Laboratorio di Calcolo per Fisici,
Prova d'esame dell'8/07/2024, A.A. 2023/2024

Nome _____ Cognome _____

Matricola _____ ☐ Ritirato/a

Lo scopo di questa esercitazione è di scrivere un programma in C e uno script in python seguendo la traccia riportata di seguito. Si tenga presente che:

1. Per svolgere il compito si hanno a disposizione 3 ore.
2. Si possono usare libri di testo, prontuari e gli appunti ma non è ammesso parlare con nessuno né utilizzare cellulari, tablet o laptop, pena l'annullamento del compito.
3. **Tutti i file vanno salvati in una cartella chiamata `ELCG24_NOME_COGNOME` nella home directory**, dove `NOME` e `COGNOME` indicano rispettivamente il tuo nome e cognome. Ad esempio lo studente *Marco Rossi* deve creare una cartella chiamata `ELCG24_MARCO_ROSSI` contenente tutti i file specificati nel testo. **Tutto ciò che non si trova all'interno della cartella suddetta non verrà valutato.** In tutti i programmi e script inserisci all'inizio un commento con il tuo nome, cognome e numero di matricola.
4. **Dovete consegnare il presente testo indicando nome, cognome e numero di matricola** (vedi sopra), barrando la casella "Ritirato/a" se ci si vuole ritirare, ovvero se non si vuole che la presente prova venga valutata.
5. **Per consegnare il compito** dovrete eseguire, all'interno della cartella creata in precedenza (come spiegato al punto 4), il seguente comando da terminale: `cp * /media/sf_esame/`

Il modello *Hydrophobic-Polar* (HP) per le proteine è un modello altamente semplificato per studiare come le proteine si ripiegano e acquisiscono la loro funzione biologica (*protein folding*). Nella versione semplificata che consideriamo, una proteina è modellizzata come una sequenza di N_A amminoacidi disposti su di un reticolo bidimensionale. Gli amminoacidi possono essere di due tipi: H (idrofobici) e P (polari). In questa prova considereremo proteine di forma quadrata, per cui $N_A = M \times M$, dove M è un numero intero che indica la dimensione del quadrato, e indicheremo come *sequenza* i-esima una specifica configurazione di caselle H e P sul reticolo quadrato. Data la forma quadrata e la sequenza i-esima, è possibile calcolare per ciascuna proteina un'energia $E_i = -N_C$, dove N_C è il numero di "contatti idrofobici", definito come il numero di coppie di amminoacidi H che sono vicini spazialmente in orizzontale o in verticale (ma **non** in diagonale).

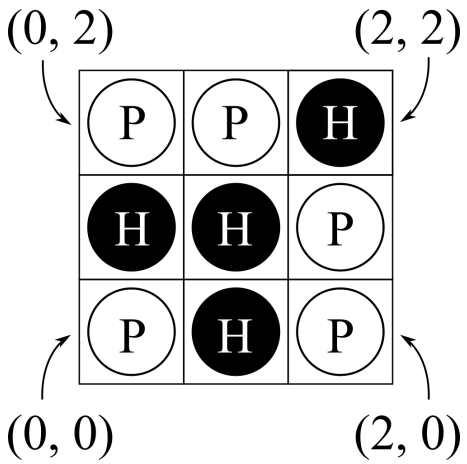

Figura 1: Un esempio di proteina su un reticolo quadrato di dimensioni 3×3. La sequenza della proteina, letta nell'ordine richiesto dal testo nella pagina seguente, cioè dal basso verso l'alto e da sinistra a destra, è `PHPHHPPPH`. I contatti idrofobici sono due (`(0, 1)` con `(1, 1)` e `(1, 1)` con `(1, 0)`), quindi l'energia vale $E = -2$.

In questa prova dovete generare proteine di forma quadrata con N_S sequenze diverse generate casualmente: ogni amminoacido sarà H con probabilità q o P con probabilità $(1-q)$, indipendentemente dal tipo degli altri amminoacidi. Lo scopo della prova è di calcolare l'energia mediata sulle N_S sequenze:

$$\langle E \rangle = \frac{1}{N_S} \sum_{i=1}^{N_S} E_i.$$

Si scriva un programma chiamato `nome_cognome.c` (tutto minuscolo, senza eventuali spazi, accenti o apostrofi) che calcoli l'energia di proteine quadrate $M \times M$, mediata su N_S sequenze, per diversi valori della probabilità q che ogni amminoacido sia idrofobico. Per fare ciò il codice deve

1. Chiedere all'utente di inserire un valore per $N_S \in [10, 10^5]$;
2. Definire un array bidimensionale di `char` di dimensione $M \times M$ (con $M = 5$ da definirsi tramite una macro opportuna) che conterrà la proteina;
3. Inizializzare la proteina con una sequenza casuale di amminoacidi H o P, dove la probabilità che un amminoacido sia H è data dal valore di q. Il primo amminoacido generato andrà a occupare la casella (0,0) in basso a sinistra, il secondo la casella immediatamente superiore, fino a riempire tutta la prima colonna, per poi passare alla seconda colonna – si veda la figura.
4. Scorrere l'array e contare le coppie N_C di vicini idrofobici H per calcolare l'energia $E_i = -N_C$. **Suggerimento:** testate il calcolo dell'energia su una sequenza di cui conoscete già E_i: ad esempio, nel caso $q = 1.0$ e $M = 5$ l'energia dovrebbe essere pari a -40, mentre nel caso $q = 0$ l'energia dovrebbe essere nulla;
5. Ripetere il procedimento N_S volte per ottenere il valore medio dell'energia $\langle E \rangle$, per i seguenti valori di q: 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 e 1.0;
6. Per ogni valore di q stampare su schermo q, la sequenza e l'energia della prima proteina generata. Stampate la sequenza scorrendo prima le colonne e poi le righe, come mostrato nella figura 1. Ad esempio, se la proteina in figura fosse la prima generata a $q = 0.4$ e $M = 3$, il codice dovrebbe stampare:

   ```
   0.4  PHPHHPPPH -2.00
   ```

7. Salvare su un file chiamato `protein.dat` i risultati. Ogni riga del file dovrà contenere due numeri: q e $\langle E \rangle$. Il primo va stampato con una cifra dopo la virgola, il secondo con due.
8. Quando si è certi del funzionamento del programma, con uno script python `nome_cognome.py`, creare un grafico che mostri l'andamento di $\langle E \rangle$ in funzione di q. Infine, salvare un'immagine di tale grafico in un file chiamato `protein.png`.

Nello scrivere il programma si richiede che vengano implementate le seguenti funzioni:

- `input()` che chieda all'utente il valore di $N_S \in [10, 10^5]$, reiterando la richiesta qualora il numero inserito non sia nell'intervallo.
- `fill_sequence()` che prenda come argomento il valore di q e l'array bidimensionale della proteina e riempia quest'ultimo con una sequenza casuale di amminoacidi.
- `energy()` che prenda come argomento l'array della proteina e restituisca la sua energia. Esistono metodi diversi per calcolare l'energia, uno di questi è spiegato nel suggerimento seguente.
  Suggerimento: All'inizio della funzione potete inizializzare un array `coord_diff` di dimensioni 4×2 che contiene le differenze di coordinate tra un amminoacido e i suoi quattro vicini, $(0, \pm 1)$ e $(\pm 1, 0)$. L'energia associata a un amminoacido H si può quindi calcolare facendo un ciclo su `coord_diff` per ottenere le coordinate dei quattro vicini e controllare quanti di questi sono anch'essi H: l'energia è la metà del numero di coppie trovate in questo modo, invertito di segno.
  Attenzione: gli amminoacidi sulla "superficie" della proteina, cioè sulle caselle al bordo, hanno meno di quattro vicini: aggiungete delle condizioni opportune per gestire questi casi.