Laboratorio di Algoritmi e Strutture Dati

Docente: V. Lonati

Progetto "Frammenti di DNA"

valido per gli appelli di gennaio e febbraio 2014

1 Il problema

Il professor Sapevatelo, scienziato di chiara fama, sta sviluppando importanti esperimenti su brevi frammenti di DNA. Un frammento di DNA è costituito da una sequenza di nucleotidi (Adenina, Timina, Citosina e Guanina) e può quindi essere rappresentato come una parola sull'alfabeto delle lettere $\{A, T, C, G\}$. Ad esempio $AATCTGATCGA$ è il frammento formato dalla sequenza dei nucleotidi Adenina, Adenina, Timina, Citosina, Timina, Guanina, e così via. Il Professore considera soltanto frammenti di piccole dimensioni, ovvero formati da all'incirca 50 al massimo. Indichiamo con $\mathcal{F}$ l'insieme di tali frammenti.

Esperimenti Il professor Sapevatelo ha sviluppato in particolare il seguente esperimento: inserisce in una provetta alcuni frammenti di DNA, poi aggiunge alcuni enzimi che possono provocare la trasformazione di questi frammenti.

Indichiamo con $\mathcal{E}$ l'insieme degli enzimi a disposizione dal Professore.

Una volta inserito nella provetta, ogni enzima può essere inibito o ri-attivato con un'opportuna preparazione del professor Sapevatelo.

Ogni enzima di $\mathcal{E}$ può attaccarsi ad un frammento di DNA e provocarne la trasformazione; se su un frammento è attaccato già un enzima, non è possibile per altri enzimi attaccarsi. Finita la trasformazione, l'enzima si stacca e può attaccarsi ad un altro frammento (o allo stesso). Non è detto che la presenza di un enzima determini necessariamente la trasformazione di un frammento: possono esserci frammenti nella provetta ai quali non si attacca nessun enzima.

Esempio 1. *Ecco alcuni esempi di enzimi che sta studiando il Professore:*

`Rovesciasi`*: rovescia l'ordine dei frammenti, ad esempio $AATC$ diventa $CTAA$.*

`Troncasi-`*n, per n pari ad un numero intero compreso tra 1 e 3: rimuove da un frammento di lunghezza almeno n la porzione formata dagli ultimi n nucleotidi, lasciando invariati i frammenti più corti. Ad esempio* `Troncasi-`*3 trasforma il frammento $GATTACA$ in $GATT$.*

`Insertasi-`*nN, per n pari ad un numero intero compreso tra 1 e 3: se il frammento ha almeno n nucleotidi, inserisce il nucleotide N dopo i primi n. Ad esempio* `Insertasi-`*$2C$ trasforma il frammento $GATTACA$ in $GACTTACA$.*

A partire dai frammenti $GATTACA$, $TAGGAT$, $CCCCTAAG$ e AC, aggiungendo gli enzimi `Troncasi-3`*,* `Rovesciasi` *e* `Insertasi-`*$2C$ si possono ottenere i seguenti frammenti*

- *$ACATTAG$ ottenuta da $GATTACA$ mediante l'enzima* `Rovesciasi`
- *ACC ottenuta da AC mediante l'enzima* `Insertasi-`*$2C$*
- *CA ottenuta da AC mediante l'enzima* `Rovesciasi`
- *$CCCCCTAAG$ ottenuta da $CCCCTAAG$ mediante l'enzima* `Insertasi-`*$2C$*

- *CCCCT ottenuta da CCCCTAAG mediante l'enzima* `Troncasi-`*3*
- *GAATCCCC ottenuta da CCCCTAAG mediante l'enzima* `Rovesciasi`
- *GACTTACA ottenuta da GATTACA mediante l'enzima* `Insertasi-`*2C*
- *GATT ottenuta da GATTACA mediante l'enzima* `Troncasi-`*3*
- *TACGGAT ottenuta da TAGGAT mediante l'enzima* `Insertasi-`*2C*
- *TAG ottenuta da TAGGAT mediante l'enzima* `Troncasi-`*3*
- *TAGGAT ottenuta da TAGGAT mediante l'enzima* `Rovesciasi`

Si noti che l'applicazione dell'enzima `Troncasi-`*3 al frammento AC non produce alcuna trasformazione.*

Durata degli esperimenti Le trasformazioni dei frammenti di DNA ad opera degli enzimi non avvengono istantaneamente ma richiedono un certo periodo di tempo Δ. Maggiore è la durata di un esperimento, maggiore è il numero delle trasformazioni che possono avvenire.

Esempio 2. *Consideriamo i frammenti e gli enzimi dell'Esempio 1. Dopo un esperimento di durata* 3Δ *sarà possibile trovare nella provetta ad esempio il frammento ACATTCCAG ottenuto dalle tre trasformazioni successive provocate, nell'ordine, dagli enzimi* `Insertasi-`*2C, di nuovo* `Insertasi-`*2C e infine* `Rovesciasi`*. Tale frammento non si sarebbe potuto produrre con un esperimento di durata* Δ*.*

Notate che se un frammento non è trasformato nel primo periodo di tempo Δ, potrebbe essere trasformato da un enzima nei periodi di tempo sucessivi.

Il Professor Sapevatelo si premura di inserire nella provetta di ogni esperimento molti duplicati di ciascun frammento di DNA usato, nonché una buona quantità di ogni enzima usato: in questo modo, anche dopo parecchio tempo dall'inizio dell'esperimento, nella provetta ci saranno ancora duplicati dei frammenti originali e dei frammenti trasformati.

Energia Le trasformazioni provocate dagli enzimi si verificano soltanto se si fornisce energia riscaldando la provetta: ogni enzima richiede una quantità specifica di energia per innescare il processo di trasformazione del frammento e naturalmente ci sono enzimi più energivori di altri. L'energia richiesta è indicata in ***`UEA`*** (*untià elementari di energia*). In caso di trasformazioni successive di un frammento di DNA ad opera di diversi enzimi, è necessario fornire energia pari alla somma dell'energia richiesta da ciascuna trasformazione.

Esempio 3. *Consideriamo i frammenti e gli enzimi dell'Esempio 1 e assumiamo che* `Rovesciasi` *richieda 5* `UEA`*,* `Troncasi-`*3 richieda 9* `UEAInsertasi-`*2C richieda 2* `UEA`*.*

Allora la trasformazione di GATTACA in TTACA mediante l'applicazione successiva degli enzimi `insertasi-`*2C,* `rovesciasi`*,* `troncasi-`*3 e di nuovo* `rovesciasi` *richiede complessivamente 21* `UEA`*.*

Consideriamo ora anche l'enzima `Prefissasi-`*n, che consumando* $4n$ `UEA` *riesce ad eliminare i primi n nucelotidi di un frammento di lunghezza almeno n, lasciando invariati i frammenti più corti. Allora il frammento TTACA si ottiene da GATTACA anche direttamente tramite l'enzima* `Prefissasi-`*2 con un consumo di sole* 8 `UEA`*.*

Similarità di frammenti Come abbiamo visto, per preparare i suoi esperimenti, il Professor Sapevatelo ha spesso bisogno di duplicare frammenti di DNA, ovvero generarne copie identiche a partire da una soluzione contenente nucleotidi liberi. Alcune volte però il processo di duplicazione non funziona perfettamente e le copie dei frammenti risultano contenere degli errori: ad esempio possono mancare dei nucleotidi, oppure esserci dei nucleotidi in più, oppure alcuni nucleotidi possono essere cambiati di posto.

Una maniera per stabilire se due frammenti sono duplicati imprecisi dello stesso frammento è misurare quanto sono simili.

Il Professor Sapevatelo ha brevettato un sistema per misurare la "similarità" tra due frammenti di DNA. E' sufficiente cospargere con una speciale mistura enzimatica i due frammenti: questa mistura enzimatica innanzitutto consente che il frammento cui è applicata si possa espandere creando degli spazi tra un nucleotide e il successivo; inoltre fa sì che i due frammenti da confrontare si allineino in modo che due nucleotidi uguali risultino appaiati e che il numero di tali coppie sia il massimo possibile. La *similarità* tra i due frammenti è data dal numero totale di coppie di nucleotidi che risultano appaiate.

Esempio 4. *Consideriamo i frammenti $GATTACA$ e $TACCA$. La similarità tra i due frammenti è pari a 4. Il sistema brevettato da Sapevatelo porta i due frammenti a posizionarsi ad esempio nel modo seguente:*

$$\begin{array}{cccccccc} G & A & T & T & A & C & & A \\ & & T & & A & C & C & A \end{array}$$

2 Specifiche di progettazione

Si deve progettare un programma che sia in grado di gestire i dati relativi ai frammenti di DNA e alle trasformazione tramite l'uso di enzimi.

Non potete fare assunzioni sul numero dei frammenti usati negli esperimenti e sulla quantità di enzimi a disposizione del Professor Sapevatelo.

Potete invece immaginare di avere a disposizione una libreria con le informazioni relative a tutti gli enzimi di $\mathcal{E}$. In particolare questa libreria implementa le trasformazioni provocate dagli enzimi, tramite la funzione

`char *enzima(char *nome_enzima, char *frammento_src)`

che restituisce il frammento ottenuto da `frammento_src` mediante l'enzima `nome_enzima`; inoltre la libreria mette a disposizione la funzione

`int energia_enzima(char *nome_enzima)`

che restituisce la quantità di energia, in *UEA*, necessaria all'enzima `nome_enzima` per provocare la trasformazione di un frammento.

La progettazione deve prevedere la scelta delle strutture dati da usare per rappresentare i dati e gli algoritmi da applicare per risolvere in maniera efficiente i problemi descritti nella traccia. Non basta limitarsi a riferimenti generici alle tecniche algoritmiche utilizzate (es: "l'operazione X si risolve con un algoritmo greedy") ma è necessario dettagliare le procedure da utilizzare, tramite pseudocodice o direttamente in linguaggio C, eventualmente facendo riferimento alla letteratura sugli algoritmi citati.

In particolare si richiede di analizzare, in funzione delle scelte di progettazione fatte, quale risulta essere il costo delle diverse operazioni richieste dalla specifica.

Si richiede inoltre di fornire una rassegna *esauriente* di esempi che potrebbero essere usati per testare il programma e che mettono in evidenza particolari caratteristiche del suo funzionamento (non solo casi tipici di input, ma anche casi limite e/o situazioni patologiche; input che evidenzino la differenza di prestazioni tra le soluzioni progettuali scelte e altre più semplicistiche).

Non si richiede un'implementazione completa del progetto; è necessario però fornire l'ossatura del programma (possibilmente suddiviso su più file) contenente in particolare: le definizioni dei tipi fondamentali, i prototipi delle funzioni che realizzano le operazioni specificate nella traccia, e tutte le porzioni di codice utili ad illustrarne il loro funzionamento e uso.

Facoltativamente, è possibile consegnare un'implementazione completa e funzionante del progetto o di alcune sue parti. In questo caso è necessario commentare il codice e fornire nella relazione indicazioni precise sul formato dell'input atteso e sul formato dell'output prodotto, assieme ad alcune coppie di file di input e relativo output da usare per testare il codice.

2.1 Operazioni da eseguire

Si noti che le operazioni richieste sono liberamente implementabili; in particolare, non vanno necessariamente intese come prototipi di funzioni.

- **nuovo_esperimento**()

  Prepara un nuovo esperimento: nella provetta non ci sono frammenti né enzimi.

- **nuovo_frammento**(f)

  Aggiunge alla provetta dell'esperimento un numero molto ampio di duplicati del frammento f costituito da una sequenza di nucleotidi.

- **aggiungi_enzima**(e)

  Se l'enzima $e \in \mathcal{E}$ non è presente nella provetta lo aggiunge, se è già presente ma è stato precedentemente inibito, lo riattiva.

- **elimina_enzima**(e)

  Inibisce l'enzima $e \in \mathcal{E}$.

- **prepara_esperimento**($frammenti\_file, enzimi\_file$)

  Legge dal file $frammenti\_file$ un elenco di frammenti di DNA e dal file $enzimi\_file$ un elenco di enzimi, quindi simula la preparazione di un esperimento con una provetta contenente questi frammenti e questi enzimi (attivati).

- **esperimento**($tempo$)

  Simula l'esecuzione di un esperimento, di durata pari a $tempo \cdot \Delta$, a partire dai frammenti di DNA e degli enzimi attualmente presenti e attivi nella provetta, e stampa l'elenco dei frammenti ottenibili con questo esperimento.

  Se $tempo$ è pari a 0, allora stampa l'elenco dei frammenti attualmente nella provetta.

- **enzimi**($f, g, tempo$)

  Stampa la più breve sequenza di enzimi di $\mathcal{E}$ da aggiungere alla provetta per ottenere il frammento g a partire dal frammento f in un esperimento di durata pari a $tempo \cdot \Delta$.

- **energia**($f, g, \mathcal{E}', tempo$)

  Calcola l'energia necessaria per ottenere il frammento g a partire dal frammento f, usando gli enzimi in $\mathcal{E}' \subset \mathcal{E}$, in un esperimento di durata pari a $tempo \cdot \Delta$. Stampa inoltre la sequenza di enzimi di $\mathcal{E}'$ da aggiungere alla provetta per ottenere questa trasformazione.

- **similarità**(f, g)

  Calcola la similarità tra i frammenti f e g e stampa, in ordine, la sequenza dei nucleotidi che risultano appaiati secondo il sistema brevettato dal Professor Sapevatelo (se più di un posizionamento è possibile, basta stamparne uno). Ad esempio, nel caso dell'esempio 4 si otterrà la sequenza $TACA$.

3 Modalità di consegna

La presente traccia è valida per gli appelli di gennaio e febbraio 2014.

La relazione (non meno di 3, non più di 10 pagine in formato pdf o rtf) va inviata per posta elettronica all'indirizzo `lonati@dsi.unimi.it` entro lunedì 17 febbraio.

Per coloro che intendono sostenere la prova orale nel mese di gennaio, la scadenza è anticipata a martedì 21 gennaio 2014.

La relazione e gli altri file aggiuntivi (file sorgenti C, esempi di input, ecc) devono essere contenuti in un unico archivio `.zip` il cui nome dovrà essere della forma `cognome_matricola.zip`. Tutti i file nell'archivio, compresa la relazione, devono riportare nome, cognome e matricola dell'autore.

In generale non è prevista una discussione orale dei progetti, ma in alcuni casi potranno essere richiesti dei chiarimenti via mail o dal vivo all'autore delle relazione.

La realizzazione del progetto è una prova d'esame da svolgersi **individualmente**. I progetti giudicati frutto di **copiatura** saranno **estromessi** d'ufficio dalla valutazione.

La versione aggiornata del progetto è pubblicata in `.pdf` sul sito: `http://lonati.dsi.unimi.it/algo/`. Si consiglia di consultare periodicamente questo sito per eventuali correzioni e/o precisazioni relative al testo del progetto. Per ogni ulteriore chiarimento potete chiedere un appuntamento scrivendo una mail all'indirizzo `lonati@dsi.unimi.it` .