA



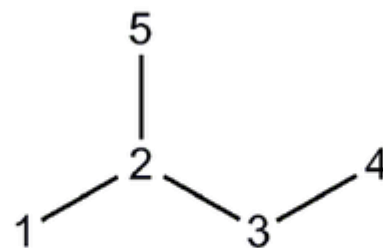
Dado un grafo, es posible construir una **matriz de adyacencia**

Es una aproximación (heurística) a la búsqueda de subestructuras: localizar coincidencias en una matriz de adyacencias

ADJACENCY MATRICES



Molecule



Graph

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

adjacency matrix

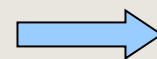
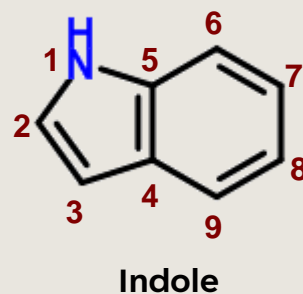
The chemical graph and adjacency matrix of the isopentane.

BÚSQUEDA DE SUBESTRUCTURAS: MATRICES DE ADYACENCIA

Indol: compuesto heterocíclico aromático, precursor de muchas drogas

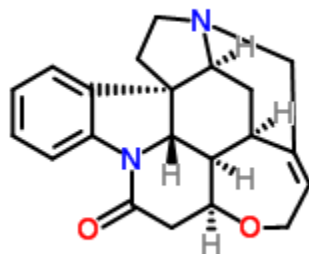
Búsqueda de compuestos que contengan el grupo **indol**

1. Calcular la matriz de adyacencia para la molécula 'query'
2. Calcular las matrices de adyacencia para todas las moléculas a testear (la base de datos)
3. Buscar coincidencias en las matrices de adyacencia

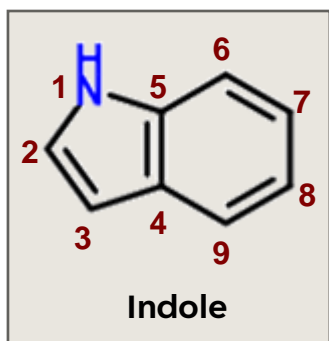


$$A(\text{indole}) = \begin{pmatrix} \textcolor{red}{1} & \textcolor{red}{2} & \textcolor{red}{3} & \textcolor{red}{4} & \textcolor{red}{5} & \textcolor{red}{6} & \textcolor{red}{7} & \textcolor{red}{8} & \textcolor{red}{9} \\ 0 & 1 & & & 1 & & & & \textcolor{red}{1} \\ 1 & 0 & 1 & & & & & & \textcolor{red}{2} \\ & 1 & 0 & 1 & & & & & \textcolor{red}{3} \\ & & 1 & 0 & 1 & & & 1 & \textcolor{red}{4} \\ 1 & & & 1 & 0 & 1 & & & \textcolor{red}{5} \\ & & & & 1 & 0 & 1 & & \textcolor{red}{6} \\ & & & & & 1 & 0 & 1 & \textcolor{red}{7} \\ & & & & & & 1 & 0 & 1 & \textcolor{red}{8} \\ & & & 1 & & & & 1 & 0 & \textcolor{red}{9} \end{pmatrix}$$

BÚSQUEDA DE SUBESTRUCTURAS: MATRICES DE ADYACENCIA



Strychnine
Database Molecule



Query Molecule

$$A(\text{strychnine}) =$$

	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
01	0	1														1									
02	1	0	1	1																					
03		1	0																						
04		1		0																					
05					0	1		1																	
06					1	0	1																		
07						1	0	1																	
08							1	0	1																
09								1	0	1															
10									1	0	1														
11										1	0	1													
12											1	0	1												
13												1	0	1											
14													1	0	1										
15														1	0	1									
16															1	0	1								
17																1	0	1							
18																	1	0	1						
19																		1	0	1					
20																			1	0	1				
21																				1	0	1			
22																					1	0	1		
23																						1	0	1	
24																							1	0	1
25																								1	0

$$A(\text{indole}) =$$

	1	2	3	4	5	6	7	8	9
1	0	1				1			
2	1	0	1						
3		1	0	1					
4			1	0	1				1
5	1			1	0	1			
6					1	0	1		
7						1	0	1	
8							1	0	1
9				1				1	0

BÚSQUEDA DE SUBESTRUCTURAS: MATRICES DE ADYACENCIA

Problema de esta estrategia (hasta acá):

Puede dar falsos positivos

Grafos que tienen el mismo número de nodos, con la misma adyacencia, pero cuyos nodos están compuestos por distintos átomos (en el caso de moléculas)

Posible solución:

Screening – realizar la búsqueda sólo sobre un subconjunto de moléculas (grafos) compatibles

Ej: (query = indol) filtrar la base de datos: seleccionar solamente moléculas que tengan al menos 1 átomo de nitrógeno

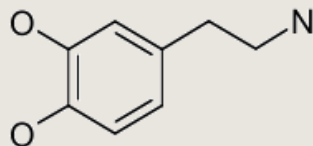
BÚSQUEDA DE SUBESTRUCTURAS

Screenings

Simple:

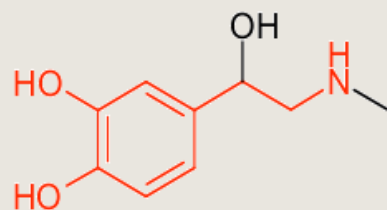
- Usa la fórmula molecular
 - La fórmula de todos los compuestos está almacenada en la base de datos
 - La fórmula de la molécula *query* se calcula al inicio de la búsqueda
 - Se descartan moléculas a las que les faltan átomos requeridos

Query:

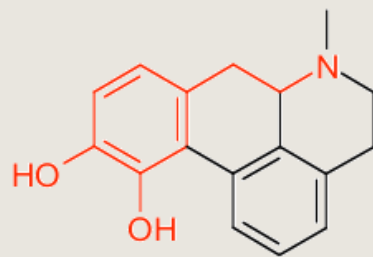


MF: C8 O2 N (H implícito)

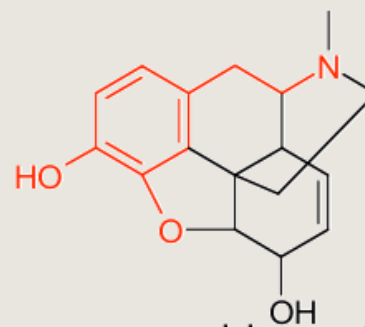
Hits:



adrenaline



apomorphine

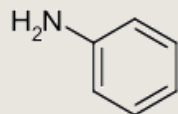
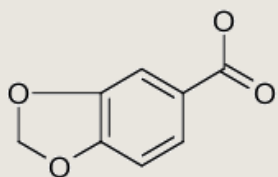
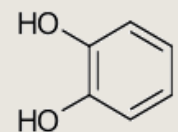


morphine

BÚSQUEDA DE SUBESTRUCTURAS: FINGERPRINTS

Fingerprint: representación abstracta de características o propiedades de una molécula (features)

- Presencia/ausencia de cada elemento
- Configuraciones electrónicas inusuales (carbono sp³, nitrógeno unido con un triple enlace)
- Anillos y sistemas de anillos (naftaleno, piridina, cyclohexano)
- Grupos funcionales (alcoholes, aminas, carboxilos, etc.)
- Se suelen utilizar tanto para búsquedas de subestructuras como para detectar similitud



1	0	0	0	1	1	0
---	---	---	---	---	---	---

Query

1	0	1	1	1	1	0
---	---	---	---	---	---	---



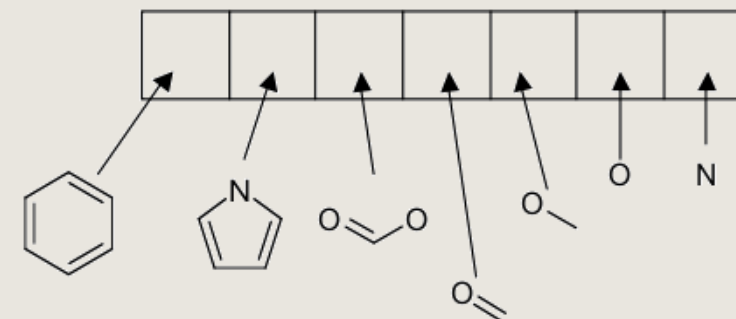
passes

1	0	0	0	0	0	1
---	---	---	---	---	---	---

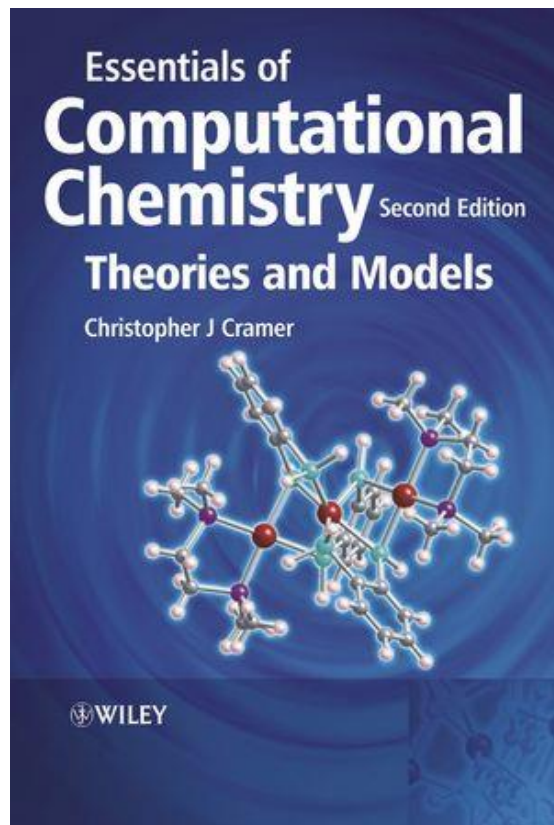


does not pass

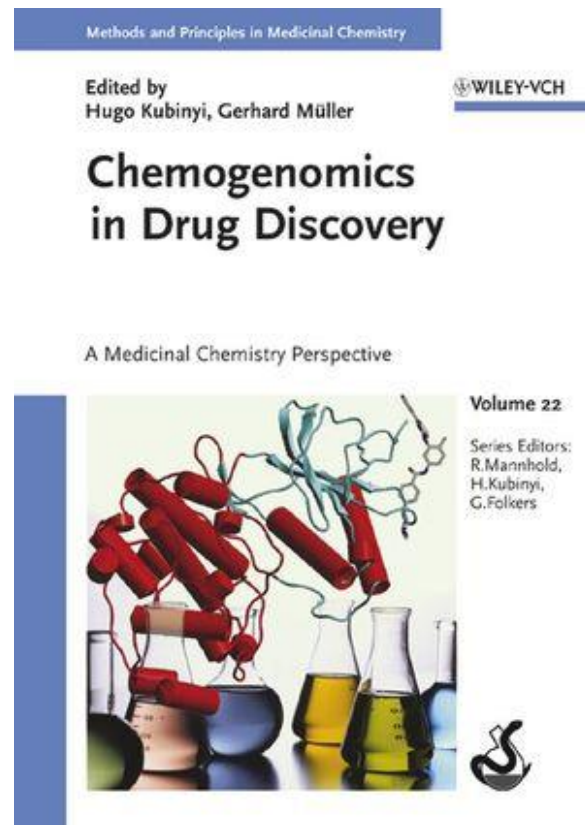
Un fingerprint



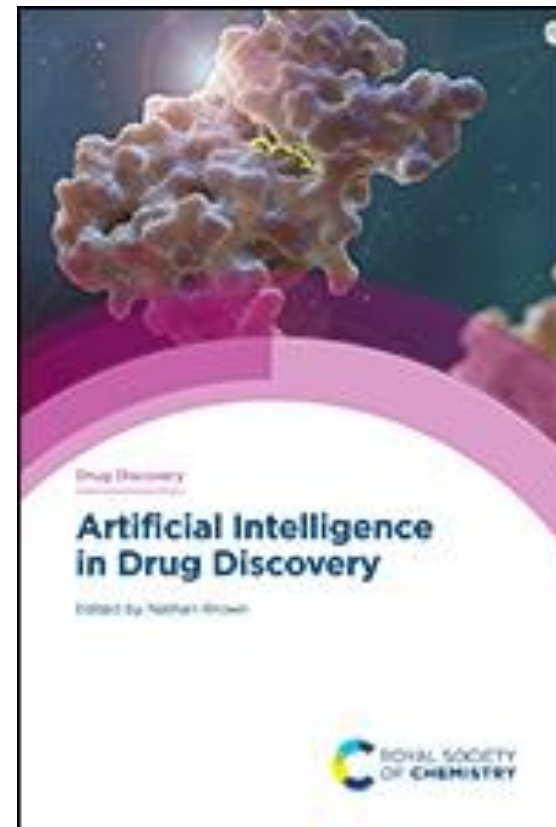
BIBLIOGRAFÍA | MATERIAL DE LECTURA



Essentials of Computational Chemistry (2004), 2nd Ed, CJ Cramer. Wiley.
<https://www.wiley.com/en-us/Essentials+of+Computational+Chemistry:+Theories+and+Models,+2nd+Edition-p-9780470091821>



Chemogenomics in Drug Discovery: A Medicinal Chemistry Perspective (2006). Edited by Hugo Kubinyi & Gerhard Müller, Wiley-VCH.
<https://www.wiley.com/en-us/Chemogenomics+in+Drug+Discovery%3A+A+Medicinal+Chemistry+Perspective-p-9783527604029>



Artificial Intelligence in Drug Discovery (2020). Edited by Nathan Brown. Royal Society of Chemistry.
<https://doi.org/10.1039/9781788016841>



Open-Source Cheminformatics and Machine Learning

The RDKit Book (2023).
https://www.rdkit.org/docs/RDKit_Book.html

PREGUNTAS?

Fernán Agüero

fernan@iib.unsam.edu.ar

