Software disegno strutture e nomenclatura:

ACDLabs ChemSketch
http://www.acdlabs.com/resources/freeware/chemsketch/
Solo per Windows e Linux

Chemaxon MarvinSketch
http://www.chemaxon.com/products/marvin/marvinsketch/
Windows, Linux, MacOSX

Freeware per università, usare email univr per registrarsi

Indicizzazione molecole, notazioni di struttura, database molecolari

➢CAS: Simplified molecular-input line-entry system

E' un identificativo numerico che individua in maniera univoca una sostanza chimica. Il Chemical Abstracts Service, una divisione della American Chemical Society, assegna questi identificativi a ogni sostanza chimica descritta in letteratura. Attualmente oltre 63 milioni di composti hanno ricevuto un numero CAS e circa 7.000 vengono aggiunti ogni giorno. La maggior parte dei database molecolari permettono al giorno d'oggi di fare ricerche in base al numero CAS.

Il numero CAS è costituito da tre sequenze di numeri separati da trattini. Il primo gruppo è costituito da un numero variabile di cifre, fino a sei, il secondo da due cifre, mentre il terzo e ultimo gruppo è costituito da una singola cifra che serve da codice di controllo. I numeri sono assegnati in ordine progressivo e non hanno quindi nessun significato chimico. Il codice di controllo viene calcolato moltiplicando ciascuna cifra da destra a sinistra per un numero intero progressivo (la cifra più a destra va moltiplicata per 1, quella immediatamente a sinistra per 2 e così via), sommando i vari termini e calcolando poi il modulo 10 della somma così ottenuta. Per esempio il numero CAS dell'acqua è 7732-18-5 e il codice di controllo (5) è dato da (8×1 + 1×2 + 2×3 +3×4 + 7×5 + 7 ×6) mod 10 = 105 mod 10 = 5.

Se una molecola ha più isomeri a ciascun isomero sarà assegnato un numero CAS differente. Per esempio il numero CAS del D-glucosio è 50-99-7 mentre quello del L-glucosio è 921-60-8. In alcuni casi particolari a una intera classe di composti è stato assegnato un unico numero CAS; per esempio tutte le alcol deidrogenasi hanno come numero CAS 9031-72-5.

➢ SMILES: Simplified molecular-input line-entry system

E' una specifica in forma di una notazione lineare per descrivere la struttura di molecole chimiche con brevi stringhe ASCII. Stringhe SMILES possono essere importate dalla maggior parte degli editor di molecole per la riconversione in rappresentazioni bidimensionali o modelli tridimensionali delle molecole.

In termini di procedure computazionali basate sui grafi, SMILES è una stringa ottenuta per stampa dei simboli dei nodi presenti sul grafo che rappresenta la formula di struttura. Dal grafo vengono prima rimossi gli atomi di idrogeno, quindi i cicli vengono aperti per convertire il grafo in un albero aperto. Dove i cicli sono stati aperti, vengono aggiunti dei suffissi numerici per indicare quali sono i nodi connessi. Le ramificazioni dell'albero sono indicate attraverso l'uso di parentesi.

Gli atomi sono rappresentati utilizzando il loro simbolo chimico chiuso tra parentesi quadre, come [Au] per oro. L'anione idrossido è [OH-]. Le parentesi quadre possono essere omesse per gli atomi "organici" C, N, O, P, S, Br, Cl e I. Tutti gli altri elementi devono essere racchiusi tra parentesi quadre. Se si omettono le parentesi quadre, si presume che il numero degli atomi di idrogeno sia implicito; per esempio lo SMILES per l'acqua è semplicemente O e per l'etanolo è CCO.

Il doppio legame del biossido di carbonio è rappresentato come O=C=O e il triplo legame dell'acido cianidrico come C#N.

Il cicloesano è rappresentato come C1CCCCC1, l'idea è che i due uno indicano la stessa posizione nella molecola, formando così un anello con sei atomi di carbonio. Da notare che è il numerale (in questo caso 1) che rappresenta la posizione piuttosto che la combinazione "C1". Ecco la notazione espansa per chiarire: (C1)-(C)-(C)-(C)-(C)-(C)-1 piuttosto che (C1)-(C)-(C)-(C)-(C)-(C)-(C1).

Gli atomi di C, O, S e N aromatici vengono rappresentati con i loro caratteri minuscoli, rispettivamente 'c', 'o', 's' e 'n'.

Le ramificazioni sono rappresentate da parentesi tonde, ad esempio CCC(=O)O per l'acido propionico e C(F)(F)F per il fluoroformio, che potrebbe anche essere descritto con la formula non canonica: FC(F)F.

La configurazione del carbonio tetraedrico è specificato da @ o @@. L-alanina, l'enantiomero più comune dell'amminoacido alanina può essere scritta come N[C@@H](C)C(=O)O. L'identificatore @@ indica che, quando visti dall'azoto lungo il legame al centro chirale, la sequenza di sostituenti idrogeno (H), metile (C) e carbossilato (C(=O)O) appaiono in senso orario. D-alanina può essere scritta come N[C@H](C)C(=O)O

➢ InChI: Interational Chemical Identifier

E' un identificatore testuale per le sostanze chimiche, progettato per fornire un modo standard e leggibile per codificare le informazioni molecolari e per facilitare la ricerca di tali informazioni nei database e sul web.

Gli identificatori descrivono le sostanze chimiche in termini di livelli di informazione - gli atomi e il loro legame di connettività, le informazioni tautomeriche, informazioni isotopiche, la stereochimica e le informazioni di carica elettronica. Non tutti gli strati devono essere forniti, per esempio, lo strato tautomero può essere omesso se questo tipo di informazioni non è rilevante per la particolare applicazione.

InChIs differiscono dagli ampiamente utilizzati numeri di registro CAS in tre aspetti:

sono liberamente utilizzabili e non di proprietà;
possono essere calcolati dalle informazioni strutturali e non devono essere assegnati da qualche organizzazione;
maggior parte delle informazioni in un InChI è leggibile (con la pratica).

InChIs possono esprimere più informazioni rispetto alla semplice notazione SMILES e si differenziano nel fatto che ogni struttura ha una stringa univoca InChI che è importante nelle applicazioni di database.

L'algoritmo InChI converte le informazioni strutturali di input in un identificatore univoco InChI in un processo in tre fasi: la normalizzazione (per rimuovere le informazioni ridondanti), canonicizzazione (per generare un unico numero per ogni atomo), e serializzazione (per dare una stringa di caratteri) .

InChIKey è una rappresentazione digitale condensata di lunghezza fissa (25 caratteri) del InChI che non è comprensibile dall'uomo. La specifica InChIKey è stata rilasciata nel settembre 2007 al fine di agevolare le ricerche sul Web per i composti chimici, problematici da identificare con il full-length InChI.

Ogni InChI inizia con la stringa "InChI =" seguito dal numero di versione, attualmente 1. Questo è seguito dalla lettera S per InChIs standard. L'informazione rimanente è strutturata come una sequenza di strati e sotto-strati, con ciascuno strato che fornisce un tipo specifico di informazioni. Gli strati e sotto-strati vengono separati dal delimitatore "/" e iniziano con una lettera prefisso caratteristica (eccetto per la formula chimica sub-strato dello strato principale). I sei strati con sottolivelli importanti sono:

1 strato principale
- Formula chimica (senza prefisso). Questa è l'unico sottolivello che deve esserci in ogni InChI.
- Connessioni tra atomi (prefisso: "c"). Gli atomi nella formula chimica (eccetto gli idrogeni) sono numerati in sequenza, questo sottolivello descrive quali atomi sono collegati da legami con altri.
- Gli atomi di idrogeno (prefisso: "h"). Descrive quanti atomi di idrogeno sono collegati a ciascuno degli altri atomi.

2 livello di carica
- sottolivello di carica positiva (prefisso: "p" di "protoni")
- sottolivello di carica negativa (prefisso: "q")

3 strato di stereochimica
- doppi legami (prefisso: "b")
- stereochimica tetraedrica (prefissi: "t", "m")
- tipo di informazione stereochimica (prefisso: "s")

4 Strato isotopico (prefissi: "i", "h", nonché "b", "t", "m", "s" per la stereochimica isotopica)
- strato H-fissi
- strato ricollegato

Il formato delimitatore-prefisso ha il vantaggio che un utente può facilmente utilizzare una ricerca con caratteri jolly per trovare identificatori che corrispondono solo in certi livelli.

➢ MDL Molfile

E' un formato di file creato dalla MDL Information Systems e adesso di proprietà della Symyx Technologies, per la gestione dell'informazione riguardante gli atomi, legami, connettività e coordinate di una molecola. Molfile è costituito da una intestazione, una tabella di connessione (Ctab, Connection Table) contenente informazioni sugli atomi, quindi sui legami e sulla tipologia di legame, seguita da sezioni che forniscono una informazione più completa.

Molfile è abbastanza comune per la maggior parte, se non tutte, le applicazioni software di chemoinformatica.

```
benzene
ACD/Labs0812062058

  6  6  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
$$$$
```

Lines	Section	Description
1-3	Header	
1		Molecule name ("benzene")
2		User/Program/Date/etc information
3		Comment (blank)
4-17	Connection table (Ctab)	
4		Counts line: 6 atoms, 6 bonds, ..., V2000 standard
5-10		Atom block (1 line for each atom): x, y, z, element, etc
11-16		Bond block (1 line for each bond): 1st atom, 2nd atom, type, etc
17		Properties block (empty)
18	$$$$	See note

Database molecolari

pubchem
http://pubchem.ncbi.nlm.nih.gov

e-molecules
http://www.emolecules.com