Introduzione alla chemioinformatica

Loriano Storchi

loriano@storchi.org

http::://www.storchi.org/

Definizione

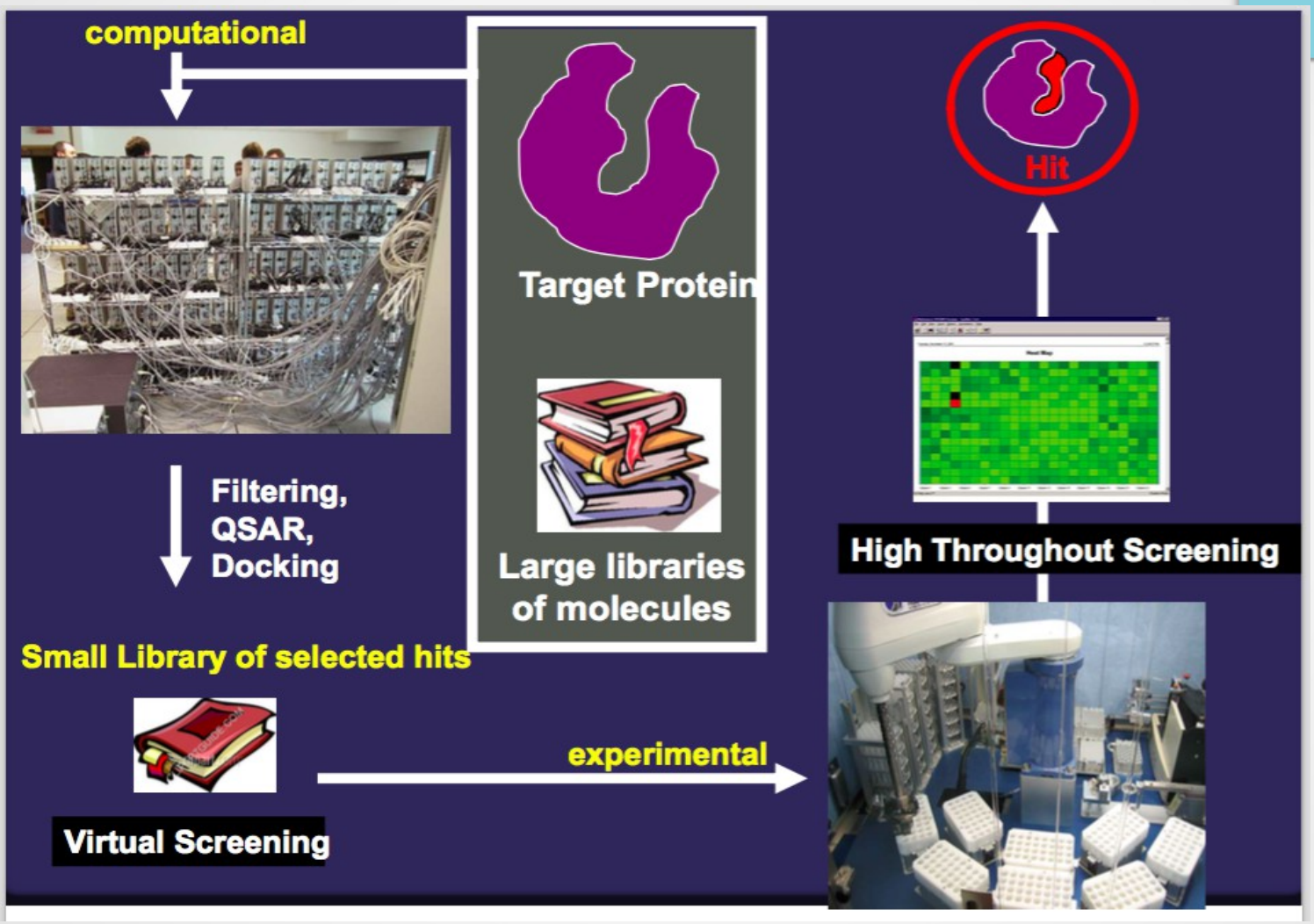
Chemioinformatic:a riguarda l'applicazione di metodi computazionali per affrontare i problemi chimici di varia naturai, con particolare attenzione per la manipolazione delle informazioni strutturali. Il termine è stato introdotto alla fine del 1990 ed è così nuovo che non c'è nemmeno un accordo universale sulla ortografia corretta. Diversi tentativi sono stati fatti per definire la chemioinformatica; tra i più ampiamente citato sono i seguenti :

The mixing of information resources to transform data into information, and information into knowledge, for the intended purpose of making better decisions faster in the arena of drug lead identification and optimisation. [Brown 1998]

Chem(o)informatics is a generic term that encompasses the design, creation, organisation, management, retrieval, analysis, dissemination, visualisation and use of chemical information. [Paris 2000]

Due parole su Chimica Computazionale e Chimica Quantistica

Definizione

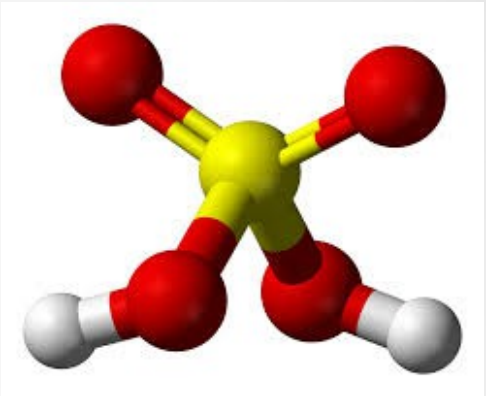


Rappresentazione della struttura molecolare

- L'informazione strutturale deve essere memorizzata in modo tale da poter essere utilizzata da applicazioni software.
- Si deve poter ad esempio poter visuailzzare le strutture, manipolarle, inserirle in un daatbase dove poter poi fare ricerche di strutture o sottostrutture. E poi fare predizione di proprieta' chimico-fisiche
- La rappresentazione deve essere non-ambigua e unica

IUPAC

- La nomenclatura IUPAC certamente e':
  - Standard
  - Include la stereochimica
  - Diffusa e non ambigua
  - Dal nome si puo' ricostruire il composto
- Svantaggi:
  - Nomi non unici
  - Set di regole complesso (da implementare)
  - Nomi lunghi e complicati



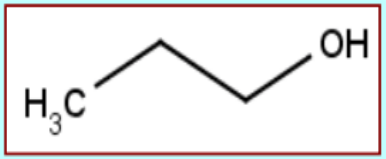
acido tetraossosolforico(VI)

Notazione lineare

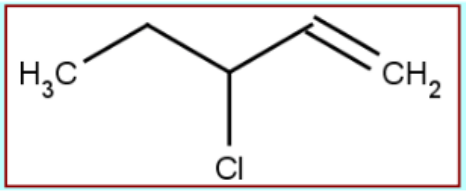
- La notazione lineare (ad esempio IUPAC) ha diversi vantaggi
- E' compatta e quindi occupa poco spazio quando deve essere memorizzata, ad esempio in un computer (Database)
- E' molto facile trasmettere le strutture via e-mail, oppure e' molto facile da usare ad esempio nella ricerca via motore (Google) o DB

SMILES

- Gli atomi sono rappresentati dai loro simboli
- Gli idrogeni sono omessi
- Gli atomi legati sono massi semplicemente l'uno accanto all'altro
- Legami doppi =, legami tripli #
- Le diramazioni sono rappresentate mediante parentesi tonde



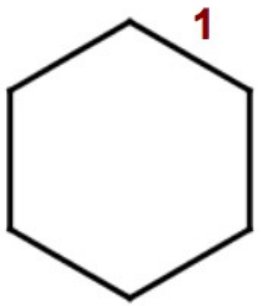
SMILES representation : **CCCO**



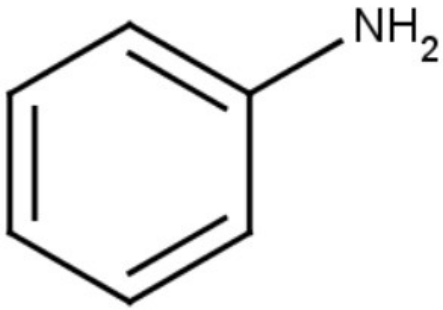
SMILES: **CCC(Cl)C=C**

SMILES

- Anelli mettendo numeri accanto a i due atomi connessi
- Anelli aromaticiusando lettere minuscole
- Si devono usare algoritmi che garantiscano una rappresentazione univoca



SMILES: **C1CCCCC1**



SMILES: **Nc1ccccc1**

SMARTS

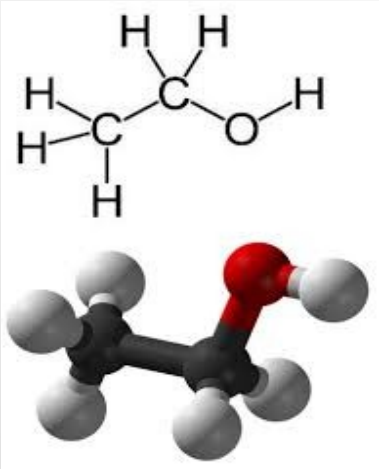
- SMARTS: e’ un linguaggio per descrivere pattern molecolari anche qui usando una stringa ASCII

- Ad esempio: la definizione di accettori o donatori di legame idrogeno usata nell’applicazione della rule of five di Lipinski' puo’ essere codificata usando una SMARTS come quella seguente. I donatori sono definiti come atomi di azoto o ossigeno che hanno almeno un atomo di idrogeno direttamente legato:

[N,n,O;!H0] or [#7,#8;!H0]

InChi

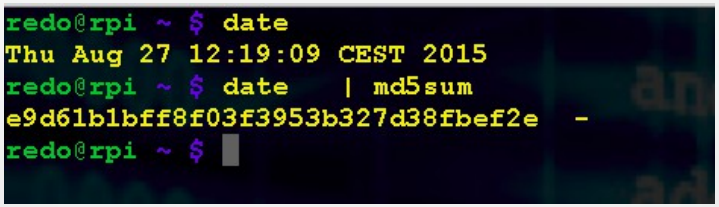
- IUPAC International Chemical Identifier
- Equivalente digitale del nome IUPAC
- La notazione comprende 5 (6) layers che contengono informazioni su: connettivita', tautomerismo, steriochimica, carica ed isotopi
- C'e' un algoritmo che genera il codice InChi che e' unico
- E' stato disegnato per essere compatto, e' poco leggibile ma puo' essere interpretato manualmente non solo automaticamente



Etanolo: InChI=1S/C2H6O/c1-2-3/h3H,2H2,1H3

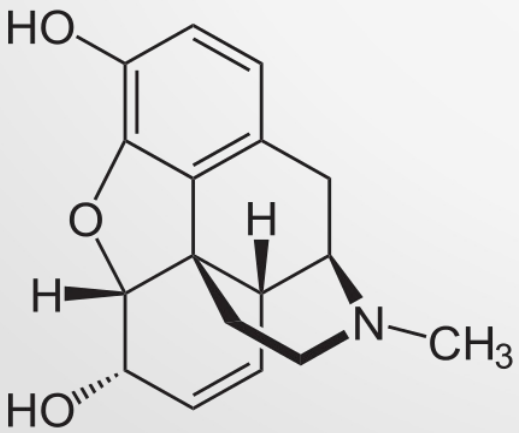
InChiKey

- Visto che un codice InChi puo' risultare troppo lungo, si puo' condensare il codice usando funzioni di hash
- Algoritmo di hash:
  - Una funzione che dato un flusso di bit di dimensione variabile restituisce una stringa di lettere o numeri
  - La stringa e' un identificativo univoco (di dimensioni fisse)
  - Non e' invertibile quindi a partire dalla stringa restituita non e' possibile determinare il flusso originale



InChiKey

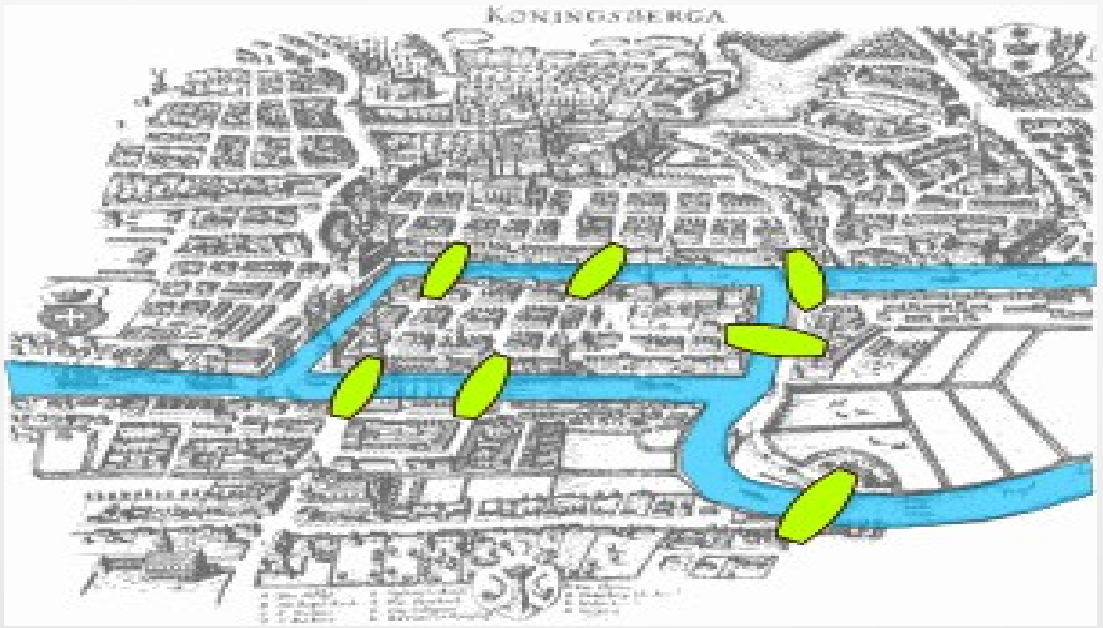
- InChiKey e' dunque ottenuta applicando al codice InChi l'algoritmo di hashing SHA-256
- Il risultato (digest) e' una stringa 14 caratteri risultati dall'hashing della connettivita' piu' altri 10 caratteri risultati dall'hashing del resto delle informazioni (ultimo carattere versione InChi usata)
- Ovviamente non e' possibile dall'InChiKey risalire alla struttura (ovviamente e' possibile farlo partendo dl codice InChi stesso)



Morfina: BQJCRHHNABKAKU-KBQPJGBKSA-N

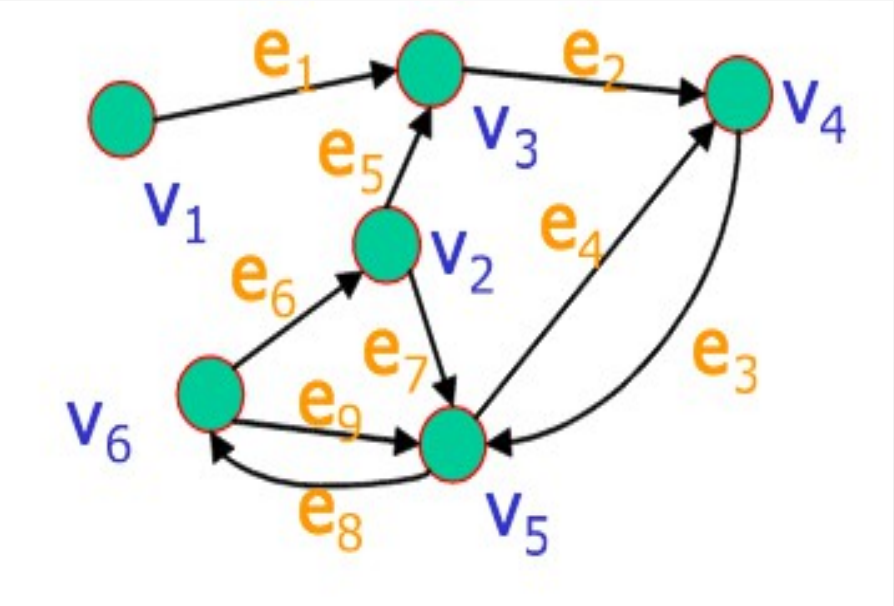
Teoria dei Grafi

- Eulero ed il problema dei sette ponti di Königsberg
- E' possibile fare una passeggiata che permetta di attraversare ogni ponte una ed una sola volta tornando alla fine al punto di partenza ? Eulero dimostro' che non era possibile 1736



Teoria dei grafi

- Studio dei grafi,oggetti che permettono di schematizzare una certa varieta' di problemi.
- Grafo formalmente e' una coppia di insiemi (N, A) , dove N = {v1, v2, v3, … } e' insieme finito di elementi detti nodi , mentre A = {e1, e2, e3, e4,...} ⊆N X N e' un sotto-insieme finito di coppie ordinate di nodi detti archi

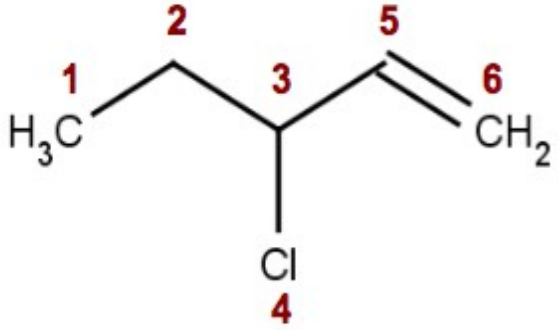


Grafi e molecole

- In teoria dei grafi ci interessiamo solo della connettivita' quindi non ci interessano le posizioni relative dei nodi, ma solo come sono connessi
- E facile immaginare di usare gli algoritmi e la teoria dei grafi in ambito chimico vedendo gli atomi delle molecole come nodi del grafo e gli archi come i legami
- Facile memorizzare strutture in un calcolatore some grafi e d usare algoritmi di ricerca di sottografi ad esempio e tanti altri

Grafi, Molecole e matrici

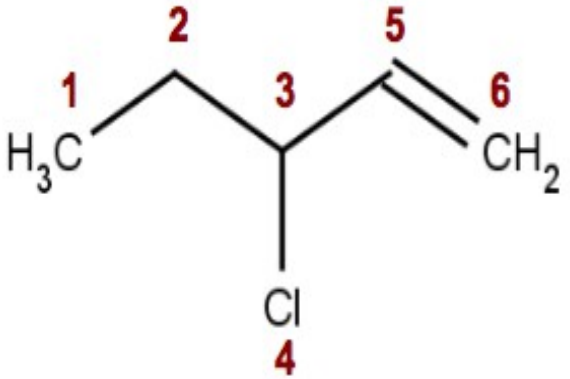
- Per rappresentare e lavorare con i grafi si possono usare diverse strutture
- Matrice di adiacenza, indica gli atomi (nodi) che sono legati
- Posso non memorizzare gli zeri e memorizzare solo meta' matrice essendo questa simmetrica



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

Grafi, Molecole e matrici

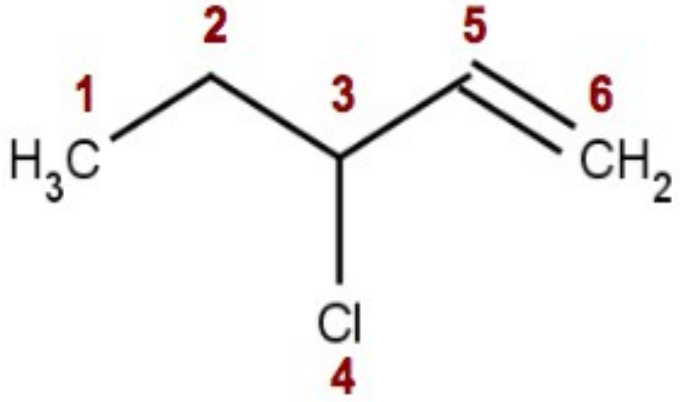
- Matrice delle distanze, ogni elemento della matrice memorizza la distanza tra atomi (vertici). Distanza definita come il numero di legami tra due atomi lungo il cammino piu' breve



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

Grafi, Molecole e matrici

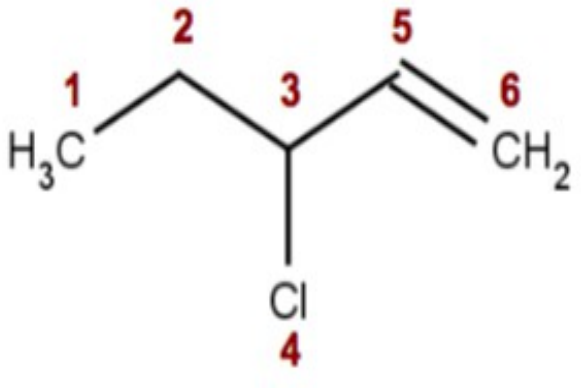
- Matrice dei legami, in questo caso si idicano non solo gli atomi legati ma anche la molteplicita' di lagame fra essi



	1	2	3	4	5	6
1	0	1	0	0	0	0
2	1	0	1	0	0	0
3	0	1	0	1	1	0
4	0	0	1	0	0	0
5	0	0	1	0	0	**2**
6	0	0	0	0	**2**	0

Connection Table

- Uno dei problemi dell'uso delle matrici nella memorizzazione dei grafi e' senza dubbio che le dimensioni crescono come **n^2** dove **n** e' il numero di atomi
- La connection table non e' altro che una lista degli atomi assime ai legami



List of atoms

1	C
2	C
3	C
4	Cl
5	C
6	C

List of bonds

1st	2nd	order
1	2	1
2	3	1
3	4	1
3	5	1
5	6	2

Il formato MDL

number of bonds

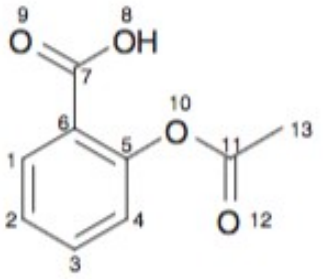
number of atoms

the first atom is a carbon

```
  -ISIS-  09270222202D

 13 13  0  0  0  0  0  0  0  0  0999 V2000
   -3.4639   -1.5375    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -3.4651   -2.3648    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -2.7503   -2.7777    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -2.0338   -2.3644    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -2.0367   -1.5338    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -2.7521   -1.1247    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -2.7545   -0.2997    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -2.0413    0.1149    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   -3.4702    0.1107    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   -1.3238   -1.1186    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   -0.6125   -1.5292    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.6167   -2.3542    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    0.1000   -1.1125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  2  0  0  0  0
  6  7  1  0  0  0  0
  3  4  2  0  0  0  0
  7  8  1  0  0  0  0
  7  9  2  0  0  0  0
  4  5  1  0  0  0  0
  5 10  1  0  0  0  0
  2  3  1  0  0  0  0
 10 11  1  0  0  0  0
  5  6  2  0  0  0  0
 11 12  2  0  0  0  0
  6  1  1  0  0  0  0
 11 13  1  0  0  0  0
M  END
```

the first three numbers are the *x*, *y* and *z* coordinates of the atom

the first bond is between atoms 1 and 2 and has order 2



SDFile

- Sono una serie di strutture in formao MDL separate da un separatore standard ed ogni struttura puo' avere dati extra associati

```
first
     test

  4  3  0  0  0  0  0  0  0  0999 V2000
    0.9392    1.1022   -0.4875 N   0  0  0  0  0  0  0  0  0  0  0  0
    2.0192    1.1021   -0.4866 H   0  0  0  0  0  0  0  0  0  0  0  0
    0.5795    0.5926   -1.3692 H   0  0  0  0  0  0  0  0  0  0  0  0
    0.5792    2.1204   -0.4875 H   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0
  1  3  1  0
  1  4  1  0
M  END
>  <data>
testdatavale

$$$$
second
     test 3D

  3  2  0  0  0  0  0  0  0  0999 V2000
    0.0000   -0.2750    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    0.7140    0.1370    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
   -0.7140    0.1370    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
  2  1  1  0
  1  3  1  0
M  END


$$$$
```

Un paio di esempi….

git clone https://bitbucket.org/lstorchi/teaching.git

```
[redo@banquo teaching (master)]$ cd xyzviwer/
[redo@banquo xyzviwer (master)]$ python xyzview.py methane.xyz ^C
[redo@banquo xyzviwer (master)]$ cd ../ringperception/
[redo@banquo ringperception (master)]$ python ./ring_per.py
1.sdf        2.sdf        3.sdf        4.sdf        mols.smi     ring_per.py  test.sdf
[redo@banquo ringperception (master)]$ python ./ring_per.py mols.smi
Molecule number :  1
1  -->  False 3
   1
   2
   3
2  -->  False 4
   6
   7
   8
   9
3  -->  True 6
   11
   12
   13
   14
   15
   16
Molecule number :  2
```

Unpaio di esempi...

- Visualizzatore basato su VTK

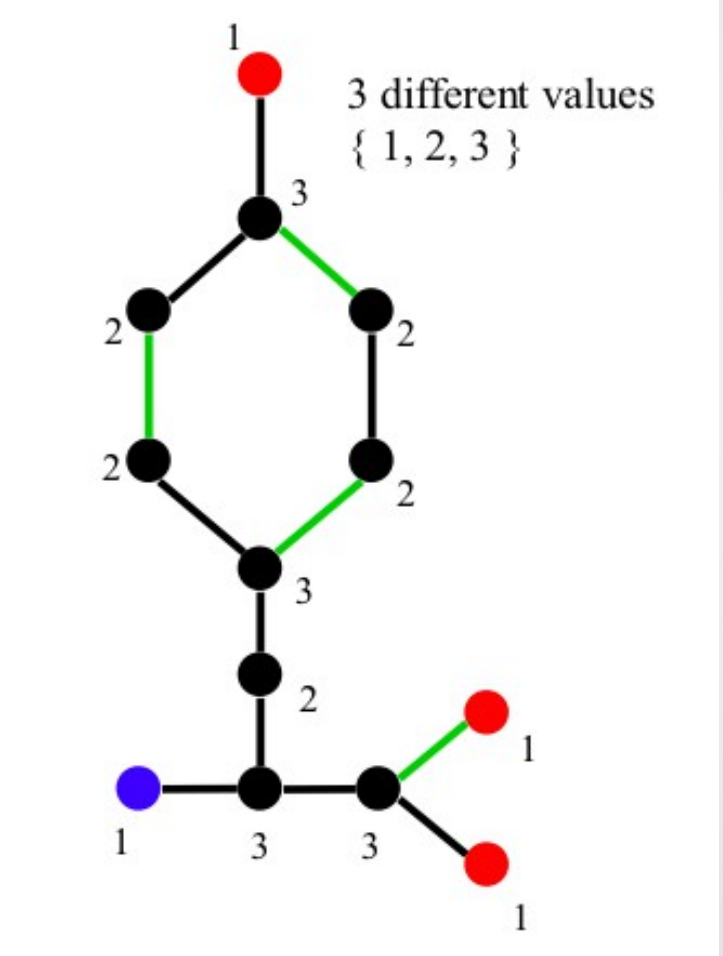
```
import vtk
import sys
import re

########################################################################

def get_color (atom):

  if (atom == 'C'):
    return 1.0, 0.0, 0.0
  elif (atom == 'H'):
    return 1.0, 1.0, 1.0

  return 0.0, 0.0, 0.0

########################################################################

filename = ""

radius = {'H':1.2, 'C':1.7}

if (len(sys.argv)) == 2:
  filename = sys.argv[1]
else:
  print "usage :", sys.argv[0] , " xyzfile"
  exit(1)

filep = open(filename, "r")

filep.readline()
filep.readline()

actors = []

for line in filep:
  p = re.compile(r'\s+')
  line = p.sub(' ', line)
  line = line.lstrip()
  line = line.rstrip()
```

Esempio la canonicalizzazione

- Una struttura molecolare o un grafo puo' essere scritta in diversi modi, ad esempio ordine degli atomi
- E' facile immaginare che sia molto utile, ad esempio nel confrontare due strutture in modo semplice avere lo stesso ordine di atomi e legami
- Per questo sono stati sviluppati algoritmi utili allo scopo
- Un esempio e' l'algoritmo di Morgan (1965)
- Di seguito ne faremo una descrizione semplice e sommaria essenzialmente allo scopo di introdurre all'idea di algoritmo

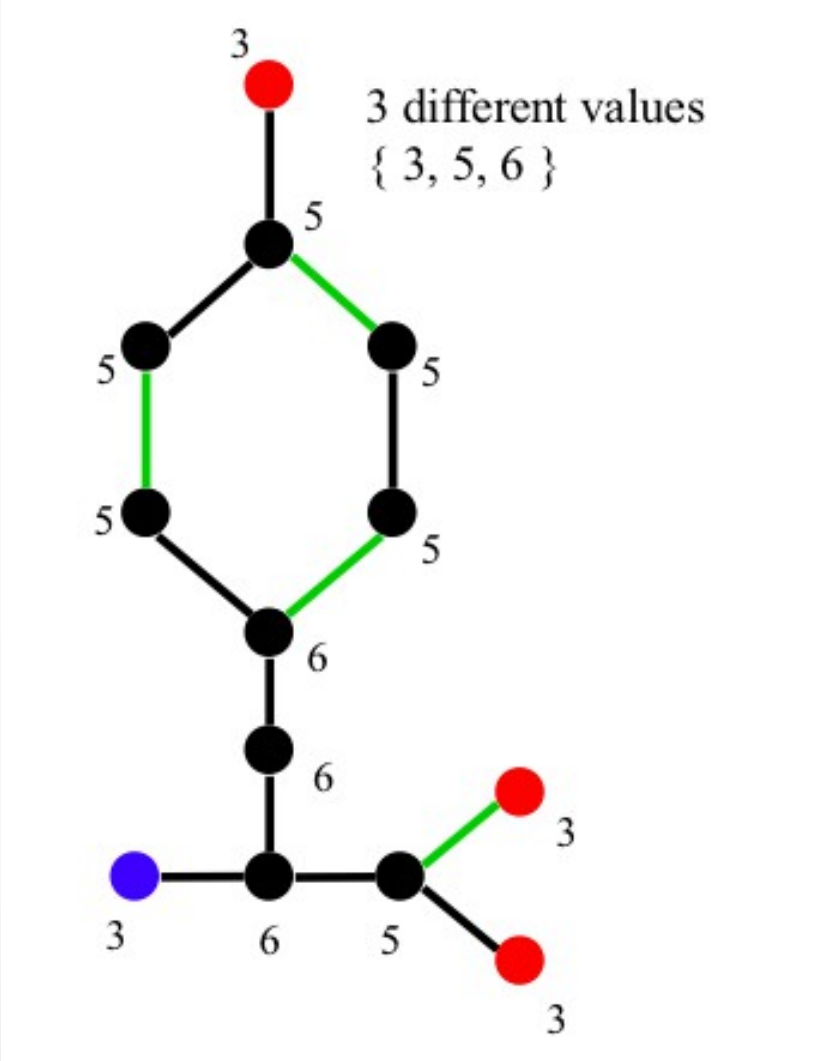
Algoritmo di Morgan

- (1) Associamo ad ogni atomo (nodo) un'etichetta uguale al numero di legami che forma
- (2) Conto quante classi ottengo. In questo caso ottengo tre diverse classi



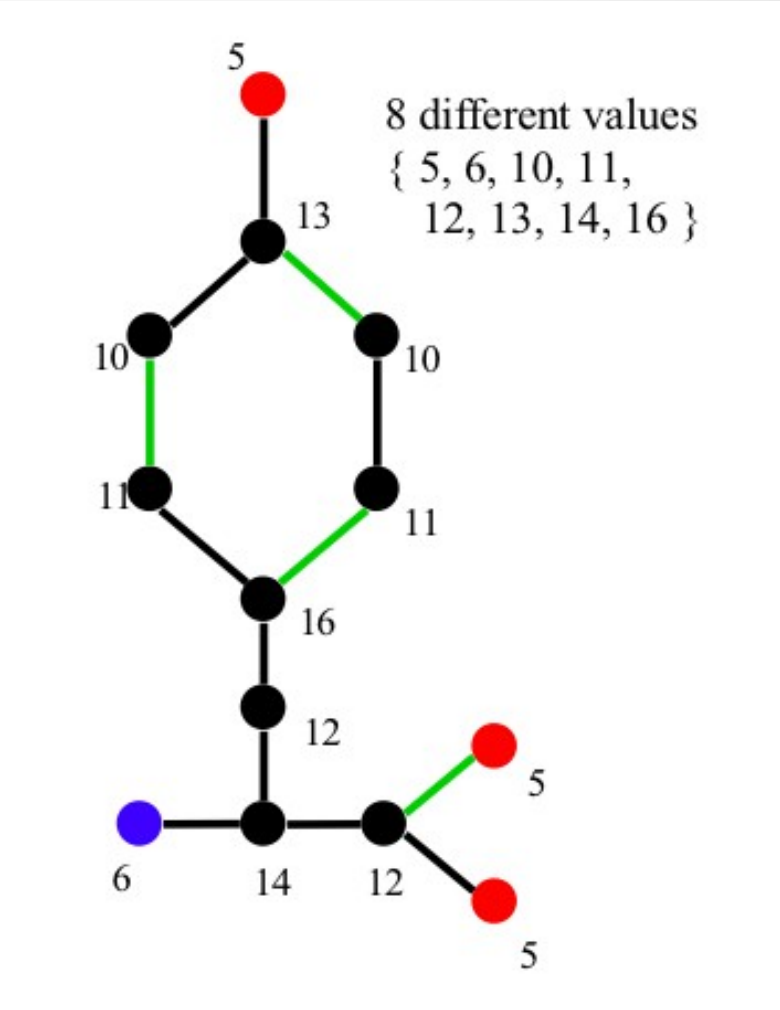
Algoritmo di Morgan

- (3) Ricalcolo le etichette sommando i valori di atomi legati
- (4) E conto quante classi differenti ottengo. In questo caso sempre tre



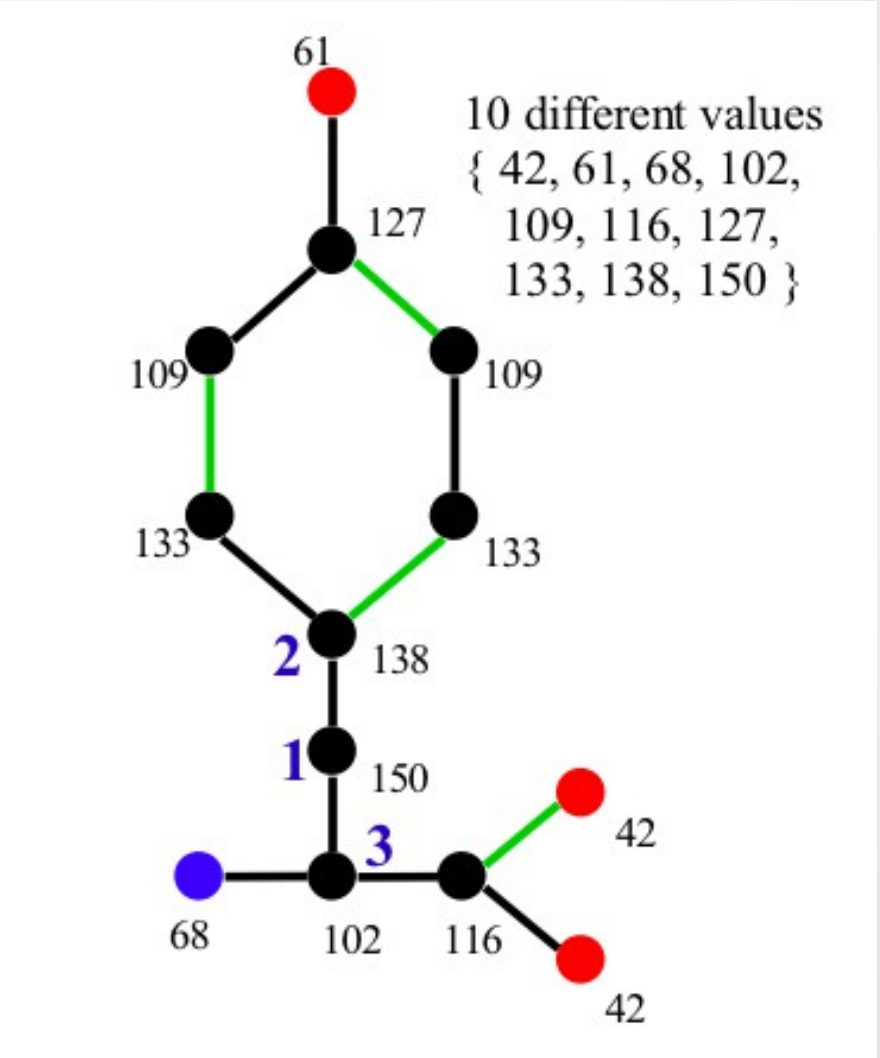
Algoritmo di Morgan

- (5) Ripeto dal punto tre fino che il numero delle classi che ottengo non rimane costante



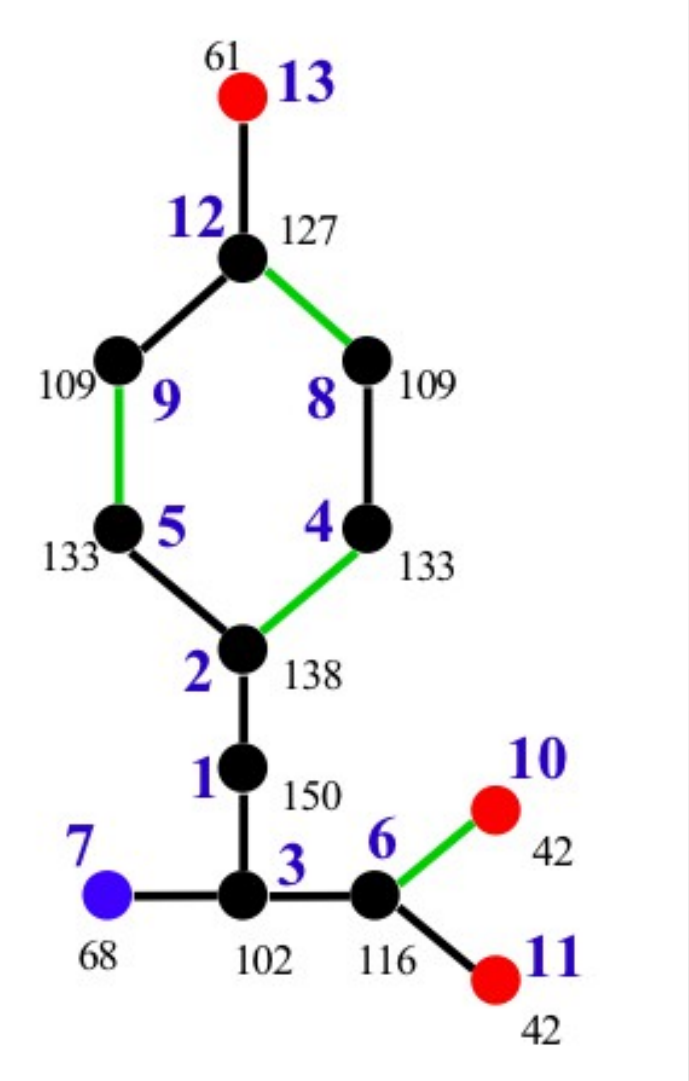
Algoritmo di Morgan

- (5) Ripeto dal punto tre fino che il numero delle classi che ottengo non rimane costante
- Numeriamo adesso a partire dal nodo con l'etichetta a valore maggiore (150 in questo caso)
- Poi numero i vicini (vedi 2 e 3 in figura)



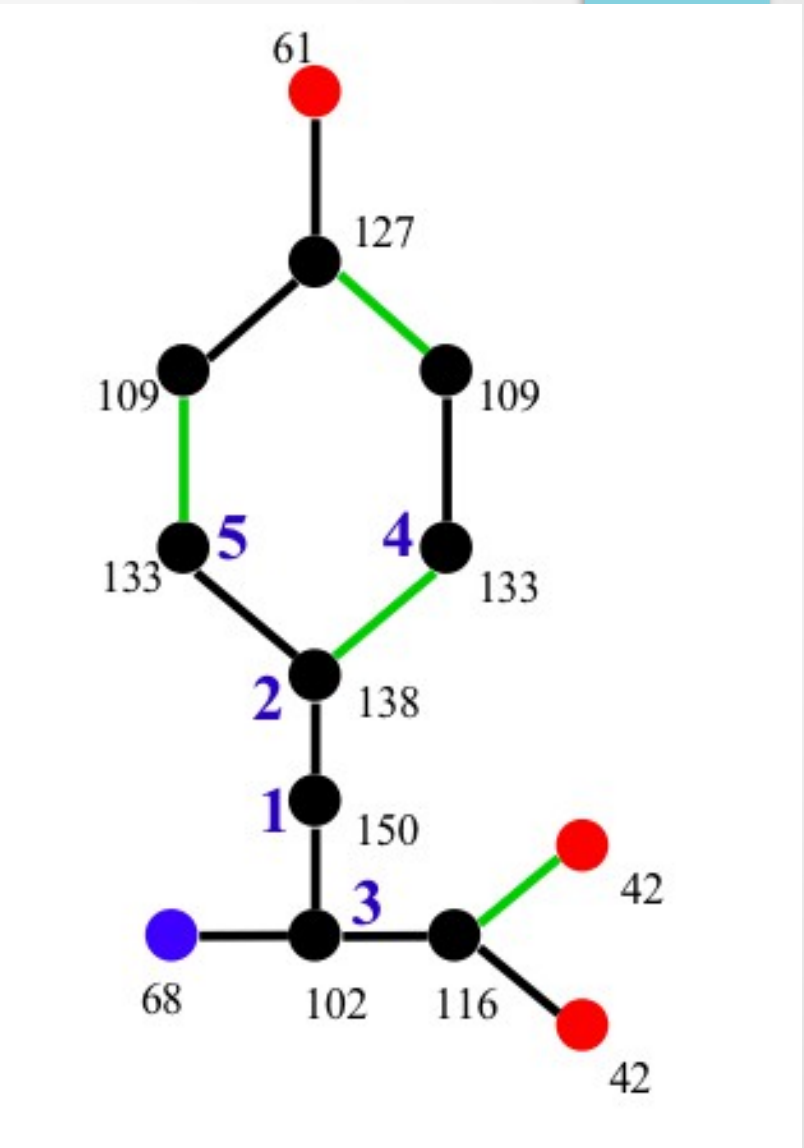
Algoritmo di Morgan

- Posso in questo modo arrivare a dare un'ordine a tutti i nodi (atomi) canonico.
- Questo algoritmo non e' perfetto nel tempo si sono trovati ovviamente algoritmi “migliori”



Algoritmo di Morgan

- Quando passo al nodo 2, in questo case ho due etichette uguali. Assegno per primo il nodo (atomo) con ordine di legame maggiore (vedi in figura)



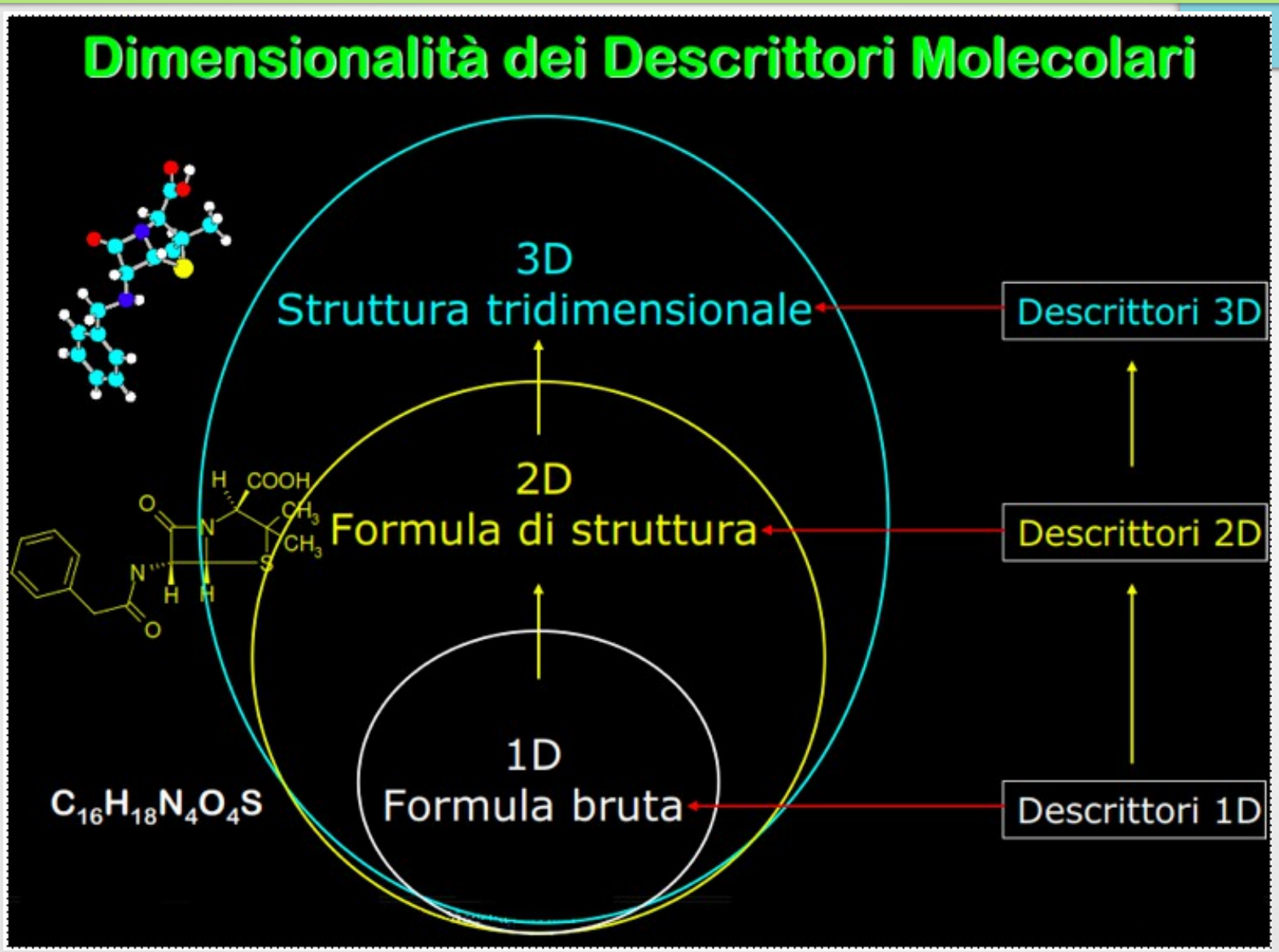
...

- Complessivamente sotto il nome di chemioinformatica vanno ad essere definite numerose tecniche che appunto sono la combinazione di chimica, informatica e teoria dell'informazione. Farne una classificazione o ance solamente lista completa e' al di fuori della durata del presente corso
- Prenderemo invece in considerazione solo alcuni aspetti di base

Descrittori molecolari

- Uno degli aspetti fondamentali della chimica e' la ricerca della relazione fra la struttura di una molecola e le sue caratteristiche o proprieta'
- Dove per proprieta' vogliamo intendere le proprieta' fisiche, come punto di ebollizione o pKa ad esempio, piuttosto che reattive e soprattutto biologiche, ad esempio tossicita' piuttosto che attivita' farmacologica
- I descrittori molecolari possono essere grandezze chimico-fisiche (peso molecolare, punto di ebollizione, punto di fusione, pressione di vapore, coefficiente di ripartizione e di distribuzione, indice di rifrazione entalpia di formazione...) o valori che si ottengono dall'applicazione di un algoritmo alla struttura molecolare (1D, 2D o 3D).
- Possiamo poi trovare il "modo" di mettere in relazione questi descrittori, od una combinazione di essi, ad un qualche tipo di proprieta' di nostro interesse. Questo ci permette ade sempio di "predire" queste proprieta' anche senza esperimenti o addirittura senza avere effettivamente sintetizzato la struttura.

Descrittori molecolari



Descrittori

- L'uso di descrittori determinati sperimentalmente non e' generalmente conveniente in funzione del fatto che per poterli usare questo richiede la sintesi delle strutture e la determinazione sperimentale generalmente “lenta”
- Esempio di descrittore 1D, ad esempio il peso molecolare
- Descrittori 2D hashed fingerprints, indice di Wiener (somma di tutte le distanze interatomiche):

$$W = \frac{1}{2} \sum_{i,j}^{N} d_{ij}$$

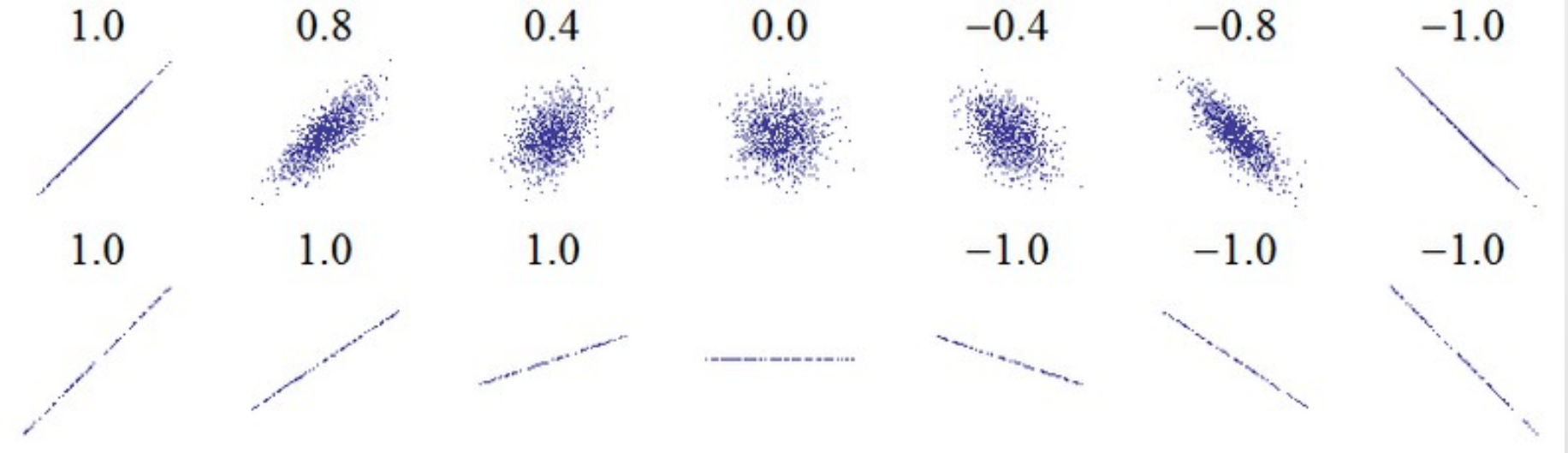
- Descrittori 3D ad esempio area superficiale

Verifica e manipolazione dei dati

- Dopo aver “generato” i descrittori per un set di molecole e' generalmente utile una fase di verifica
- Verificare che i dati siano distribuiti secondo una distribuzione normale
- Verificare quanto sono dispersi di dati. Se una dato descrittore mostra una scarsa variabilita' all'interno del set delle molecole non ha molto senso includerlo nel modello
- Scalatura dei dati, se i dati sono numericamente molto diversi fra di loro e' utile scalarli (normalizzarli)
- Verificare se e quanto sono correlati i descrittori, se due descrittori sono molto correlati non ha senso includerli entrambi

Correlazione

$$-1 \leq \rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i - \mu_x)^2}\sqrt{\sum_{i=1}^{n}(y_i - \mu_y)^2}} \leq +1$$



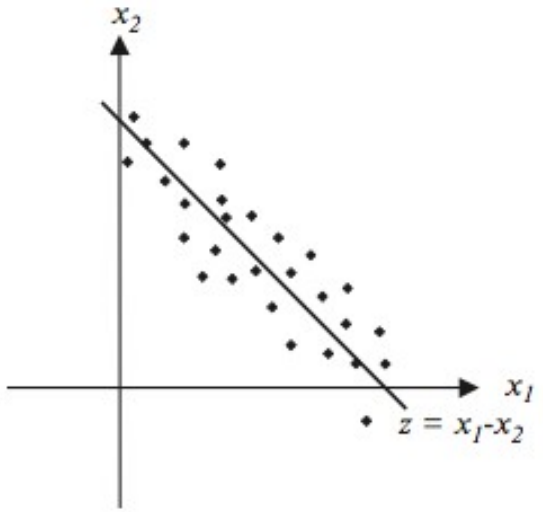
PCA

- Per ridurre la dimensionalita' delle variabili si puo' usare l'analisi delle componenti principali (PCA)

$$PC_1 = c_{1,1}x_1 + c_{1,2}x_2 + \cdots c_{1,p}x_p \quad (3.24)$$

$$PC_2 = c_{2,1}x_1 + c_{2,2}x_2 + \cdots c_{2,p}x_p \quad (3.25)$$

$$PC_i = c_{i,1}x_1 + c_{i,2}x_2 + \cdots c_{i,p}x_p = \sum_{j=1}^{p} c_{i,j}x_j \quad (3.26)$$



Equazione QSAR

- Come si ottiene l'equazione QSAR. Ci sono diversi approcci possibili, citiamone alcuni:
- algoritmi genetici e di evoluzione
- reti neurali artificiali
- regressione delle componenti principali (PCR, *principal component regression*)
- metodo dei minimi quadrati parziali (PLS, *partial least squares*) o proiezione delle strutture latenti

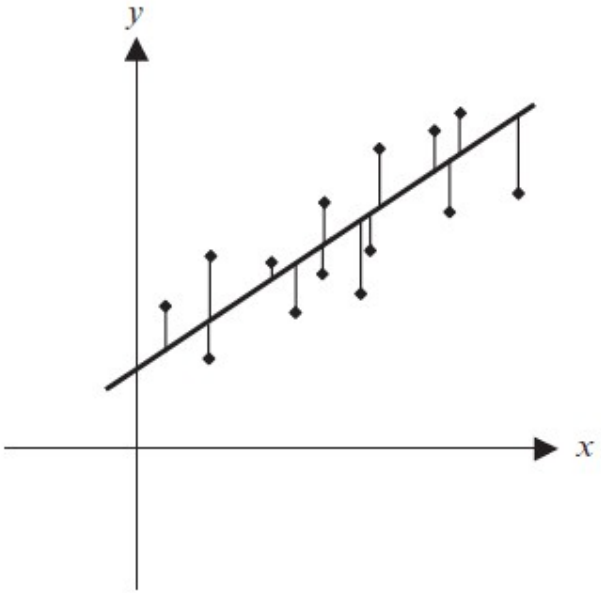
Regressione Lineare

- Ad esempio consideriamo x variabile indipendente (il nostro descrittore, ad esempio il logP = coefficiente di ripartizione ottanolo acqua = log ([soluto]ottanolo/ [soluto]acqua) ed y la variabile dipendente (la proprieta' che vogliamo essere in grado di calcolare, ad esempio attivita' biologica)

$$m = \frac{\sum_{i=1}^{N} (x_i - \langle x \rangle)(y_i - \langle y \rangle)}{\sum_{i=1}^{N} (x_i - \langle x \rangle)^2}$$

$$c = \langle y \rangle - m \langle x \rangle$$

$$y = mx + c$$



Regressione lineare multipla, PCR, PLS

- In generale saranno pero' presenti piu' descrittori e quindi piu' variabili indipendenti. In questo caso si ricorre alla regressione lineare multipla:

$$y=b0 + b1X1 + b2X2 + b3X3 +...bpXp$$

- Nella PCR (Principal Components Regression) si usano le componenti principali nella regressione:

$$y = a_1PC_1 + a_2PC_2 + a_3PC_3 + \cdots$$

- Nella PLS (Partial Least Squares) la variabile dipendente (o le variabili dipendenti) dipendono da variabili latenti:

$$y = a_1t_1 + a_2t_2 + a_3t_3 + \cdots a_nt_n$$

Predizione pKa