Laboratorio di Algoritmi

Progetto "Molecole" (novembre 2018)

Nota: La scadenza del progetto è fissata per mercoledì 14 novembre **compreso**.

Nota: Si consiglia di consultare sulla pagina web il documento che riporta le avvertenze utili per lo svolgimento del progetto. In particolare, si consiglia di verificare di tanto in tanto gli aggiornamenti al testo del progetto, che riporteranno la correzione di eventuali errori e le risposte ai dubbi degli studenti.

Il problema Una delle tecniche più fruttuose nello studio dei farmaci è la simulazione, svolta al calcolatore in base alle leggi della Meccanica e dell'Elettrodinamica, del comportamento delle molecole investigate in un solvente (in genere, acqua). I dati di tale simulazione sono le posizioni iniziali delle molecole sotto studio. I risultati sono invece le posizioni successive da loro assunte nel tempo: si può cioè pensare di disporre, per ogni istante di tempo, di una lista di atomi, per ognuno dei quali è noto l'elemento chimico, la molecola di appartenenza (che non cambia nel tempo: le interazioni sono fisiche, non chimiche) e la posizione (assumendo di modellare l'atomo come un punto nello spazio a tre dimensioni). L'analisi di questi risultati, nei limiti entro i quali la simulazione è fedele, consente di indagare in modo qualitativo e ad alto livello il comportamento delle molecole stesse, come se si disponesse di un microscopio ad altissima risoluzione che le riprendesse, per così dire, nel loro "ambiente naturale".

Uno degli aspetti interessanti di tale analisi è il gioco costituito dal continuo rompersi e riformarsi dei *legami a idrogeno*. Ogni atomo di idrogeno del sistema appartiene a una molecola ed è in essa legato con un legame chimico covalente a un atomo *donore* di un altro elemento. L'atomo di idrogeno, però, può anche stabilire un debole legame temporaneo con un atomo *accettore*, anch'esso di un elemento chimico diverso, appartenente ad un'altra molecola. Nei casi più comuni, donore e accettore sono atomi di ossigeno, ma possono essere anche atomi di azoto, zolfo, ecc... La causa del legame è che l'unico elettrone dell'idrogeno tende a trovarsi non più genericamente intorno al nucleo dell'idrogeno, ma spostato in direzione dell'atomo donore, il che tende a creare una debole (e fluttuante) carica positiva nella direzione opposta, la quale a sua volta attira gli elettroni dell'atomo accettore. A questo punto, il nucleo dell'idrogeno risulta legato sia al donore (dal forte legame covalente, di natura chimica) sia all'accettore (dal debole legame a idrogeno, di natura elettrostatica). I legami a idrogeno sono molto più deboli rispetto a quelli chimici, e si rompono e riformano di continuo. Nonostante ciò, determinano molte proprietà fisiche importanti dell'acqua, delle proteine e degli acidi nucleici. Per convenzione, si ritiene che si stabilisca un legame a idrogeno quando

1. donore e accettore sono vicini, cioè la distanza fra loro è inferiore a una data soglia $d_m = 3.5$ Å;
2. la terna donore-idrogeno-accettore forma una linea quasi retta, cioè l'angolo formato da essa supera una data soglia $\alpha_m = 150°$.

La descrizione fine del sistema attraverso le coordinate di tutti gli atomi racchiude tutte le informazioni utili sulle interazioni fra le molecole del soluto e quelle del solvente, ma in modo troppo implicito. Si vogliono rendere esplicite alcune di tali informazioni estraendole dai risultati grezzi della simulazione. In particolare, si ipotizza che la topologia dei legami a idrogeno, cioè le coppie di molecole interagenti, forniscano tali informazioni in modo più leggibile. In questo progetto ci concentreremo su:

1. il rapporto fra il numero di legami a idrogeno e il numero di atomi di idrogeno disponibili;
2. l'esistenza e le caratteristiche dei *cluster* di molecole legate fra loro direttamente o indirettamente da legami a idrogeno;
3. l'esistenza e le caratteristiche delle sequenze di legami che connettono fra loro atomi lontani;
4. l'esistenza e le caratteristiche di "scheletri" di legami a idrogeno che tengono insieme i *cluster*.

Ovviamente, le semplificazioni che faremo renderanno lo studio piuttosto un pretesto per un esercizio di algoritmica. Altri temi, che non indagheremo in questo progetto sono:

- le interazioni fra molecole di soluto piuttosto che fra molecole di soluto e di solvente (cioè quanto le molecole di soluto tendono ad aggregarsi o a disperdersi nel solvente);
- la distribuzione spaziale del soluto nel solvente;
- l'esistenza e le caratteristiche delle "sfere di influenza" (ovvero *shell*), cioè delle catene di molecole di solvente che risultano abbastanza stabilmente legate al soluto e ne propagano gli effetti;
- l'analisi comparata di *framework* successivi, per valutare la durata dei legami a idrogeno e delle strutture che essi inducono.

Il progetto È possibile costruire un grafo ausiliario non orientato $G(V, E)$, i cui vertici corrispondono alle molecole del sistema. Si noti che i vertici non corrispondono agli atomi, ma alle molecole, perché lo studio investiga l'influenza fisica dei legami a idrogeno, che riguarda le molecole, e non i singoli atomi. I lati del grafo corrispondono invece alle coppie di molecole legate da legami a idrogeno, e non ai singoli legami. Questo per semplificare la trattazione, ed evitare di dover considerare un multigrafo, potenzialmente con più lati fra gli stessi due vertici (è infatti possibile in generale che si stabiliscano più legami a idrogeno fra le stesse due molecole).

Il progetto richiede la stesura di un programma che legga da un file l'immagine di una soluzione acquosa in un determinato istante di tempo. Il file rispetta il formato PDB (Protein Data Bank), secondo il quale ogni riga descrive le caratteristiche di un atomo. Ogni riga è introdotta dalla parola chiave `ATOM`, seguita da[1]:

- l'indice numerico dell'atomo nel file;
- il suo elemento chimico[2];
- una stringa di caratteri che identifica la sostanza chimica della molecola;
- l'indice numerico della molecola nel file;
- le coordinate rispetto agli assi x, y e z dell'atomo, in Ångström[3].

[1] Il formato PDB prevede altre informazioni, che omettiamo per semplicità.

[2] In realtà, nello standard PDB qui compare una stringa di caratteri che identifica l'atomo entro la molecola di cui fa parte, mentre l'elemento chimico comparirebbe più avanti, ma stiamo semplificando.

[3] Per semplicità, omettiamo le informazioni seguenti richieste dal formato PDB.

Esempio Quindi, per esempio:

```
...
ATOM 32 H URE 4 13.800 14.055 14.470
ATOM 33 O WAT 5 11.596 7.495 20.457 ...
```

indica che il trentaduesimo atomo è di idrogeno, appartiene alla quarta molecola, che è di urea, e ha coordinate $(13.800, 14.055, 14.470)$, mentre il trentatreesimo atomo è di ossigeno, appartiene alla quinta molecola, che è di acqua, e ha coordinate $(11.596, 7.495, 20.457)$.

Quindi, il programma deve ricostruire un grafo, nel quale

- i vertici sono le molecole;
- i lati sono le coppie di molecole legate da almeno un legame a idrogeno.

Si assuma che un legame a idrogeno sussista per ogni terna di atomi (d, h, a) dove:

- d è un atomo non di idrogeno;
- h è un atomo di idrogeno della stessa molecola di d;
- a è un atomo non di idrogeno di una molecola diversa da quella di d e h;
- la distanza fra d e a non supera $d_m = 3.5$ Å;
- l'angolo $\hat{dha}$ non è inferiore ad $\alpha_m = 150°$.

A questo punto, il programma deve stampare il numero l dei legami a idrogeno e il numero n degli atomi di idrogeno, nel formato:

`Legami:` l `su` n atomi

Si calcoli quindi il grado di ogni molecola, cioè il numero di altre molecole alle quali essa è legata da legami a idrogeno (N.B.: non il numero dei legami!) e si stampi il numero di molecole che hanno ciascun grado, in ordine crescente da zero al grado massimo, secondo il formato:

`Grado 0:` n_0 `molecole`
`Grado 1:` n_1 `molecole`
`...`

Dato il grafo, si costruiscano i *cluster* di molecole legate da legami a idrogeno, direttamente o indirettamente. Per ciascuno, si calcoli la cardinalità, cioè il numero di molecole in esso contenute. Quindi, si stampino i *cluster*, riportandone cardinalità e molecole componenti. I *cluster* vanno stampati in ordine di cardinalità decrescente e (a pari cardinalità) di indice minimo crescente, dove con indice minimo si intende l'indice numericamente più basso fra le molecole componenti. Infine, le molecole di ogni *cluster* vanno ordinate per indici crescenti. Il formato deve essere il seguente:

`Cluster 1:` $n^{(1)}$ `molecole -` $m_1^{(1)}$ $m_2^{(1)}$ $m_3^{(1)}$ `...`
`Cluster 2:` $n^{(2)}$ `molecole -` $m_1^{(2)}$ $m_2^{(2)}$ $m_3^{(2)}$ `...`
`...`

Si calcoli quindi il diametro del primo *cluster* (quello di cardinalità massima). Con diametro si intenda la massima distanza fra due molecole del *cluster*. Con distanza fra due molecole si intenda il numero di lati del cammino più breve (rispetto, appunto, al numero dei lati) che abbia le due molecole come estremi. Si stampi quindi il diametro nel formato seguente:

`Diametro:` δ

Infine, si attribuisca ad ogni lato un costo, determinato dalla distanza in Å fra donore e accettore del corrispondente legame a idrogeno. Se vi sono più legami fra le stesse molecole, si consideri come costo del lato la distanza minima. Si calcoli quindi il costo dell'albero ricoprente minimo per il primo *cluster* e si riporti il risultato nel formato seguente, indicando anche il numero di lati che formano l'albero ricoprente stesso:

`Costo MST:` ℓ `(` m `legami)`

Esempio Si consideri il seguente esempio, costituito da cinque molecole di acqua, cioè 15 atomi:

```
ATOM 1 O WAT 1 20.807 4.147 17.051
ATOM 2 H WAT 1 21.623 3.888 16.622
ATOM 3 H WAT 1 20.190 4.281 16.331
ATOM 4 O WAT 2 20.382 0.571 16.989
ATOM 5 H WAT 2 20.902 -0.232 16.955
ATOM 6 H WAT 2 20.841 1.127 17.619
ATOM 7 O WAT 3 21.115 2.271 19.201
ATOM 8 H WAT 3 21.528 2.692 19.955
ATOM 9 H WAT 3 21.095 2.951 18.528
ATOM 10 O WAT 4 24.474 22.754 20.464
ATOM 11 H WAT 4 25.093 22.226 19.962
ATOM 12 H WAT 4 25.024 23.296 21.030
ATOM 13 O WAT 5 23.357 21.041 22.507
ATOM 14 H WAT 5 22.483 21.426 22.567
ATOM 15 H WAT 5 23.741 21.441 21.726
```

L'analisi delle posizioni spaziali rivela che i dieci atomi di idrogeno sono impegnati in tre legami a idrogeno. Delle cinque molecole, una ha due legami, mentre le altre ne hanno uno solo a testa. Le molecole formano due *cluster*, rispettivamente di cardinalità 3 e 2. Il primo *cluster* ha diametro 2 e l'albero minimo che lo ricopre è formato da due lati, con un costo complessivo di 5.754 Å. Il risultato è quindi il seguente:

```
Legami:  3 su 10 atomi
Grado 0:  0 molecole
Grado 1:  4 molecole
Grado 2:  1 molecole
Cluster 1:  3 molecole - 1 2 3
Cluster 2:  2 molecole - 4 5
Diametro:  2
Costo MST: 5.754 (2 legami)
```

Chiarimenti

In questa sezione saranno riportate le risposte a domande e dubbi.