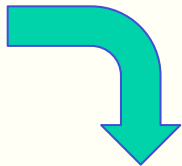




**Dal sequenziamento del pangenoma agli  
algoritmi per la medicina di precisione**

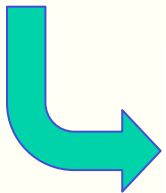
**21 Dicembre 2018**

# Viaggio al centro del DNA



**+ algoritmi**

## Sequenziamento DNA

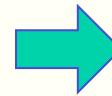


**file**

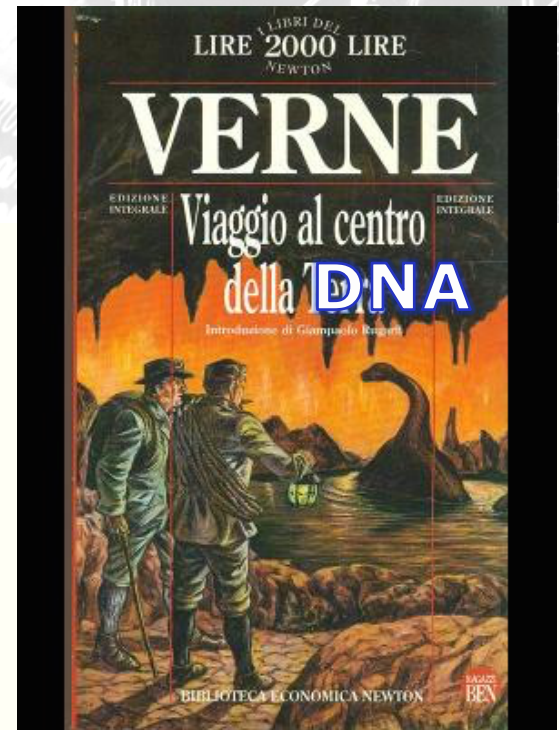
```

12854400 tcaaagtaagttagataaacatgatcattcacaggt
12854500 tggactaccagaattgagttacctagtagtctc
12854600 taggaaaagttaattgttacggccaatcactttt
12854700 attgtccactcaaaacgtgacaaatggaggtctaa
12854800 cttgtaaatgtattcacatttcattcccaagaaaa
12854900 aggttttaggggtctcaatataaacacacaaagcag
12855000 TCTTTTCTTCATCGTCTTTCCAACCTTCACGTTTTC
12855100 tcacttattgggtttctttcaattgtgaaacagAGT
12855200 ACATAGCCAACGCTGGAATCACTCATCTTTGGCTTC
    
```

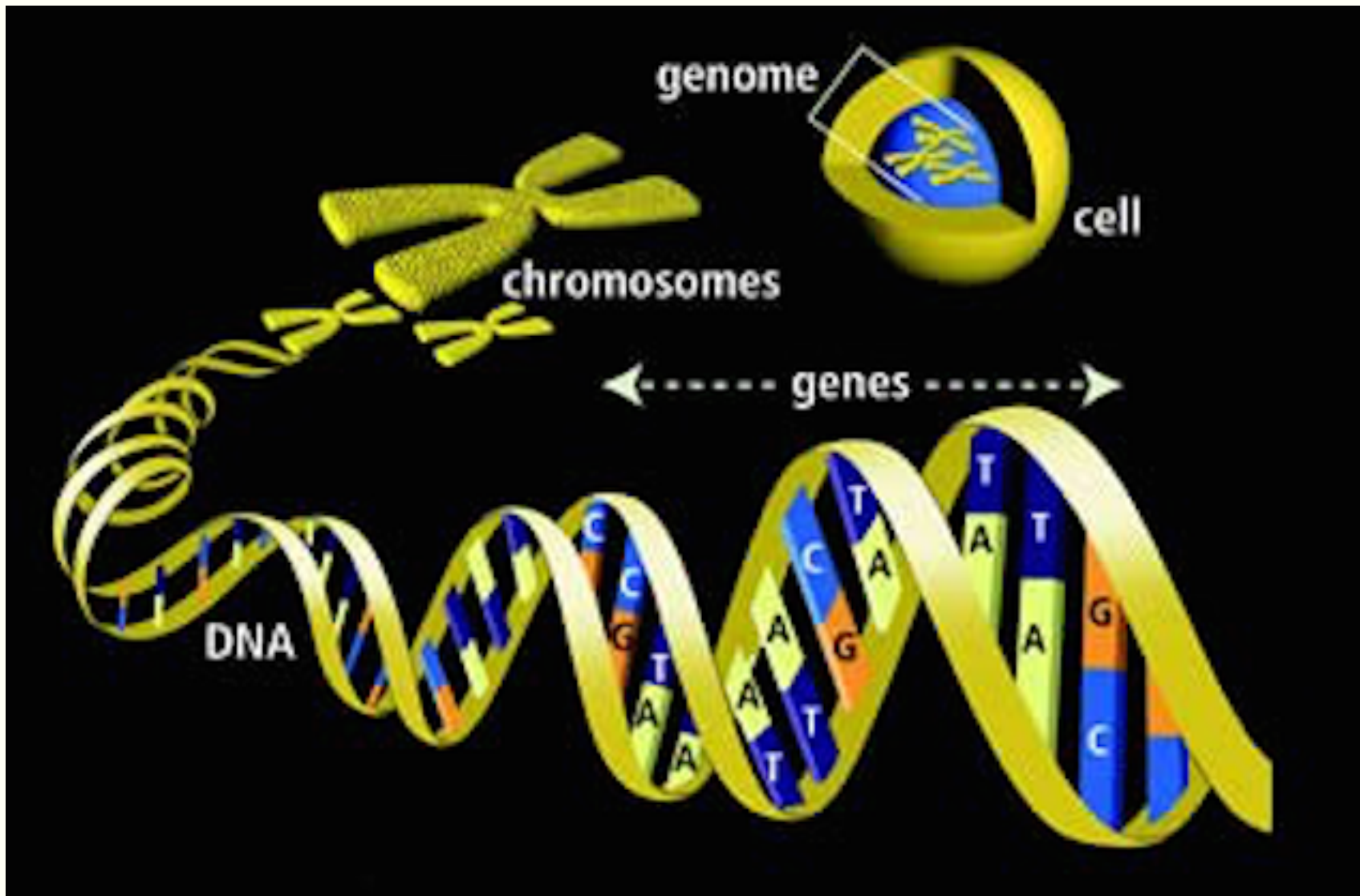
**+ algoritmi ...**



**Terapia medica**



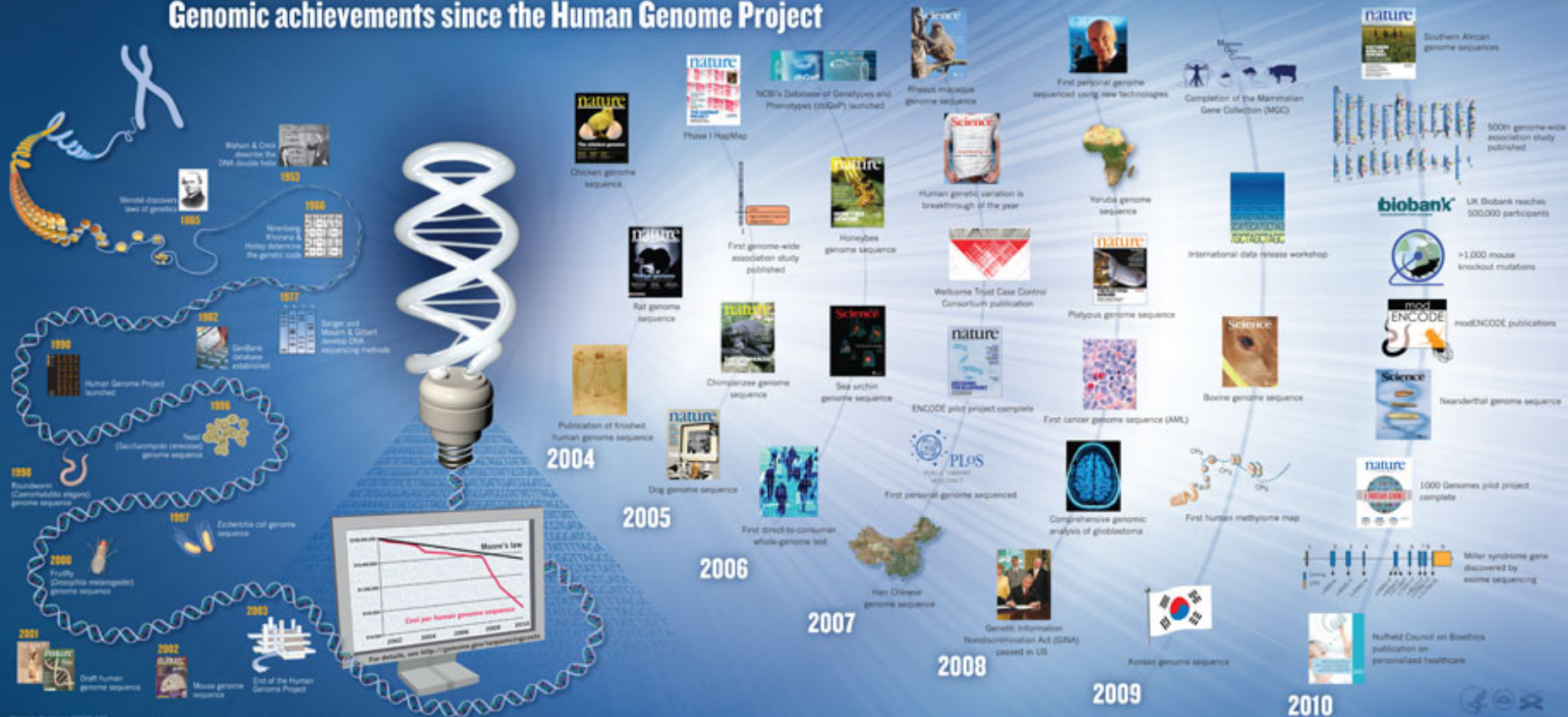
# Il genoma umano





# Bio -informatica ?

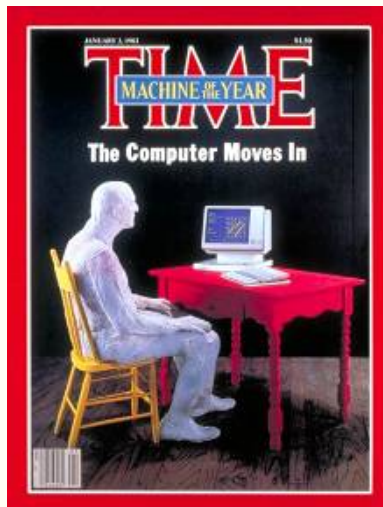
## Genomic achievements since the Human Genome Project





# The Human Genome project

- **Obiettivi: identificare i 30000 geni nel DNA umano e determinare la sequenza di 3 miliardi di basi->**
- **iniziato nel 1990 concluso nel 2003...**



**1982**



**2012**



# Sequenziamento DNA?

video DNA sequencing

[://www.youtube.com/watch?v=sn3\\_FIEbe0U](https://www.youtube.com/watch?v=sn3_FIEbe0U)



# The Newspaper Problem



stack of NY Times, June 27, 2000

# The Newspaper Problem



stack of NY Times, June 27, 2000



stack of NY Times, June 27, 2000  
on a pile of dynamite

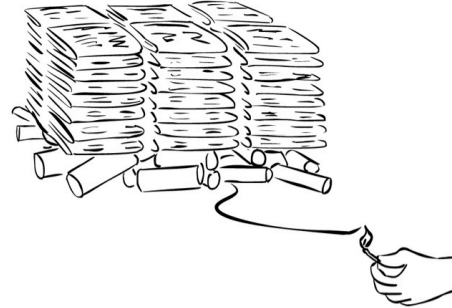
# The Newspaper Problem



stack of NY Times, June 27, 2000



stack of NY Times, June 27, 2000  
on a pile of dynamite



this is just hypothetical



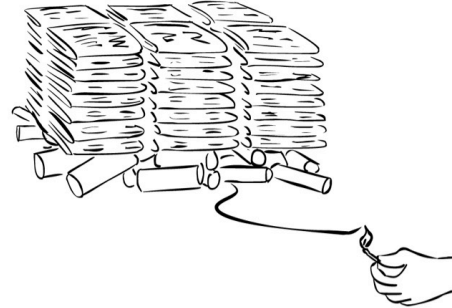
# The Newspaper Problem



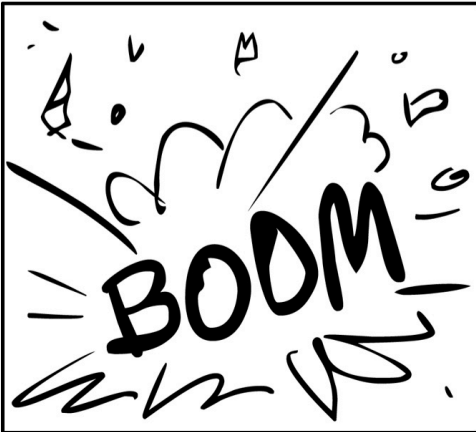
stack of NY Times, June 27, 2000



stack of NY Times, June 27, 2000  
on a pile of dynamite



this is just hypothetical



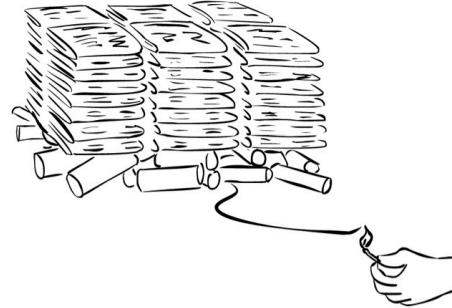
# The Newspaper Problem



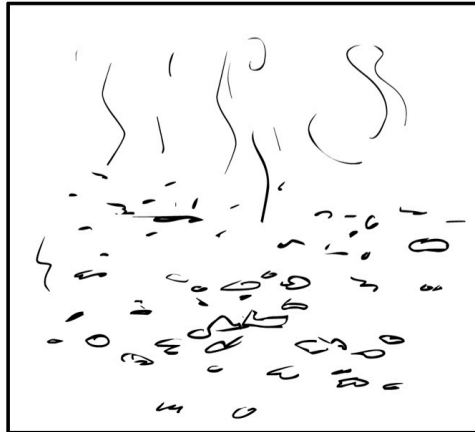
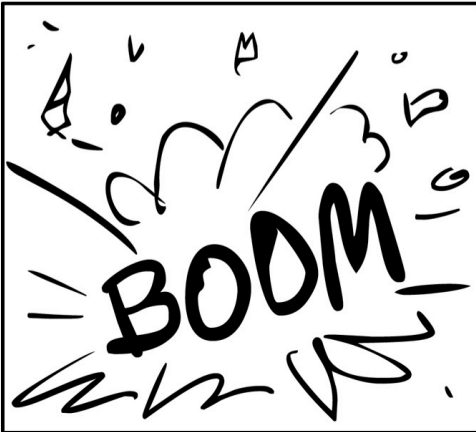
stack of NY Times, June 27, 2000



stack of NY Times, June 27, 2000  
on a pile of dynamite



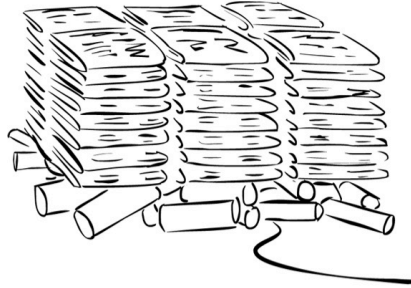
this is just hypothetical



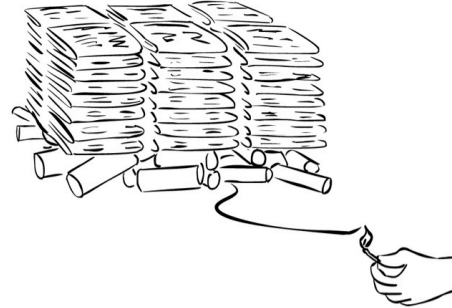
# The Newspaper Problem



stack of NY Times, June 27, 2000



stack of NY Times, June 27, 2000  
on a pile of dynamite



this is just hypothetical



so, what did the June 27, 2000 NY  
Times say?

## DNA Sequencing è più difficile del newspaper problem, per almeno 3 motivi

- Nel newspaper problem ci sono le regole di Inglese
- Non c'è un “linguaggio” *conosciuto* per il DNA
- **Esempio:** “suspect” and “murder” sono correlati.

...e murder occurred at approximately 5:2

...hoddie, approximately 5:2  
...we have not yet named any suspects, alt  
...information is welcome. The ca



## DNA Sequencing è più difficile del newspaper problem, per almeno 3 motivi

- Alfabeto inglese ha 26 lettere, mentre il DNA ha 4 lettere!
- **Esempio:** uso della “z” per connettere i frammenti

mentalists have noted low levels of oz

ne world's me  
zone as a contributing facto  
t the e as a

## DNA Sequencing è più difficile del newspaper problem, per almeno 3 motivi

- Alfabeto inglese ha 26 lettere, mentre il DNA ha 4 lettere!
- **Esempio:** uso della “z” per connettere i frammenti

...the world's me  
mentalists have cited low levels of ozone as a contributing factor  
...t the ... as a r

# DNA Sequencing è più difficile del newspaper problem, per almeno 3 motivi

Ci sono molte ripetizioni nel DNA che rendono la ricostruzione della sequenza originale un Puzzle



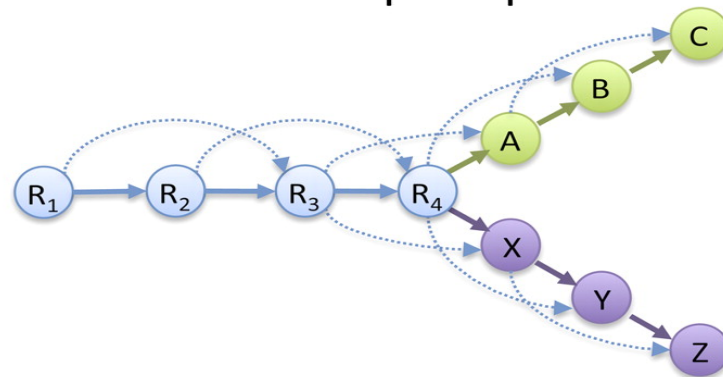
# sequenziamento ... algoritmi su grafi

## Sequenziamento DNA o assemblaggio grafi

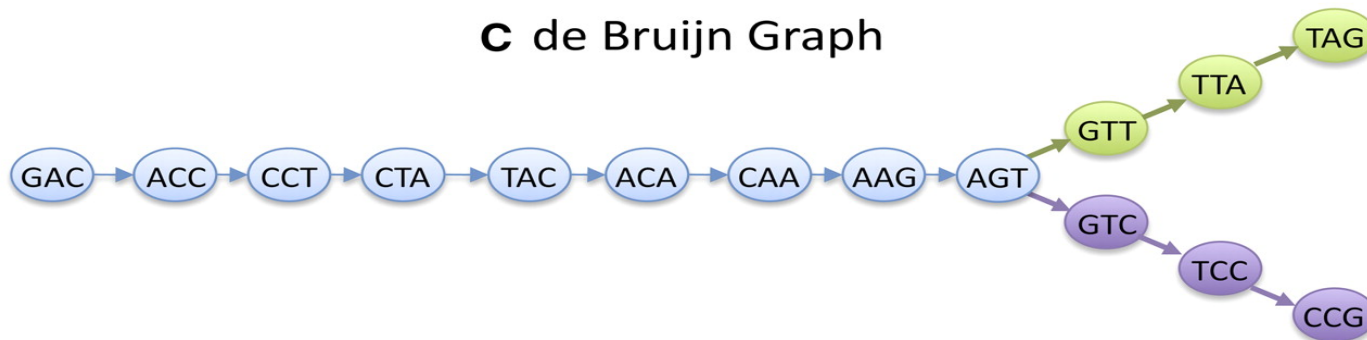
**A Read Layout**

$R_1$ : GACCTACA  
 $R_2$ : ACCTACAA  
 $R_3$ : CCTACAAG  
 $R_4$ : CTACAAGT  
**A**: TACAAGTT  
**B**: ACAAGTTA  
**C**: CAAGTTAG  
**X**: TACAAGTC  
**Y**: ACAAGTCC  
**Z**: CAAGTCCG

**B Overlap Graph**



**C de Bruijn Graph**



# Grafo di overlap

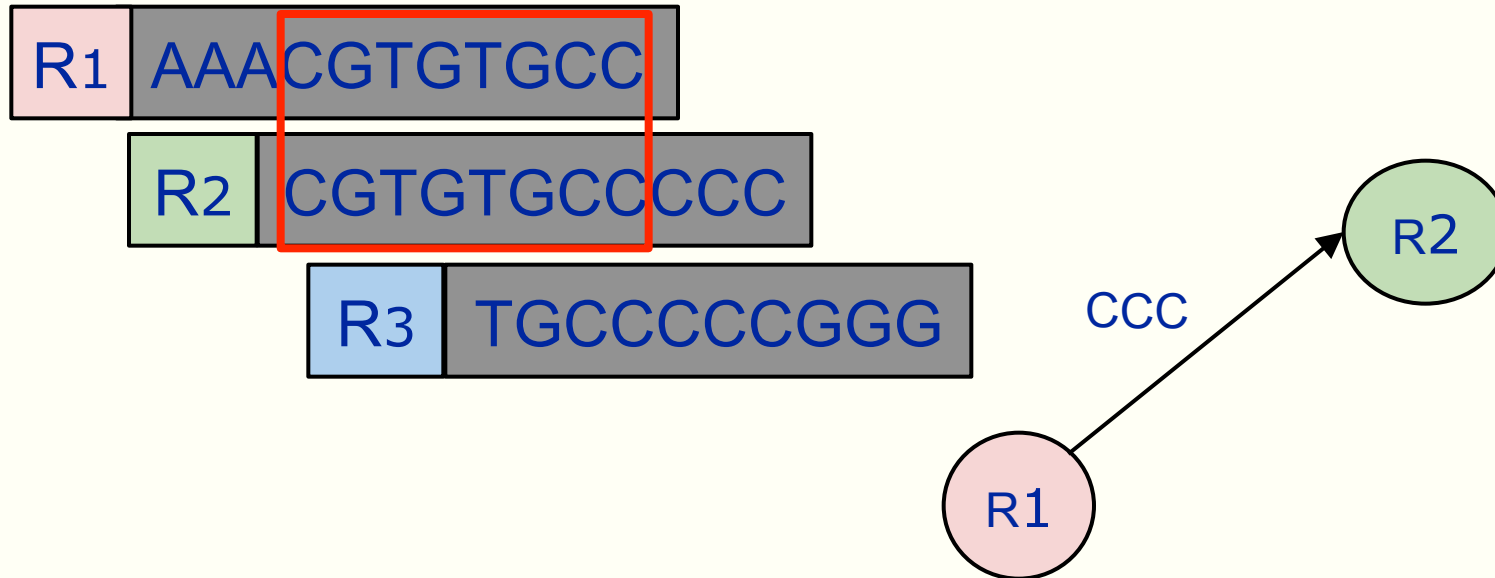
R1 AAACGTGTGCC

R2 CGTGTGCCCCC

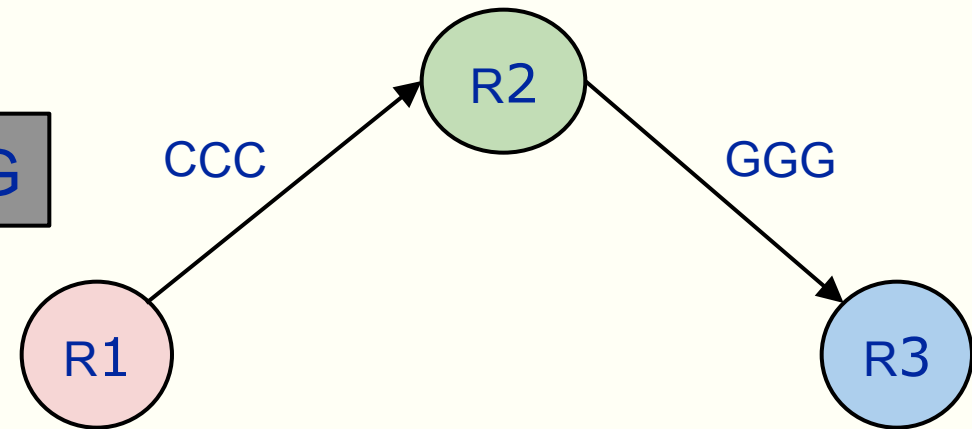
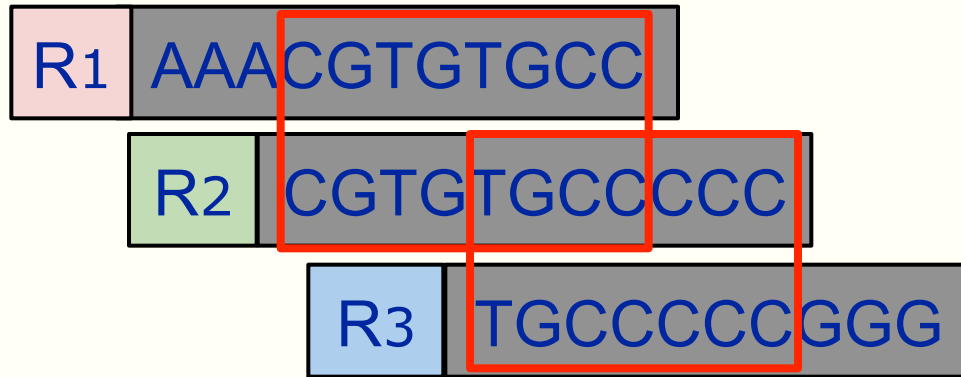
R3 TGCCCCCGGG



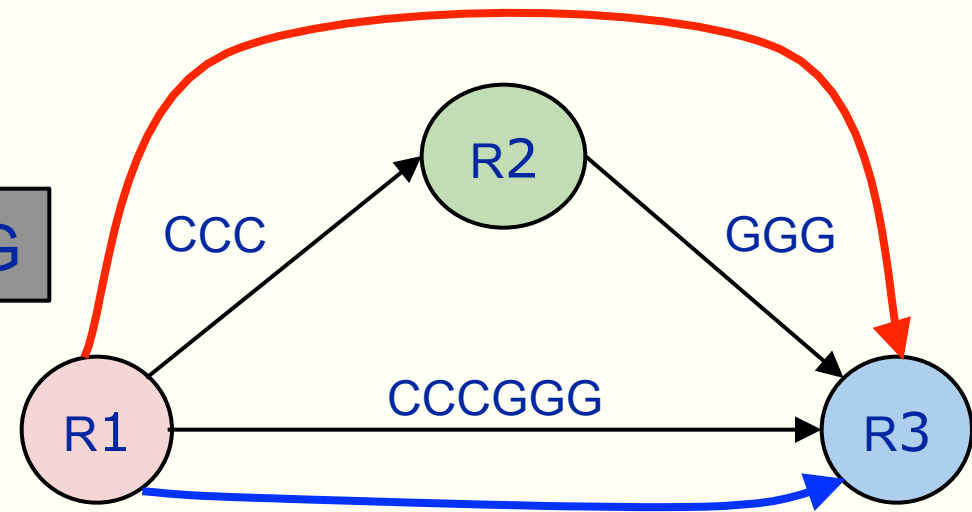
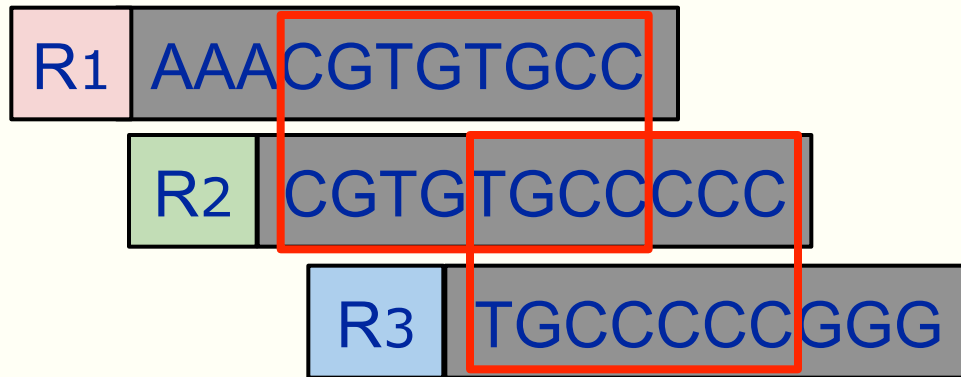
# Grafo di overlap



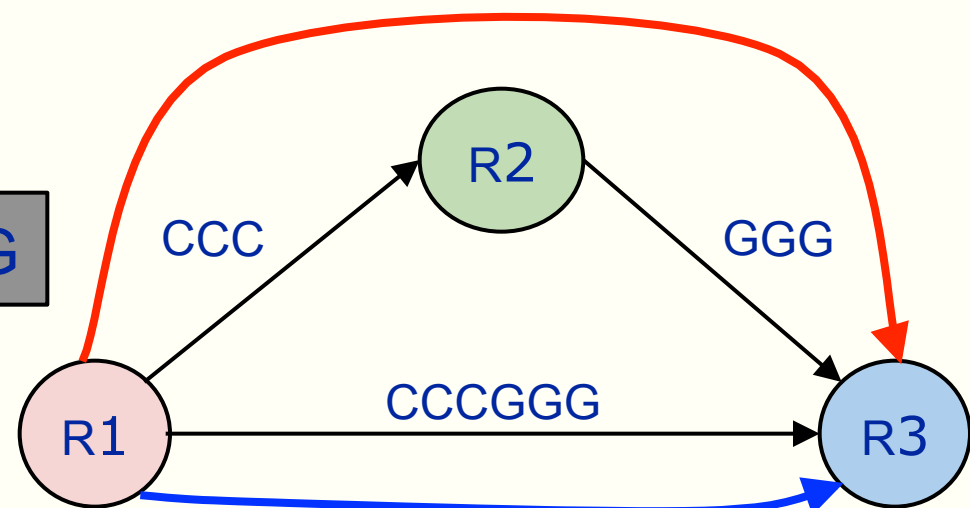
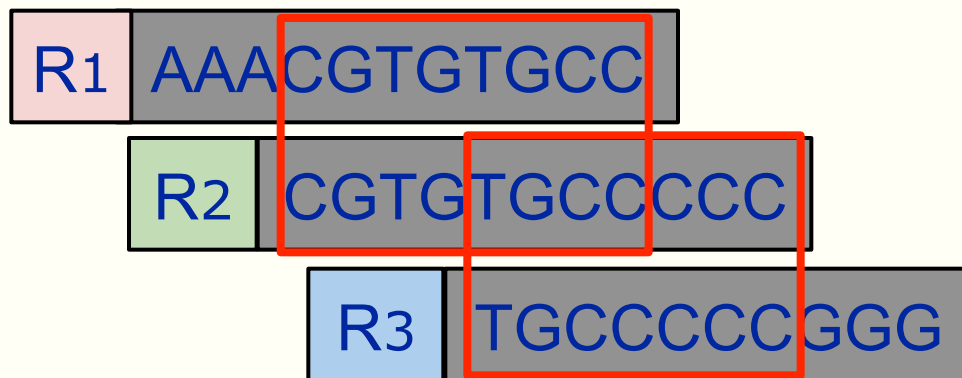
# Grafo di overlap



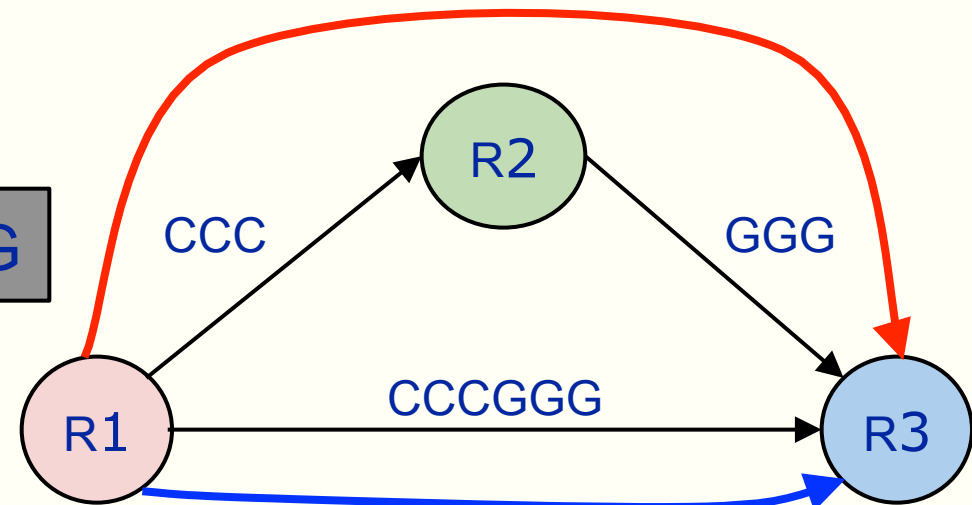
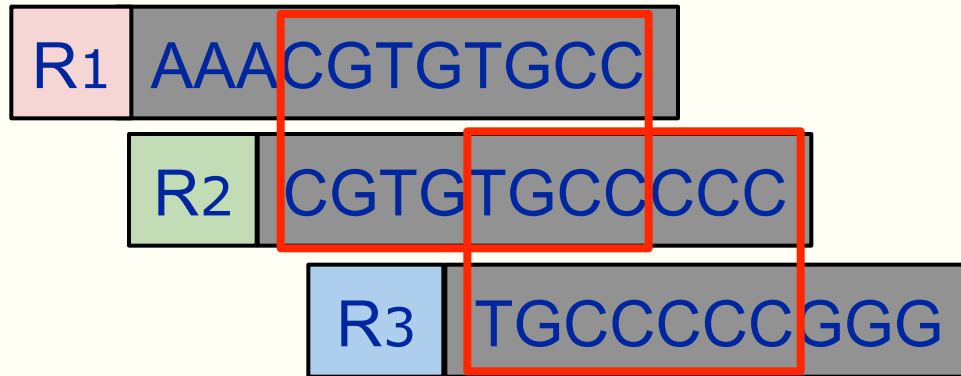
# Grafo di overlap



# Grafo di overlap



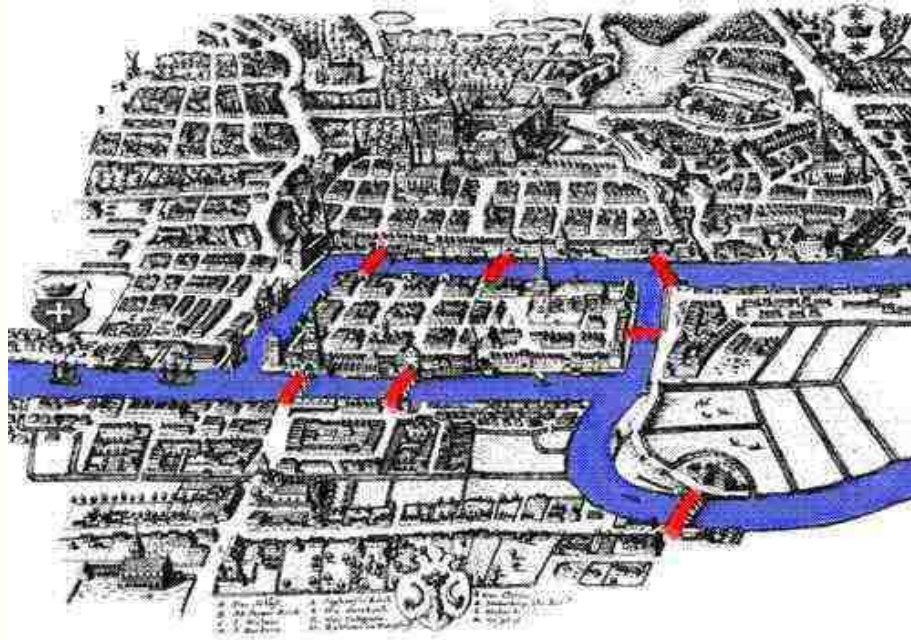
# Grafo di overlap





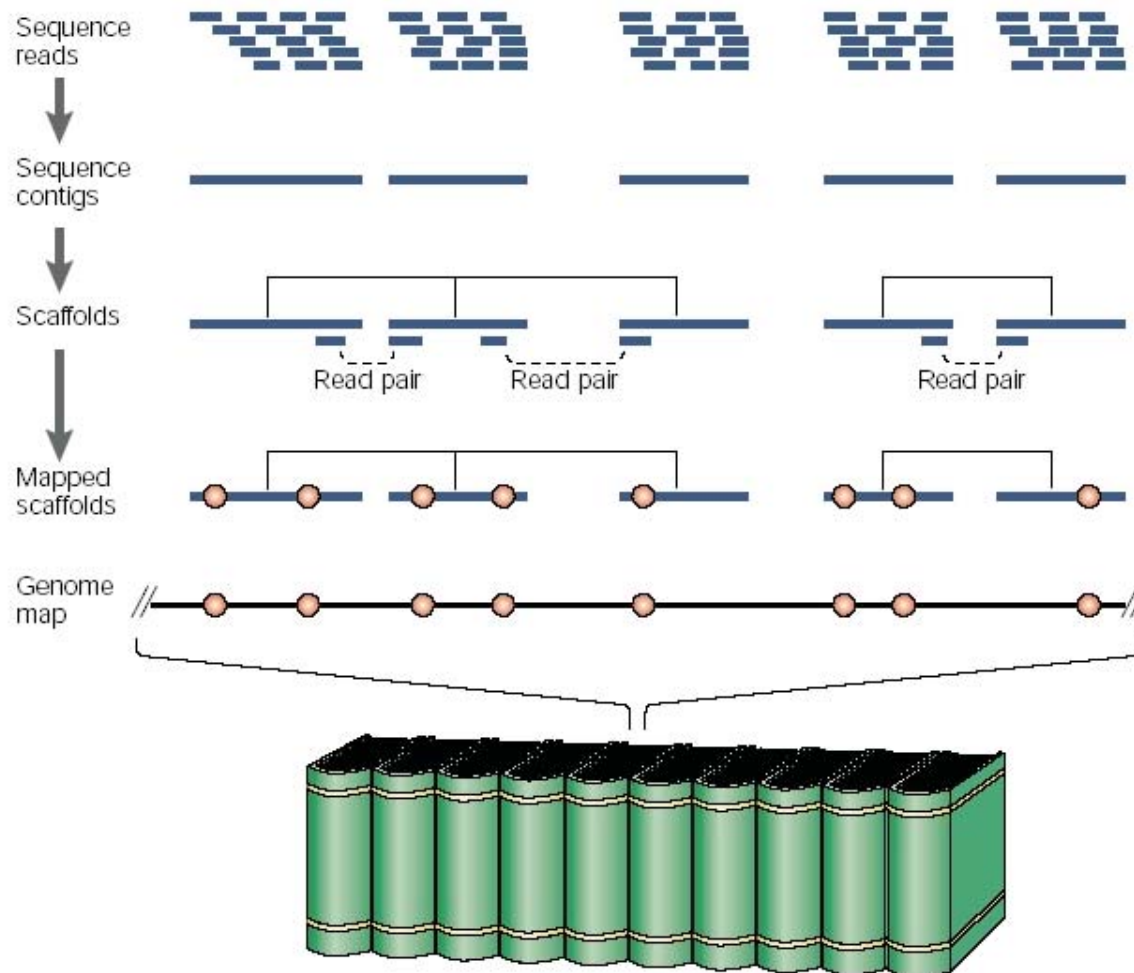
# sequenza DNA = cammino di Eulero

Trova un ciclo che attraversa ogni ponte una sola volta Leonhard Euler, 1735 -> **cammino di Eulero**



Ponti di Königsberg

# Assemblaggio

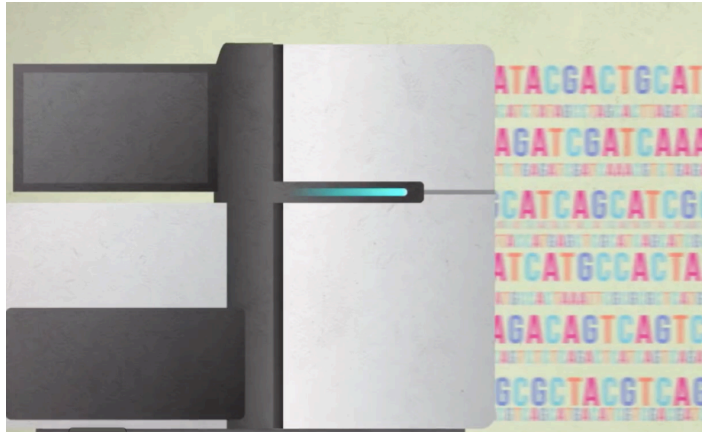


Courtesy of Nature Education. Used with permission.

Source: Green, Eric D. "[Strategies for the Systematic Sequencing of Complex Genomes](#)." *Nature Reviews Genetics* 2, no. 8 (2001): 573-83.

Adams, J. (2008) Complex genomes: Shotgun sequencing. Nature Educati

# Sequenziatori... oggi



HiSeq 2500  
**60 billion** bp per day  
(2012)



Tecnologia MinION

# Sequenziatori... oggi



HiSeq 2500  
**60 billion** bp per day  
(2012)



Tecnologia MinION



all'inizio

**15 YEARS**  
**\$3 BILLION**

AAGCTT  
 TAGCCG  
 TCGATTG  
 TAGTCA  
 TATAG  
 ATAGCTCGATTGAAC  
 AGCCTAGTCAGTTACA  
 TCGATTGAACCTGA  
 TAGTCAGTTA  
 ATCATGATCA  
 AGACTATAGCTCG  
 GATCATTAGCTA  
 ATAGCTCGATTGA  
 GGATCATGATTAG  
 GACTATAGCTCG  
 TGA TAGC  
 TAGC TCGATT  
 TCAT TCGCT  
 AGCT TATTC  
 CAGT TCAAG  
 GTAA TGGAT  
 ACAA TCTAT  
 CGCC TGTAC  
 AGTT TAAAG  
 AACT TATCA  
 AAG TATA  
 CC TGC

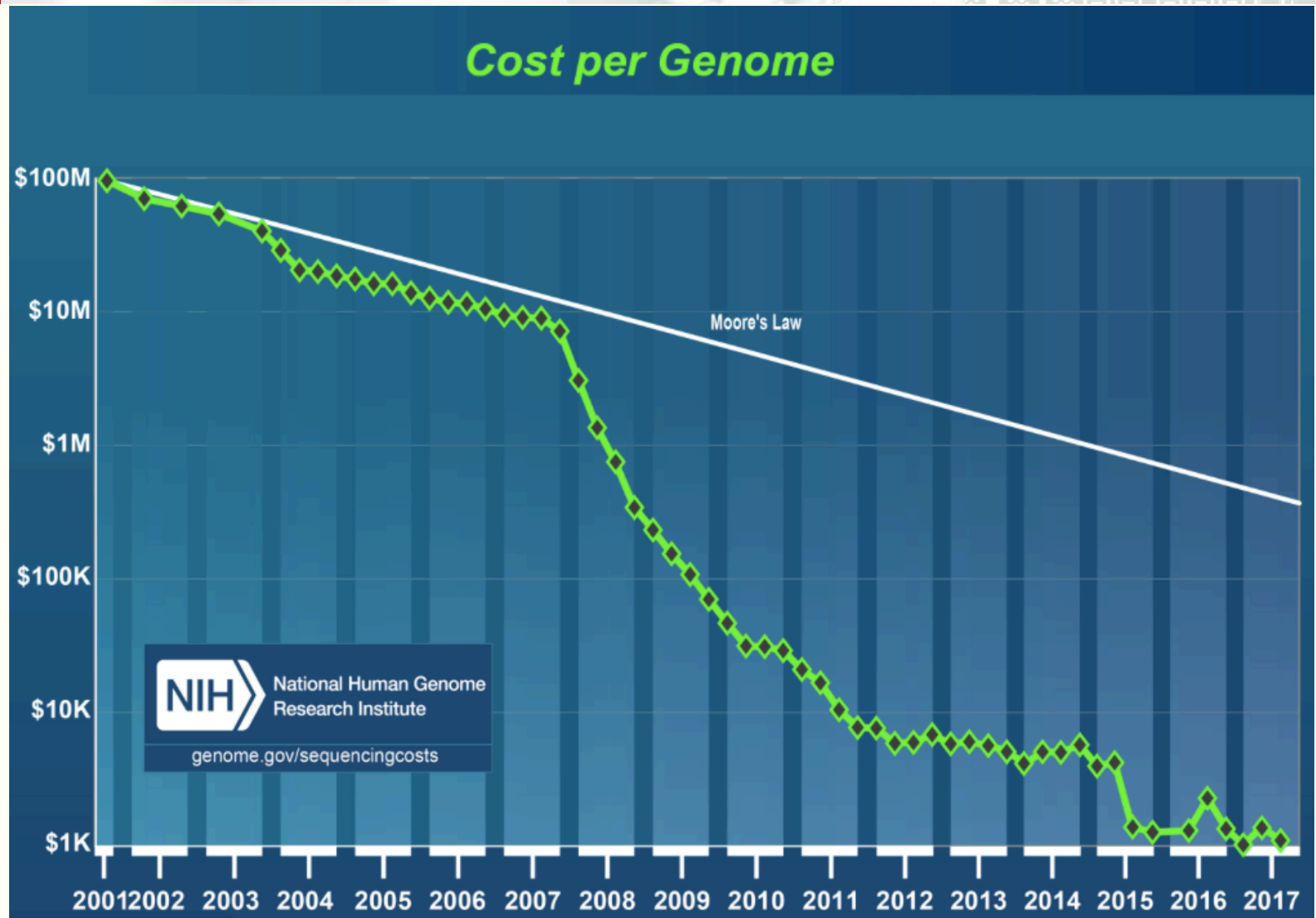
...nel 2014

.AGC.  
 .TAGCC)  
 (CGATTG/  
 .TAGTCA  
 .TATAC'  
 .GATTAGCCTAGC.  
 ATAGCTCGATTGAAC,  
 AGCCTAGTCAGTTACA  
 (GC\*CGATTGAACTSGA  
 (GC TAGTCAGTTA)AA  
 .TG ATCATGATCA/ITA  
 (AC AGACTATAGC (CG  
 (CA GATCATTAGC)TA  
 .GAC ATAGCTCGAT (GA  
 AAC GGATCATGAT (AG  
 ACA (GACTATAGC' CG/  
 TGA A TAGC  
 TAGC CGATT  
 TCAT (GCCT  
 AGCT GATTE  
 CAGT (CAAG  
 GTAA (GGA)  
 ACAA (CTAT  
 CGCC (GTAC  
 AGTT (AAGA  
 AACT GATC/  
 'AAG/ TATA  
 .CC .TTC

**40 HOURS**  
**\$5,000**



# Today...



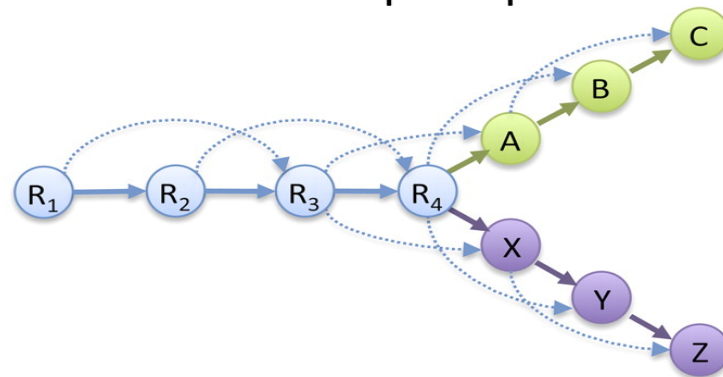
# sequenziamento ... algoritmi su grafi

## Sequenziamento DNA o assemblaggio grafi

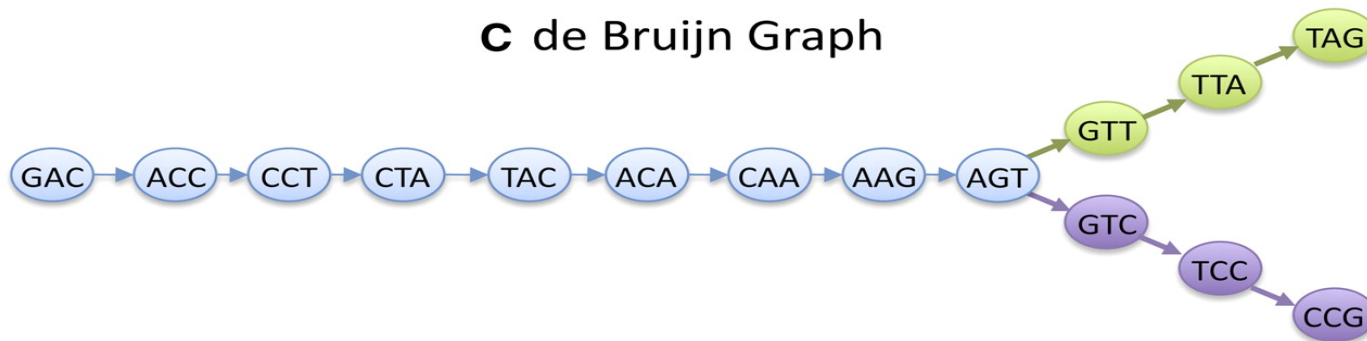
**A Read Layout**

$R_1$ : GACCTACA  
 $R_2$ : ACCTACAA  
 $R_3$ : CCTACAAG  
 $R_4$ : CTACAAGT  
**A**: TACAAGTT  
**B**: ACAAGTTA  
**C**: CAAGTTAG  
**X**: TACAAGTC  
**Y**: ACAAGTCC  
**Z**: CAAGTCCG

**B Overlap Graph**



**C de Bruijn Graph**



# Bio-Informatica ... le questioni principali

Riscaldare **muscoli algoritmici** per risolvere problemi di biologia **senza** indossare il **camice di laboratorio**



# Bio-Informatica ... le questioni principali

Riscaldare **muscoli algoritmici** per risolvere problemi di biologia **senza** indossare il **camice di laboratorio**

How Do We Assemble Genomes? (*Graph Algorithms*)

How Do We Compare Biological Sequences? (*Dynamic Programming Algorithms*)

Are There Fragile Regions in the Human Genome? (*Combinatorial Algorithms*)

Which Animal Gave us SARS? (*Evolutionary Trees*)

How Do We Locate Disease-Causing Mutations? (*Combinatorial Pattern Matching*)

How do we find mutations to "stop a tumor"? (*Linear Integer Programming and tree evolution*)

<http://bioinformaticsalgorithms.com>

# IL genoma

- Il genoma è formato da milioni di *basi* anche in organismi semplici:

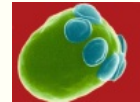
$$1 \text{ Gbp} = 10^9$$

Batterio (E. Coli)



4.8 Mbp

Lievito



14.4 Mbp

Verme (C. elegans)



100.0 Mbp

Moscerino (Drosophila)



165.0 Mbp

Uomo (Homo sapiens)



3.3 Gbp

- **Per memorizzare ed analizzare un intero genoma occorrono grandi risorse computazionali**

# Bio-Informatica ... e i BIG DATA

Si impara a trattare dati di **grandi dimensioni e come risparmiare sulle risorse computazionali (memoria e tempo di calcolo)**

How do we analyze multiple genomes in a few days?  
(*Hashing, Burrows Wheeler Transform*)

How do we discover interesting genes for a disease in an individual in few hours?  
(*Indexing by Bloom Filters, Bloom trees, self-indexes*)

<http://bioinformaticsalgorithms.com>



# IL Viaggio continua...

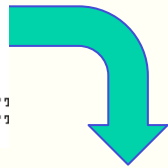
**genoma è una sequenza?**

Unaligned sequences

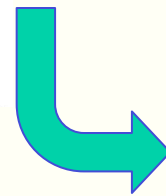
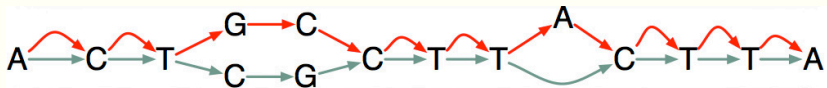
|            |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|------------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| human      | A | C | A | T | A | T | G | G | A | C | A | G | T | A | A | A | A | A | C | A | T | A |
| Chimpanzee | A | C | A | T | A | T | G | G | A | C | A | G | T | A | A | A | A | A | C | A | T | A |
| Vacaque    | A | T | A | C | A | T | T | A | C | G | C | A | G | T | A | A | A | A | C | A | T | A |

Aligned sequences

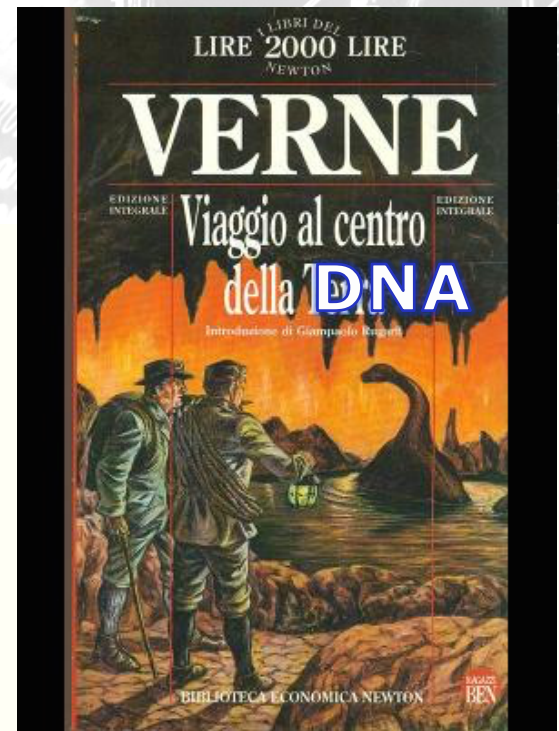
|            |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|------------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| human      | A | C | A | T | A | T | G | G | A | C | A | G | T | A | A | A | A | A | C | A | T | A |
| Chimpanzee | A | C | A | T | A | T | G | G | A | C | A | G | T | A | A | A | A | A | C | A | T | A |
| Vacaque    | A | T | A | C | A | T | T | A | C | G | C | A | G | T | A | A | A | A | C | A | T | A |



**grafo - Pan-genome**



**+ algoritmi .. ➡ Terapia ad hoc**



|            |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|------------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -human     | A | C | A | T | T | A | T | G | G | A | C | A | G | G | T | A | A | G | T | A | A | A | A | A | C | A | T | A | T | T |   |
| Chimpanzee | A | C | A | T | T | A | T | G | G | A | C | A | G | G | T | A | A | G | T | A | A | A | A | A | C | A | T | A | T | T |   |
| Macaque    | A | T | A | T | A | C | A | T | T | A | C | G | G | A | C | A | G | G | T | A | A | G | T | A | A | A | A | A | C | A | T |

|            |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|------------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Human      | A | C | A |   | T | T | A | T | G | G | A | C | A | G | G | T | A | A | G | T | A | A | A | A | A | A | C | A | T | A | T | T |   |
| Chimpanzee | A | C | A |   | T | T | A | T | G | G | A | C | A | G | G | T | A | A | G | T | A | A | A | A | A | A | A | C | A | T | A | T | T |
| Macaque    | A | T | A | T | A | C | A | T | T | A | C | G | G | A | C | A | G | G | T | A | A | G | T | A | A | A | A | A | C | A | T |   |   |

# algoritmi su sequenze

## Edit Distance: Example

TGCATAT → ATCCGAT in 5 passi

TGCATAT<sup>T</sup> → (delete last <sup>T</sup>)

TGCAT<sup>A</sup> → (delete last <sup>A</sup>)

TGCAT → (insert <sup>A</sup> at front)

<sup>A</sup>T<sup>G</sup>CAT → (substitute <sup>C</sup> for 3<sup>rd</sup> <sup>G</sup>)

AT<sup>C</sup>CAT → (insert <sup>G</sup> before last A)

ATCC<sup>G</sup>AT (Done)

## Edit Distance: Example

TGCATAT → ATCCGAT in 5 steps

TGCATAT<sup>T</sup> → (delete last <sup>T</sup>)

TGCAT<sup>A</sup> → (delete last <sup>A</sup>)

TGCAT → (insert <sup>A</sup> at front)

<sup>A</sup>T<sup>G</sup>CAT → (substitute <sup>C</sup> for 3<sup>rd</sup> <sup>G</sup>)

AT<sup>C</sup>CAT → (insert <sup>G</sup> before last A)

ATCC<sup>G</sup>AT (Done)

**What is the edit distance? 5?**

## Edit Distance: Example (cont' d)

TGCATAT → ATCCGAT in 4 steps

TGCATAT → (insert A at front)

ATGCATAT → (delete 6<sup>th</sup> T)

ATGCATA → (substitute G for 5<sup>th</sup> A)

ATGCGTA → (substitute C for 3<sup>rd</sup> G)

ATCCGAT (Done)



## Edit Distance: Example (cont' d)

TGCATAT → ATCCGAT in 4 steps

TGCATAT → (insert A at front)

ATGCATAT → (delete 6<sup>th</sup> T)

ATGCATA → (substitute G for 5<sup>th</sup> A)

ATGCGTA → (substitute C for 3<sup>rd</sup> G)

ATCCGAT (Done)

**Can it be done in 3 steps???**

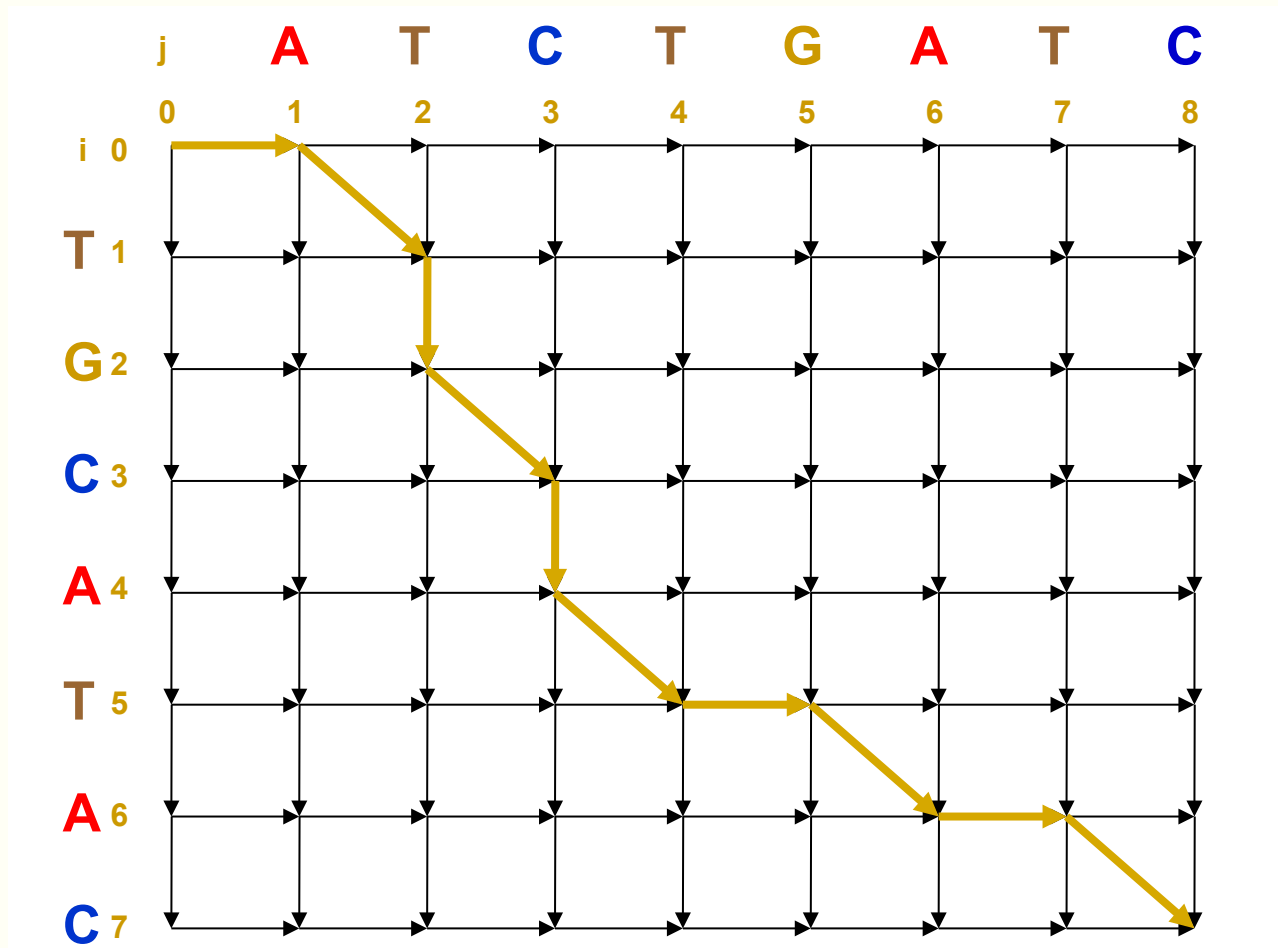
# Allineamento

Date 2 sequenze di DNA **v** and **w**:

allineamento :  $2 * k$  matrix (  $k > m, n$  )

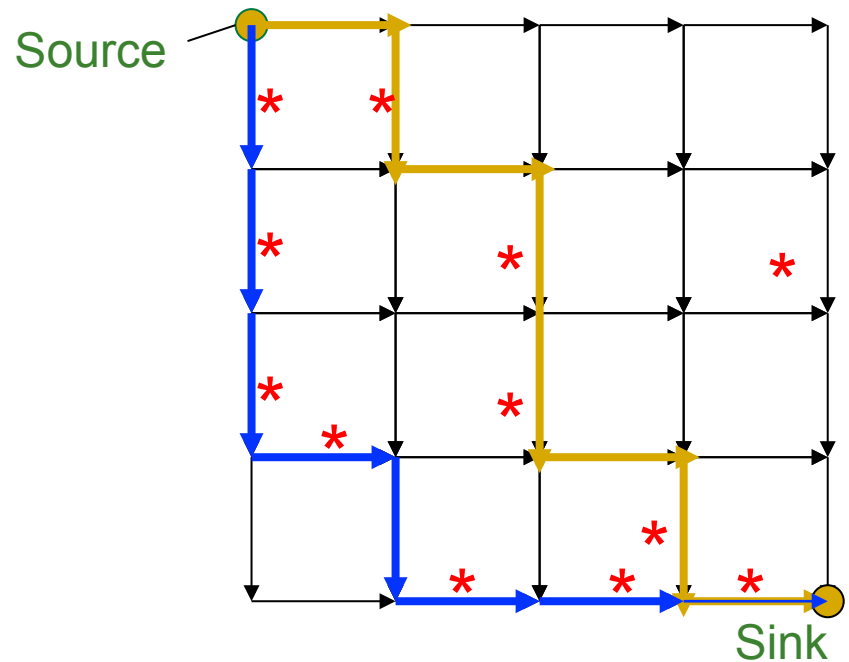
|                     |           |          |              |          |          |             |          |          |          |
|---------------------|-----------|----------|--------------|----------|----------|-------------|----------|----------|----------|
| Lettere di <b>v</b> | <b>A</b>  | <b>T</b> | -            | <b>G</b> | <b>T</b> | <b>T</b>    | <b>A</b> | <b>T</b> | -        |
| Lettere di <b>w</b> | <b>A</b>  | <b>T</b> | <b>C</b>     | <b>G</b> | <b>T</b> | -           | <b>A</b> | -        | <b>C</b> |
|                     |           |          |              |          |          | -           |          | -        |          |
|                     | 4 matches |          |              |          |          | 2 deletions |          |          |          |
|                     |           |          | 2 insertions |          |          |             |          |          |          |

# Allineamento



# Manhattan Tourist Problem (MTP)

Imagine seeking a path (from source to sink) to travel (only eastward and southward) with the most number of attractions (\*) in the Manhattan grid



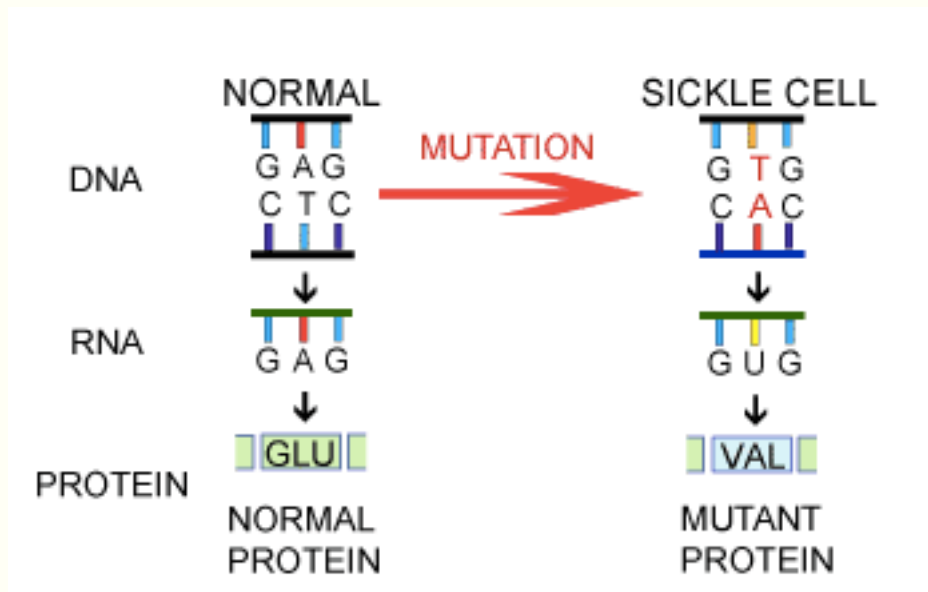
# mutazioni nel DNA

## Malattie **monogenetiche**

- anemia falciforme

**HBB gene** -> omoglobina

[https://evolution.berkeley.edu/evolibrary/article/mutations\\_06](https://evolution.berkeley.edu/evolibrary/article/mutations_06)



# mutazioni nel DNA

## Malattie **monogenetiche**

- Fibrosi cistica

**CFTR gene** almeno **2000** mutazioni diverse nello stesso gene  
del cromosoma 7q31.2



# aplotipi ed ereditarietà



Haplotype

Sequence of genetic characters (alleles)

Paternal  
Haplotype

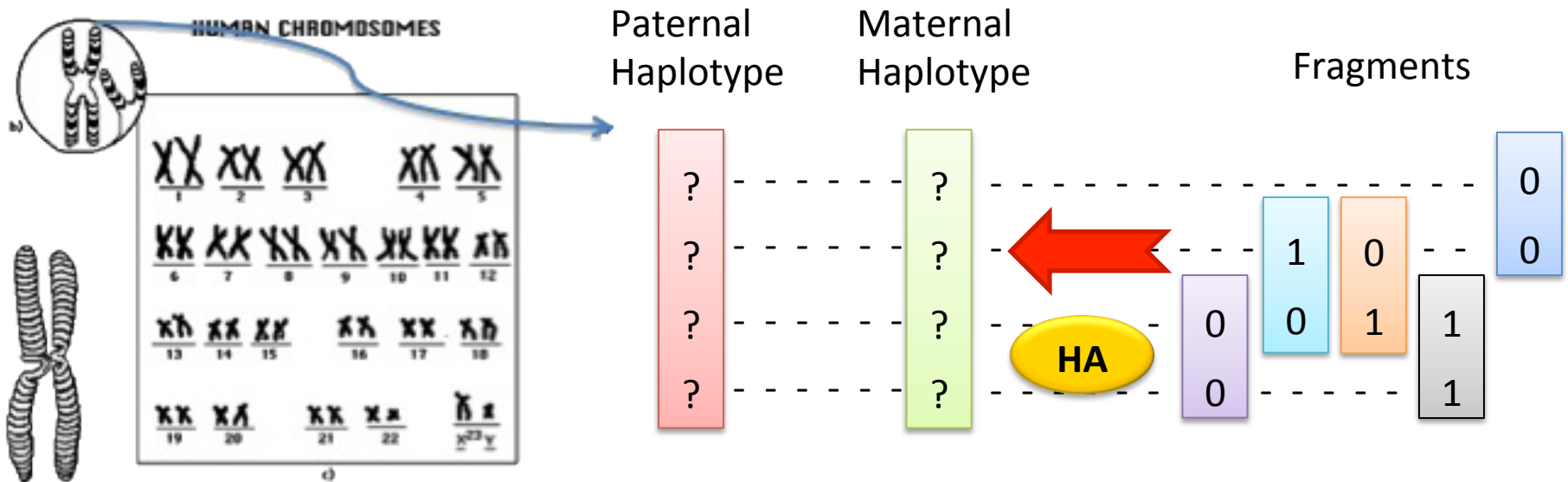
Maternal  
Haplotype

|   |           |   |
|---|-----------|---|
| 0 | - - - - - | 0 |
| 1 | - - - - - | 0 |
| 0 | - - - - - | 1 |
| 0 | - - - - - | 1 |



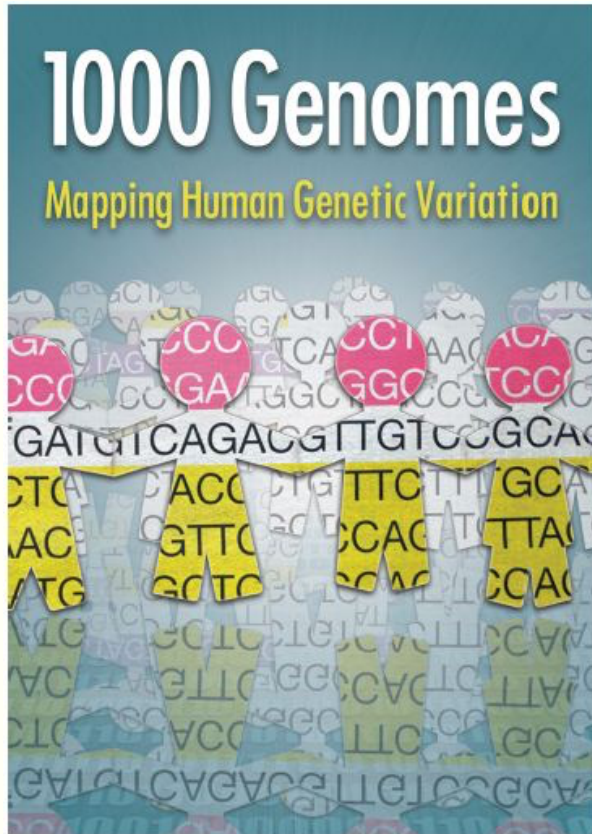
The Haplotypes are fundamental in the study of genetic variations  
(Browning and Browning, 2011)

# aplotipi ed ereditarietà



➔ Haplotype Assembly

# variazioni del DNA in popolazioni



© EMBL-EBI/ S. Phillips

## CORRIERE DELLA SERA / SALUTE



ANTROPOLOGIA

### Gli italiani non esistono. Siamo un grande mix genetico. Tranne i sardi

La penisola dal punto di vista genetico è divisa da una linea che separa più Est da Ovest che Nord da Sud. L'unica che fa storia a sé è la Sardegna

di Luigi Ripamonti



The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature, 1 November 2012.

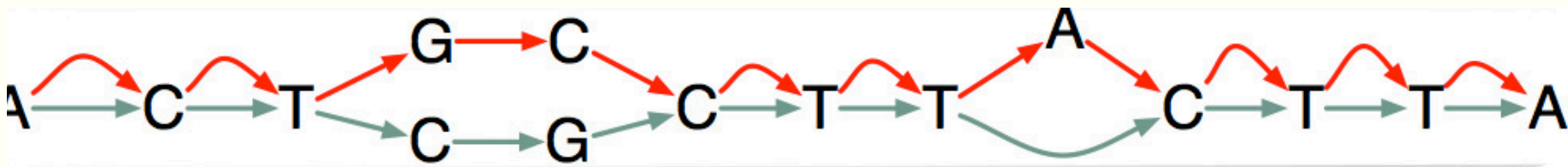


# Il genome è una sequenza?



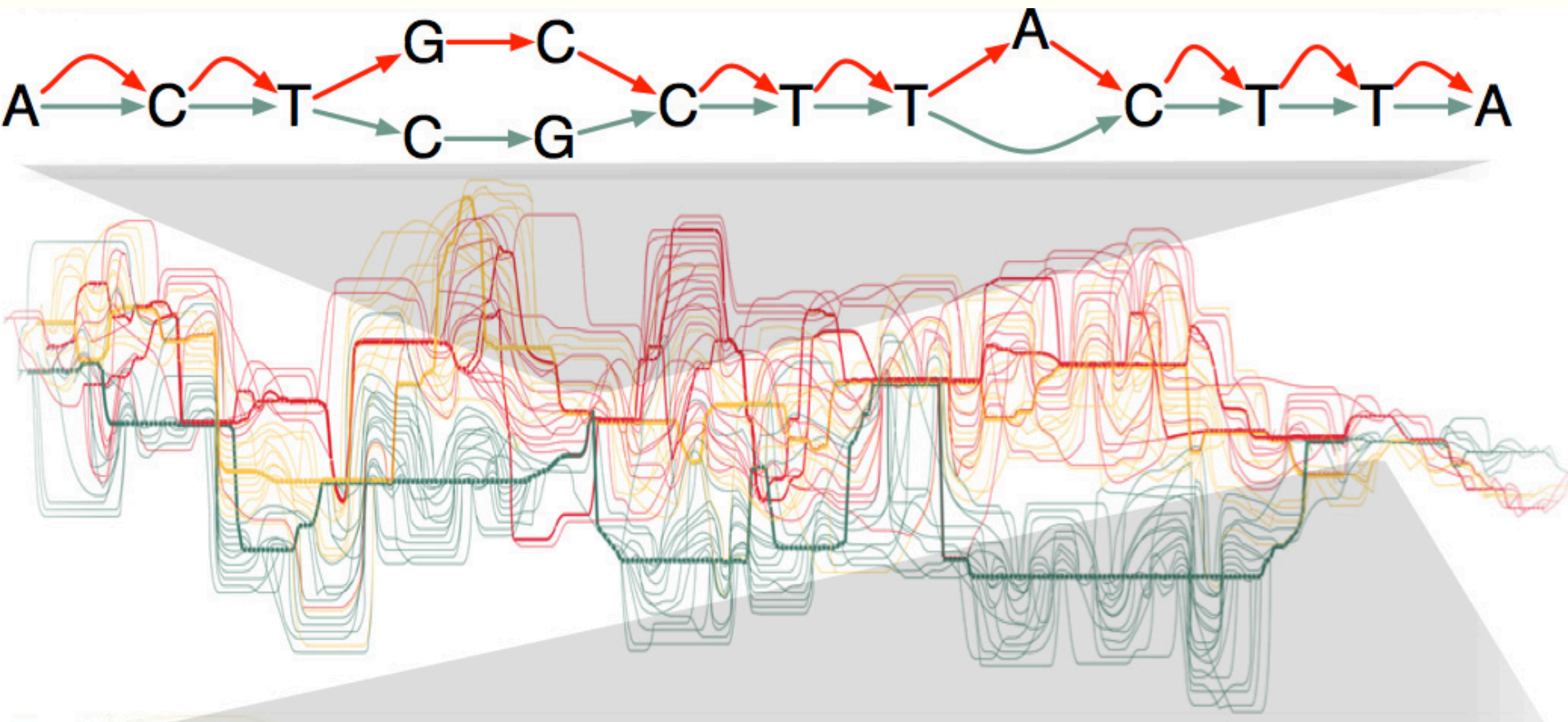
# E' un grafo...pan-genoma

Il grafo rappresenta più sequenze o genomi:



Esempio: dopo una **T** può esserci una **G** oppure una **C**

# E' un grafo...pan-genoma





# IL Viaggio continua...

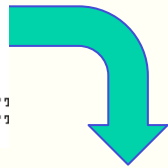
**genoma è una sequenza?**

Unaligned sequences

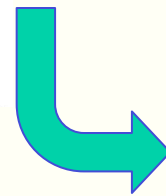
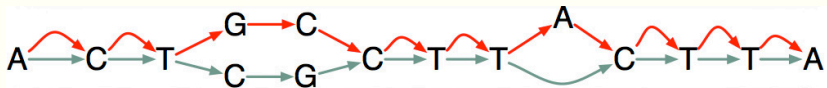
|            |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|------------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| human      | A | C | A | T | A | T | G | G | A | C | A | G | T | A | A | A | A | A | C | A | T | A |
| Chimpanzee | A | C | A | T | A | T | G | G | A | C | A | G | T | A | A | A | A | A | C | A | T | A |
| Vacaque    | A | T | A | C | A | T | T | A | C | G | C | A | G | T | A | A | A | A | C | A | T | A |

Aligned sequences

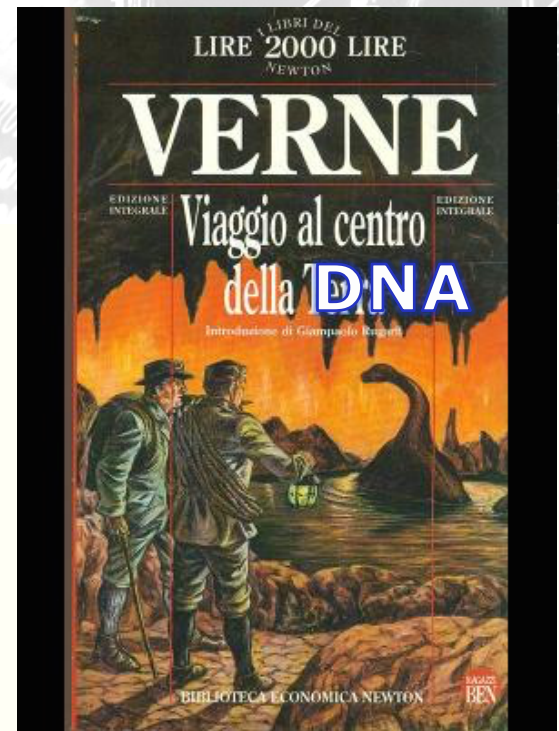
|            |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|------------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| human      | A | C | A | T | A | T | G | G | A | C | A | G | T | A | A | A | A | A | C | A | T | A |
| Chimpanzee | A | C | A | T | A | T | G | G | A | C | A | G | T | A | A | A | A | A | C | A | T | A |
| Vacaque    | A | T | A | C | A | T | T | A | C | G | C | A | G | T | A | A | A | A | C | A | T | A |



**grafo - Pan-genome**

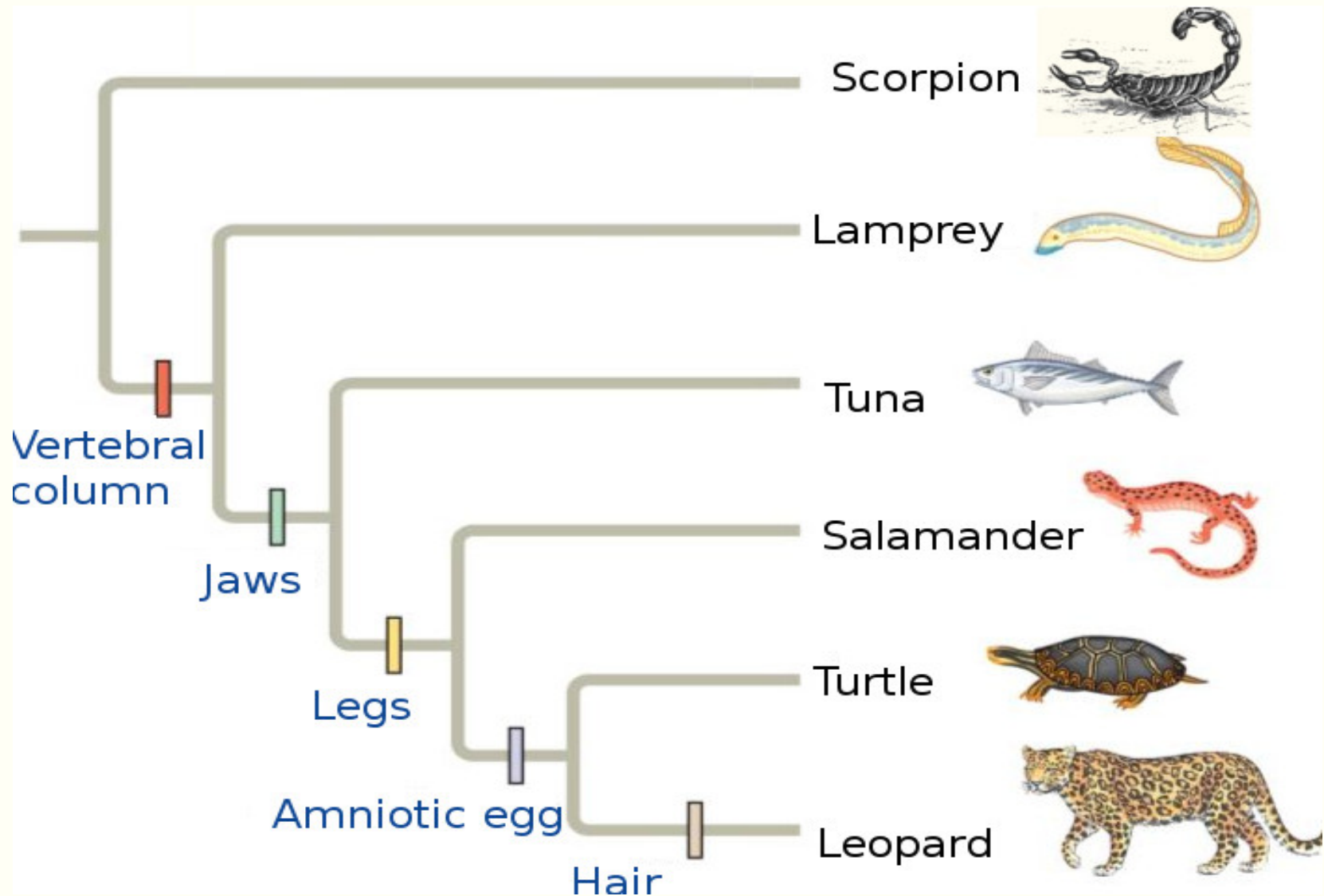


**+ algoritmi .. ➡ Terapia ad hoc**

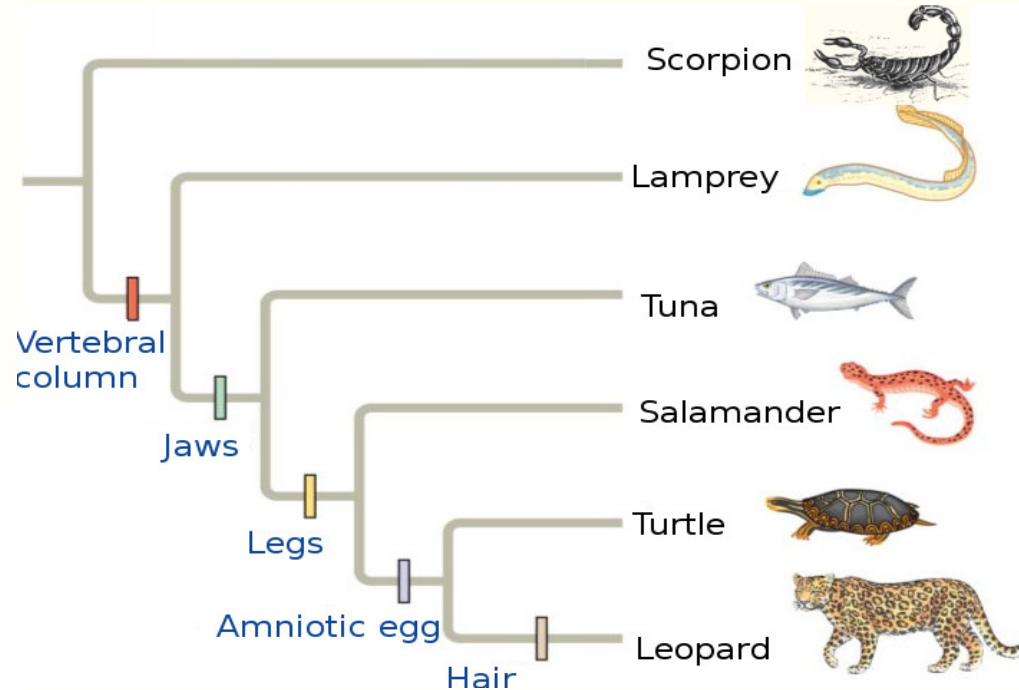




# algoritmi su alberi (evoluzione)



# algoritmi su alberi (evoluzione)

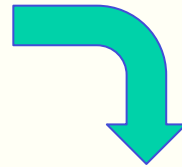


|            | A | J | H | L | V |
|------------|---|---|---|---|---|
| Scorpion   | 0 | 0 | 0 | 0 | 0 |
| Lamprey    | 0 | 0 | 0 | 0 | 1 |
| Tuna       | 0 | 1 | 0 | 0 | 1 |
| Salamander | 0 | 1 | 0 | 1 | 1 |
| Turtle     | 1 | 1 | 0 | 1 | 1 |
| Leopard    | 1 | 1 | 1 | 1 | 1 |

V = vertebral column    L = legs    A = amniotic egg    H = hair  
J = jaws

# Una tappa del viaggio ...

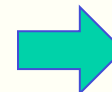
cause genetiche + **sintomi** +  
**stile di vita**  
= **diverso trattamento**



## Esempio: Fibrosi cistica



F508del genotype

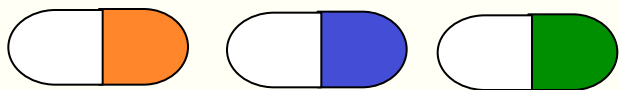


Varietà di  
risposta al  
farmaco



# Medicina di precisione

pazienti con variazione genetica:



farmaco target per variazione

## UNDERSTANDING PRECISION MEDICINE

In precision medicine, patients with tumors that share the same genetic change receive the drug that targets that change, no matter the type of cancer.

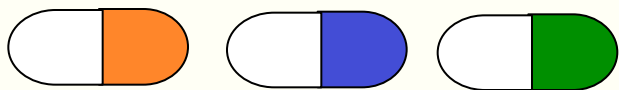


Using the genetic changes in a patient's tumor to determine their treatment is known as precision medicine.

Credit: National Cancer Institute

# Medicina di precisione

pazienti con variazione genetica:



farmaco target per variazione

## UNDERSTANDING PRECISION MEDICINE

In precision medicine, patients with tumors that share the same genetic change receive the drug that targets that change, no matter the type of cancer.



Using the genetic changes in a patient's tumor to determine their treatment is known as precision medicine.

Credit: National Cancer Institute

# pan-genomica informatica: obiettivi

- rappresentare e costruire genomi di popolazione:

- confrontare pan-genomi per:

determinare variazioni geniche,  
**associare** variazioni a patologie,  
studiare **evoluzione** degli individui

Algoritmi e strutture dati per grafi di pan-genomi



# pan-genoma: le strutture dati

Sequenze ripetute o simili  $\approx$  **terabytes** =  $10^{12}$   
?  
memorizzazione, indicizzazione



# pan-genoma: gli algoritmi e le strutture dati

Sequenze ripetute o simili  $\approx$  **terabytes** =  $10^{12}$

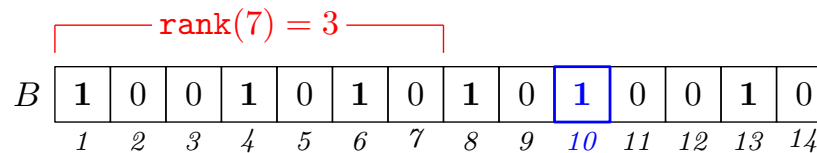
?

memorizzazione, indicizzazione

Algoritmi su  
strutture dati **comprese**

| Pos | Suffix    | BWT | GSA    | LCP |
|-----|-----------|-----|--------|-----|
| 0   | \$        | T   | (0, 1) | 0   |
| 1   | \$        | C   | (0, 2) | 0   |
| 2   | \$        | A   | (0, 3) | 0   |
| 3   | \$        | G   | (0, 4) | 0   |
| 4   | A\$       | A   | (1, 2) | 0   |
| 5   | AA\$      | C   | (2, 2) | 1   |
| 6   | AAGG\$    | C   | (4, 3) | 2   |
| 7   | ACGTGT\$  | \$  | (6, 0) | 1   |
| 8   | AGG\$     | A   | (3, 3) | 1   |
| 9   | C\$       | C   | (1, 1) | 0   |
| 10  | CAAS\$    | C   | (3, 2) | 1   |
| 11  | CAAGG\$   | C   | (3, 1) | 3   |
| 12  | CC\$      | T   | (2, 1) | 1   |
| 13  | CCAA\$    | T   | (4, 2) | 2   |
| 14  | CCAAGG\$  | \$  | (6, 3) | 4   |
| 15  | CGTGT\$   | A   | (5, 0) | 1   |
| 16  | G\$       | G   | (1, 3) | 0   |
| 17  | GG\$      | A   | (2, 3) | 1   |
| 18  | GT\$      | T   | (2, 0) | 1   |
| 19  | GTCC\$    | T   | (4, 1) | 2   |
| 20  | GTCCAAS\$ | \$  | (6, 2) | 4   |
| 21  | GTGT\$    | C   | (4, 0) | 2   |
| 22  | GTGTCC\$  | \$  | (6, 1) | 4   |
| 23  | T\$       | G   | (1, 0) | 0   |
| 24  | TCC\$     | G   | (3, 1) | 1   |
| 25  | TCCAAS\$  | G   | (5, 2) | 3   |
| 26  | TGT\$     | G   | (3, 0) | 1   |
| 27  | TGTCC\$   | G   | (5, 1) | 3   |

Efficient data structure used to store a set of bits



select(5) = 10



# pan-genoma: un' esplosione di dati

"L' European Bioinformatics Institute (EBI) in UK memorizza **20 petabyte** (1 petabyte =  $10^{15}$  bytes) di **dati biologici**"  
... acceleratore CERN produce **15 petabytes** di **dati fisici**  
tratto da *The big challenge of big data* 13 JUNE 2013 -**NATURE** | 255

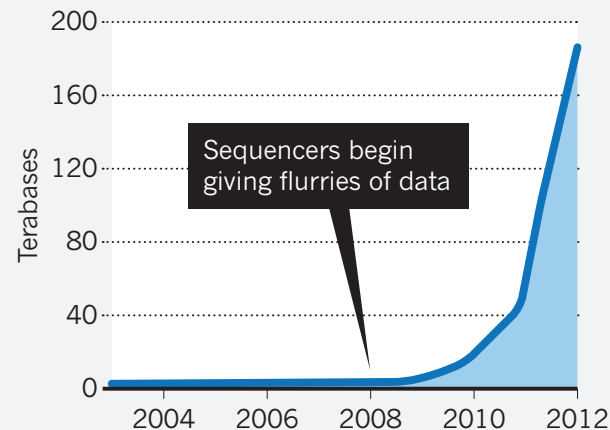


**CERN: RICERCA Nucleare**

SOURCE: EMBL-EBI

## DATA EXPLOSION

The amount of genetic sequencing data stored at the European Bioinformatics Institute takes less than a year to double in size.



# Il viaggio in futuro ...

degli  
ALGORIMI



DNA del virus EBOLA?  
Controlla il file...



# Il viaggio in futuro ...

degli  
ALGORIMI



**EBI**



Aspetta che  
calcolo?



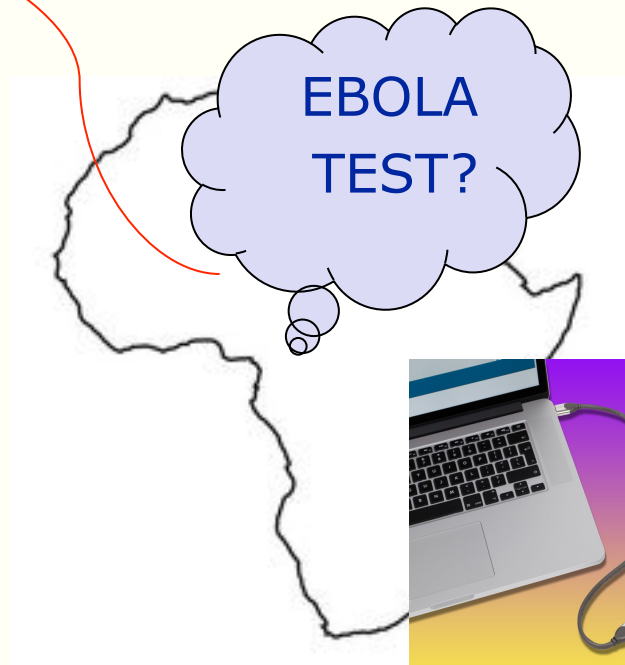
DNA del virus EBOLA?  
Controlla il file...





# Il viaggio in futuro ...

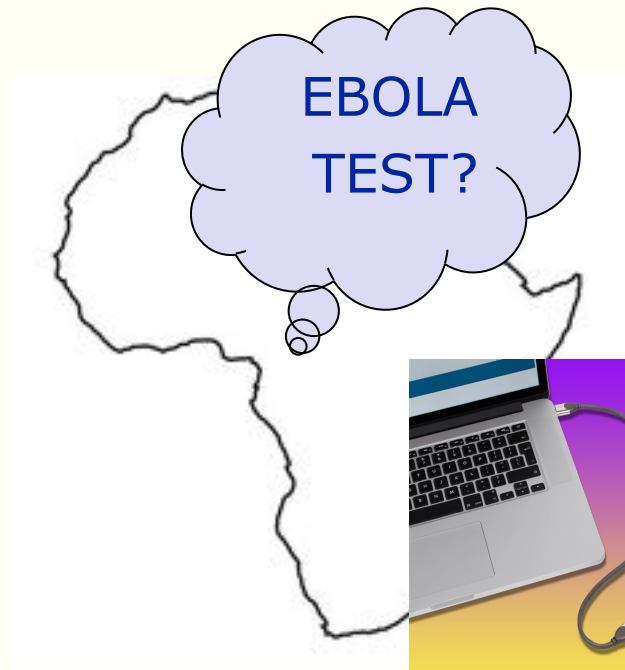
degli  
ALGORIMI





# Il viaggio in futuro ...

degli  
ALGORIMI



# Il viaggio in futuro ...

“GLI ALGORITMI SONO **PIÙ SAGGI E BRAVI**  
DI NOI nel prendere decisioni” – IL PREMIO  
NOBEL DANIEL KAHNEMAN

**PIÙ VELOCI ED EFFICIENTI**

