

Introducción a la Bioinformática

Cheminformatics

Fernán Agüero
Instituto de Investigaciones Biotecnológicas, UNSAM

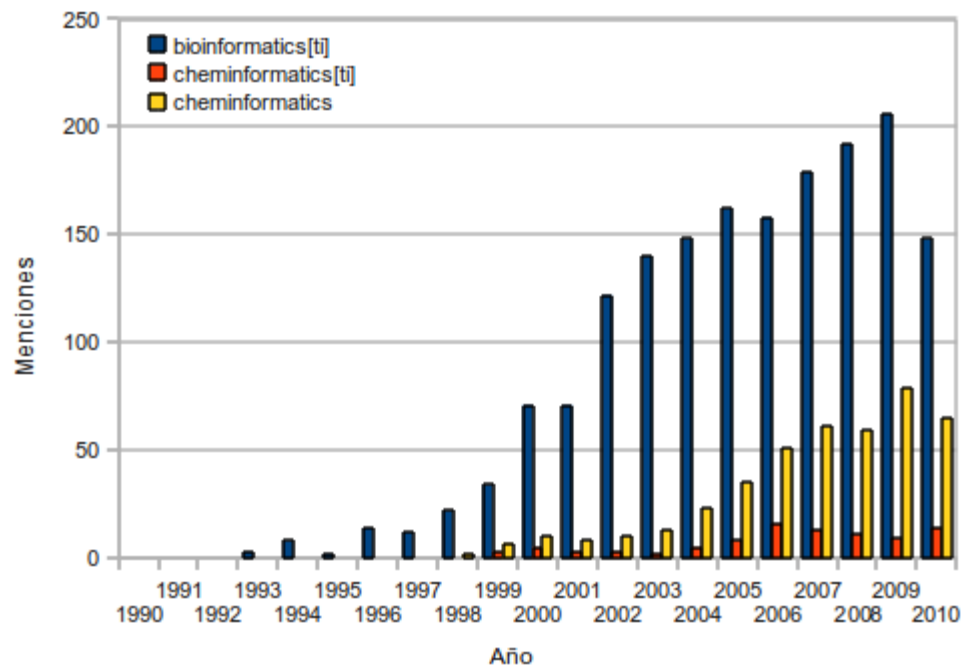
Cheminformatics – Qué es?

- Se la conoce como:
 - Computational chemistry
 - Theoretical chemistry
 - Molecular modeling
- Nace con el desarrollo de la mecánica cuántica a principios del siglo XX
 - Pero parece haber pasado desapercibida en la revolución “ómica”
- En activo desarrollo y expansión a partir de la introducción de las computadoras

“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)

Menciones en PubMed
Bioinformatics vs Cheminformatics



Term	Google	Google Scholar	Web of Knowledge	Scopus
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*

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

Google:

- **“Bioinformatics” (2010): ~ 50 millones de páginas**
- **“Cheminformatics” (2010): ~ 1 millón de páginas**

Cuestiones químicas

La química se ocupa de

- **Compuestos**
 - **Propiedades**
 - Físicas
 - Energía
 - Químicas
 - Estructura
 - Reactividad
 - Biológicas
 - Actividad
 - **Separaciones de mezclas de compuestos**
 - **Aspectos estáticos**
- **Transformaciones**
 - **Reacciones químicas**
 - **Aspectos dinámicos**

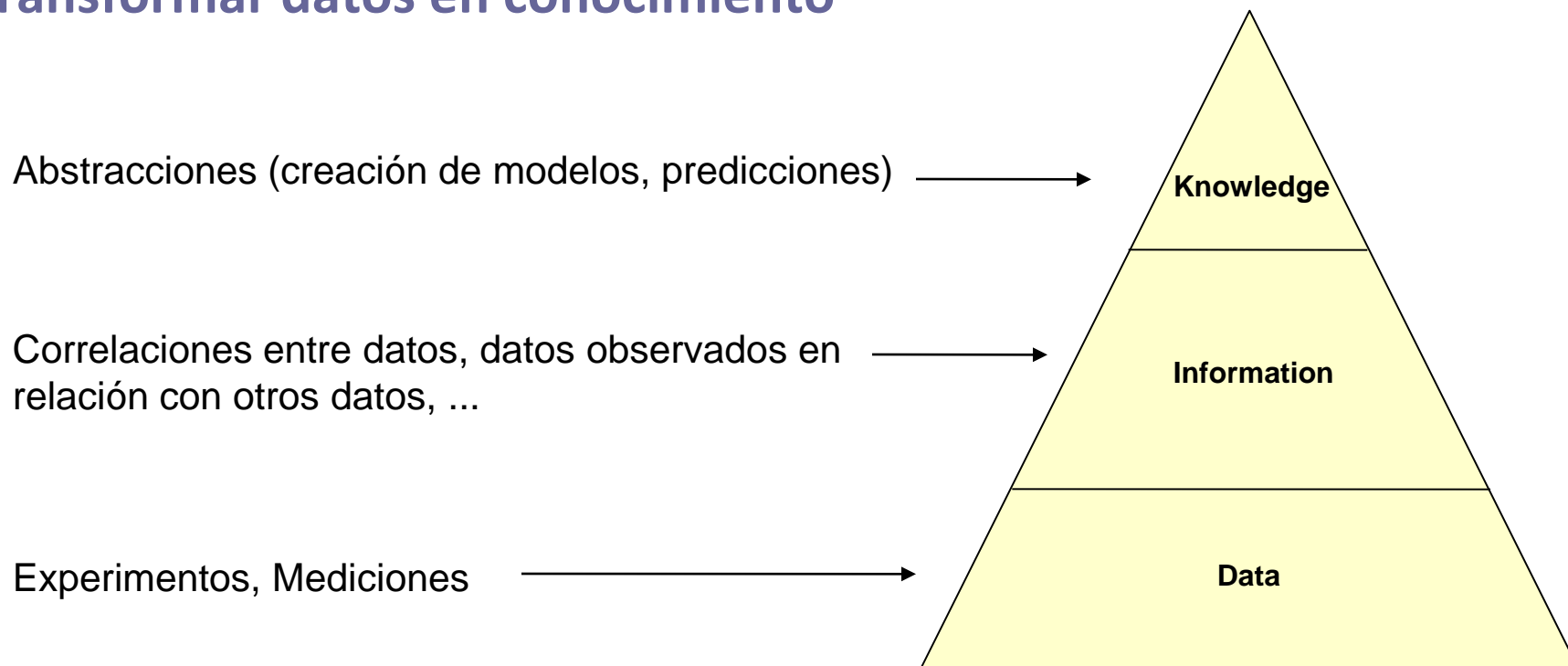
estructura  propiedades

Desafíos fundamentales de la química

- **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 (costo/beneficio) para la síntesis de estos compuestos**
 - **Reacciones**
 - **Materiales iniciales**
- **Dilucidar estructuras**
 - **A partir de datos experimentales (ej NMR)**
 - **Compuestos desconocidos**

propiedades  estructura

Transformar datos en conocimiento



Predecir (ejemplos):

- 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 determinado

“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?”

Molecular shape and medicinal chemistry: a perspective. 2010.

A Nicholls *et al.* J Med Chem 53: 3862

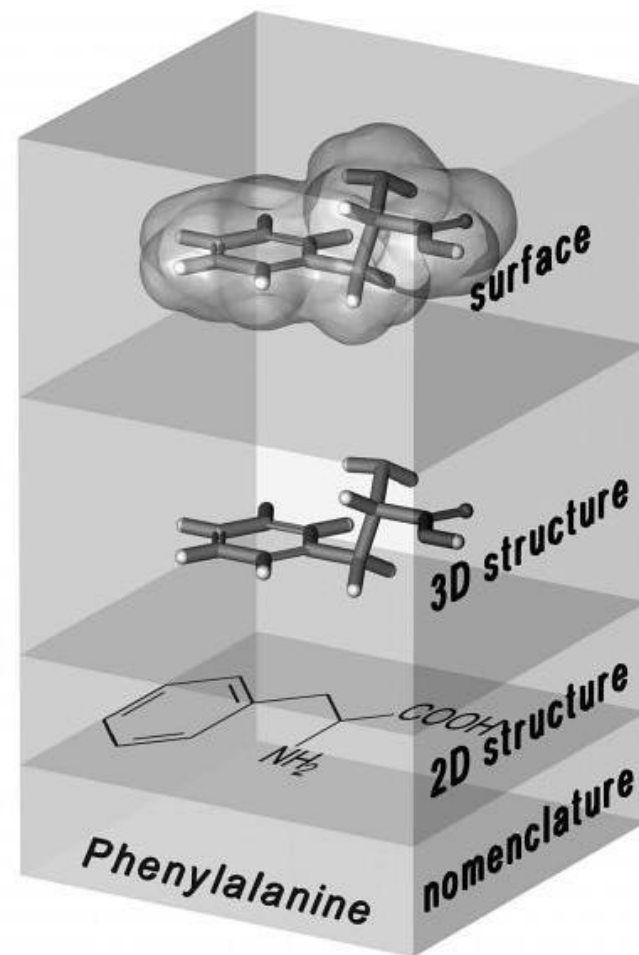
Teoría vs Modelos

- En esencia son lo mismo, pero
 - Una teoría suele ser general
 - Mientras que los modelos introducen particularidades para facilitar la interpretación y el entendimiento
- Mecánica cuántica
 - Teoría fundamental de la química
 - Permite describir un sistema (por ej una molécula) en forma completa, usando funciones de onda, formación y ruptura de enlaces, reacciones químicas, etc.
- Modelo de valencia, capas de electrones y repulsión
 - Todos estudiamos esto en cursos básicos de química
 - Es un modelo o aproximación
 - Permite entender los mismos sistemas fácilmente
 - Pero tiene problemas para describir comportamientos de algunos sistemas químicos

Representación de compuestos químicos

El formato más conocido y difundido:

- **Representación 2D de moléculas**
 - **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**
- **Representación 3D de moléculas**
 - **Para esto se requieren datos adicionales**
 - Posición de los átomos en el espacio
 - Ángulos y distancias de los enlaces
- **Representaciones más complejas**
 - **Mapear propiedades (ej potencial electrostático) sobre la superficie de una molécula**



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

Tomado de J Gasteiger & T Engel (2003).

Representación de compuestos químicos: nomenclatura

- **Trivial**
 - **Fenilalanina**
 - Popular en química, pero difícil de sistematizar
- **IUPAC**
 - **2-amino-3-phenylpropanoic acid**
 - Sistemático, pero los nombres pueden ser largos!
- **Fórmula empírica**
 - **$\text{C}_9 \text{H}_{11} \text{N O}_2$**
 - Ambiguo: varios compuestos pueden tener la misma fórmula

Representación de compuestos químicos: SMILES

SMILES (Simplified Molecular Intput Line Entury System)

- Introducido en 1986
- Representación lexicográfica de una molécula
- Usa conceptos de *grafos*
- *Nodos* conectados a través de *aristas* o *arcos*
- Reglas:
 - Los átomos se representan con sus respectivos símbolos
 - Los hidrógenos son implícitos
 - Los átomos vecinos aparecen juntos
 - Se usan paréntesis cuando hay más de un vecino
 - Enlaces dobles se representan usando '='
 - Enlaces triples se representan usando '#'
 - Quiralidad tetrahédrica se especifica usando '@' (contrario a las agujas del reloj) o '@@' (en el sentido de las agujas del reloj)
 - Un anillo se representa con números a continuación de los átomos que cierran el anillo
- Más información y reglas en:
 - http://www.daylight.com/dayhtml_tutorials/languages/smiles/index.html

Otros ejemplos:

Benceno: C1CCCCC1

Etanol: CCO

Piridina: C1CNCCC1

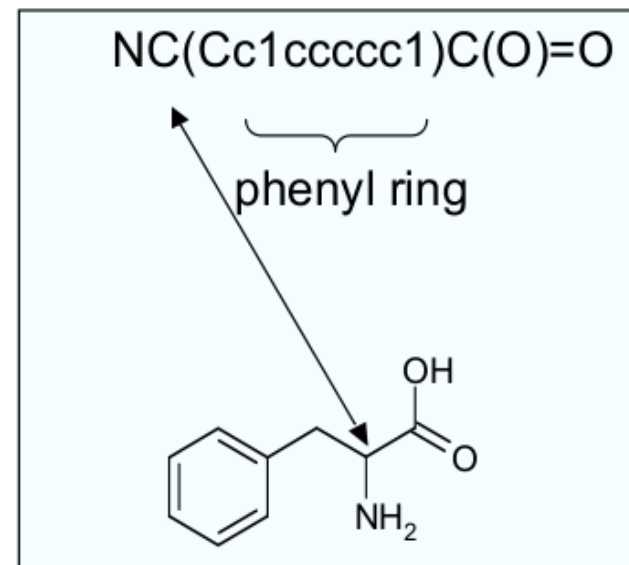
Acido acético: CC(=O)O

Acido cianhídrico: C#N

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

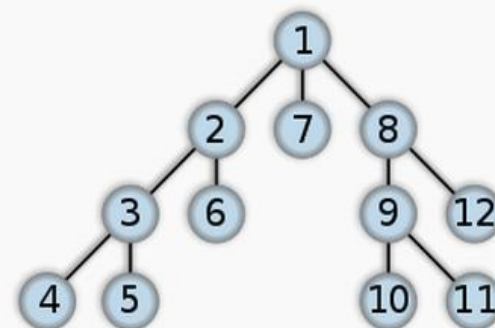
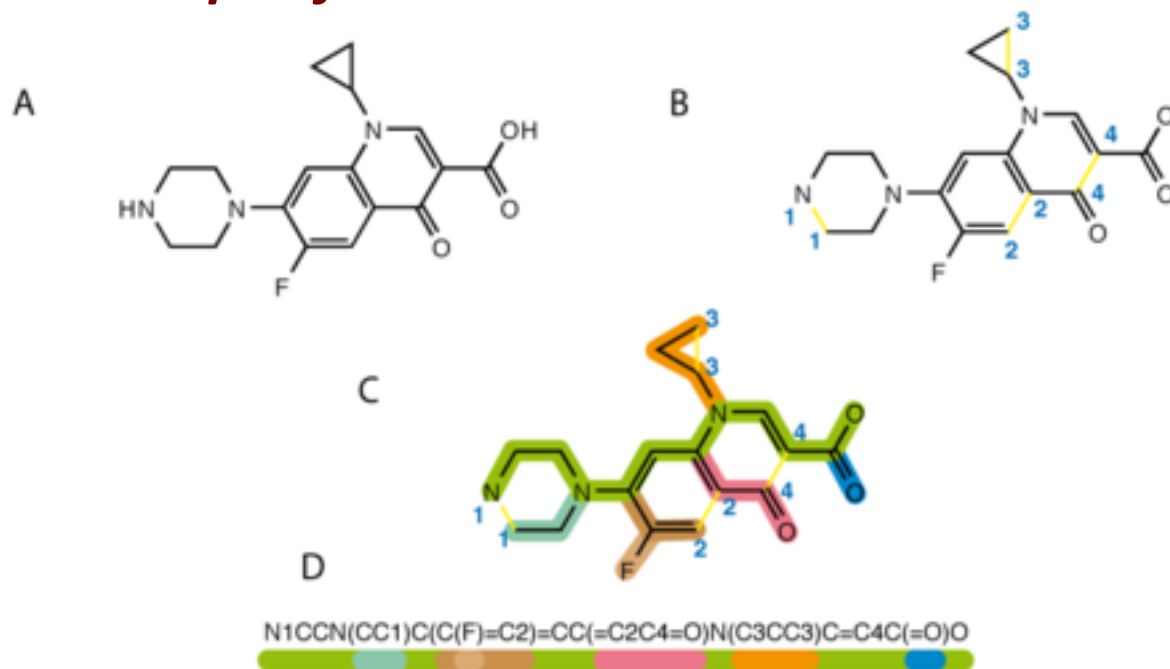
L-alanina (sin especificar quiralidad): N[CH](C)C(=O)O

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



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: Representar la conectividad de una molécula de manera uniforme

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

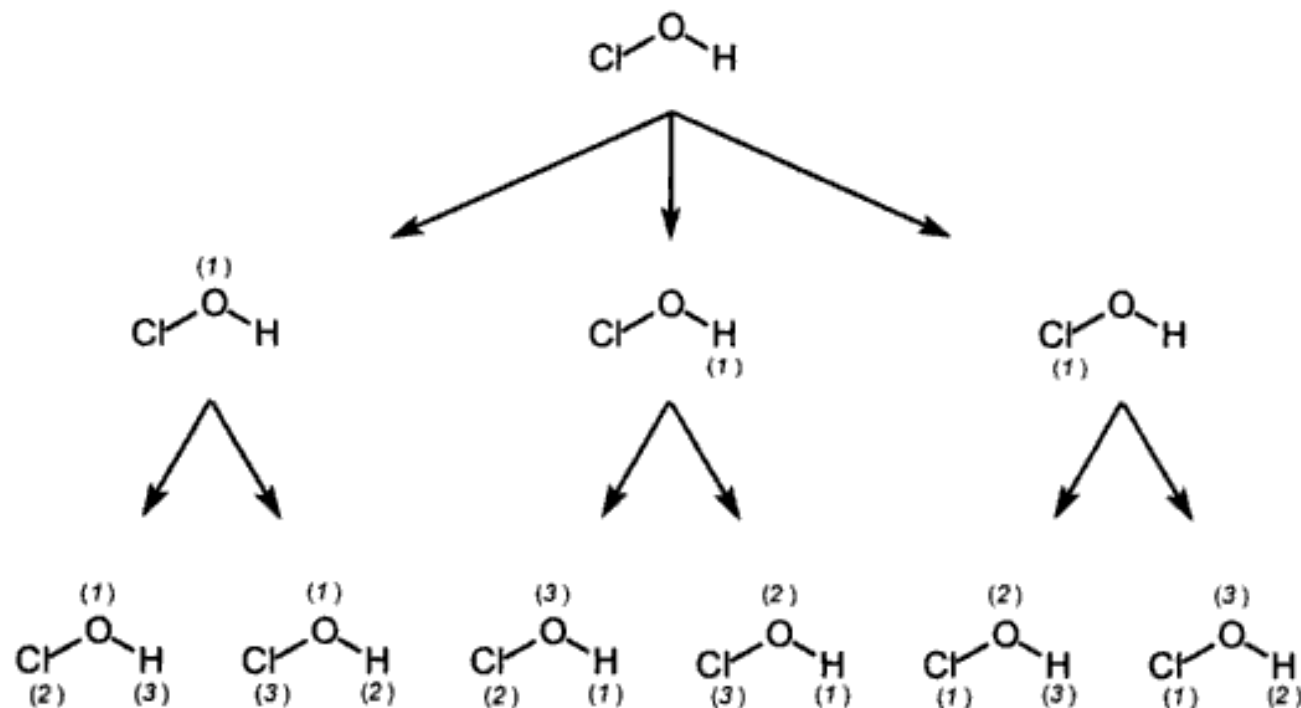


Figure 2-41. Six different possibilities for numbering the atoms in a hypochlorous acid molecule.

Canonización de moléculas: el algoritmo de Morgan

El algoritmo:

- **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

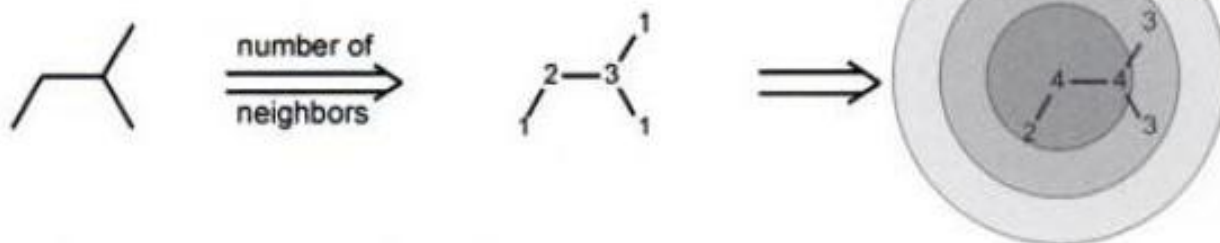


Figure 2-43. The EC value or the atom classification of each atom, respectively, is calculated by summing the EC values of the directly connected neighboring atoms of the former sphere (relaxation process).

Canonización de moléculas: el algoritmo de Morgan

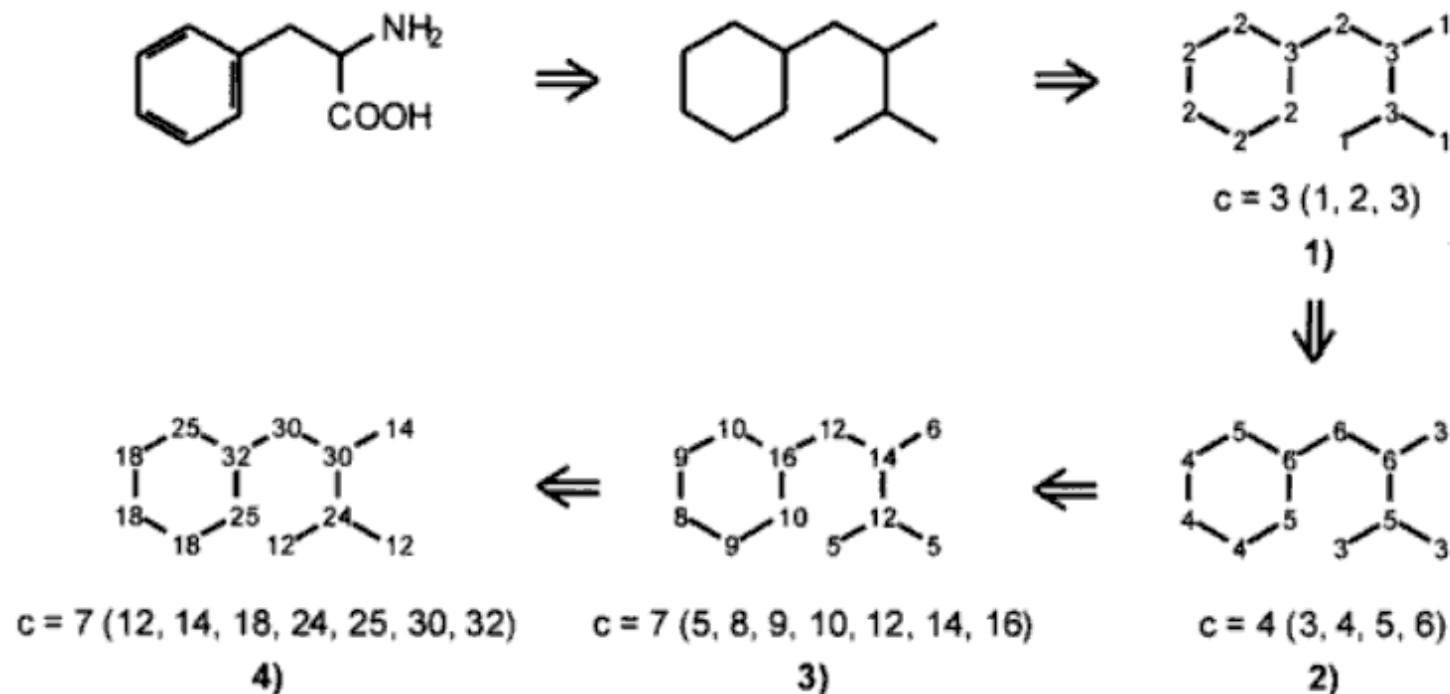


Figure 2-44. The EC values of the atoms of phenylalanine (without hydrogens) are calculated by considering the class values of the neighboring atoms. After each relaxation process, c , the number of equivalent classes (different EC values), is determined.

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.

Paso 3: Asignación de números de átomos únicos

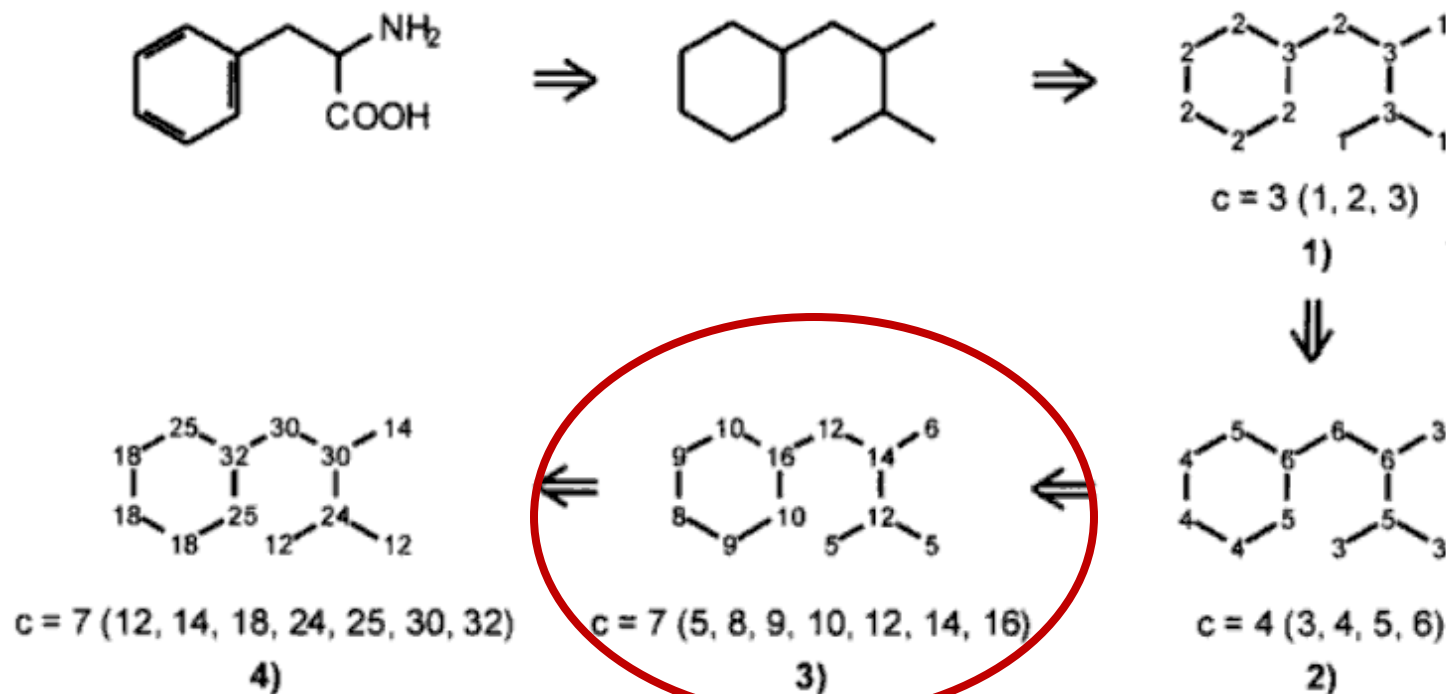


Figure 2-44. The EC values of the atoms of phenylalanine (without hydrogens) are calculated by considering the class values of the neighboring atoms. After each relaxation process, c , the number of equivalent classes (different EC values), is determined.

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 algoritmo data de 1965!
- 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

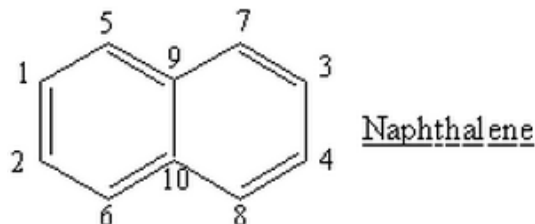
InChI – IUPAC International Chemical Identifier

- Introducido recientemente (2005)
- 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

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.

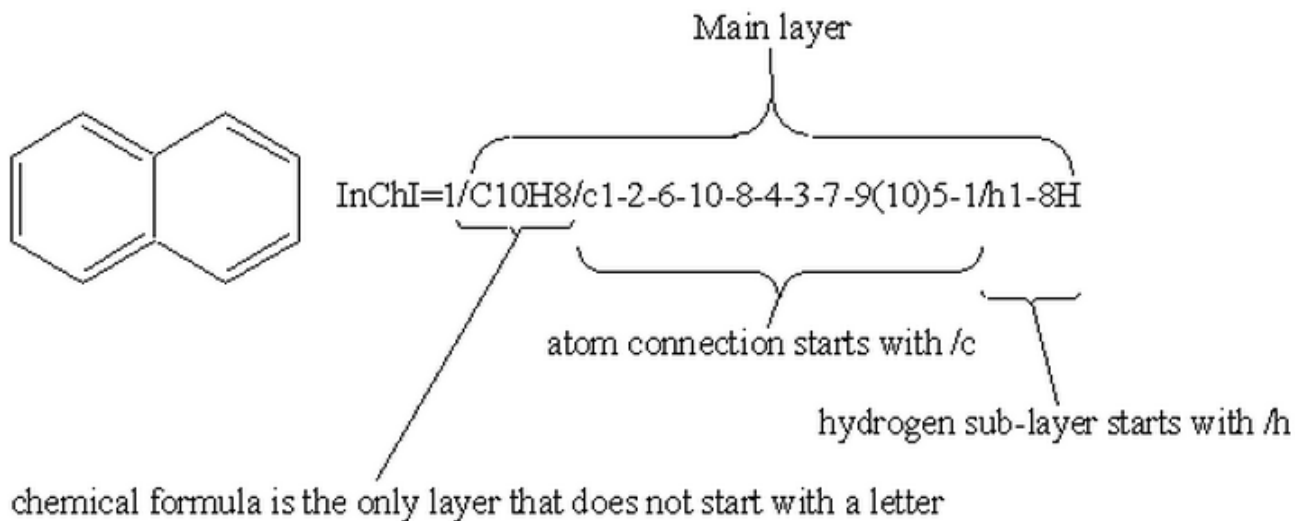


Ejemplos:

Agua: InChI = 1/H2O/h1H2

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

Naftaleno: InChI = 1/C10H8/c1-2-6-10-8-4-3-7-9(10)5-1/h1-8H



Tomado de:

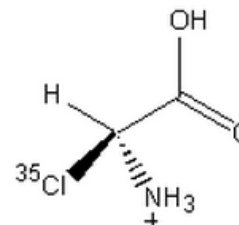
Unofficial InChI FAQ,

<http://wwwmm.ch.cam.ac.uk/inchifaq/>

Representación de compuestos químicos: InChI

Ventajas de representar la información en *capas*

- Permiten elegir el nivel de detalle que uno quiere representar
 - Sólo la capa principal es mandatoria
 - El resto de las capas agregan detalles en forma incremental
- Hay 6 capas posibles:
 - **Main layer – Mandatoria**
 - Tiene subcapas: fórmula, conectividad (c), hidrógenos (h)
 - **Charge layer (q/p) – opcional**
 - **Stereochemical layer (b/t/s/m) – opcional**
 - **Isotopic layer (i) – opcional**
 - **Fixed-H layer (f) – opcional**
 - **Reconnected Layer –**
 - permite especificar union a metales



InChI=1/C2H4ClNO2/c3-1(4)2(5)6/h1H,4H2,(H,5,6)/p+1/t1-/m1/s1/i3+0/fC2H5ClNO2/h4-5H/q+1

Main layer

Charge layer

Stereochemical layer

Isotopic layer

Fixed-H layer

Tomado de:

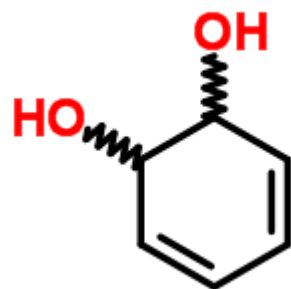
Unofficial InChI FAQ,

<http://wwwmm.ch.cam.ac.uk/inchifaq/>

Representación de compuestos químicos: InChI

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

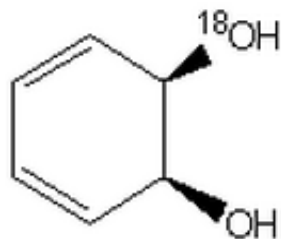
Pero: Los compuestos pueden estar representados con diferente nivel de detalle



3,5-Cyclohexadien-1,2-diol

InChI=1/C6H8O2/c7-5-3-1-2-4-6(5)8/h1-8H

Identical Main Layers



InChI=1/C6H8O2/c7-5-3-1-2-4-6(5)8/h1-8H/t5-,6+/i7+2/m1/s1

Extra Stereo Layer

Extra Isotopic Layer

Tomado de:

Unofficial InChI FAQ,

<http://wwmm.ch.cam.ac.uk/inchifaq/>

Representación de compuestos químicos: InChI

Uso de InChI

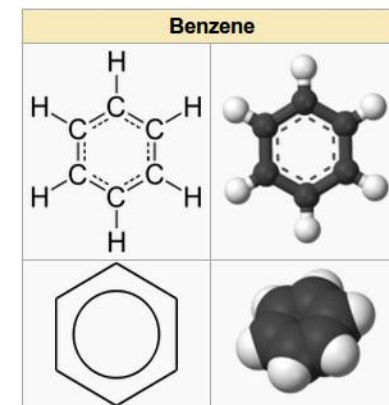
- **InChI se está incorporando en una amplia variedad de bases de datos públicas y comerciales:**
- US National Institute of Standards and Technology (NIST) - 150,000 structures
- [NIH/NCBI/PubChem](#) project - >3.2 million structures
- [EBI/EMBL/ChEMBL](#) project - > 600,000 structures (bioactive compounds)
- Thomson ISI - 2+ million structures
- US National Cancer Institute(NCI) Database - 23+ million structures
- US Environmental Protection Agency(EPA)-DSSToX Database - 1450 structures
- Kyoto Encyclopaedia of Genes and Genomes (KEGG) database - 9584 structures
- University of California at San Francisco ZINC - >3.3 million structures
- BRENDA enzyme information system (University of Cologne) - 36,000 structures
- EBI/Chemical Entities of Biological Interest (ChEBI) - 5000 structures
- University of California Carcinogenic Potency Project - 1447 structures
- Compendium of Pesticide Common Names - 1437 structures
- [Royal Society of Chemistry / ChemSpider](#) – millions
- **TDR Targets! (UNSAM)** – ~ 400,000 structures (bioactive)

- Journals that have adopted InChI:
 - Nature Chemical Biology.
 - Beilstein Journal of Organic Chemistry.
- Software that have incorporated InChI generation:
 - ACD/Labs ACD/ChemSketch.
 - ChemAxon Marvin.
 - SciTegic Pipeline Pilot.
 - CACTVS Chemoinformatics Toolkit by Xemistry, GmbH.

Representación de compuestos químicos: molfiles, SDF

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

```
benzene
ACD/Labs0812062058

6 6 0 0 0 0 0 0 0 0 0 1 V2000
  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
  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
M  END
$$$$
```

1. Header → benzene
ACD/Labs0812062058

2. Comment →

3. General information → 6 atoms, 6 bonds, ..., V2000 standard

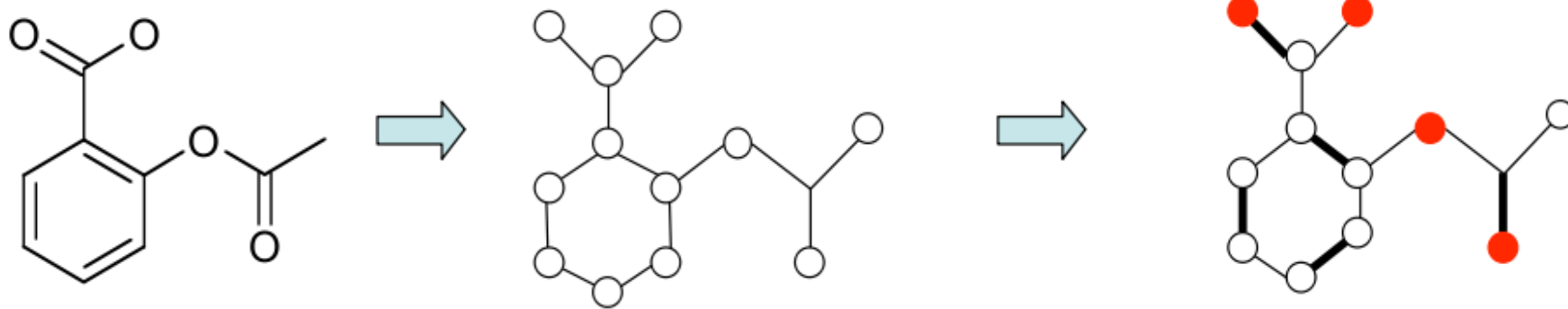
4. Spatial coordinates → X, Y, Z, element, extra information

5. Bonding information → 1st atom, 2nd atom, bond type, extra information

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

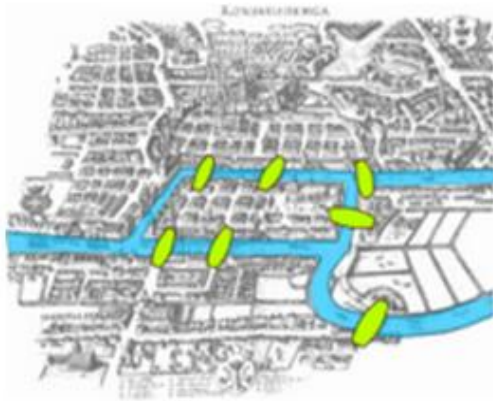


Un desvío: historia de los grafos

- 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:
 - Euler notó que 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
 - Abstracción del problema
 - En una lista de porciones de tierra (nodos)
 - Y una lista de puentes (aristas)
 - **Sólo la información de conectividad era relevante!**



Leonard Euler (1707-1783)



Tomado de Wikipedia
http://en.wikipedia.org/wiki/Seven_Bridges_of_K%C3%B6nigsberg

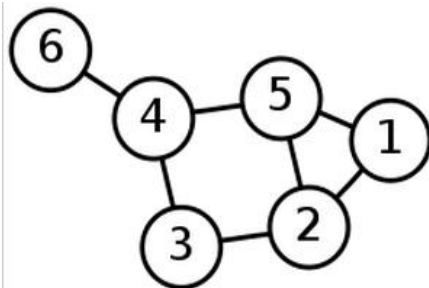
Un desvío: grafos: propiedades y operaciones

Propiedades de los grafos:

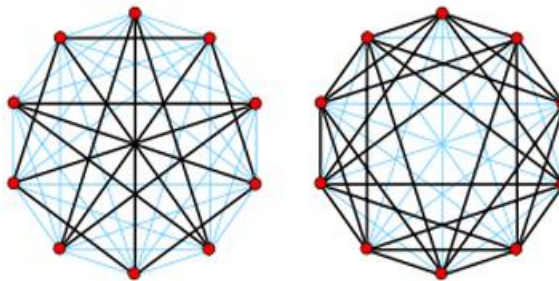
- Grado de conectividad de los nodos (degree)
- Direccionalidad (**asimetría**) o no-direccionalidad (**simetría**) de las aristas
- Intensidad (en el sentido vectorial) de cada arista
 - Las aristas pueden tener asociado un valor numérico (peso, largo, costo)
- Posibilidad de identificar los nodos
 - Son elementos de un conjunto
 - Grafos **etiquetados** (labeled) vs no-etiquetados (unlabeled)
- Muchas propiedades de los grafos son **heredables**
 - Un grafo tiene una propiedad X si todos sus subgrafos la tienen

Operaciones con grafos (algunos ejemplos):

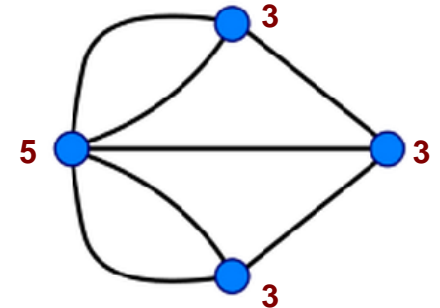
- Unarias
 - Complementación
- Binarias
 - Unión
 - Producto cartesiano



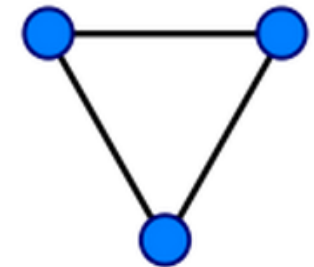
Un grafo etiquetado



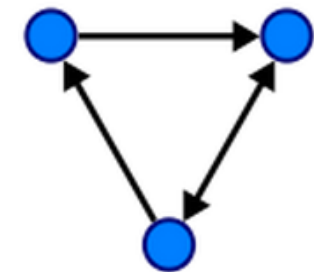
Un grafo y su complemento



Grados de los nodos



Grafo simple o regular



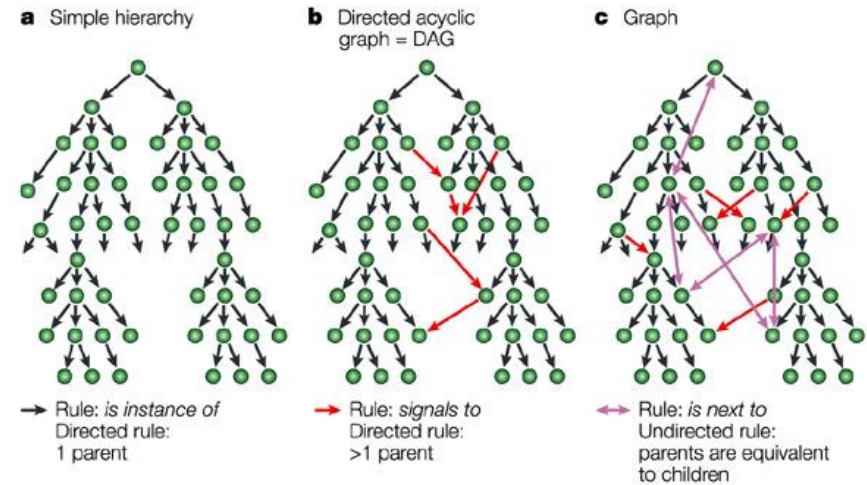
Grafo dirigido (red)

Entender cómo representar un problema

... usualmente es parte de la solución

Punto de vista computacional (aplicaciones)

- Es importante poder **reconocer** un problema y **clasificarlo**
 - Ej: es un problema de conectividad, tengo que representar el problema con grafos
- Es importante poder reconocer características salientes (**propiedades**) en el problema
 - Ej: es un grafo dirigido acíclico (directed acyclic graph, DAG)
 - Ej: Gene Ontology (GO) está estructurada como un DAG
 - Cada término tiene relaciones (aristas) con uno o más términos en el mismo grafo
- Esto nos puede llevar directamente a encontrar una solución
 - Ej: tengo que escribir un programa que recorra el grafo en forma eficiente
 - Buscar una solución desarrollada por alguien que ya estudió el problema
 - Un *algoritmo* que recorra DAGs
 - Busco en Google (o compro un libro) sobre 'Directed acyclic graph algorithms'

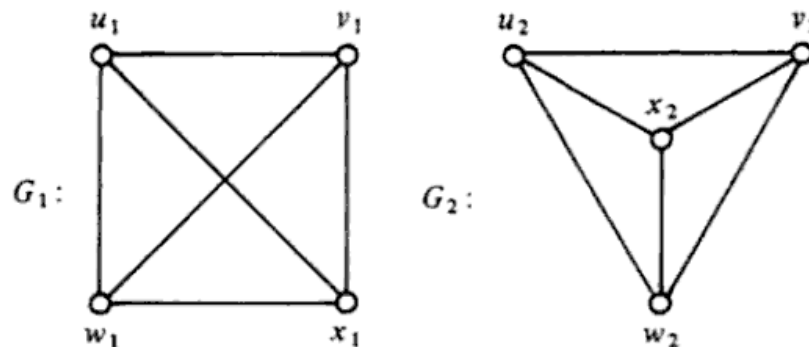


Nature Reviews | Genetics

Problema: encontrar moléculas iguales

Problema común en química

- Es importante saber si dos objetos en estudio son el mismo (en algún sentido) o diferentes.
- Si representamos moléculas como **grafos** las 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 si vemos como se definen **matemáticamente** estos grafos
 - **G1:** **nodos** = {u1, v1, w1, x1}; **aristas** = { {u1,v1}, {u1,w1}, {u1,x1}, {v1,x1}, {v1,w1}, {x1,w1} }
 - **G2:** **nodos** = {u2, v2, w2, x2}; **aristas** = { {u2,v2}, {u2,w2}, {u2,x2}, {v2,x2}, {v2,w2}, {x2,w2} }
- Computacionalmente tratables (usualmente)
 - Soluciones dependen del tipo de grafo
 - Árboles
 - Grafos planos
 - Grafos de intervalos
 - Grafos de permutaciones



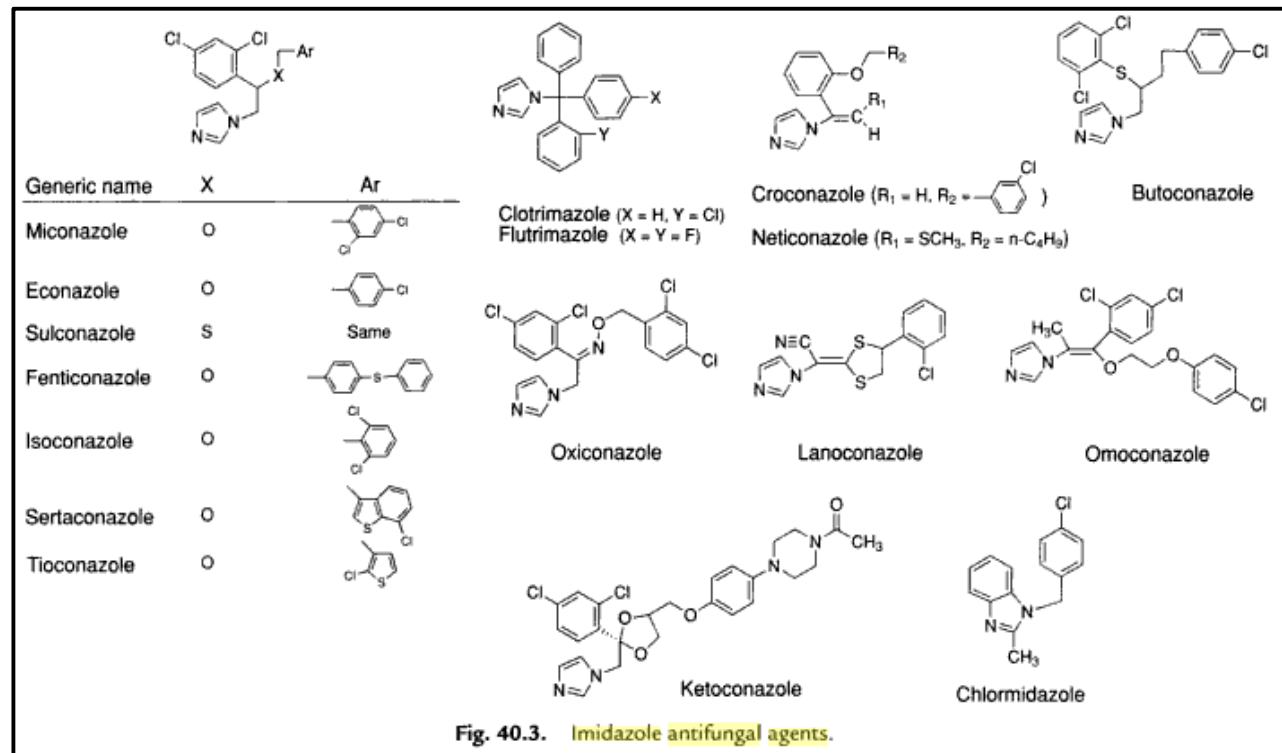
Two isomorphic graphs. Tomado de "Introductory Graph Theory".
G. Chartrand (1977). Dover Publications.

Problema más difícil: encontrar moléculas con grupos similares

Otro problema común en química

- Identificar compuestos que comparten grupos químicos similares
 - Farmacóforos – grupos químicos responsables de actividad farmacológica
 - Grupos reactivos
- Aplicaciones
 - Agrupar compuestos químicos en familias
 - Desarrollo de nuevas drogas
 - Inferencia

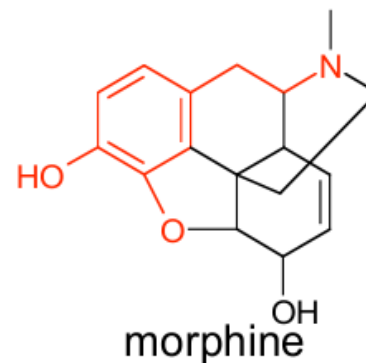
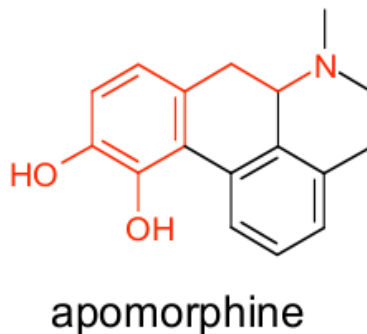
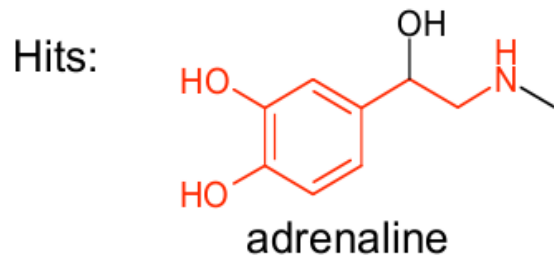
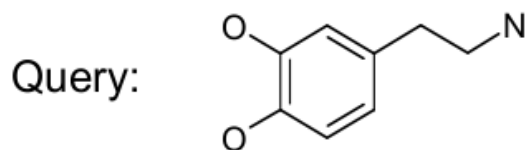
Foye's Principles of Medicinal Chemistry (2008).
T Lemke, DA Williams. Wolters Kluwer



Problema más difícil: encontrar moléculas con grupos similares

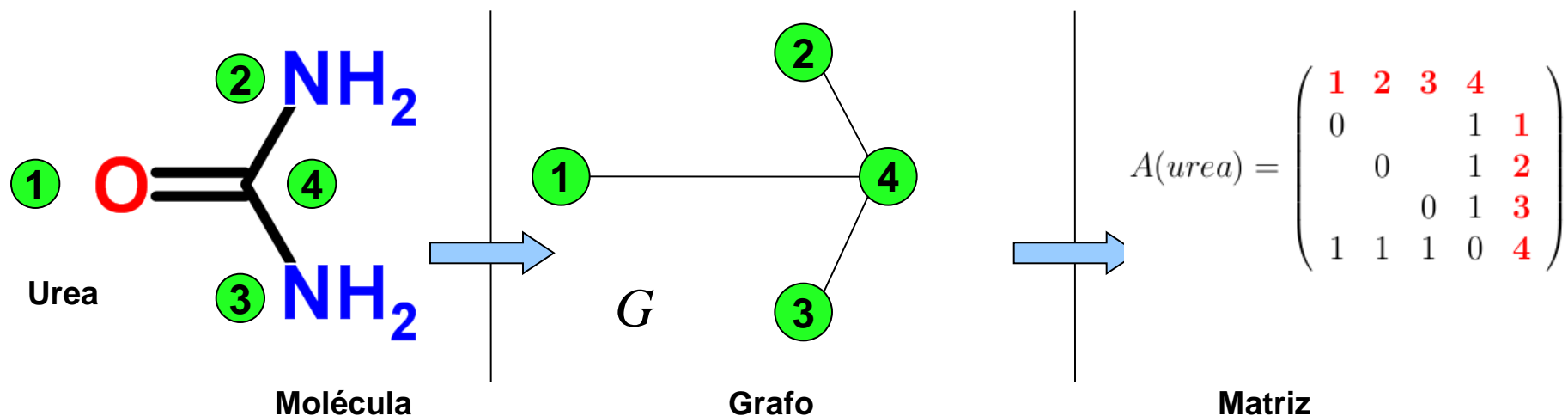
Computacionalmente: *subgraph isomorphism problem*

- Encontrar un grafo determinado (fijo) dentro de otro grafo
- Encontrar el subgrafo máximo compartido entre dos grafos
- Es un problema computacionalmente difícil
 - **NP-complete**, tiempo se incrementa exponencialmente con el tamaño del problema (en este caso el número de nodos del grafo)
- Existen *heurísticas*
 - Detección y descarte de grafos que no cumplen ciertas reglas
 - *Screening* para reducir el número de moléculas sobre las que se detecta el isomorfismo



Búsqueda de subestructuras: matrices de adyacencia

Dado un grafo, es posible construir una matriz de adyacencia

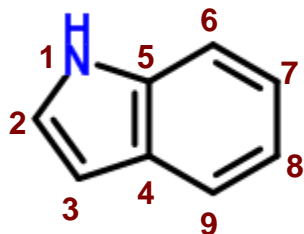


Una aproximación (heurística) a la búsqueda de subestructuras

- Localizar coincidencias en una matriz de adyacencias

Búsqueda de subestructuras: matrices de adyacencia

Indole



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

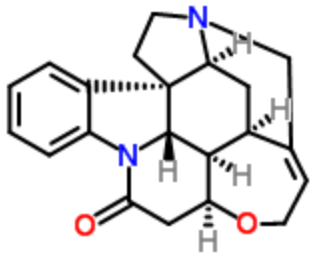
Indol: compuesto heterocíclico aromático

- 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

Búsqueda de subestructuras: matrices de adyacencia



strychnine

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

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

Se pueden también usar otras matrices

- Matrices de Distancia
 - Distancia entre un nodo y el resto de los nodos en el grafo

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:

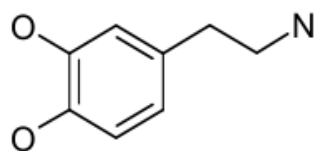
- **Screening** – realizar la búsqueda sólo sobre un subconjunto de moléculas (grafos) compatibles
- Ej: (query = indol) que tengan al menos 1 átomo de **nitrógeno**

Screenings

Simple:

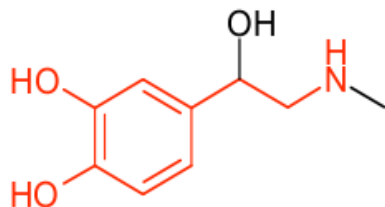
- Usa la fórmula molecular
 - La fórmula de todos los compuestos está almacenada en la base de datos
 - La fórmula de la molécula *query* se calcula al inicio de la búsqueda
 - Se descartan moléculas a las que les faltan átomos requeridos

Query:

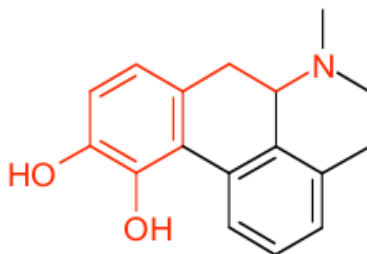


MF: C8 O2 N (H implícito)

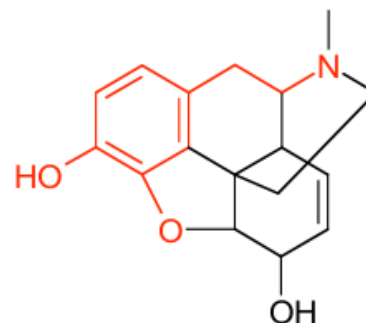
Hits:



adrenaline



apomorphine

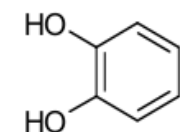
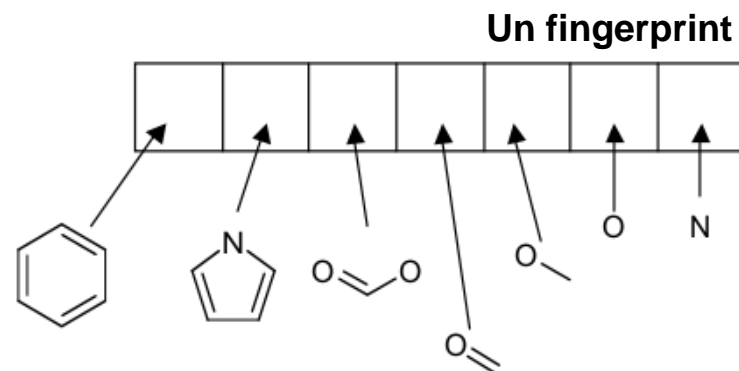


morphine

Búsqueda de subestructuras: fingerprints

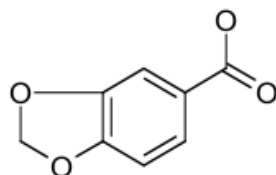
Screenings más complejos usan fingerprints

- Representación abstracta de características o propiedades de una molécula
- Qué características se suelen incluir en un fingerprint?
 - 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
---	---	---	---	---	---	---

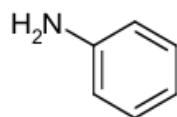
Query



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



passes



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



does not pass

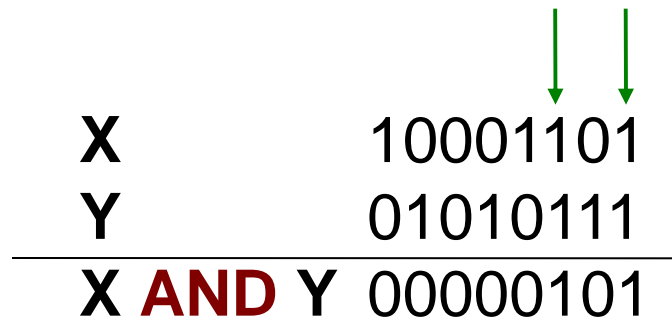
Cuestiones a tener en cuenta

- **Desventajas**

- **El fingerprint debe ser definido de antemano**
- **Distintas aplicaciones pueden generar distintos tipos de fingerprints**
 - OpenBabel (<http://openbabel.org>)
 - Daylight (diferentes aplicaciones, <http://www.daylight.com>)
 - Matchmol (N Haider (2010) Molecules 15: 5079-5092)
- **Algunas moléculas pueden no estar bien definidas por el fingerprint en la base de datos**
- **Gran parte de la base de datos puede ser irrelevante para un problema dado**
- **El tamaño del fingerprint (en bits) es importante**
- **Costo / Beneficio ; Sensibilidad / Especificidad**

Búsqueda de subestructuras/similitud: fingerprints

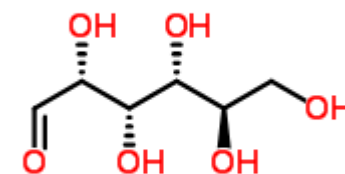
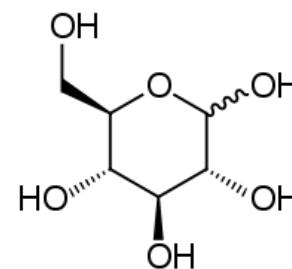
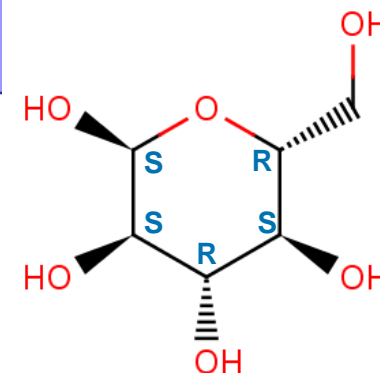
- Ventajas
- El screening es extremadamente rápido
 - Se evalúa equivalencia entre conjuntos de bits usando el operador AND binario

$$\begin{array}{r} \text{X} \quad \quad 10001101 \\ \text{Y} \quad \quad 01010111 \\ \hline \text{X AND Y} \quad 00000101 \end{array}$$


Representación de compuestos químicos

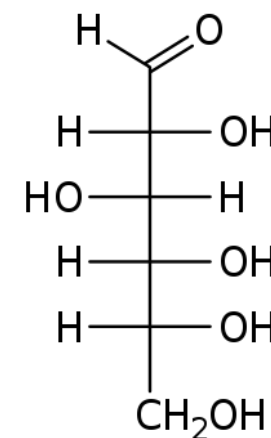
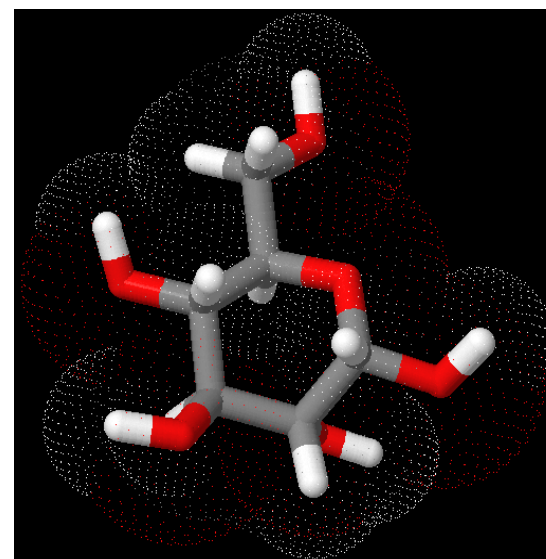
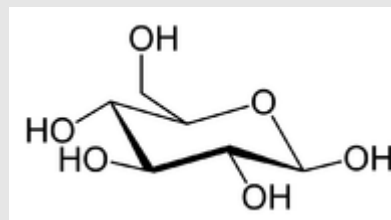
• Formatos

- **Fórmula:** $C_6H_{12}O_6$
- **Canonical Smiles:** C(C1C(C(C(C(O1)O)O)O)O)O
- **Smiles:** O=C[C@H](O)[C@@H](O)[C@H](O)[C@H](O)CO
- **InCHI:** 1/C6H12O6/c7-1-3(9)5(11)6(12)4(10)2-8/h1,3-6,8-12H,2H2/t3-,4+,5+,6+/m0/s1
- **SDF:**



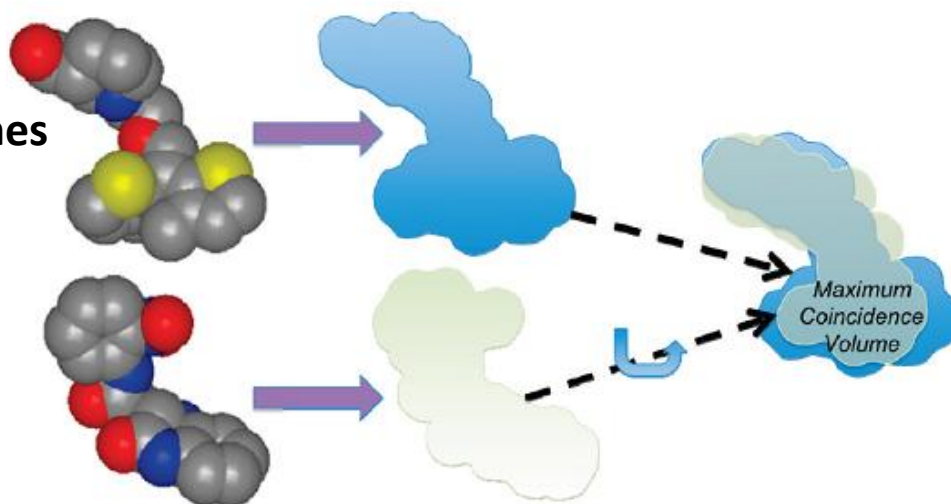
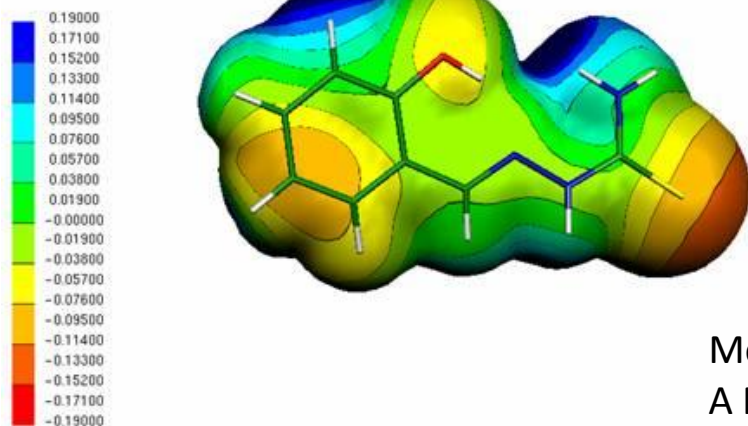
D-glucose
Chemexper 492-62-6

```
12 12 0 0 1 0 0 0 0 0 0999 v2000
3.4641 -1.4999 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
3.4641 -0.5000 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1.7320 -1.5000 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2.5980 -1.9999 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2.5980 0.0000 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1.7320 -0.5000 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0
4.3301 -1.9999 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0
4.3301 0.0000 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2.5980 1.0000 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2.5980 -2.9999 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0.8660 0.0000 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0.0000 -0.5000 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```



Una representación tridimensional de la molécula requiere no sólo especificar coordenadas espaciales de átomos

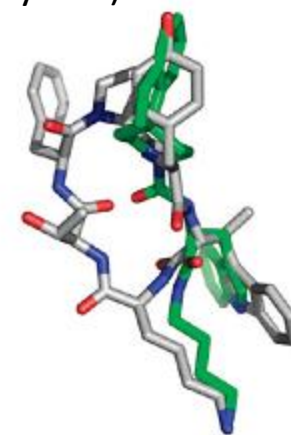
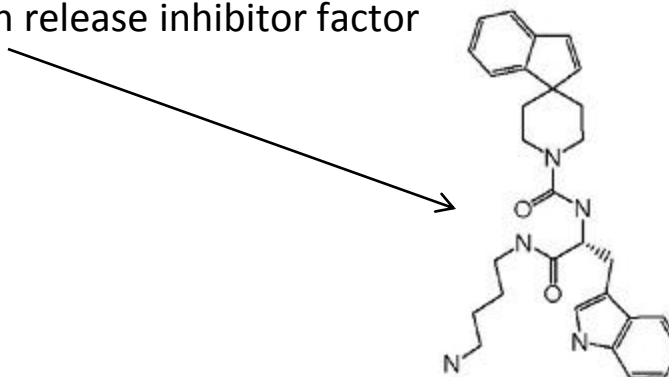
- También hay que especificar
 - **Volumen**
 - Fused spheres
 - Atom-centered Gaussians
 - **Superficie**
 - **Forma**
 - Coincidencia de volumen



Molecular shape and medicinal chemistry: a perspective. 2010.
A Nicholls *et al.* J Med Chem 53: 3862

Varias aplicaciones posibles:

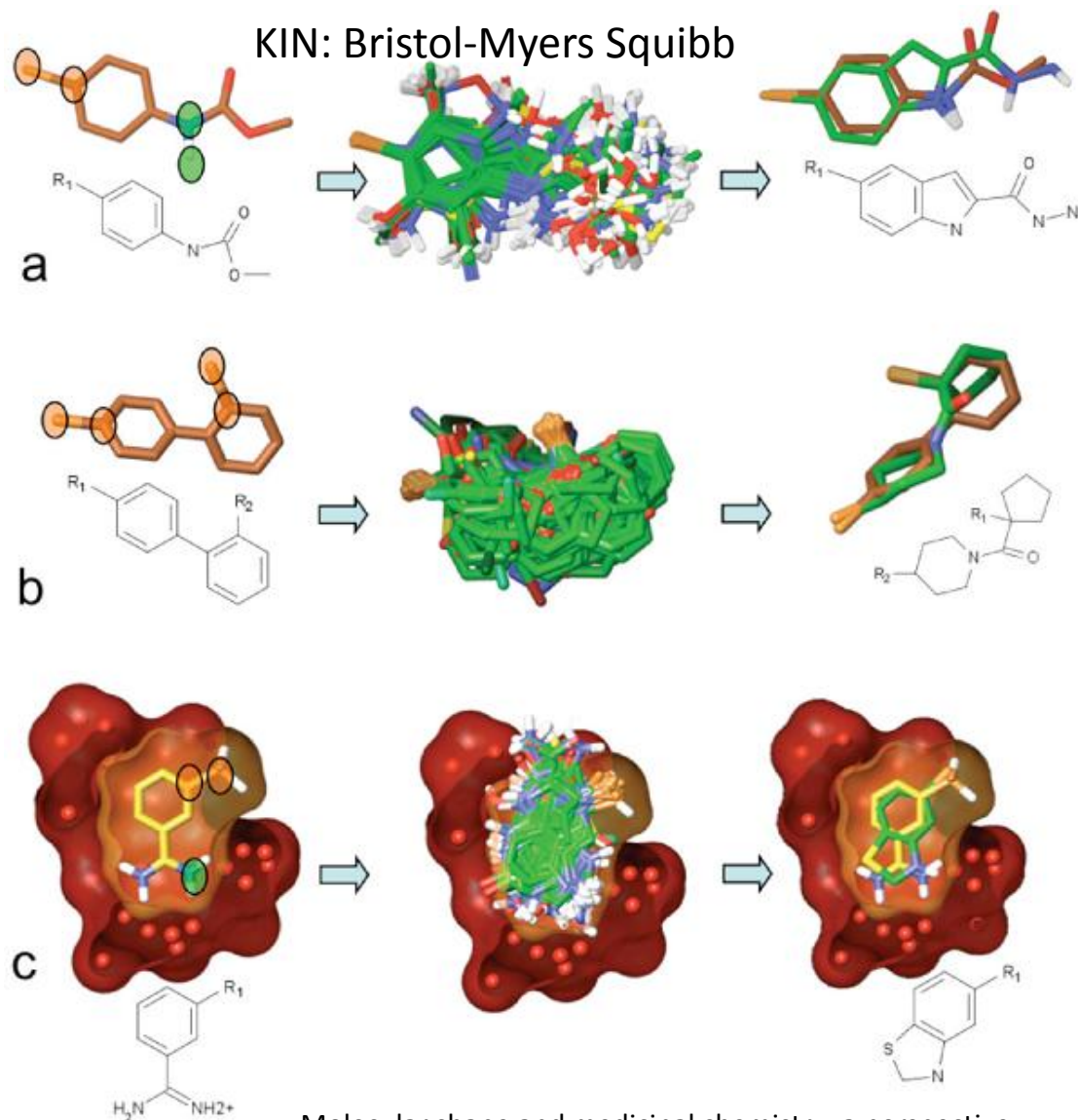
- **Búsqueda de moléculas similares**
 - En este caso la similitud es a nivel de **forma**
 - Se pueden agregar adicionalmente limitaciones
 - **Varias implementaciones en la industria farmacéutica**
 - **Virtual screening**
 - Varios casos de éxito conocidos
 - Merck, primer aplicación publicada del método
 - Identificación de análogos no-peptídicos de:
 - antagonista endógeno del receptor de fibrinógeno (Arg-Gly-Pro)
 - Somatotrophin release inhibitor factor



Representación de forma (shape)

Varias aplicaciones posibles: Lead optimization

- Uno cuenta con una molécula activa que quiere optimizar
- Síntesis de compuestos que exploren el espacio químico alrededor del compuesto *lider*
- Alto costo y labor
- Fácilmente explorable utilizando métodos computacionales



Molecular shape and medicinal chemistry: a perspective.
2010. A Nicholls *et al.* J Med Chem 53: 3862

- **Essentials of Computational Chemistry, 2nd Ed**
 - **CJ Cramer, 2004. Wiley**
- **Chemoinformatics: a textbook, 1st Ed**
 - **J Gasteiger & T Engel, 2003. Wiley**