Docente: **Giorgio Valentini**
Tutor: **Matteo Re**

UNIVERSITÀ DEGLI STUDI DI MILANO

Insegnamento: **Bioinformatica**
A.A. 2011-2012 semestre II

C.d.l. **BIOTECNOLOGIE DEL FARMACO**

Cheminformatica in R

Giorgio Valentini
e –mail: *valentini@dsi.unimi.it*
http://homes.dsi.unimi.it/~valenti

Matteo Re
e –mail: *re@dsi.unimi.it*
http://homes.dsi.unimi.it/~re

DSI – Dipartimento di Scienze dell' Informazione
Università degli Studi di Milano

Cheminformatica

- **Cheminformatica** è una disciplina definita recentemente (**1998**) che si pone come obiettivo quello di integrare informazioni disponibili in banche dati pubbliche e riguardanti farmaci, molecole e processi patologici in modo da produrre nuove conoscenze che possano essere di supporto durante il processo di **sviluppo di nuovi farmaci**.

Cheminformatica

- In **R** esistono diverse librerie contenenti strumenti utilizzabili in esperimenti cheminformatici. In particolare noi utilizzeremo:

- rcdk (CRAN)

- ChemmineR (Bioconductor)

Cheminformatica

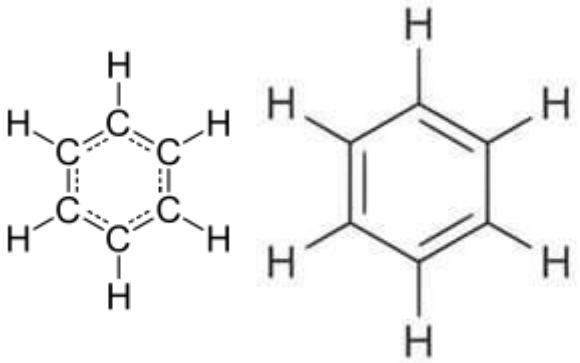
La realizzazione di una generica analisi cheminformatica pone alcuni problemi generali:

- **Rappresentazione** di molecole in modo che esse possano essere analizzate con un calcolatore.
- Definizione di nozioni di **similarità** tra molecole.
- Progettazione/implementazione di test che possano essere utilizzati per **predire un effetto della molecola su un sistema biologico**

1 Cheminformatica : rappresentazione di molecole

Molecola: «gruppo elettricamente neutro di atomi tenuti insieme da legami chimici di tipo covalente»

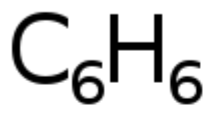
benzene:



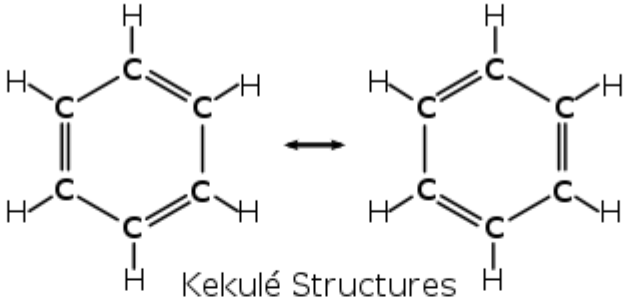
C_6H_6

Sono tutte rappresentazioni valide ma <u>incomprensibili</u> per un computer

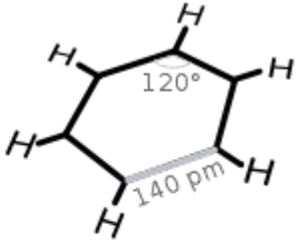
Cheminformatica : rappresentazione di molecole



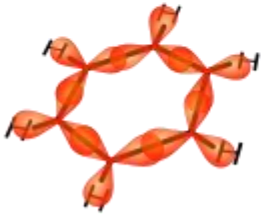
Benzene
Molecular formula



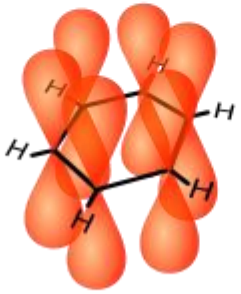
Kekulé Structures
(Isomers)



Planar Hexagon
Bond Length 140 pm

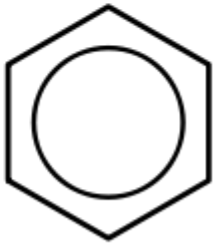


Sigma Bonds
sp^2Hybridized orbitals



6 p_z orbitals

delocalized pi
system



Benzene ring
Simplified depiction

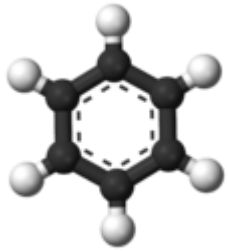
E' un **problema di codifica**. Ci sono molti modi di rappresentare la struttura di questa molecola. Dobbiamo trovarne uno che sia adatto ad essere «compreso» da una macchina.

Cheminformatica : rappresentazione di molecole

Sono stati proposti diversi formati di file adatti a rappresentare la struttura delle molecole. Alcuni permettono di rappresentare solo la struttura, altri permettono l'inclusione di informazioni aggiuntive.

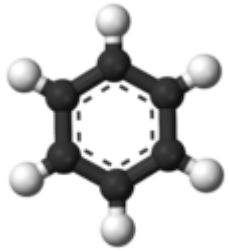
- MDL **MOL** (*.mol) : permette di codificare atomi, legami tra di essi, coordinate atomiche. Il file MOL (o molfile) contiene alcune righe di intestazione, la **Connection Table** (CT) contenente informazioni sugli atomi, una sezione dedicata ai **legami tra atomi** e sezioni aggiuntive adatte a contenere eventuali informazioni più complesse.
- Structure Data Format (**SDF**) è una estensione del formato MOL adatta a rappresentare informazioni aggiuntive più complesse e a gestire insiemi di molecole.
- Simplified Molecular Input Line Entry Specification (**SMILES**) rappresenta ogni molecola utilizzando una sola riga di testo

MOL file



```
1   benzene
2    ACD/Labs0812062058
3
4      6  6  0  0  0  0  0  0  0  0  1 V2000
5        1.9050   -0.7932    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
6        1.9050   -2.1232    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
7        0.7531   -0.1282    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
8        0.7531   -2.7882    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
9       -0.3987   -0.7932    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
10      -0.3987   -2.1232    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
11     2  1  1  0  0  0  0
12     3  1  2  0  0  0  0
13     4  2  2  0  0  0  0
14     5  3  1  0  0  0  0
15     6  4  1  0  0  0  0
16     6  5  2  0  0  0  0
17    M  END
18    $$$$
...
```

MOL file



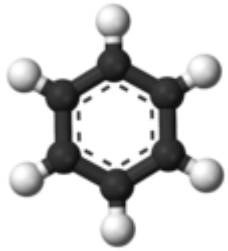
intestazione

Connection Table (CT)

```
1   benzene
2    ACD/Labs0812062058
3
4     6  6  0  0  0  0  0  0  0  0  1 V2000
5       1.9050   -0.7932    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
6       1.9050   -2.1232    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
7       0.7531   -0.1282    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
8       0.7531   -2.7882    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
9      -0.3987   -0.7932    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
10     -0.3987   -2.1232    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
11    2  1  1  0  0  0  0
12    3  1  2  0  0  0  0
13    4  2  2  0  0  0  0
14    5  3  1  0  0  0  0
15    6  4  1  0  0  0  0
16    6  5  2  0  0  0  0
17   M  END
18   $$$$
...
```

Conteggio: 6 atomi, 6 legami, … , standard V2000

MOL file



intestazione

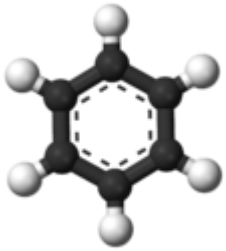
Connection Table (CT)

```
1   benzene
2    ACD/Labs0812062058
3
4      6  6  0  0  0  0  0  0  0  0  1 V2000
5       1.9050   -0.7932    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
6       1.9050   -2.1232    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
7       0.7531   -0.1282    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
8       0.7531   -2.7882    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
9      -0.3987   -0.7932    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
10     -0.3987   -2.1232    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
11     2  1  1  0  0  0  0
12     3  1  2  0  0  0  0
13     4  2  2  0  0  0  0
14     5  3  1  0  0  0  0
15     6  4  1  0  0  0  0
16     6  5  2  0  0  0  0
17    M  END
18    $$$$
...
```

Atomi e coordinate

Tipo di atomo

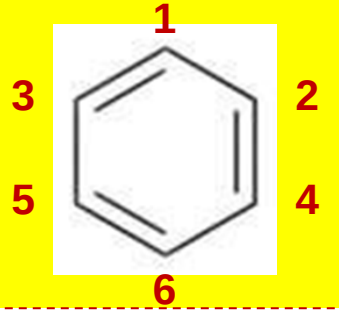
MOL file



intestazione

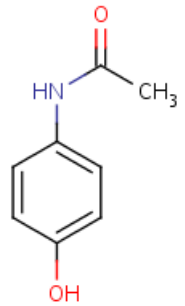
Connection Table (CT)

```
1   benzene
2    ACD/Labs0812062058
3
4     6  6  0  0  0  0  0  0  0  0  1 V2000
5       1.9050   -0.7932    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
6       1.9050   -2.1232    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
7       0.7531   -0.1282    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
8       0.7531   -2.7882    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
9      -0.3987   -0.7932    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
10     -0.3987   -2.1232    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
11    2  1  1  0  0  0  0
12    3  1  2  0  0  0  0
13    4  2  2  0  0  0  0
14    5  3  1  0  0  0  0
15    6  4  1  0  0  0  0
16    6  5  2  0  0  0  0
17  M  END
18  $$$$
...
```



Definizione **legami** (da, a, tipo,…)

MOL file




```
Marvin  03190821382D

11 11  0  0  0  0            999 V2000
    0.7145    0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.7145   -0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.8250    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000   -0.8250    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.7145   -0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.7145    0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.4288               0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.7145               0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.7145               0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    1.6500    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000   -1.6499    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
  2  1  1  0  0  0  0
  1  3  2  0  0  0  0
 10  3  1  0  0  0  0
  4  2  2  0  0  0  0
  4 11  1  0  0  0  0
  4  5  1  0  0  0  0
  6  5  2  0  0  0  0
  3  6  1  0  0  0  0
  7  8  1  0  0  0  0
  8  9  2  0  0  0  0
  8 10  1  0  0  0  0
M  END
```

Atom count

Coordinates

Elements

Line numbers of bonded atoms in above atom table

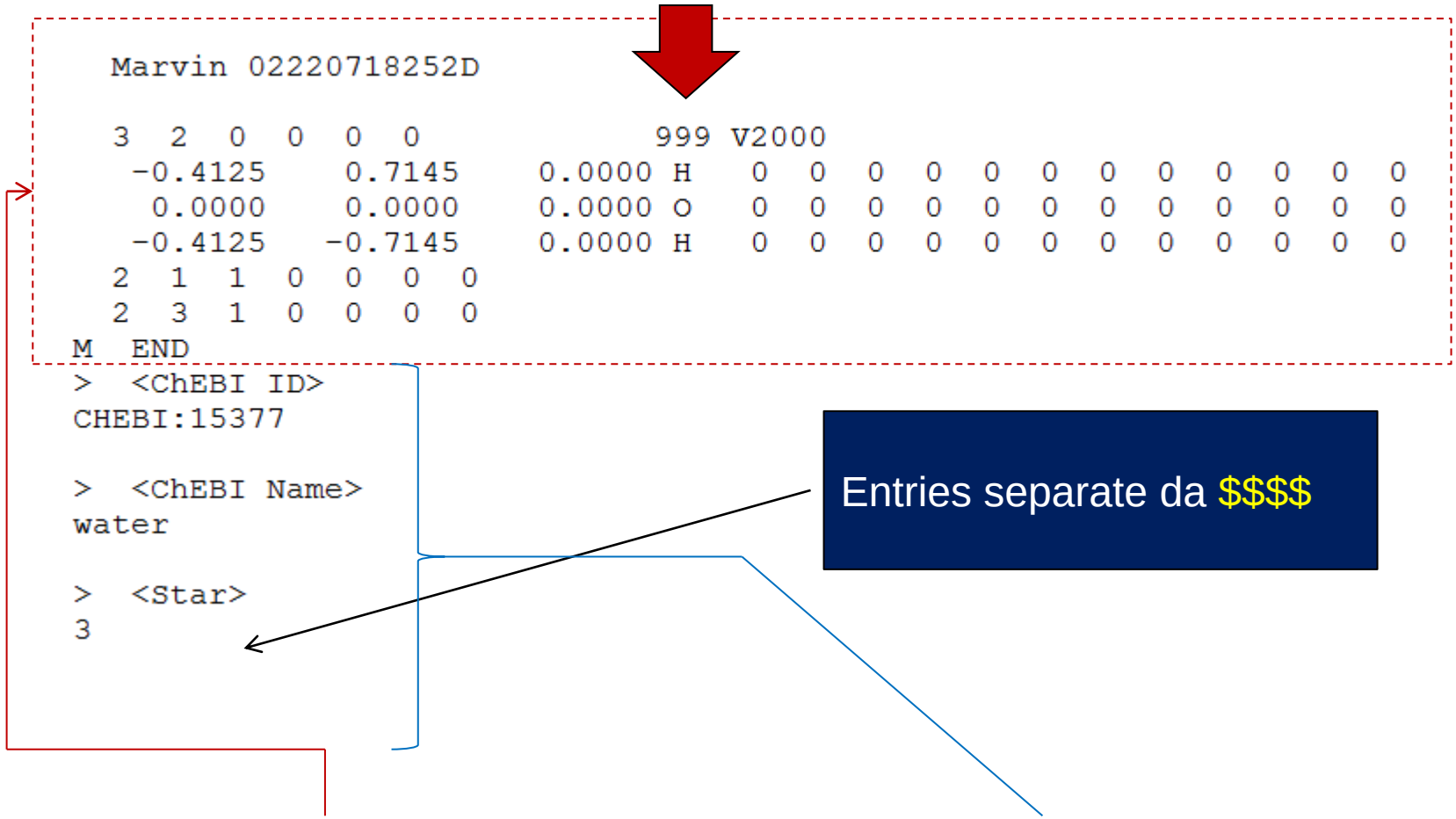
Bond order

Atoms

Bonds

SDF file

Che molecola è ?



SDF: **formato MOL** arricchito con **informazioni aggiuntive**

SDF file

```
  Marvin 02220718252D

  3  2  0  0  0  0            999 V2000
   -0.4125    0.7145    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.0000    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   -0.4125   -0.7145    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
  2  1  1  0  0  0  0
  2  3  1  0  0  0  0
M  END
>  <ChEBI ID>
CHEBI:15377

>  <ChEBI Name>
water

>  <Star>
3

>  <Secondary ChEBI ID>
CHEBI:10743
CHEBI:13352
...
CHEBI:44819

>  <SMILES>
[H]O[H]

>  <InChI>
InChI=1/H2O/h1H2

>  <InChIKey>
InChIKey=XLYOFNOQVPJJNP-UHFFFAOYAF

>  <Formulae>
H2O

>  <Charge>
0

>  <Mass>
18.01528

>  <IUPAC Names>
oxidane
water

>  <Synonyms>
H(2)O
```

Entries separate da $$$$

```
>  <Beilstein Registry Numbers>
3587155

>  <CAS Registry Numbers>
7732-18-5

>  <Gmelin Registry Numbers>
117

>  <ArrayExpress Database Links>
E-TOXM-12
E-TOXM-14

>  <BioModels Database Links>
BIOMD0000000090

>  <IntEnz Database Links>
EC 1.1.1.115

>  <IntEnz Database Links>
EC 1.1.1.115
...
EC 6.6.1.2

>  <KEGG COMPOUND Database
C00001

>  <MolBase Database Links>
1

>  <MSDchem Database Links>
HOH

>  <Patent Database Links>
EP0769531
...
WO2008157552

>  <PubChem Database Links>
8145132

>  <Reactome Database Links>
REACT_10000
...
REACT_9996
```

Link a banche dati pubbliche

SMILES file

Simplified Molecular Input Line Entry Specification (**SMILES**) è un sistema di codifica per la struttura delle molecole in grado di convertire una struttura chimica **in una stringa di testo** seguendo un set di regole predefinite. le stringhe SMILES contengono tipi di atomi e legami tra di essi ma non contengono coordinate 2D o 3D.

Gli atomi H non sono rappresentati. Altri atomi vengono raprpesentati mediante il loro simbolo chimico, ad es. B, C, N, O, F, P, S, Cl, Br, e I. Il simbolo "=" sappresenta il doppio legame ed il simbolo "#" rappresenta il triplo legame. Gruppi di atomi (es, CH3 per il metile) vengono racchiusi tra parentesi. I cicli sono espressi da coppie di numeri (ad es. la rappresentazione SMILES dell'anello del benzene inizia e finisce con **1**: C**1** ... **1** .

SMILES file

Esempi :

Nome	Formula	stringa SMILES
Metano	CH4	C
Acqua	H2O	O
Etanolo	C2H6O	CCO
Benzene	C6H6	C**1**=CC=CC=C**1** oppure c**1**ccccc**1**
Etilene	C2H4	C=C

Idrogeni <u>non</u> sono rappresentati

anello

SMILES file

```
DB00116      NC1=NC(=O)C2=C(NCC(CNC3=CC=C(C=C3)C(=O)N[C@@H](CCC(O)=O)C(O)=O)N2)N1
DB00117      N[C@@H](CC1=CN=CN1)C(O)=O
DB00118      C[S+](CC[C@H](N)C(O)=O)C[C@H]1O[C@H]([C@H](O)[C@@H]1O)N1C=NC2=C1N=CN=C2N
DB00119      CC(=O)C(O)=O
DB00120      N[C@@H](CC1=CC=CC=C1)C(O)=O
DB00121      [H][C@]12CS[C@@H](CCCCC(O)=O)[C@@]1([H])NC(=O)N2
DB00122      C[N+](C)(C)CCO
DB00123      NCCCC[C@H](N)C(O)=O
DB00125      N[C@@H](CCCNC(N)=N)C(O)=O
DB00126      [H][C@@]1(OC(=O)C(O)=C1O)[C@@H](O)CO
DB00127      NCCCNCCCCNCCCN
DB00128      N[C@@H](CC(O)=O)C(O)=O
DB00129      NCCC[C@H](N)C(O)=O
DB00130      N[C@@H](CCC(N)=O)C(O)=O
DB00131      NC1=NC=NC2=C1N=CN2[C@@H]1O[C@H](COP(O)(O)=O)[C@@H](O)[C@H]1O
DB00132      CC\C=C/C\C=C/C\C=C/CCCCCCCC(O)=O
DB00133      N[C@@H](CO)C(O)=O
DB00134      CSCC[C@H](N)C(O)=O
DB00135      N[C@@H](CC1=CC=C(O)C=C1)C(O)=O
DB00136      [H][C@@]1(CC[C@@]2([H])C(CCC[C@]12C)=CC=C1C[C@@H](O)C[C@H](O)C1=C)[C@H](C)CCCC(C)(C)O
DB00137      CC(\C=C\C=C(C)\C=C\C1C(C)=CC(O)CC1(C)C)=C/C=C/C=C(C)/C=C/C=C(C)/C=C/C1=C(C)CC(O)CC1(C)C
DB00138      N[C@@H](CSSC[C@H](N)C(O)=O)C(O)=O
DB00139      OC(=O)CCC(O)=O
DB00140      CC1=CC2=C(C=C1C)N(C[C@H](O)[C@H](O)[C@H](O)CO)C1=NC(=O)NC(=O)C1=N2
DB00141      CC(=O)N[C@H]1C(O)O[C@H](CO)[C@@H](O)[C@@H]1O
DB00142      N[C@@H](CCC(O)=O)C(O)=O
DB00143      N[C@@H](CCC(=O)N[C@@H](CS)C(=O)NCC(O)=O)C(O)=O
DB00144      CCCC(=O)O[C@H](COC(=O)CC)COP(O)(=O)OC[C@H](N)C(O)=O
DB00145      NCC(O)=O
...
```

ID molecola
(Drugbank ID)

stringa SMILES

Cheminformatica : librerie R

Ora proveremo ad importare alcune collezioni di molecole in R. Per riuscirci dovremo installare due package R : **rcdk e ChemmineR**.

Installazione:

rcdk :

```
> install.packages('rcdk', dependencies=TRUE)
```

ChemmineR :

```
> source("http://bioconductor.org/biocLite.R")
```

```
> biocLite("ChemmineR")
```

Cheminformatica : librerie R

rcdk è una collezione di funzioni R basate sul Chemistry Development Kit (CDK). La libreria CDK è scritta in Java. rcdk non fa altro che «tradurre» le chiamate R in chiamate comprensibili per CDK, aspetta che CDK completi le sue elaborazioni, prende l'output di CDK e lo restituisce all'utente sottoforma di variabili R.

Per avere più informazioni sulle funzioni disponibili:

```
>vignette('rcdk')        # tutorial
```

rcdk (e CDK in generale) permette di effettuare molte operazioni tra cui:

- Manipolazione di molecole (es. aggiungere annotazioni)
- I/O molecole
- Visualizzazione molecole
- Calcolo di descrittori molecolari
- Calcolo di fingerprints

Cheminformatica : librerie R

Per i nostri test noi utilizzeremo principalmente ChemminerR. Dopo aver installato ChemmineR carichiamo la libreria in R :

```
> library("ChemmineR")
```

Ora scaricate i file SDF e SMILES dalla pagina web del corso (sono nella sezione «File esercizi programmazione» e salvateli sul desktop. In R cambiate directory corrente (posizionandovi nel folder Desktop).

Proviamo a importare le molecole contenute nel file SDF:

```
> sdfset <- read.SDFset("approved.sdf")
```

NB: Se ChemmineR incontra problemi durante l'importazione (in questo esempio trova 4 molecole mal formattate) non le carica. Controllo consistenza.

Cheminformatica : librerie R

Cosa abbiamo importato ? E che tipo di variabile è sdfset ?

```
> length(sdfset)
[1] 1412
> str(sdfset)
```

Output più complesso. Ci informa che l'oggetto è di tipo «SDF» e che contiene 2 slot (due variabili). Le variabili si chiamano SDF e ID e contengono, rispettivamente, le molecole e gli identificativi. Per accedere agli slot si utilizza il simbolo @ :

```
> sdfset@SDF
> sdfset@ID
```

Gli slot sono delle LISTE.

Accesso alle singole molecole

```
> sdfset[[1]]
An instance of "SDF"

<<header>>
                          Molecule_Name                                    Source
                                     ""        "  Mrv0541 09201116592D          "
                                Comment                                Counts_Line
                                     "" "182193  0  0  1  0            999 V2000"

<<atomblock>>
           C1        C2  C3  C5  C6  C7  C8  C9 C10 C11 C12 C13 C14 C15 C16
C_1   -0.6472 -1.5655   0   0   0   0   0   0   0   0   0   0   0   0   0
N_2   -0.7591  -0.762   0   0   0   0   0   0   0   0   0   0   0   0   0
...       ...     ... ... ... ... ... ... ... ... ... ... ... ... ... ...
H_181 -4.1602 -7.1395   0   0   0   0   0   0   0   0   0   0   0   0   0
H_182 -2.7824 -6.6797   0   0   0   0   0   0   0   0   0   0   0   0   0

<<bondblock>>
     C1  C2  C3  C4  C5  C6  C7
1    46   1   2   0   0   0   0
2     1   2   1   0   0   0   0
... ... ... ... ... ... ... ...
192  90 181   1   0   0   0   0
193  90 182   1   0   0   0   0

<<datablock>> (18 data items)
```

Notare i **blocchi**: sono gli stessi del MOL esteso (SDF)

```
                DRUGBANK_ID
                  "DB00115"
                DRUG_GROUPS
"approved; nutraceutical"
                GENERIC_NAME
              "Cyanocobalamin"
                    SYNONYMS
```

Informazioni estese

nome generico molecola

Funzioni di calcolo delle proprietà delle molecole

I) Conteggio atomi (singola molecola)

```
> atomcount(sdfset[[1]])

 C Co  H  N  O  P
63  1 89 14 14  1
```

II) Peso molecolare, molecular weight (singola molecola)

```
> MW(sdfset[[1]])
     CMP
1356.373
```

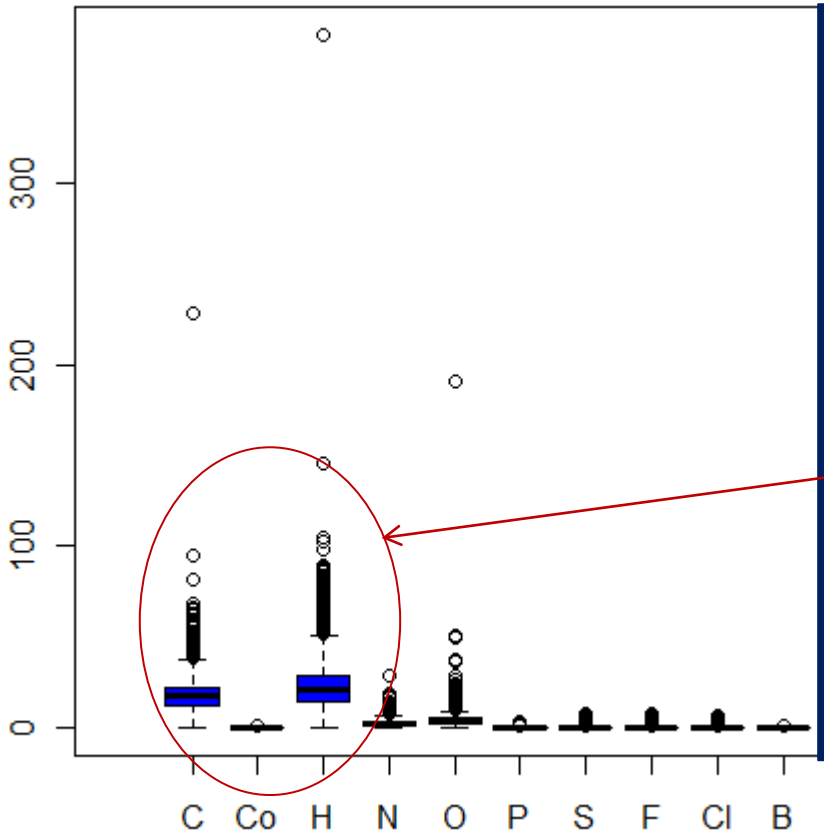
III) Molecular formula

```
> MF(sdfset[[1]])
               CMP
"C63H89CoN14O14P"
```

Funzioni di calcolo delle proprietà di SET di molecole

IV) grafico della frequenza delle specie atomiche (set di molecole)

```
> boxplot(atomcountMA(sdfset),col="blue",main="Atom Frequency")
```



- Conteggio di TUTTI gli atomi di di TUTTE le molecole in sdfset
- Ci sono specie atomiche più frequenti di altre: **C** e **H**

Matrice dei legami

V) sfdset[[3]] è L-istidina . Grazie alla funzione conMA (connection matrix) possiamo visualizzare i legami tra i suoi atomi in forma di matrice:

```
> conMA(sdfset[[3]], exclude=c("H")) # notare che escludo gli H
```

	O_1	O_2	N_3	N_4	N_5	C_6	C_7	C_8	C_9	C_10	C_11
O_1	0	0	0	0	0	0	0	0	1	0	0
O_2	0	0	0	0	0	0	0	0	2	0	0
N_3	0	0	0	0	0	0	0	1	0	0	1
N_4	0	0	0	0	0	0	1	0	0	0	0
N_5	0	0	0	0	0	0	0	0	0	1	2
C_6	0	0	0	0	0	0	1	1	0	0	0
C_7	0	0	0	1	0	1	0	0	1	0	0
C_8	0	0	1	0	0	1	0	0	0	2	0
C_9	1	2	0	0	0	0	1	0	0	0	0
C_10	0	0	0	0	1	0	0	2	0	0	0
C_11	0	0	1	0	2	0	0	0	0	0	0

O_1 e C_9 sono connessi da un singolo legame

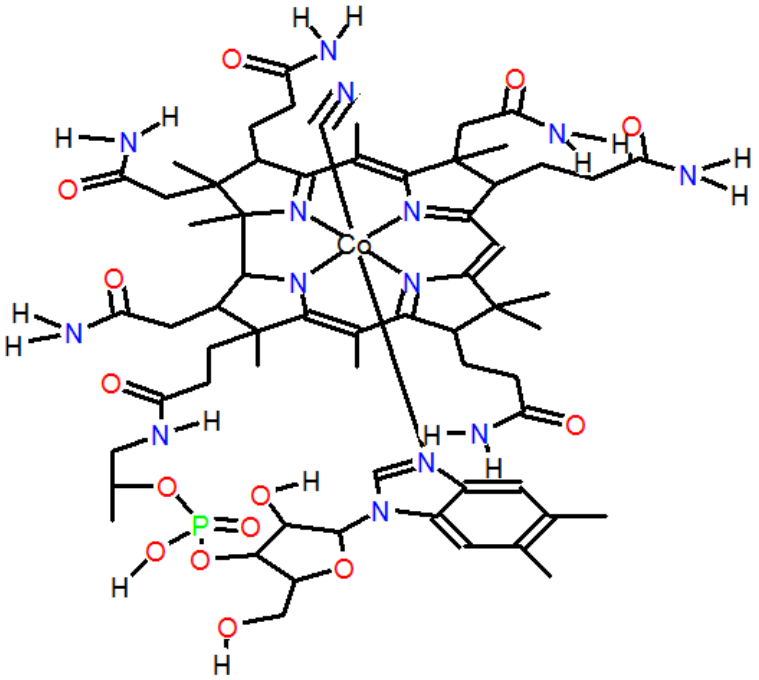
C_8 e C_10 sono connessi da un DOPPIO legame

la matrice è simmetrica

Visualizzazione di molecole

VI) Per visualizzare una molecola è possibile utilizzare la funzione plot

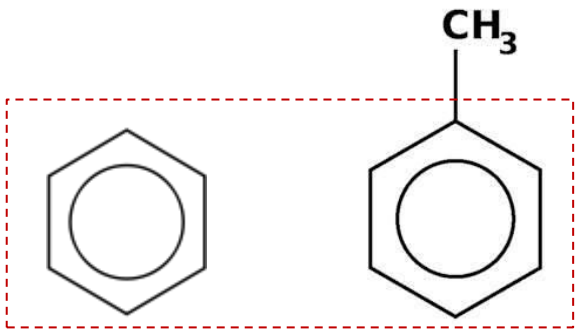
```
> plot(sdfset[[1]])
```



(R stampa informazioni aggiuntive riguardanti la molecola nella console)

2 Cheminformatica : confronto tra molecole

Ora sappiamo caricare e manipolare set di molecole in R. Ma dobbiamo riuscire calcolare in modo automatico quanto sono «simili» due molecole.

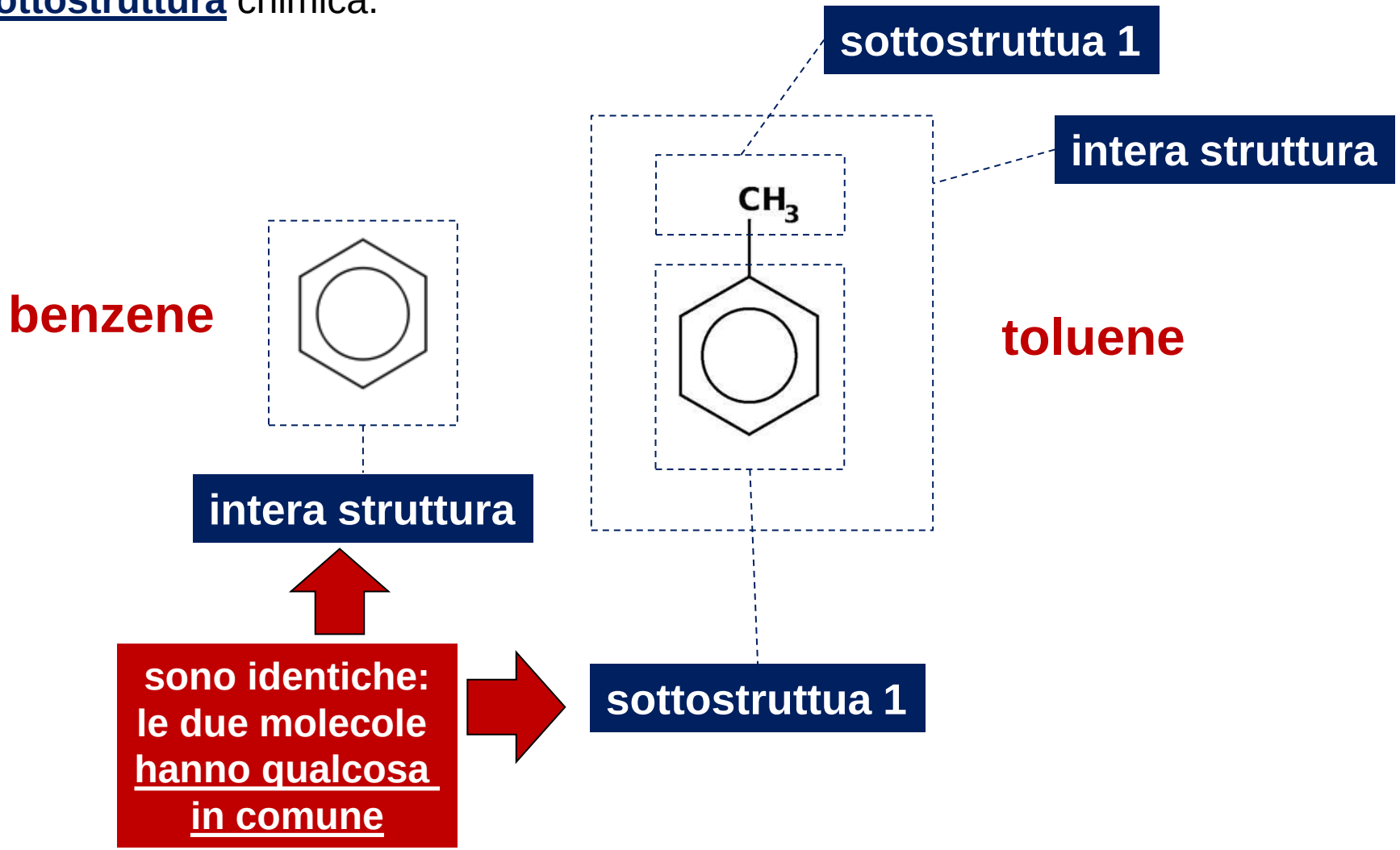


benzene toluene

«A occhio» le loro strutture sono simili … ma «a occhio» per un calcolatore non ha senso

Il concetto di fingerprint molecolare

Le fingerprint molecolari sono strettamente correlate ai concetti di **struttura** e **sottostruttura** chimica.

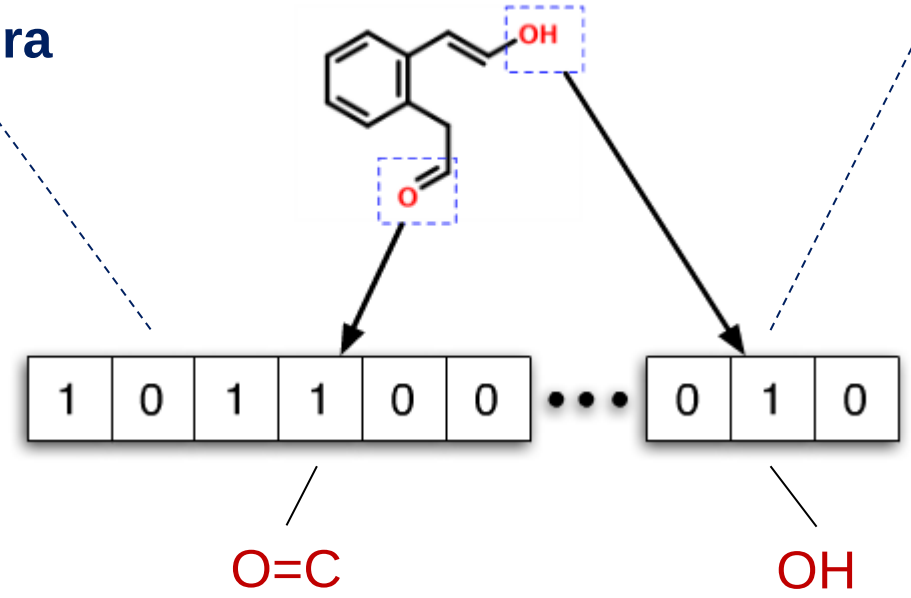


Il concetto di fingerprint molecolare

Le fingerprint molecolari sono strettamente correlate ai concetti di **struttura** e **sottostruttura** chimica.

0 : sottostruttura assente nella molecola

1 : sottostruttura presente nella molecola



Codifica **vettoriale** (vettore valori logici) : comprensibile per il calcolatore

Confronto di molecole mediante fingerprint

Una volta codificata la struttura delle molecole mediante fingerprint il loro confronto è relativamente semplice. L'obiettivo è quello di capire quante sottostrutture sono condivise da due molecole ed utilizzare questa informazione per calcolare uno score di similarità.

Date die fingerprint I e II che codificano le sottostrutture di due molecole che vogliamo confrontare definiamo:

a : numero di sottostrutture presente in fingerprint I **MA NON** in fingerprint II
b : numero di sottostrutture pesente in fingerprint I **E** in fingerprint II
c : numero di sottostrutture presente in fingerprint II **MA NON** in fingerprint I

Coefficiente di TANIMOTO :

$$\frac{b}{a + b + c}$$

Numero compreso tra 0 e 1 . 1 indica identità strutturale, 0 indica diversità totale delle strutture.

Calcolo di fingerprint e similarità

rcdk rende molto semplice il calcolo delle fingerprint molecolari.

```
> dtapp <- read.table("approved.sdf.SMILES", as.is=T);
> mols <- parse.smiles(dtapp[,2]);
> fps <- lapply(mols,get.fingerprint, type="extended");
> fp.sim <- fp.sim.matrix(fps,method="tanimoto");
> str(fp.sim)
 num [1:1408, 1:1408] 1 0.2519 0.1231 0.3683 0.0373 ...

> rownames(fp.sim)<-dtapp[,1];
> colnames(fp.sim)<-dtapp[,1];

> fp.sim[1:5,1:5]
           DB00115    DB00116   DB00117    DB00118    DB00119
DB00115 1.00000000 0.25192802 0.1231231 0.36828645 0.03728814
DB00116 0.25192802 1.00000000 0.1752137 0.21111111 0.05612245
DB00117 0.12312312 0.17521368 1.0000000 0.15357143 0.13253012
DB00118 0.36828645 0.21111111 0.1535714 1.00000000 0.04938272
DB00119 0.03728814 0.05612245 0.1325301 0.04938272 1.00000000
```

Calcolo di fingerprint e similarità

Le similarità che abbiamo visto finora sono molto basse, a parte i valori sulla diagonale (sapreste dire perchè?). Proviamo a selezionare le similarità tra alcune molecole selezionate mediante **identificativo drugbank**:

```
> drugset <- c("DB00417","DB01053","DB01163","DB01603","DB00895")
> fp.sim[drugset,drugset]
          DB00417   DB01053   DB01163   DB01603   DB00895
DB00417 1.0000000 0.8109453 0.6977778 0.7808219 0.4930876
DB01053 0.8109453 1.0000000 0.7090909 0.7695853 0.6307692
DB01163 0.6977778 0.7090909 1.0000000 0.6475410 0.4867257
DB01603 0.7808219 0.7695853 0.6475410 1.0000000 0.4763948
DB00895 0.4930876 0.6307692 0.4867257 0.4763948 1.0000000
```

Queste molecole sono strutturalmente molto simili (hanno score Tanimoto molto più alti di quelli della slide precedente). Come mai? Provate a indagare su **Drugbank**, una banca dati pubblica dedicata ai farmaci:

http://www.drugbank.ca/

Cosa hanno in comune queste molecole?