Introduzione alla Chimica Computazionale e simili

Loriano Storchi

loriano@storchi.org

http://www.storchi.org/

Definizione Chemoinformatica

Chemioinformatic:a riguarda l'applicazione di metodi computazionali per affrontare i problemi chimici di varia natura, 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]

Definizione Bioinformatica

- **Bioinformatica:** scienza multi-disciplinare, al crocevia tra biologia, chimica, matematica, fisica ed informatica, che analizza l'informazione biologica con metodi computazionali al fine di formulare ipotesi sui processi della vita. (Anna Tramontano)
- Applicazione di tecniche computazionali nella comprensione ed organizzazione di tutta l'informazione associata alle strutture biologiche. La fisiologia di un organismo vivente e' per buona parte determinata dai suoi geni che possono essere visti e trattati come informazione digitale
- Esempio, possibile applicazione
  - Scoperta una nuova sequenza proteica, posso cercare di dedurre la sua funzionalita' , in modo approssimato, confrontandola con tutte le seguenze proteiche note al mondo.
  - Gestire questo tipo di ricerca e' impossibile per un'essere umano ma invece fattibile usando computers (algoritmi di ricerca, database, protocolli, etc etc) ed in generale strumenti informatici

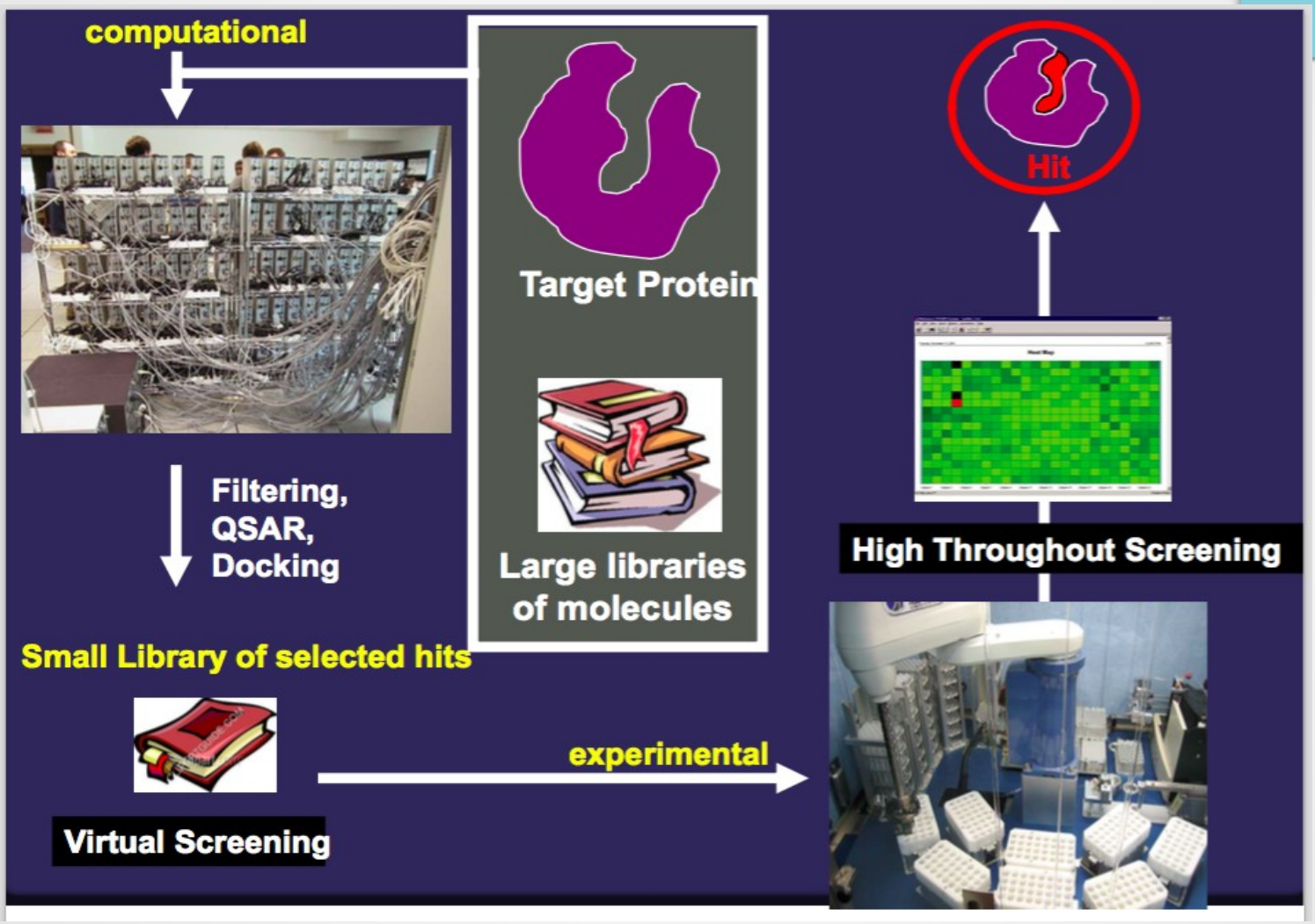
Definizione Bioinformatica

- E' necessario memorizzare analizzare ed interpretare una enorme quantita' di dati. L'intera sequenza del genoma umano, scritta in Times New Roman, dimensione 12, avrebbe una lunghezza di 5000 km!
- DNA (memoria) l'RNA (comunicazione) ed in fine le proteine (esecuzione)
- Quali parti del DNA sono importanti e controllano determinati processi ?
- Quale e' la funzione di certe proteine ?
- Come posso confrontare due sequenze proteiche o genomiche ?
- Molto di piu'.....

Definizione Chimica Computazionale , quantistica

- Parte della chimica teorica che sviluppa modelli matematici basati sulla meccanica classica e sulla meccanica quantistica utili a simulare sistemi chimici ed a determinarne le caratteristiche
- Chiaramente lo sviluppo della chimica computazionale e' fortemente legato al fatto di avere a disposizione macchine con grandi quantita' di memoria ed in grado di eseguire molte operazioni in vigola mobile nell'unita' di tempo (FLOPS)

Definizioni



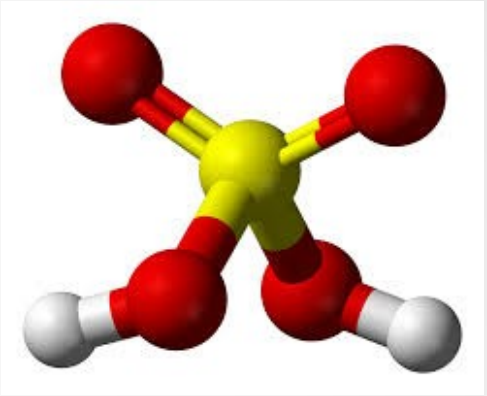
RAPPRESENTAZIONE DELLE STRUTTURE MOLECOLARI

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



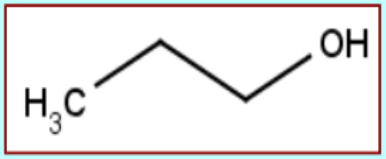
acido tetraossosolforico(VI)

Notazione lineare

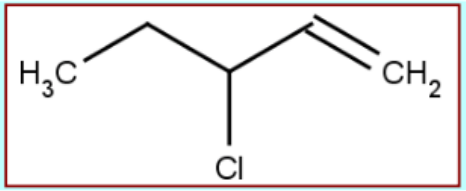
- La notazione lineare (ad esempio IUPAC) ha diversi vantaggi
- E' compatta e quindi occupa poco spazio quando deve essere memorizzata, ad esempio in un computer (Database)
- E' molto facile trasmettere le strutture via e-mail, oppure e' molto facile da usare ad esempio nella ricerca via motore (Google)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

SMILES

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



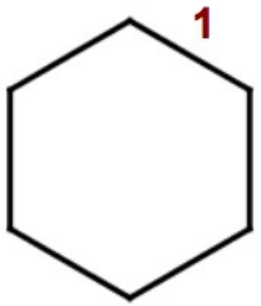
SMILES representation : **CCCO**



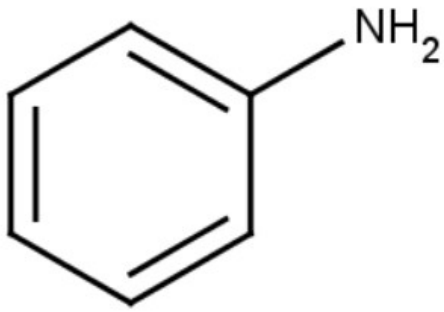
SMILES: **CCC(Cl)C=C**

SMILES

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



SMILES: **C1CCCCC1**



SMILES: **Nc1ccccc1**

SMARTS

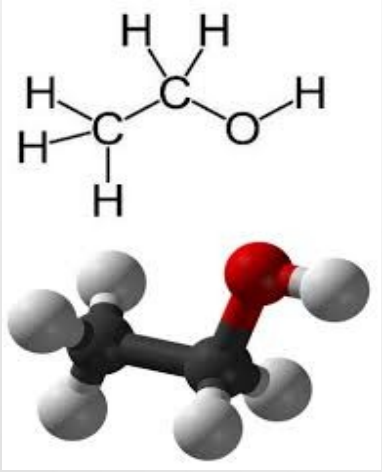
- SMARTS: e’ un linguaggio per descrivere pattern molecolari anche qui usando una stringa ASCII

- 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

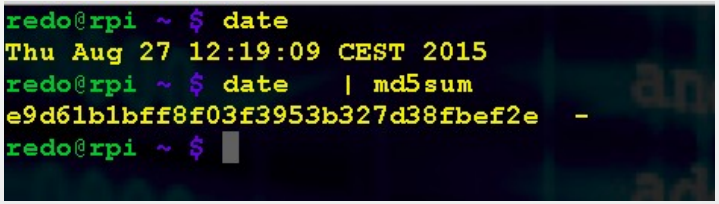
- IUPAC International Chemical Identifier
- Equivalente digitale del nome IUPAC
- La notazione comprende 5 (6) layers che contengono informazioni su: connettivita', tautomerismo, steriochimica, carica ed isotopi
- C'e' un algoritmo che genera il codice InChi che e' unico
- E' stato disegnato per essere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



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

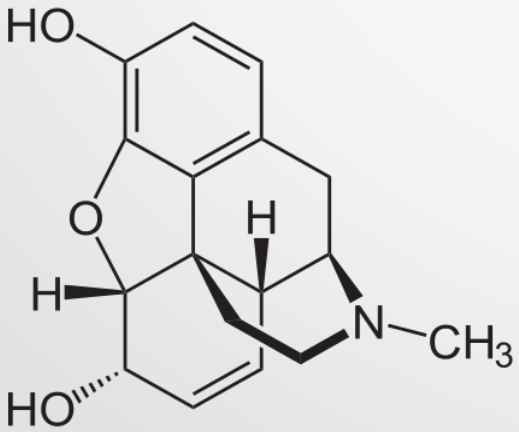
InChiKey

- 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

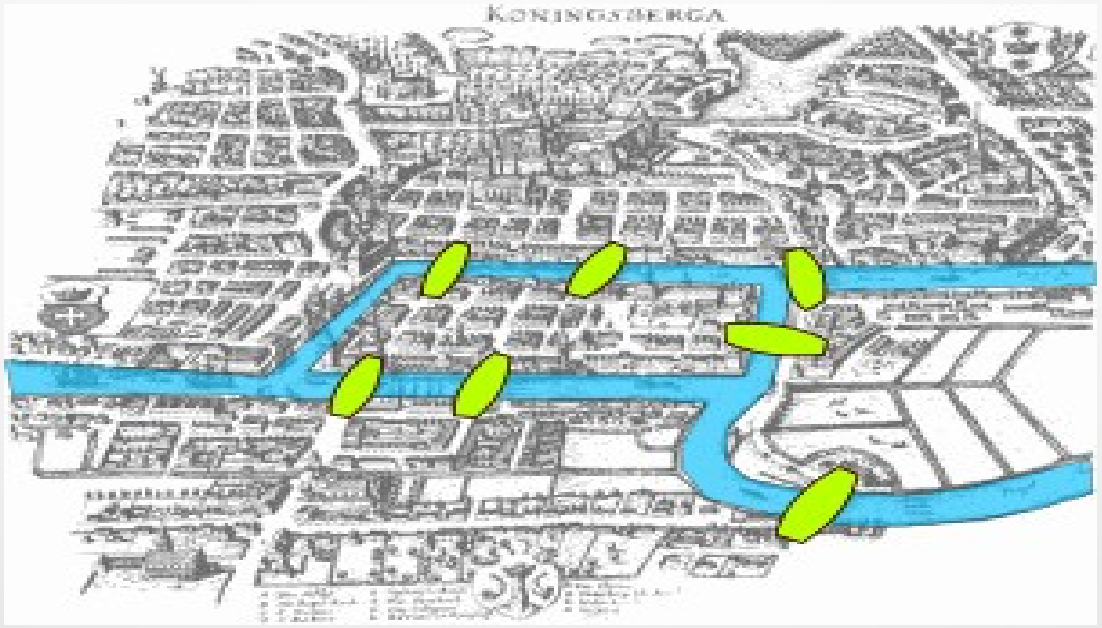
- InChiKey e' dunque ottenuta applicando al codice InChi l'algoritmo di hashing SHA-256
- Il risultato (digest) e' una stringa 14 caratteri risultati dall'hashing della connettivita' piu' altri 10 caratteri risultati dall'hashing del resto delle informazioni (ultimo carattere versione InChi usata)
-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



Morfina: BQJCRHHNABKAKU-KBQPJGBKSA-N

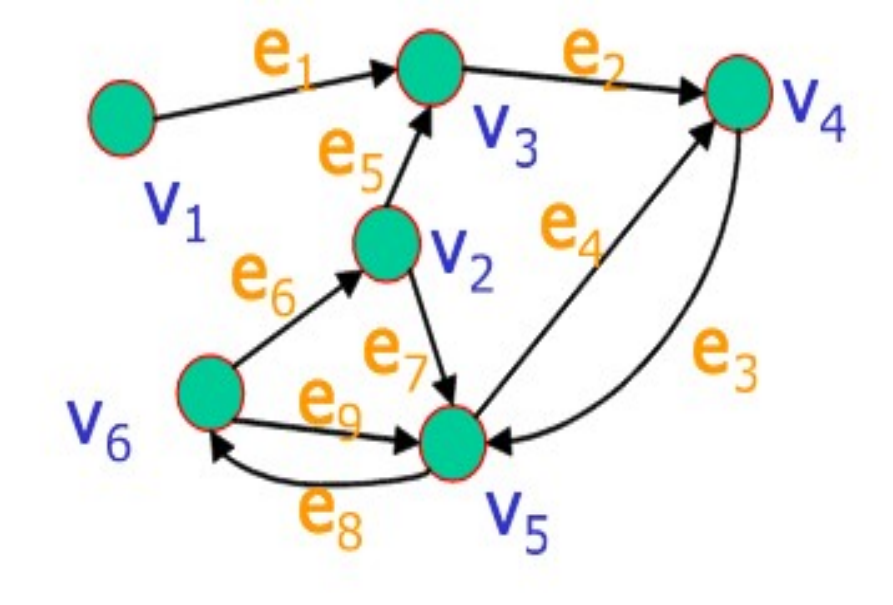
Teoria dei Grafi

- 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

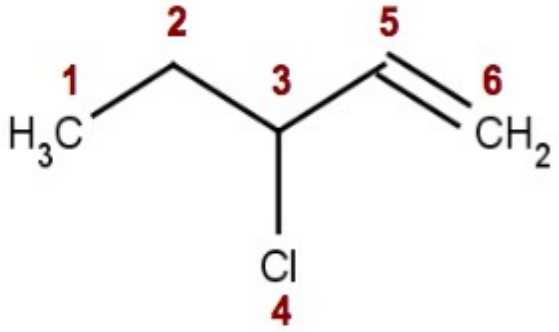


Grafi e molecole

- 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 ed usare algoritmi di ricerca di sottografi ad esempio e tanti altri

Grafi, Molecole e matrici

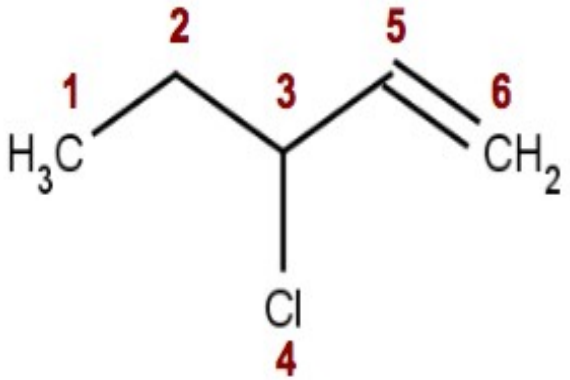
- Per rappresentare e lavorare con i grafi si possono usare diverse strutture
- **Matrice di adiacenza**, indica gli atomi (nodi) che sono legati
- Posso non memorizzare gli zeri e memorizzare solo meta' matrice essendo questa simmetrica



	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

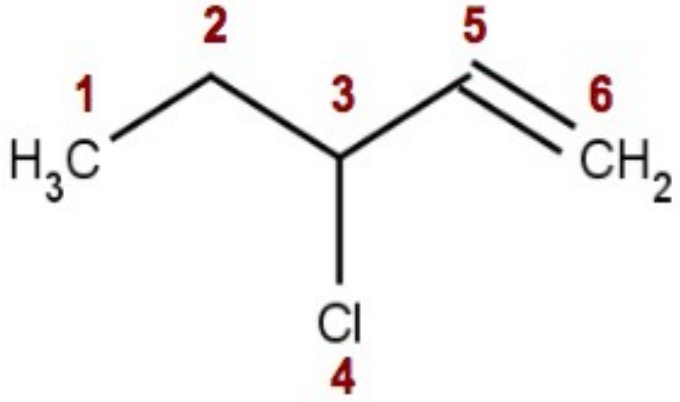
- **Matrice delle distanze**, ogni elemento della matrice memorizza la distanza tra atomi (vertici). Distanza definita come il numero di legami tra due atomi lungo il cammino piu' breve



	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

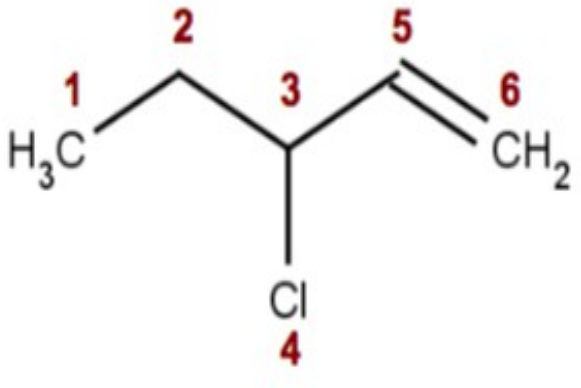
- **Matrice dei legami**, in questo caso si indicano non solo gli atomi legati ma anche la molteplicita' di legame fra essi



	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 **n²** dove **n** e' il numero di atomi
- La **connection table** non e' altro che una lista degli atomi assieme ai legami



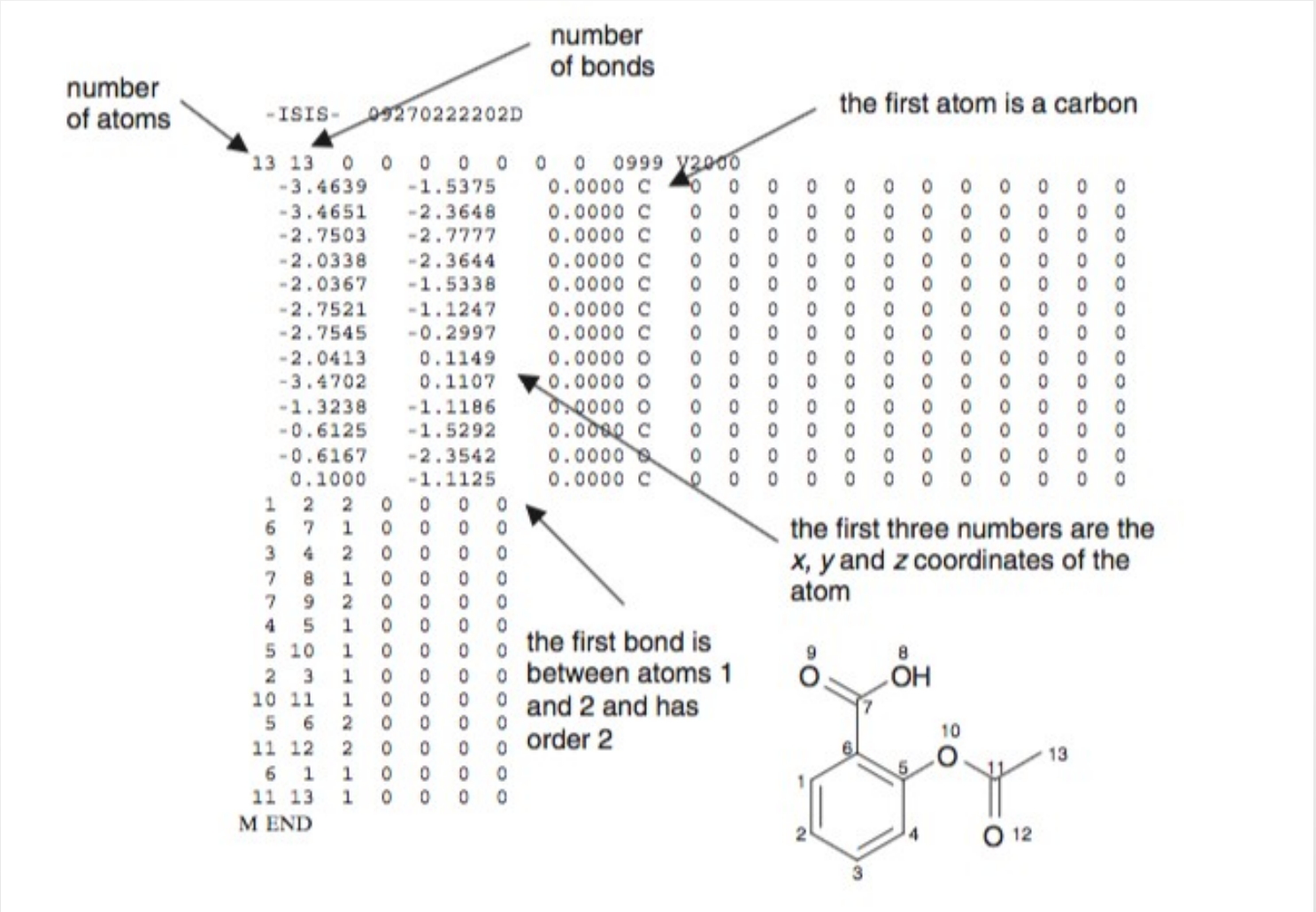
List of atoms

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



SDF

- 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


$$$$
```

ESERCITAZIONE

Convertire SMILE a 3D

- Convertiamo SMILE e InChI a 3D
  - CN1CCC23C4C1CC5=C2C(=C(C=C5)O)OC3C(C=C4)O
  - InChI=1S/C17H19NO3/c1-18-7-6-17-10-3-5-13(20)16(17)21-15-12(19)4-2-9(14(15)17)8-11(10)18/h2-5,10-11,13,16,19-20H,6-8H2,1H3/t10-,11+,13-,16-,17-/m0/s1

```
[redo@buchner 2]$ cat > 1.smi
CN1CCC23C4C1CC5=C2C(=C(C=C5)O)OC3C(C=C4)O
^C
```

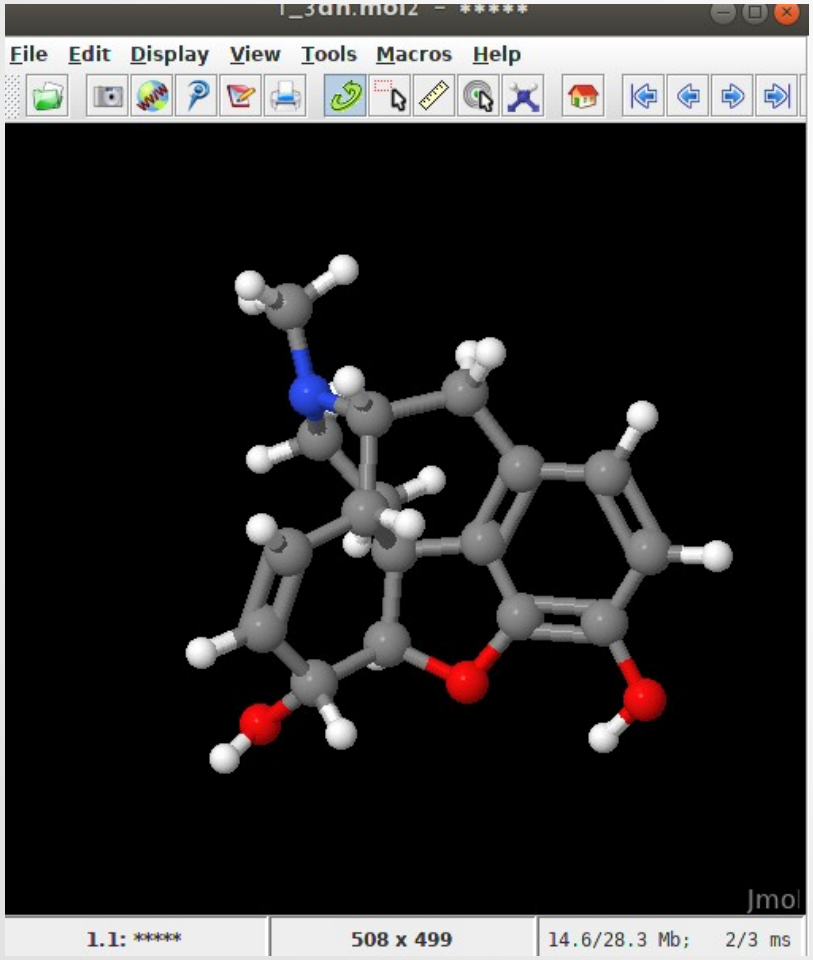
```
[redo@buchner 2]$ cat > 2.inchi
InChI=1S/C17H19NO3/c1-18-7-6-17-10-3-5-13(20)16(17)21-15-12(19)4-2-9(14(15)17)8-11(10)18/h2-5,10-11,1
,16-,17-/m0/s1
^C
```

Convertire SMILE a 3D

- Usiamo Openbabel command line:
  - time babel -ismi 1.smi -omol2 1.mol2
  - time babel -iinchi 2.inchi -omol2 2.mol2
- Visualizzate i mol2 prodotti adesso
- Adesso produciamo i 2D ed i 3D:
  - man babel
  - babel -?
  - time babel --gen2D -ismi 1.smi -omol2 1_2d.mol2
  - time babel --gen3D -ismi 1.smi -omol2 2_3d.mol2
- Adesso visualizziamo ad esempio usando Jmol

Convertire SMILE a 3D

- Adesso trovate il modo di aggiungere gli idrogeni espliciti usando babel



Volendo pulsante destro →
Computation → Optimize structure

Usando il python

- git clone https://github.com/lstorchi/teaching.git
- 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
```

Usando il python

- 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()
```

==========

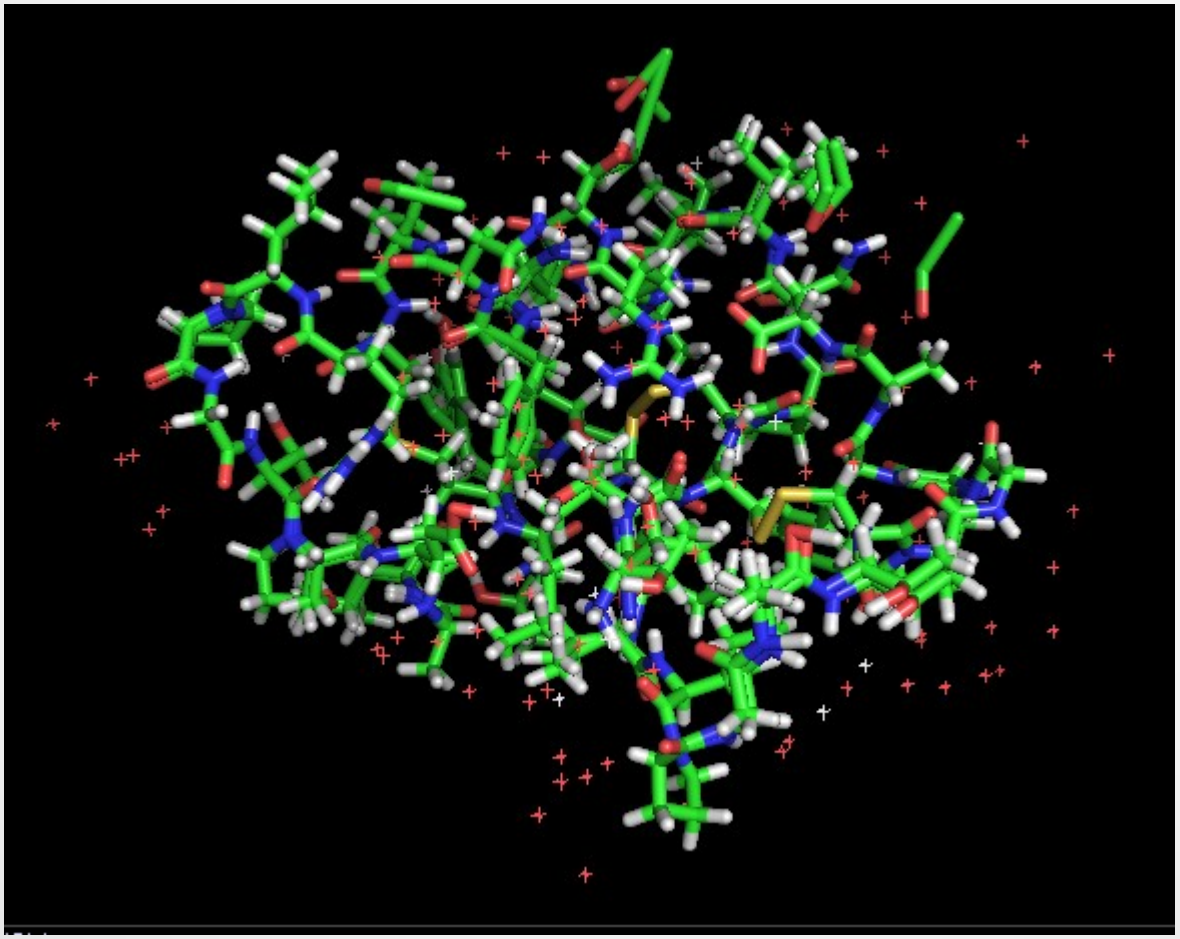
ESERCITAZIONE

Visualizzazione di strutture

- https://www.rcsb.org/structure/3NIR

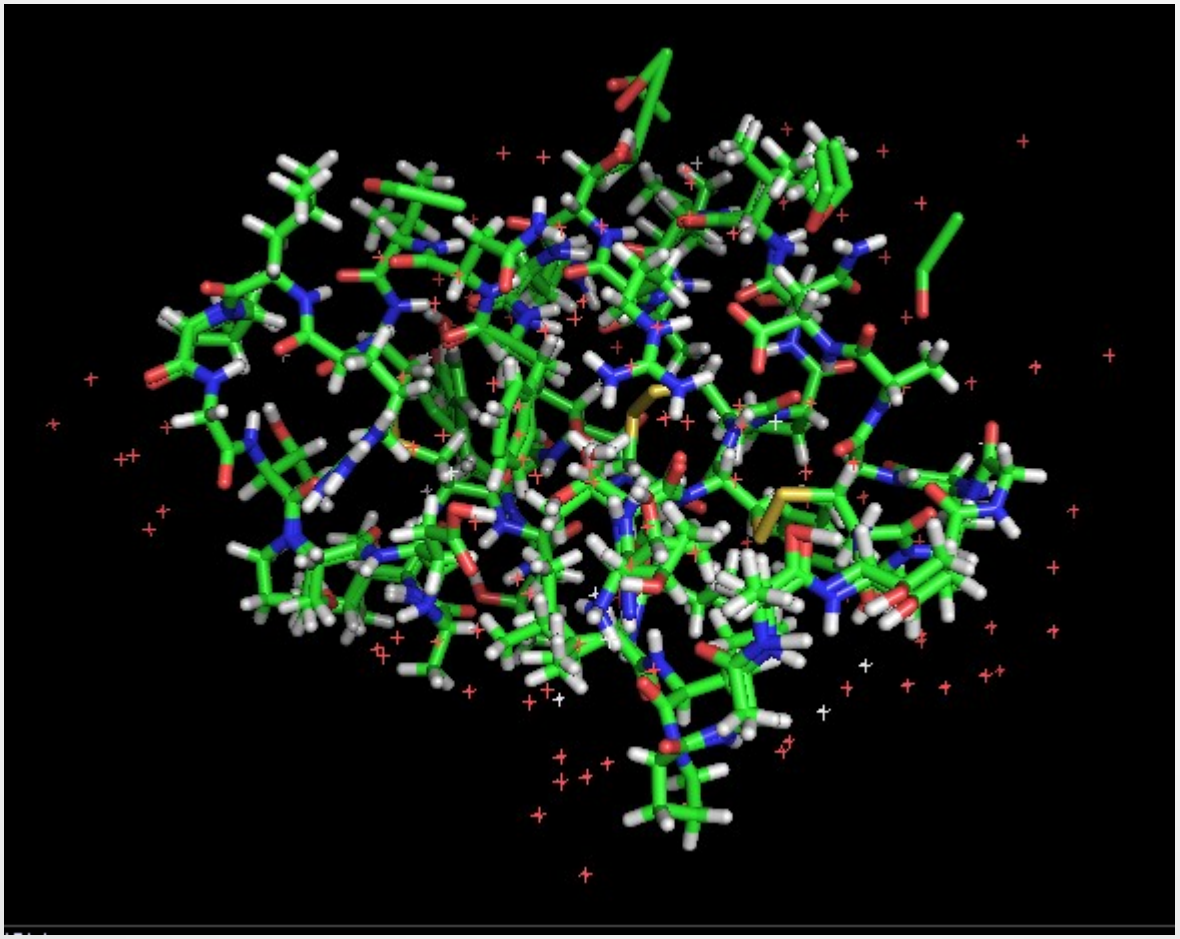
Visualizzazione di strutture

- Aprire usando pymol Stick (S → stick)



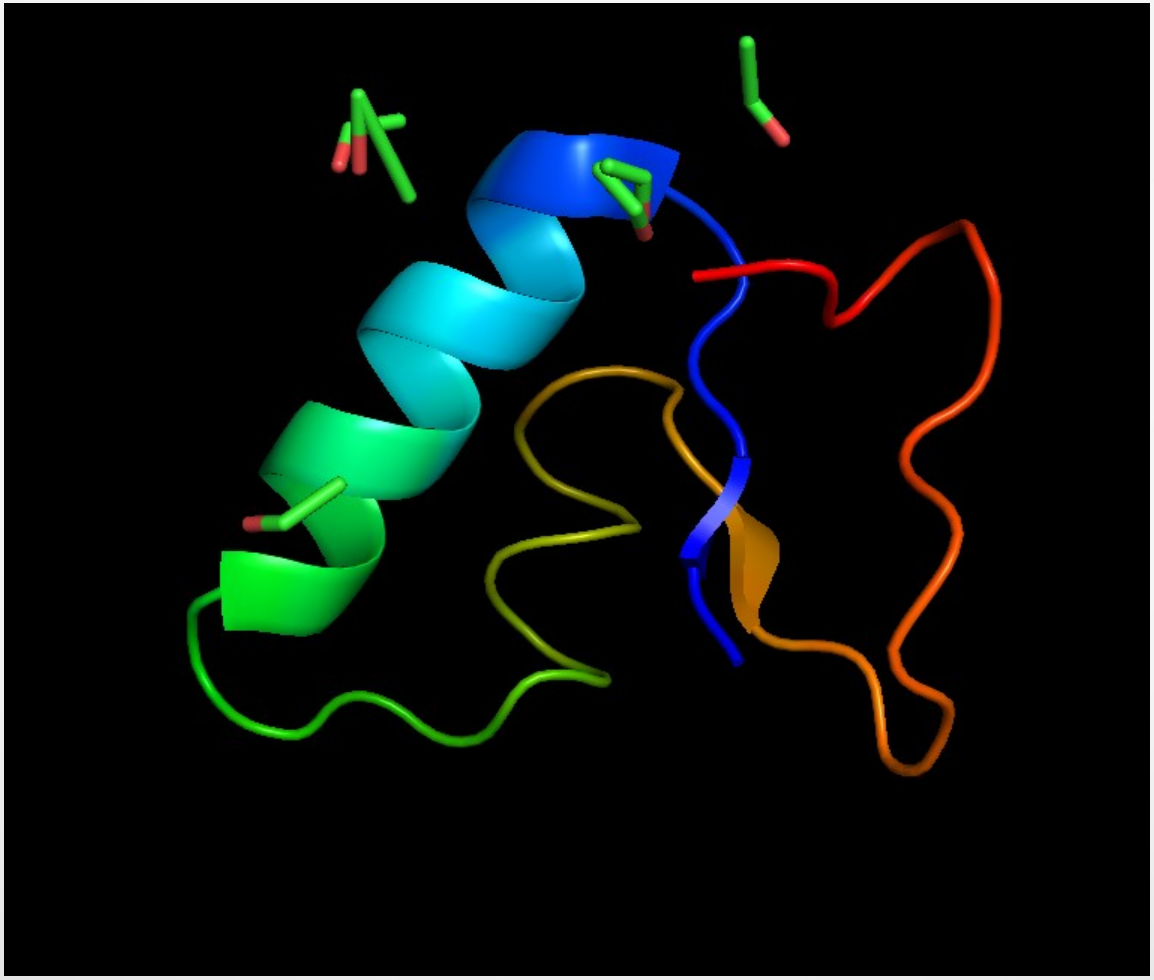
Visualizzazione di strutture

- Ball and Stick (A → preset → ball and stick)



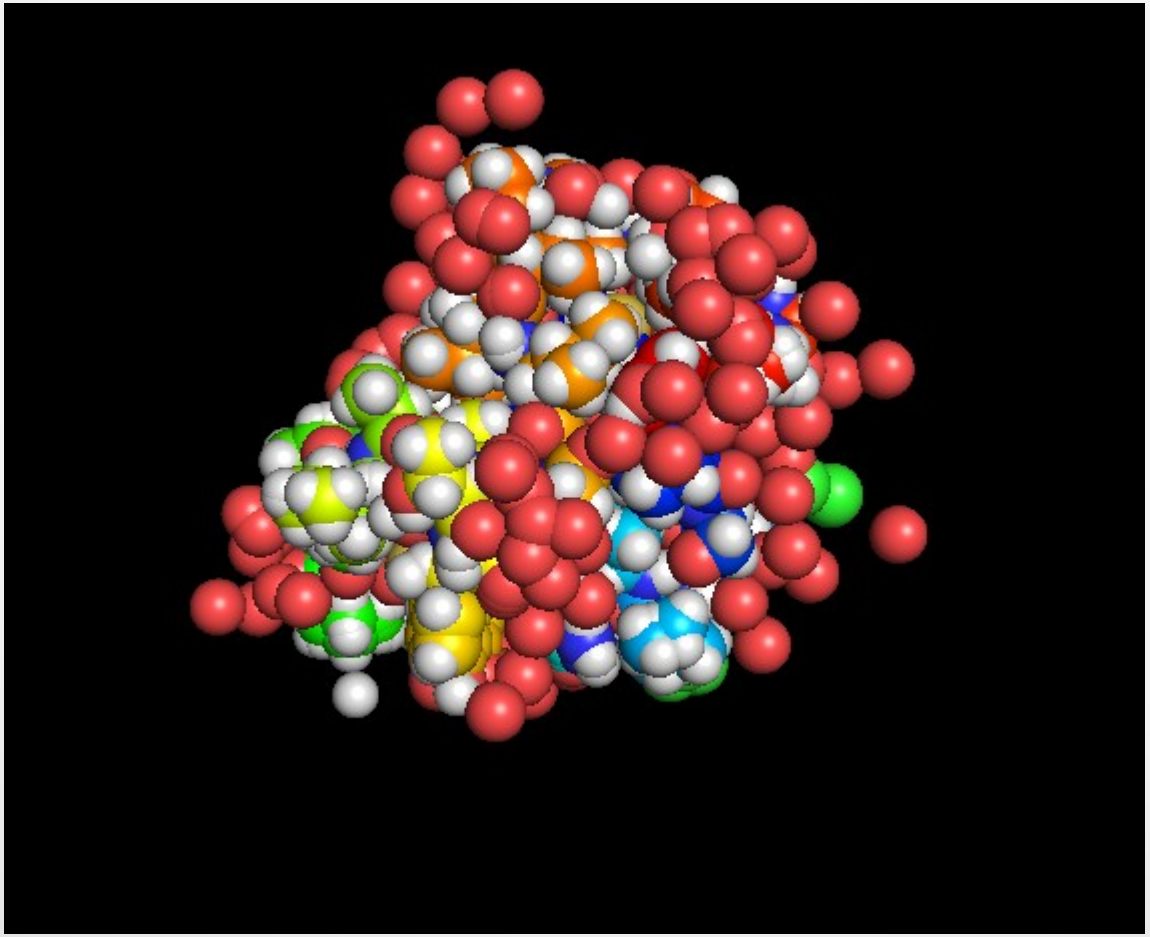
Visualizzazione di strutture

- Ribbon (A → preset → pretty)



Visualizzazione di strutture

- Ribbon (S → spheres)



- Convertiamo le due smiles in smiles.txt ad esempio in un file sdf e poi mol2. Aggiungendo opportunamente gli idrogeni e generando un 3D

==========