Algoritmi di Sequenziamento

Zsuzsanna Lipták

Università di Verona
Dipartimento di Informatica
7 marzo 2018

Che cos'è la bioinformatica?

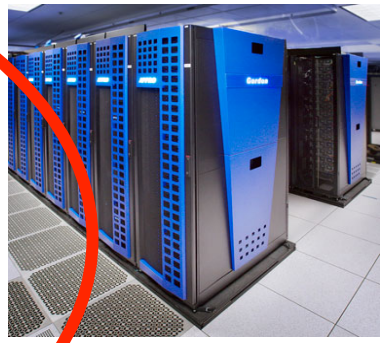
Per **bioinformatica** si intende l'uso delle tecniche matematiche e informatiche per risolvere problemi provenienti dalla biologia, tipicamente creando o usando programmi, modelli matematici o entrambi.

(Wikipedia)

dati biologici

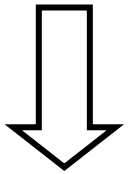
+

il bioinformatico



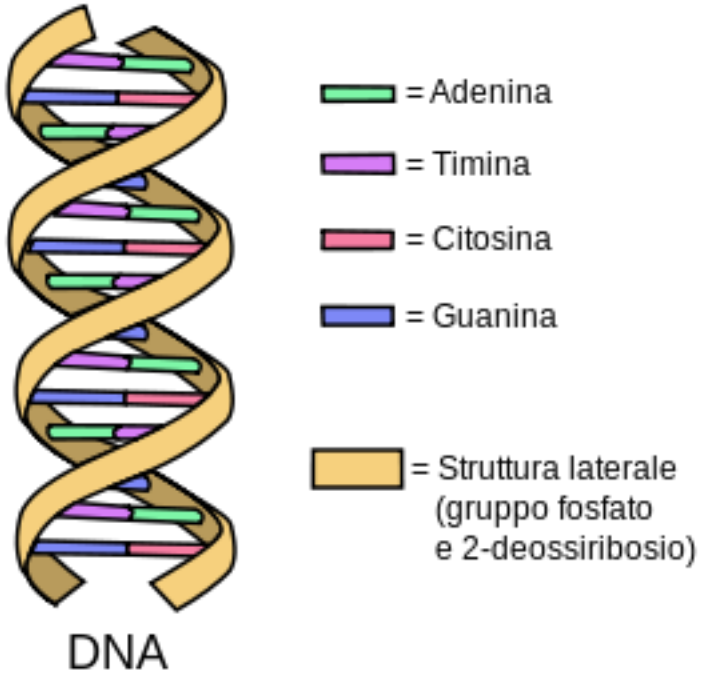
computer

Un esempio:
Il sequenziamento del genoma umano



...AACAGTACCATGCTAGGTCAATCGAG...
...TTGTCATGGTACGATCCAGTTAGCTC...

Un po‘ di biologia molecolare



DNA

- 4 lettere: **A C T G** (basi)
- **A–T** e **C–G** complementari
- direzione (5′ → 3′)
- due filamenti, uno complementare inverso dell‘altro *(reverse complement)*: **(ACCTG)^{rc} = CAGGT**
- quindi sufficiente conoscerne uno

5′...ACCATGCTAGGTCAATCGAG...3′
3′...TGGTACGATCCAGTTAGCTC...5′

Vincoli delle techniche biochimiche

- Il problema del **sequenziamento** del DNA è quello di determinare la sequenza delle basi
- si usano dei sequenziatori (macchine che applicano diversi processi biochimici)
- vengono sequenziati pezzi lunghi 200-700 basi (Sanger sequencing, usato nel Human Genome Project)
- molecole di DNA sono lunghe tra 100 000 e qualche millione di basi, quindi molto più lunghe
- il DNA deve essere frammentato (per es. metodo *shotgun*)

Riconstruzione della sequenza dai frammenti

Un esempio. Dati questi 7 frammenti, qual'è la sequenza di partenza?

1. `e nel pa`
2. `A quel pu`
3. `nico.`
4. `ry lotta`
5. `n cader`
6. `va per no`
7. `nto, Har`

Soluzione:

`2-7-4-6-5-1-3`

`A quel punto, Harry lottava per non cadere nel panico.`

Esempio: Adesso una sequenza DNA:

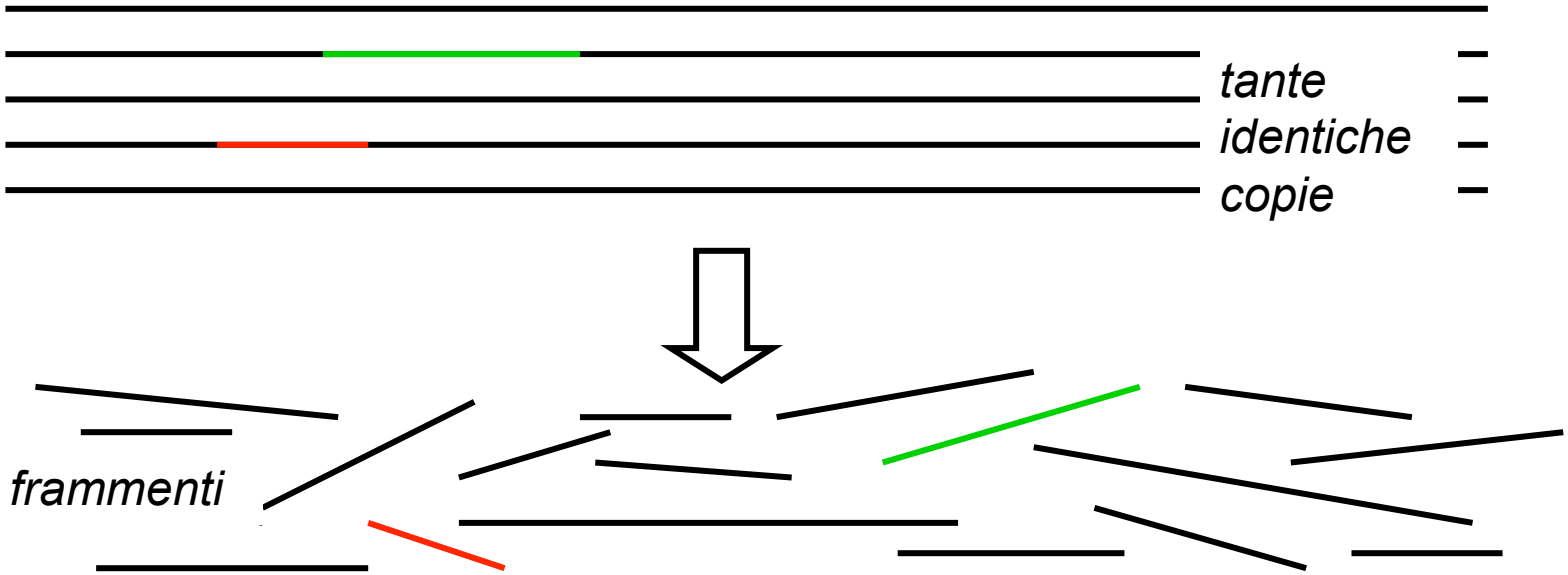
1. `ACCT`
2. `CACAT`
3. `GAGT`

Soluzioni:

- `ACCTCACATGAGT, ACCTGAGTCACAT, ...`
- `1-2-3,1-3-2,2-1-3,2-3-1,3-1-2,3-2-1`

In generale, la ricostruzione non è possibile! → Facciamo tante copie della sequenza originaria.

Sequenziamento Shotgun

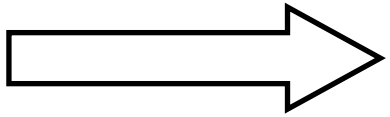


Sequenziamento Shotgun

Scopo: ricostruzione della sequenza dai frammenti

Un esempio

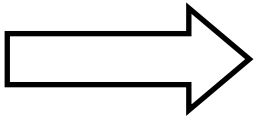
```
1.ACCGT
2.CGTGC
3.TTAC
4.TACCGT
```



```
--ACCGT--
----CGTGC
TTAC-----
-TACCGT--
TTACCGTGC
```

Un altro esempio

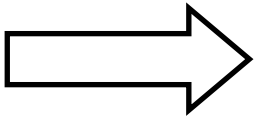
1. TACC
2. ACTAC
3. CGGACT
4. ACGGA



```
-------TACC
----ACTAC-
-CGGACT---
ACGGA-----
ACGGACTACC
```

Un altro esempio

1. TACC
2. ACTAC
3. CGGACT
4. ACGGA



```
TACC----------
---CGGACT-----
------ACTAC---
---------ACGGA
TACCGGACTACGGA
```

Le domande da fare

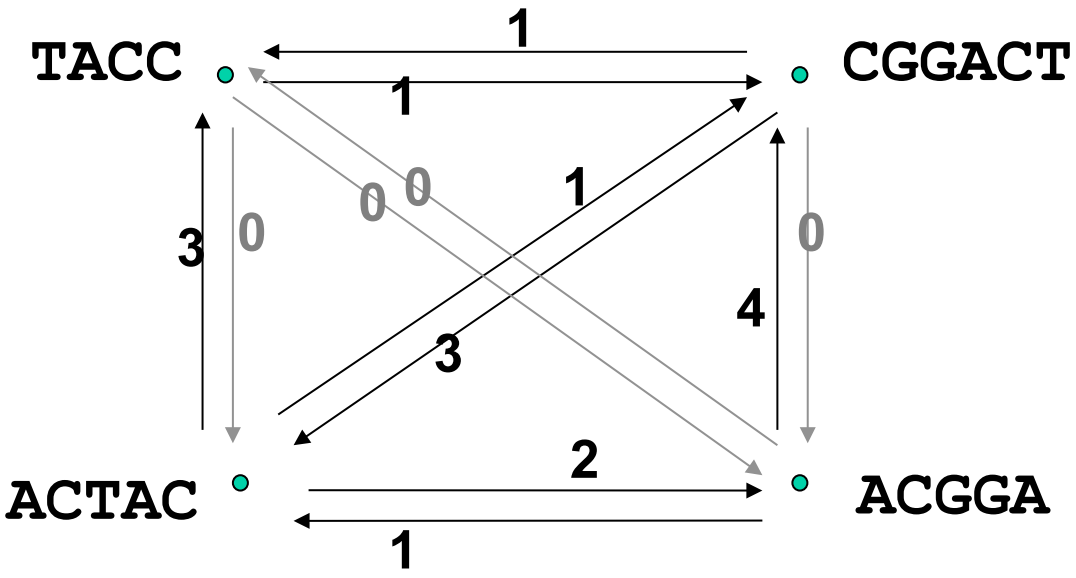
```
TACC----------
---CGGACT-----
------ACTAC---
----------ACGGA
TACCGGACTACGGA
```

```
------TACC
----ACTAC-
-CGGACT---
ACGGA-----
ACGGACTACC
```

1. Quale soluzione è la migliore?
2. Come si trovano le buoni soluzioni?

Formalizzazione del problema

Overlap-grafo (grafo orientato, pesato):



Formalizzazione del problema

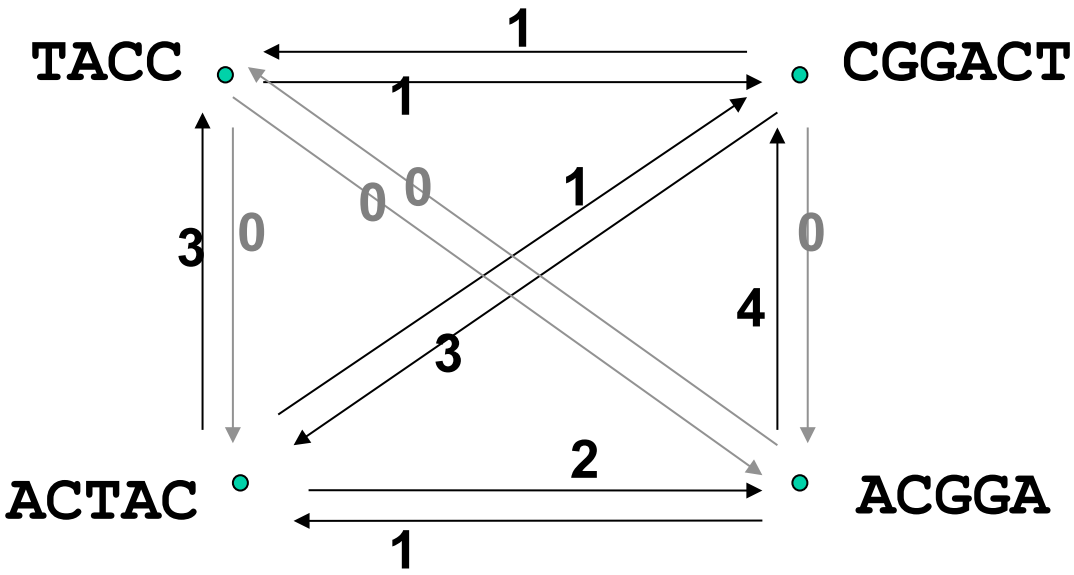
input: insieme di frammenti *F*

output: un cammino nell'overlap-grafo, che visita ogni nodo esattamente una volta *(un cammino Hamiltoniano)*

- cammino con il peso maggiore = shortest superstring
- assumiamo che F *substring-free*
  (nessuno dei frammenti è una sottostringa di un altro frammento, es. F = {`AC, TACCA`} non è substring-free)

Come trovare un cammino più pesante?

Non possiamo provare tutti, perché:



Ce ne sono troppi! in questo caso 24 = 4!, in generale n!, che cresce in modo esponenziale (5!=120, 6!=720, 7!=5040, ...)

L‘algoritmo Greedy

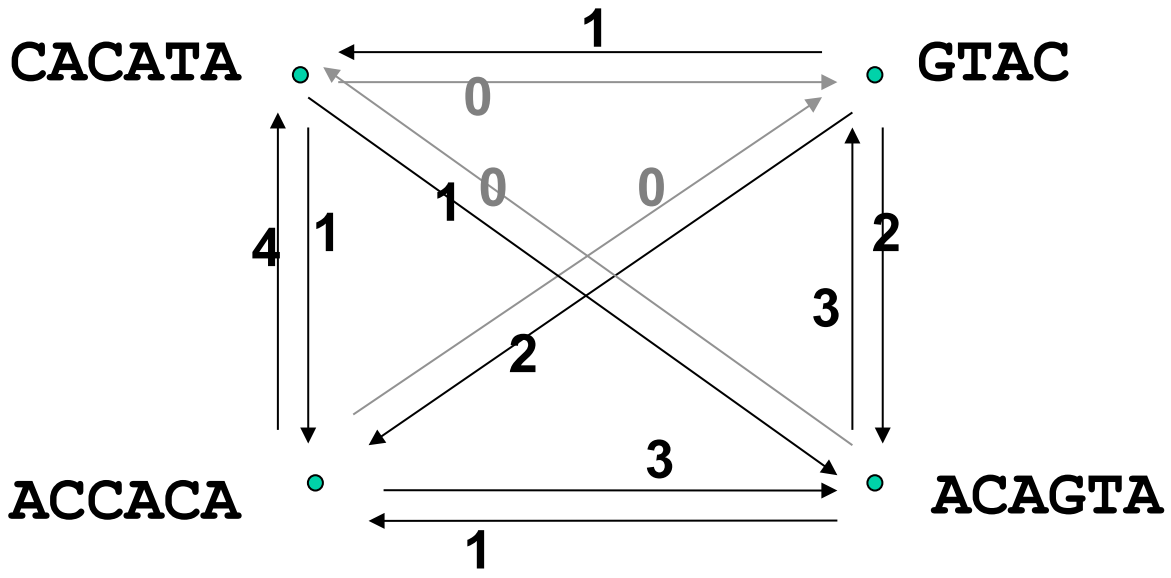
Costruisce un cammino Hamiltoniano,
aggiungendo sempre l‘arco più pesante tra quelli rimanenti, se possibile

idea:

- per ogni nodo al massimo un arco entrante e un arco uscente
- no cicli (quindi un nuovo arco deve connettere componenti diversi)

Formalizzazione del problema

Un altro esempio: **CACATA, GTAC, ACCACA, ACAGTA**



L‘algoritmo Greedy

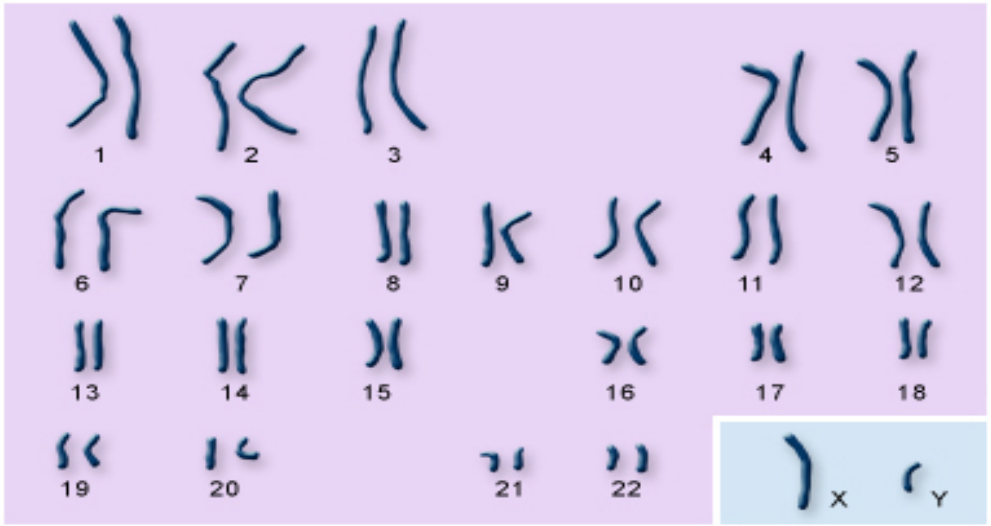
- ordina gli archi in ordine descrescente a seconda del peso
- prendi il prossimo arco (a,b) più pesante, se
  - outdegree(a) = 0 e indegree(b) = 0
  - Componente(a) ≠ Componente(b)
- aggiorna le strutture dati (elenco degree e componenti)
- termina quando il numero archi = $|F|$-1

La realtà è più complessa

- l‘algoritmo Greedy non trova sempre la soluzione ottimale, perché
- il problema è computazionalmente molto difficile *(„NP-completo“)*
- in realtà si usa un modello ancora più complesso:
  - errori di sequenziamento
  - orientamento dei frammenti (attenzione: 2^n possibilità!)
  - ripetizioni nel DNA (AXBXCXD – AXCXBXD)
- techniche moderne (Next-Generation-Sequencing, NGS) usano un altro approccio, sempre basato su grafi *(grafi de Bruijn)*

La dimensione del problema

- il genoma umano ca. 3 milliardi di basi
- 23 x 2 cromosomi, ognuno una sequenza
- tra 47 mio (cr. 21) e 248 mio (cr. 1) basi



- confronto: il primo libro Harry Potter consiste di ca. 300 pagine x 2000 caratteri = 600 000 caratteri

« A portare Dudley in ospedale » ringhiò zio Vernon. « Bisogna fargli togliere quella dannata coda, prima che vada a Smeltings ».

Il mattino dopo Harry si svegliò alle cinque, ma era troppo eccitato e nervoso per riaddormentarsi. Si alzò e si infilò i jeans, perché non voleva arrivare alla stazione con gli abiti da mago: si sarebbe cambiato poi in treno. Controllò ancora una volta l'elenco di Hogwarts per accertarsi di avere tutto quel che gli serviva, verificò che Edvige fosse ben chiusa nella sua gabbia, e cominciò a passeggiare per la stanza, in attesa che i Dursley si alzassero. Due ore dopo, il suo voluminoso e pesante baule era stato caricato sulla macchina dei Dursley, zia Petunia era riuscita a convincere Dudley a sedersi accanto a Harry, ed erano partiti.

Raggiunsero King's Cross alle dieci e mezzo. Zio Vernon mollò il baule su un carrello, spingendolo poi personalmente fin dentro la stazione. Harry si stupì per quel gesto stranamente cortese, ma si ricredette quando zio Vernon si fermò di botto, davanti ai binari, con un ghigno malevolo sul volto.

« Eccoci arrivati, ragazzo. Binario nove... binario dieci. Il tuo dovrebbe essere da qualche parte in mezzo, ma sembra che non l'abbiano ancora costruito, o sbaglio? »

Aveva ragione, era evidente. Sopra un binario torreggiava un grosso numero nove, in plastica, e su quello accanto un altrettanto grosso numero dieci, sempre in plastica; ma tra i due, niente.

« Auguri per la scuola » disse zio Vernon con un sorriso ancor più maligno. Si allontanò senza aggiungere altro. Harry si voltò e vide i Dursley ripartire in macchina. Ridevano tutti e tre. Gli si seccò la bocca. Che cosa diavolo avrebbe fatto? Intanto, stava cominciando ad attirare molti sguardi incuriositi per via di Edvige. Avrebbe dovuto chiedere a qualcuno.

Fermò un capotreno di passaggio, ma non osò fare parola del binario nove e tre quarti. Il capotreno non aveva mai sentito parlare di Hogwarts e quando si rese conto che Harry non era

in grado di dirgli neanche in che regione si trovasse, cominciò a infastidirsi, come se Harry facesse apposta a fare lo stupido. Disperato, Harry chiese del treno in partenza alle undici, ma l'uomo disse che non ce n'erano. Finì che il capotreno si allontanò imprecando contro i perditempo. A quel punto, Harry lottava per non cadere nel panico. Se il grosso orologio che sovrastava il cartellone degli arrivi funzionava, aveva ancora solo dieci minuti per prendere il treno per Hogwarts e non aveva la più pallida idea di come fare. Era lì, nel bel mezzo della stazione ferroviaria, con un baule che a stento riusciva a sollevare, le tasche piene di soldi dei maghi e una grossa civetta.

Hagrid doveva aver dimenticato di dirgli qualcosa di essenziale, come quando, per esempio, per entrare in Diagon Alley era stato necessario battere sul terzo mattone a sinistra. Si chiese se non fosse il caso di tirare fuori la bacchetta e cominciare a colpire la biglietteria tra i binari nove e dieci.

In quel momento, proprio dietro di lui, passò un gruppetto di persone e lui colse un brandello della loro conversazione.

« ...pieno zeppo di Babbani, figurarsi... »

Harry si voltò di scatto. A parlare era stata una signora grassottella, che si rivolgeva a quattro ragazzi dai capelli rosso fiamma. Ciascuno spingeva un baule come quello di Harry... e avevano anche un *gufo*.

Col cuore che gli martellava in petto, Harry li seguì, sempre spingendo il suo carrello. Quando si fermarono lui fece altrettanto, abbastanza vicino per sentire quel che dicevano.

« Allora, binario numero? » chiese la donna, che era la madre dei ragazzi.

« Nove e tre quarti! » disse con vocina stridula una ragazzina, anch'essa con i capelli rossi, che dava la mano alla madre. « Mamma, posso andare anch'io... »

« Tu sei troppo piccola, Ginny. Sta' zitta, adesso. Va bene, Percy, vai avanti tu ».

Quello che sembrava il maggiore si avviò verso i binari nove e dieci. Harry stette a guardare, bene attento a non battere ci-

Megszólított egy arra járó pályaudvari őrt, de nem merte említeni neki a kilenc és háromnegyedik vágányt. Az őr sose hallott Roxfortról, s mikor Harry azt sem tudta megmondani, hogy melyik országrészben van, még mérges is lett: azt hitte, a fiú készakarva játssza az ostobát. Harry erre megkérdezte, honnan indul a tizenegy órás vonat. Az őr azt felelte, hogy nincs tizenegy órás vonat, és szitkozódva faképnél hagyta. Harry most már kezdett valóban pánikba esni. A kijelző tábla fölötti nagy óra szerint tíz perce volt rá, hogy felszálljon a roxforti gyorsra, de egyelőre fogalma sem volt róla, miképpen tehetné meg ezt; ott állt a pályaudvar kellős közepén egy erszényre való varázslópénzzel, egy bagollyal és egy utazóládával, amit felemelni is alig tudott.

Hagrid bizonyára elfelejtett szólni, hogy itt is egy trükköt kell alkalmazni; olyasfélét, mint az Abszol út bejáratánál, ahol meg kellett kopogtatni balról a harmadik téglát. Harrynek megfordult a fejében, hogy előveszi a varázspálcáját, és megkopogtatja vele a kilences és a tízes vágány közötti jegypénztárat.

Ekkor emberek egy csoportja haladt el a háta mögött, és egy mondatfoszlány ütötte meg a fülét.

- ...persze tele van muglikkal...

Harry megperdült a tengelye körül. A beszélő, egy pufók asszonyság a négy fiának szónokolt. A fiúknak lángoló vörös hajuk volt, mind ugyanolyan ládát toltak maguk előtt, mint Harry. A csomagok tetején egy bagoly gubbasztott.

Harry kalapáló szívvel a család után indult. Mikor megálltak, ő is megállt, elég közel hozzájuk, hogy hallja, amit beszélnek.

- Hányas vágányról is indul? - kérdezte a fiúk anyja.

- A kilenc és háromnegyedikről! - sipította egy kislány. Ő is vörös hajú volt, és anyja kezét fogta. - Anyu, én miért nem...

- Mert még kicsi vagy, Ginny, és most hallgass. Jól van, Percy, te menj be elsőként.

A legidősebbnek tűnő fiú elindult a kilences és tízes vágányok felé. Harry utána meredt, pislogni se mert, nehogy lemaradjon valamiről - de épp mikor a fiú elérte a két peron közötti falat, csapatnyi turista tódult Harry orra elé. Mire az utolsó hátizsák is elhaladt, a vörös hajú fiú már sehol se volt.

- Te következel, Fred - adta ki az utasítást a pufók asszonyság.

- Nem Fred vagyok, hanem George! - felelte a fiú. - És még te nevezed magad anyánknak? Nem látod, hogy George vagyok?

- Ne haragudj, George drágám.

- Csak vicc volt, Fred vagyok - szólt a fiú, és már indult is. Ikertestvére utánaszólt, hogy igyekezzen, s a fiú bizonyára úgy is tett, hiszen egy szempillantással később már el is tűnt hogy hova, az rejtély maradt.

Most a harmadik fivér is elindult a peron felé. Már majdnem odaért, amikor hirtelen köddé vált

Che cos'è la bioinformatica?

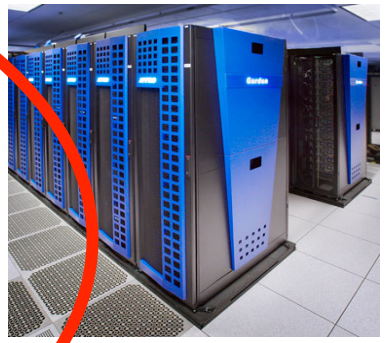
Per **bioinformatica** si intende l'uso delle tecniche matematiche e informatiche per risolvere problemi provenienti dalla biologia, tipicamente creando o usando programmi, modelli matematici o entrambi.

(Wikipedia)

dati biologici

+

il bioinformatico



computer

Perché la bioinformatica è bella

- area di ricerca giovane: tanti problemi nuovi, aperti
- fondamentale per i risultati importanti di oggi in biologia, farmacologia, medicina
- molto internazionale
- comprende sia biologia, matematica e informatica, sia teoria sia applicazioni

Speriamo di vedervi qui da noi tra qualche anno!