Corso di bioinformatica 1

(DEB, DIPARTIMENTO DI SCIENZE ECOLOGICHE E BIOLOGICHE, a.a. 2023-2024)

Organizzazione del corso e introduzione

Tiziana Castrignanò
tiziana.castrignano@unitus.it

Organizzazione del corso: Docente

Docente: Tiziana Castrignanò

e-mail: tiziana.castrignano@unitus.it

Dove trovarmi: Dip. DEB, Blocco C, Primo piano, Stanza in fondo al corridoio.

Ricevimento: scrivete una mail per fissare un appuntamento.

Le slide: saranno disponibili su drive il giorno della lezione.

Organizzazione del corso: libro di testo consigliato

- **Fondamenti di Bioinformatica**
- **Citterich, Ferrè, Pavesi, Romualdi, Pesole**
- **Editore: Zanichelli**
- E’ il testo consigliato.
- Per l'esame è necessario utilizzare sia le slide sia il libro di testo.
- E’ utile per chiarire o approfondire concetti introdotti nella lezione.

Organizzazione del corso: esami

- L'esame finale di Bioinformatica è orale

- Voto finale: è il risultato dell'interrogazione orale: 2 domande sul programma svolto a scelta del docente + la presentazione di un progetto che utilizzi gli strumenti usati nell'esercitazioni o nuovi strumenti bioinformatici.

- 9 Appelli in tutto: 3 appelli a giugno/luglio, 2 a settembre, 1 a novembre, 2 a gennaio/febbraio + 1 appello ad aprile (solo per laureandi maggio/giugno e fuori corso).

nature.com

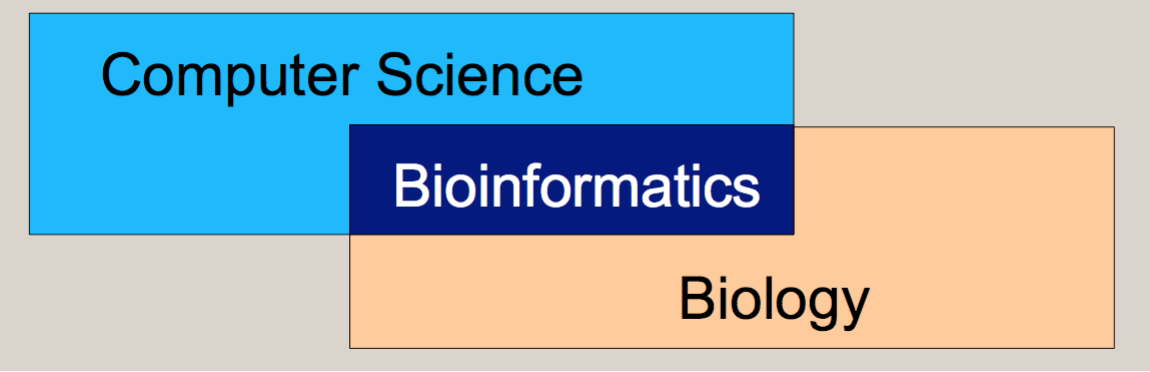
Bioinformatica

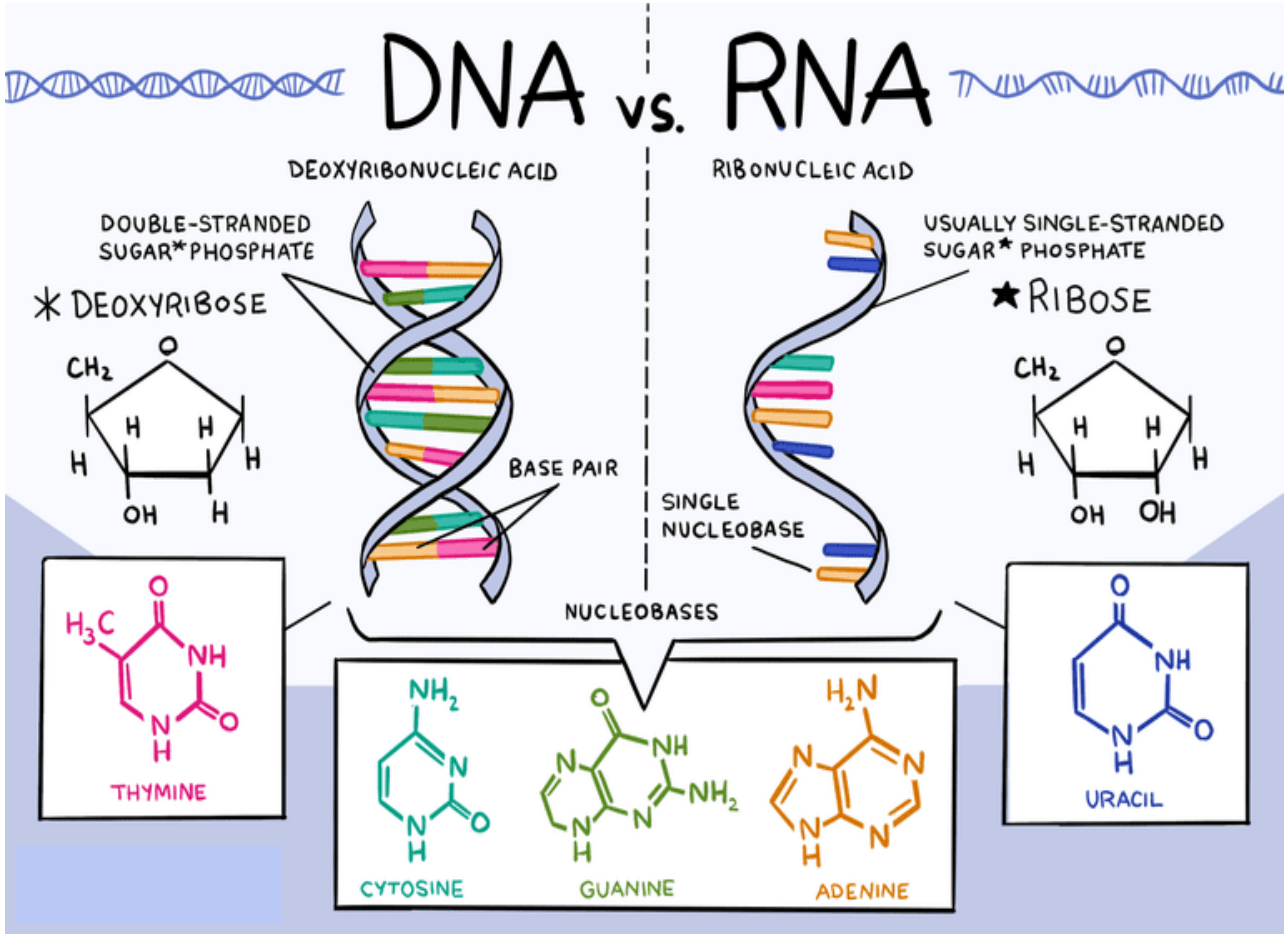
La bioinformatica è un campo di ricerca che usa il calcolo per estrarre conoscenza dai dati biologici.

La bioinformatica include la raccolta, l'archiviazione, il recupero, la manipolazione e la modellizzazione dei dati per l'analisi, la visualizzazione o la predizione attraverso lo sviluppo di algoritmi e software.

Di quali entità biologiche si occupa la bioinformatica?

Applicare le tecnologie e i metodi dell'informatica all'analisi, comprensione, organizzazione e conservazione di dati biologici, in particolare molecolari. Spesso si occupa di quelle molecole biologiche che per loro natura contengono, trasmettono e manipolano "informazione": **gli acidi nucleici (DNA, RNA) e le proteine.**

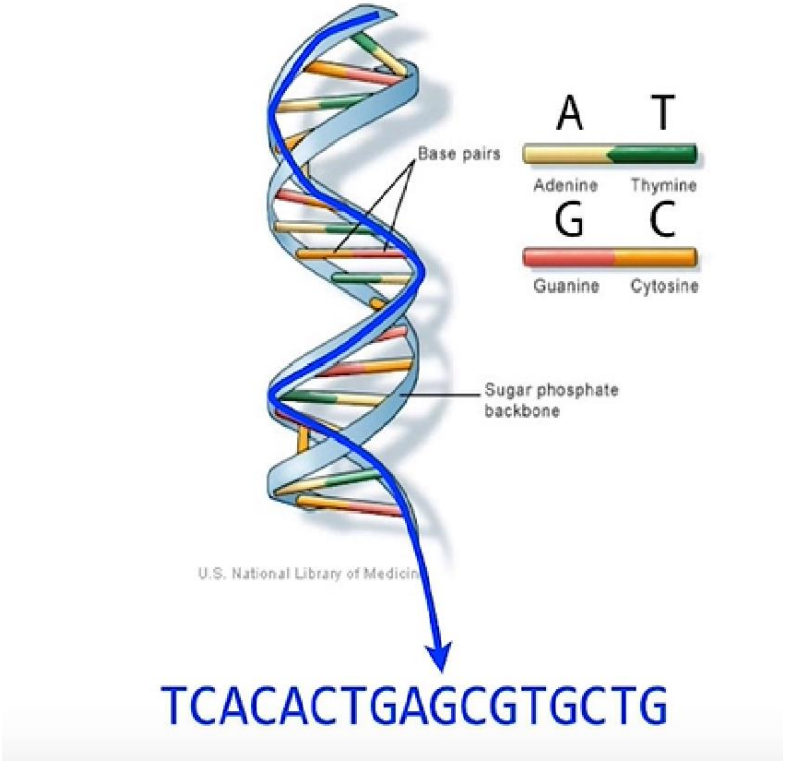



DNA vs. RNA
DEOXYRIBONUCLEIC ACID
RIBONUCLEIC ACID
DOUBLE-STRANDED
SUGAR* PHOSPHATE
* DEOXYRIBOSE
BASE PAIR
USUALLY SINGLE-STRANDED
SUGAR* PHOSPHATE
★ RIBOSE
SINGLE
NUCLEOBASE
NUCLEOBASES
THYMINE
CYTOSINE
GUANINE
ADENINE
URACIL

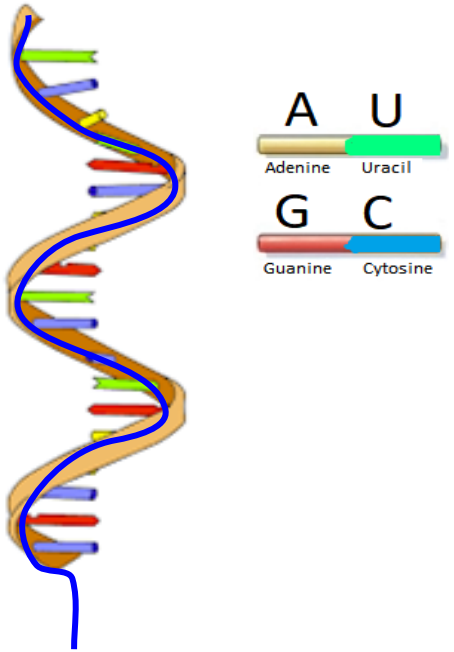
REMEMBER

Che cosa sono il DNA e l'RNA per un bioinformatico che studia l'analisi di sequenze?

DNA



RNA

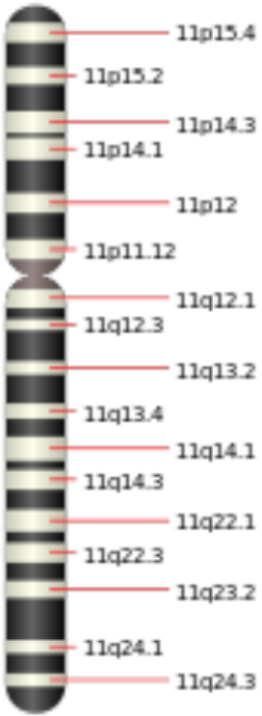


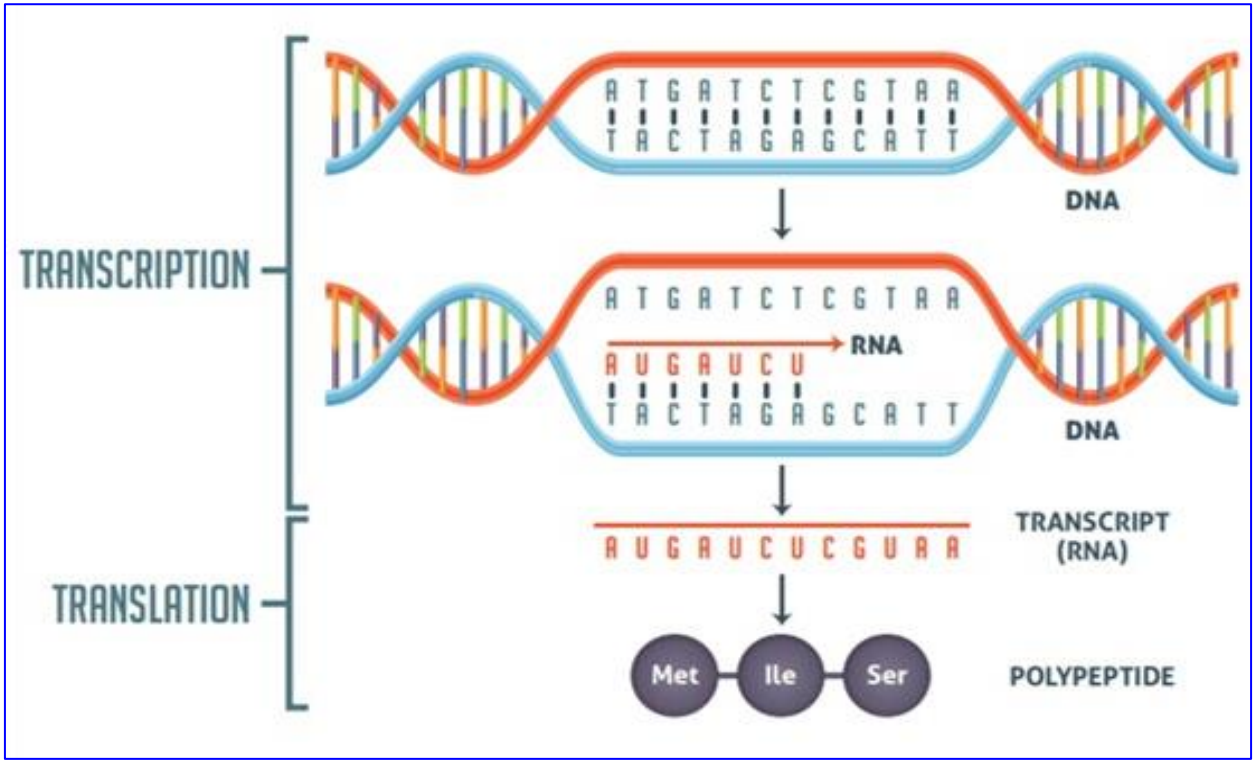
UCAUGCAUGUCCUGAACGC

Di quante lettere (nucleotidi) sono composti gli alfabeti del DNA e dell'RNA? Alfabeti a 4 lettere

Esempio di sequenza di un tratto di molecola di DNA (composta dalla successione dei 4 nucleotidi A,T,G,C):

```
atggcaattaaaattggtatcaatggttttggtcgtatcggccgtatcgtattccgtgca
gcacaacaccgtgatgacattgaagttgtaggtattaacgacttaatcgacgttgaatac
atggcttatatgttgaaatatgattcaactcacggtcgtttcgacggcactgttgaagtg
aaagatggtaacttagtggttaatggtaaaactatccgtgtaactgcagaacgtgatcca
gcaaacttaaactggggtgcaatcggtgttgatatcgctgttgaagcgactggtttattc
ttaactgatgaaactgctcgtaaacatatcactgcaggcgcaaaaaaagttgtattaact
ggcccatctaaagatgcaacccctatgttcgttcgtggtgtaaacttcaacgcatacgca
ggtcaagatatcgtttctaacgcatcttgtacaacaaactgtttagctcctttagcacgt
caaaaatagggttaatatgaatctcgatctccatttttgttcatcgtattcaacaacaagc
caaaactcgtacaaatatgaccgcacttcgctataaagaacacggcttcgagatatctct
tggaaaaactttcaagagcaactcaatcaactttctcgagcattgcttgctcacaatatt
gacgtacaagataaaatcgccatttttgcccataatatggaacgttgggttgttcatgaa
actttcggtatcaaagatggtttaatgaccactgttcacgcaacgactacaatcgttgac
attgcgaccttacaaattcgagcaatcacagtgcctatttacgcaaccaatacagcccag
caagcagaatttatcctaaatcacgccgatgtaaaaattctcttcgtcggcgatcaagag
caatacgatcaaacattggaaattgctcatcattgtccaaaattacaaaaaattgtagca
atgaaatccaccattcaattacaacaagatcctctttcttgcacttgg
```





Sequenza proteica → MKNFNTLSFETLANIVGGRNNWAANIGGVGG

- **Di quante lettere (amino acidi) è composto l'alfabeto della sequenza polipeptidica?**

1st base	2nd base								3rd base
	U		C		A		G		
U	UUU	Phenylalanine (F)	UCU	Serine (S)	UAU	Tyrosine (Y)	UGU	Cysteine (C)	U
	UUC		UCC		UAC		UGC		C
	UUA	Leucine (L)	UCA		UAA	stop codon	UGA	stop codon	A
	UUG		UCG		UAG		UGG	Tryptophan (W)	G
C	CUU		CCU	Proline (P)	CAU	Histidine (H)	CGU	Arginine (R)	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Glutamine (Q)	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Isoleucine (I)	ACU	Threonine (T)	AAU	Asparagine (N)	AGU	Serine (S)	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lysine (K)	AGA	Arginine (R)	A
	AUG	Methionine (M) & start codon	ACG		AAG		AGG		G
G	GUU	Valine (V)	GCU	Alanine (A)	GAU	Aspartic Acid (D)	GGU	Glycine (G)	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glutamic Acid (E)	GGA		A
	GUG		GCG		GAG		GGG		G

The genetic code is redundant!!!

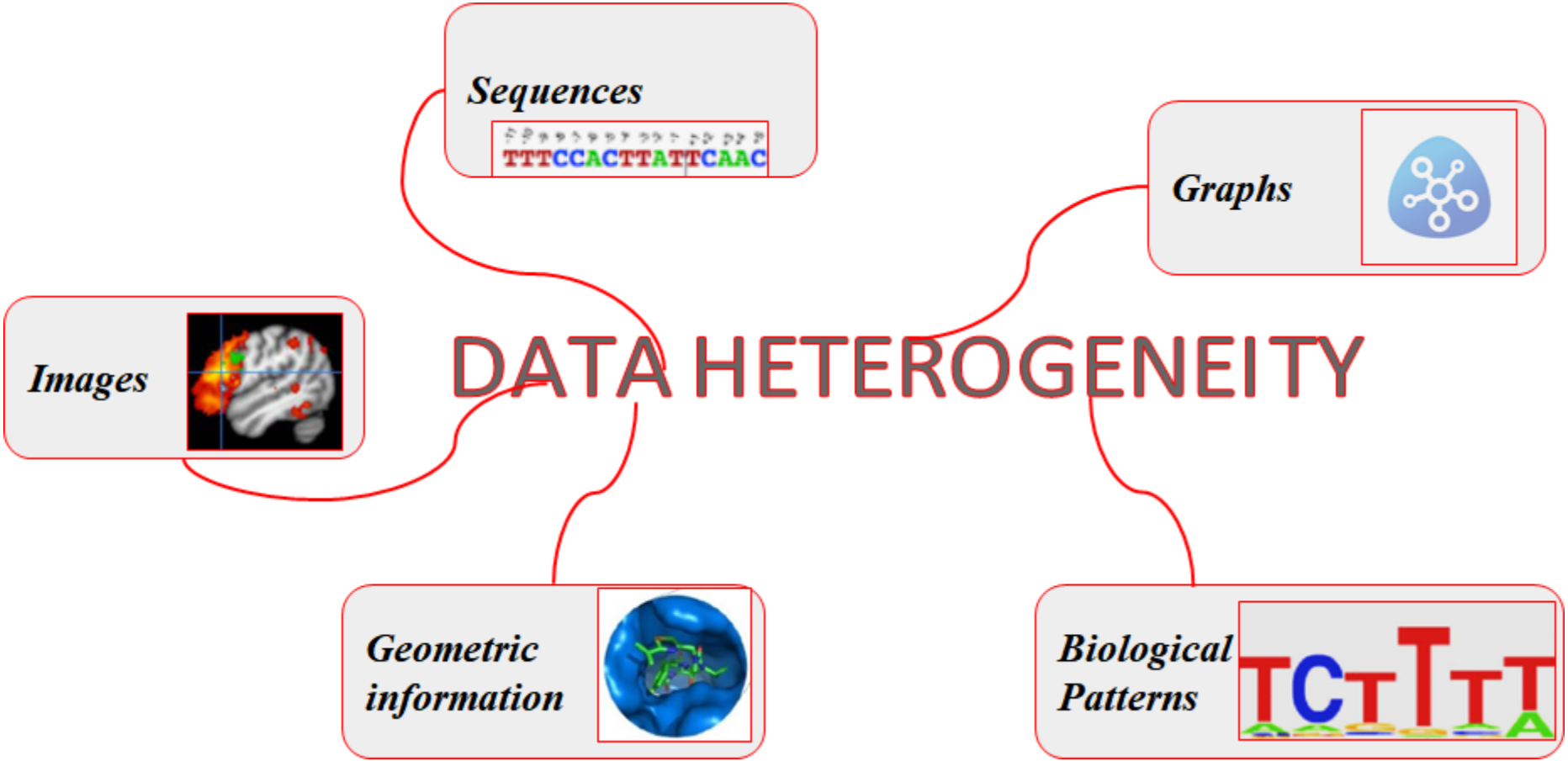
- How many possible codons do we have in the genetic code?
  (4)^3 = 4*4*4 = 64 codons
- How many different amino acids are normally found in proteins?
  20 amino acids

Amino Acid Code

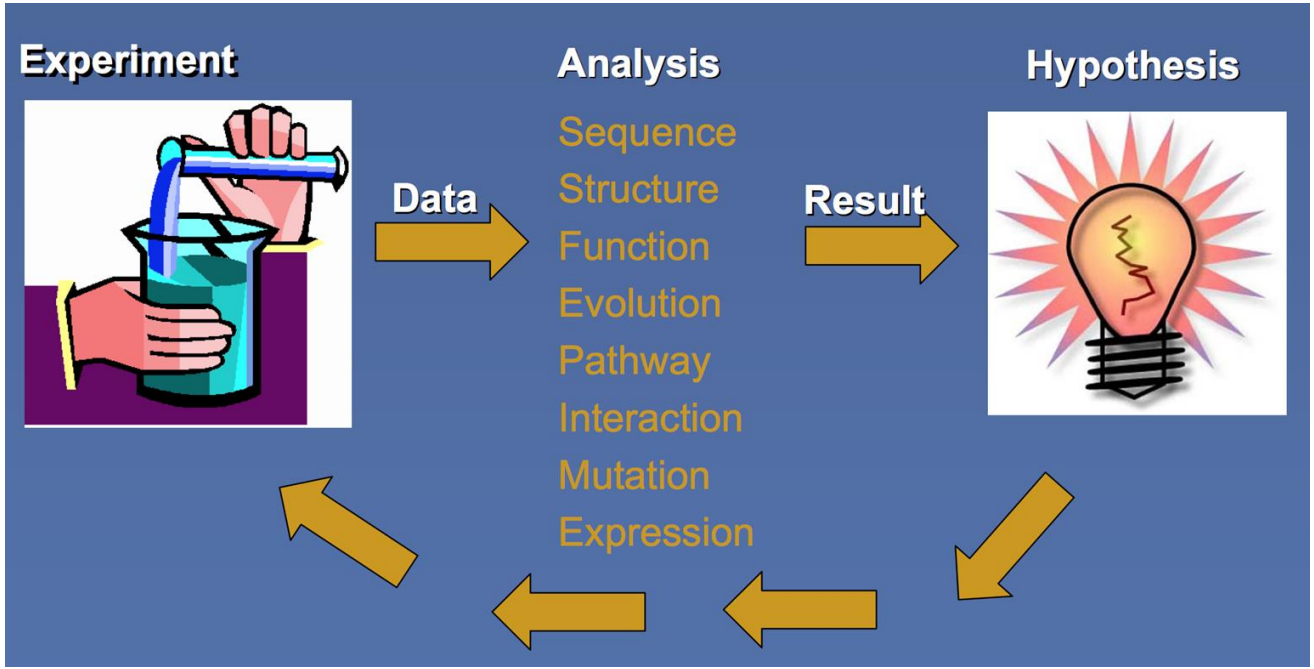
Protein sequence

MKNFNTLSFETLANIVGGRNNWAANIGGVGG

Oltre ai dati di sequenze ci sono altri tipi di dati...ad esempio:



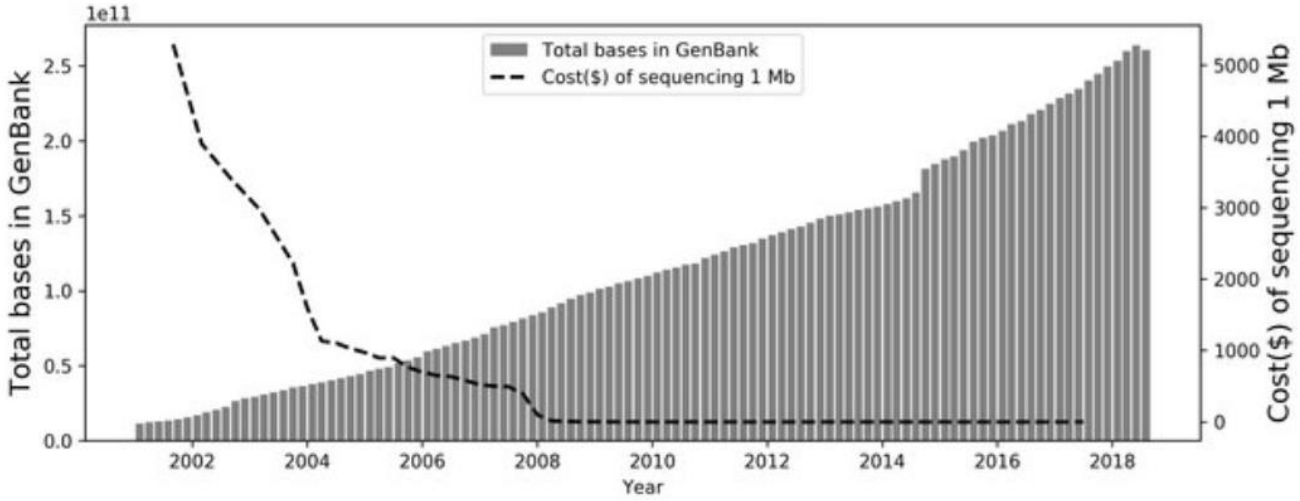
La bioinformatica



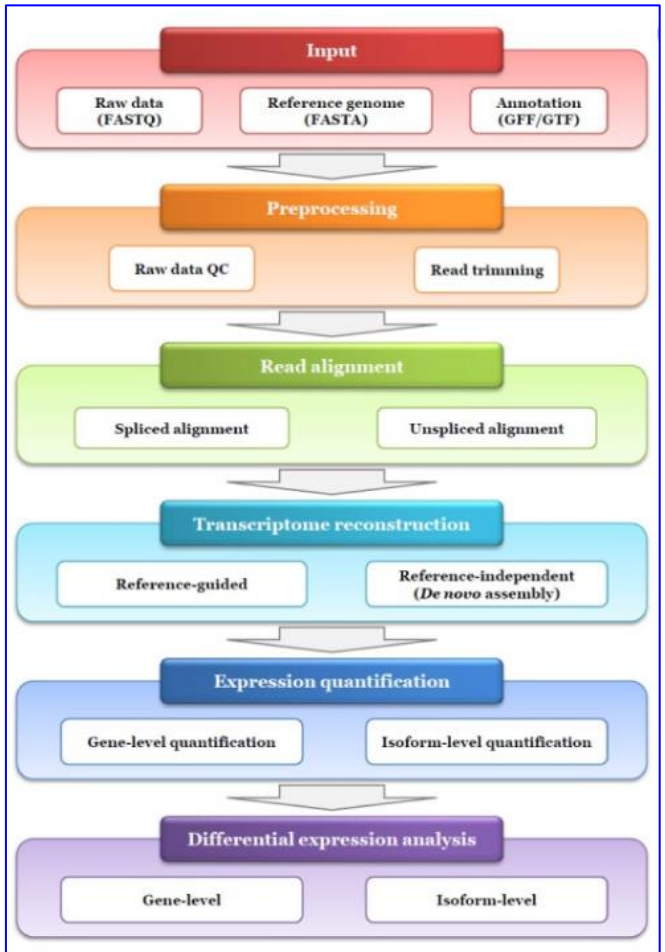
Alcuni compiti della bioinformatica..

- ✓ Organizzare i dati biologici in banche dati che siano ben strutturate, facilmente interrogabili, interoperabili e che crescano in maniera ordinata al crescere della mole di informazioni.

- ❑ La prima "sequenza" biologica risale al 1956, l'insulina bovina composta da 51 amminoacidi.



- ❑ Andamento in *size* dei dati di sequenze dentro genbank accessibile dal portale dell' ncbi (https://www.ncbi.nlm.nih.gov/genbank/) e costo del sequenziamento per ogni Mb.



Es. di pipeline di analisi di raw data di dati di RNAseq

Alcuni compiti della bioinformatica..

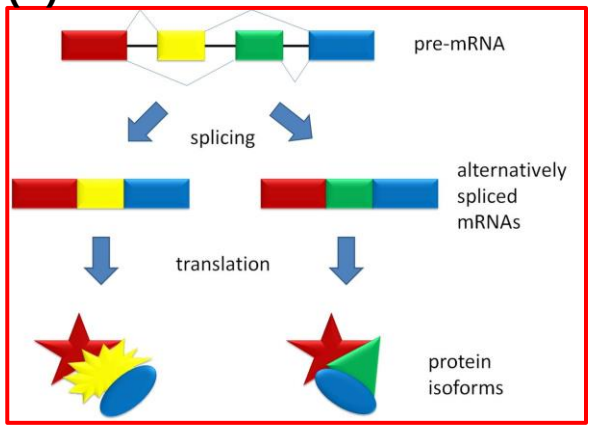
- ✓ Sviluppare strumenti di analisi (spesso pipeline bioinformatiche complesse) che consentano di integrare, visualizzare ed esplorare i dati, estraendo da essi la maggior quantità possibile di informazioni "utili".

  - ❑ Ad es. ho la sequenza di un genoma e i trascritti da esso prodotti. Posso identificare la posizione dei geni su quel genoma (ovvero le porzioni di genoma che producono quei trascritti)?

Per far questo è necessario avere una pipeline di analisi di sequenze che permetta di effettuare il mapping dei trascritti sul genoma di riferimento.

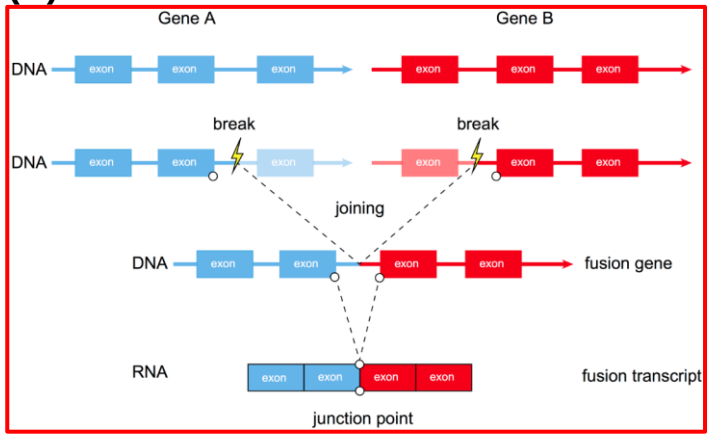
Alcuni compiti della bioinformatica..

(A)

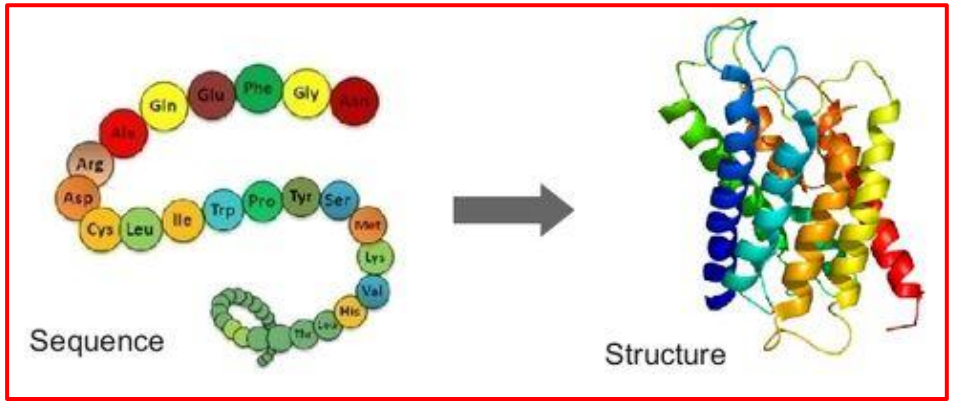


- Utilizzare i dati noti per "guidare" il ricercatore "wet" nei suoi esperimenti: ad es. predire, nei geni, (A) eventi di splicing alternativo o (B) di fusione genica o, nelle proteine, (C) la loro struttura secondaria o terziaria a partire dalla struttura primaria, etc…

(B)

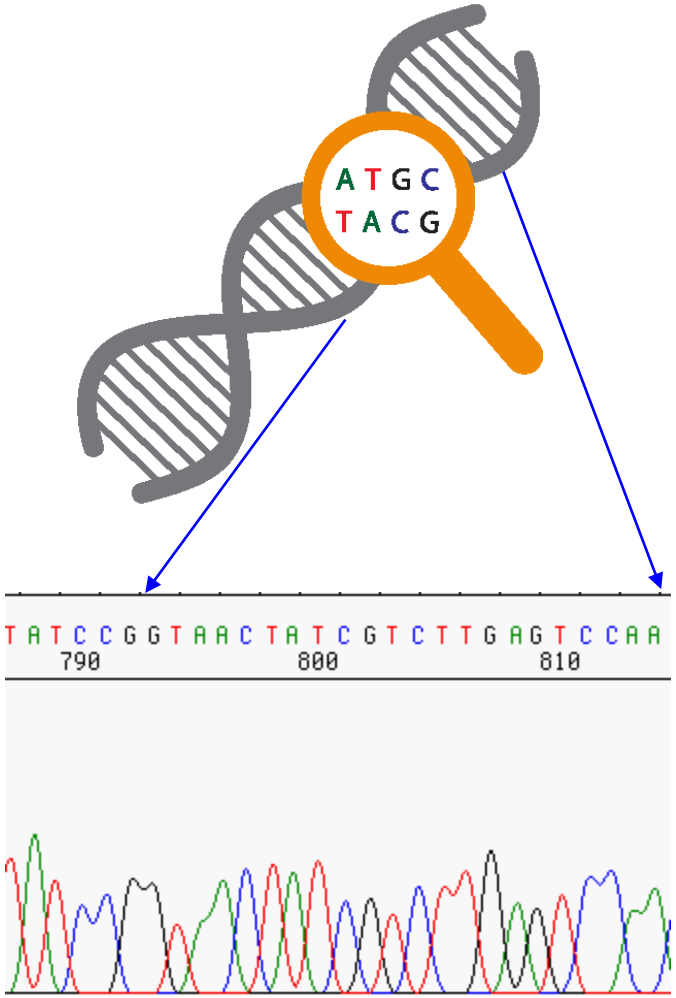


(C)



Cosa è il sequenziamento del DNA?

Sequenziamento del DNA significa determinazione della sequenza lineare delle basi che lo compongono, quindi "leggere" l'ordine in cui *A*, *T*, *C* e *G* sono disposte lungo il DNA.



Le tecnologie più usate per il sequenziamento.

Attualmente i due metodi piu' popolari di sequenziamento sono:

- il sequenziamento Sanger (**Sanger sequencing**)
- il sequenziamento di nuova generazione (**Next Generation Sequencing - NGS**)

Sebbene NGS abbia ampiamente superato il sequenziamento di Sanger, oggi è ancora utilizzato dai ricercatori per la sua semplicità e facilità d'uso.

E' inoltre usato come strumento di validazione dei dati elaborati dal sequenziamento NGS.

Ad esempio se in un determinato esperimento di sequenziamento dell'RNA (RNA-seq) si ottiene una lista di geni differenzialmente espressi in due condizioni sperimentali (ad es. un set di campioni appartenenti al gruppo A e ed altri appartenenti al gruppo B), i geni di particolare interesse per l'esperimento è bene che vengano validati con una seconda tecnica sperimentale (diversa dall'NGS). In questo caso il metodo Sanger è un ottimo candidato per la validazione dei geni di interesse.

Le tecnologie più usate per il sequenziamento.

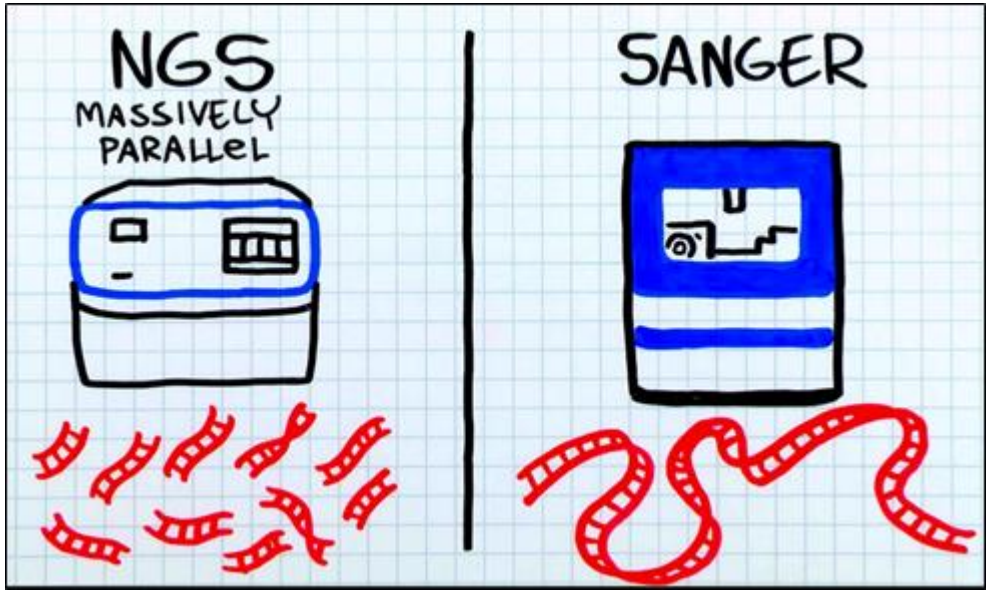
In linea di principio, i concetti alla base delle tecnologie Sanger vs. NGS sono simili.

Sia nel sequenziamento NGS sia nel Sanger, la DNA polimerasi aggiunge nucleotidi fluorescenti uno ad uno su un filamento di modello di DNA in crescita. Ogni nucleotide incorporato è identificato dal suo specifico marcatore fluorescente.

La differenza fondamentale tra il sequenziamento di Sanger e NGS è il volume del sequenziamento.

Le tecnologie più usate per il sequenziamento.

Mentre il metodo Sanger sequenzia solo un singolo frammento di DNA alla volta, NGS è enormemente parallelo, sequenziando milioni di frammenti contemporaneamente per serie. Questo processo ad alta produttività si traduce nel sequenziamento di centinaia o migliaia di geni o trascritti contemporaneamente.

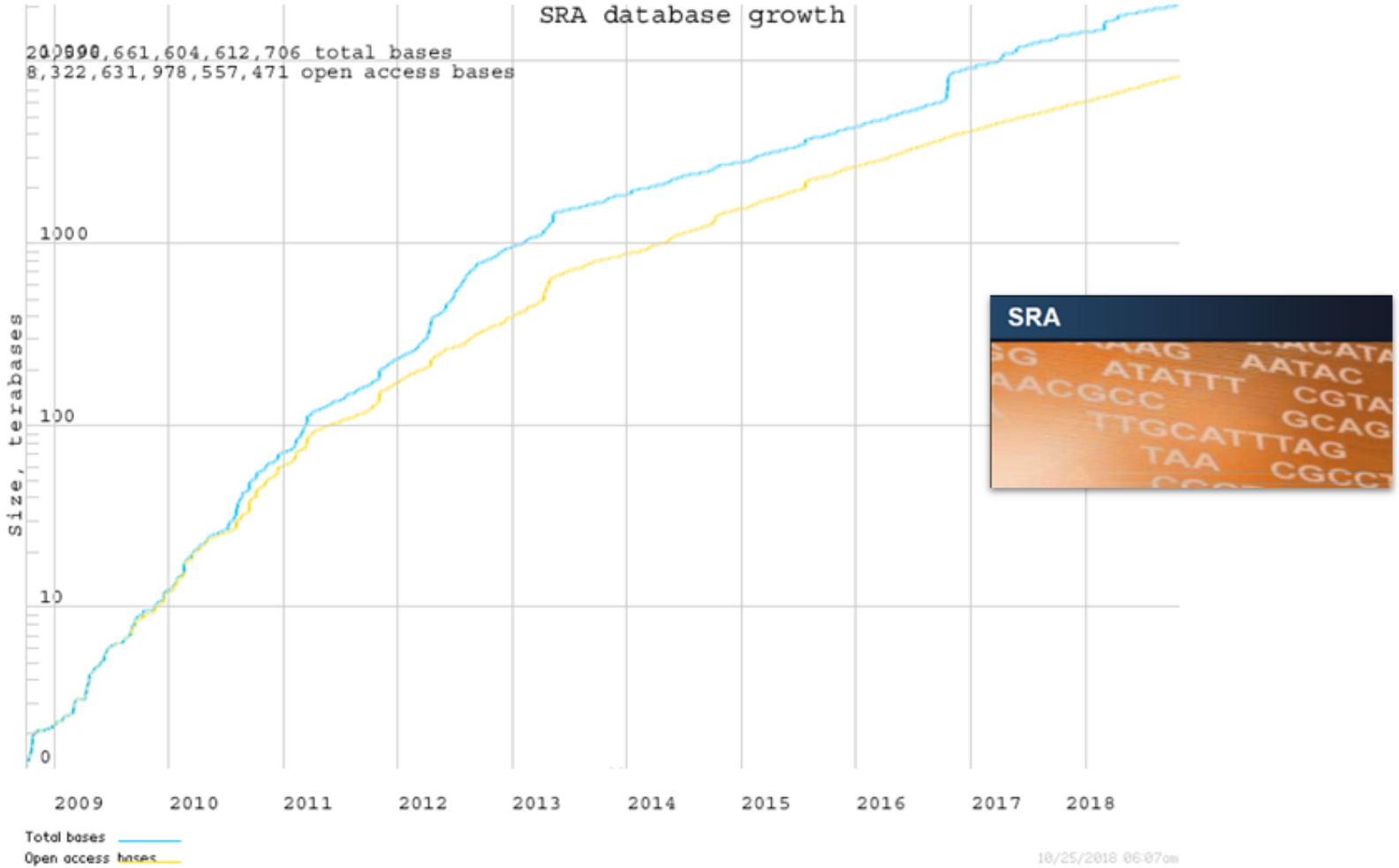


SRA: Sequence Read Archive

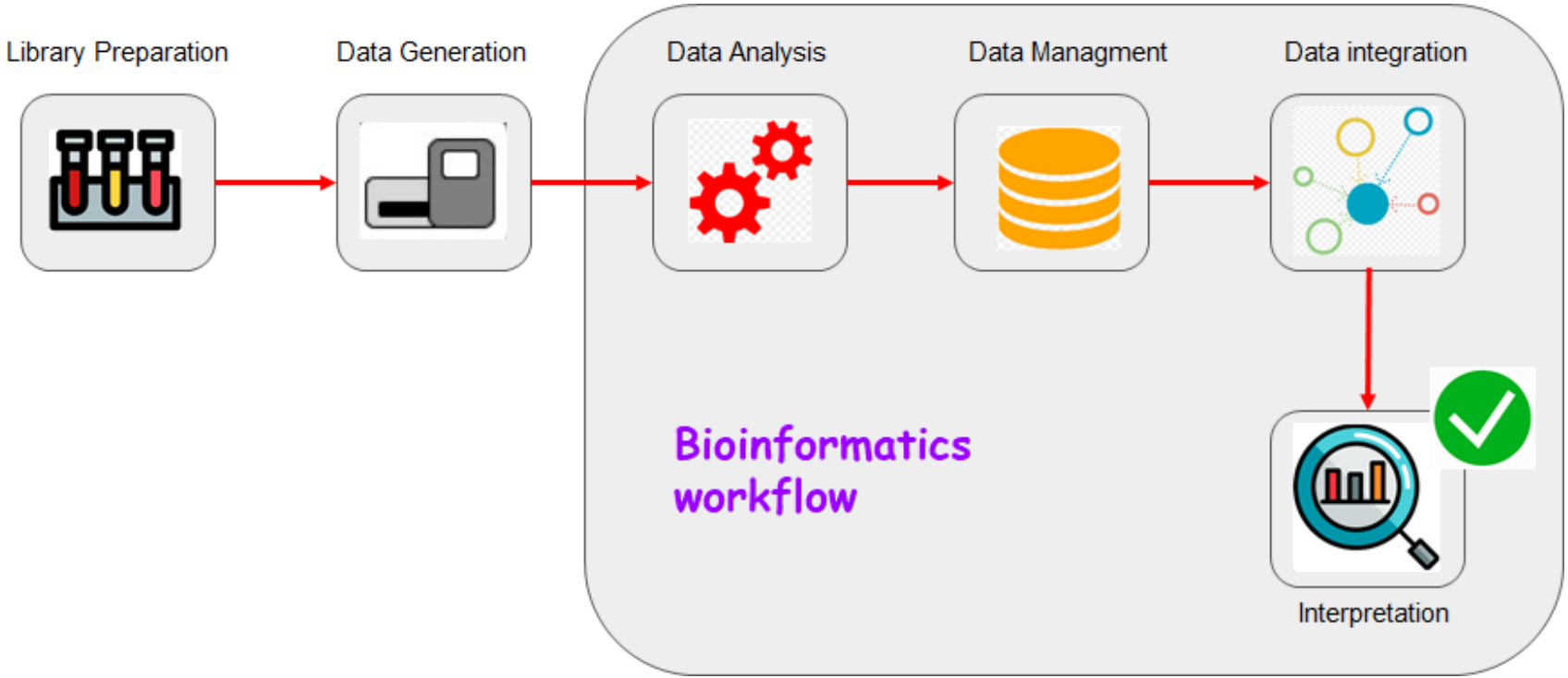
SRA (https://www.ncbi.nlm.nih.gov/sra)

Sequence Read Archive (**SRA**) makes biological sequence data available to the research community **to enhance reproducibility and allow for new discoveries by comparing data sets.**

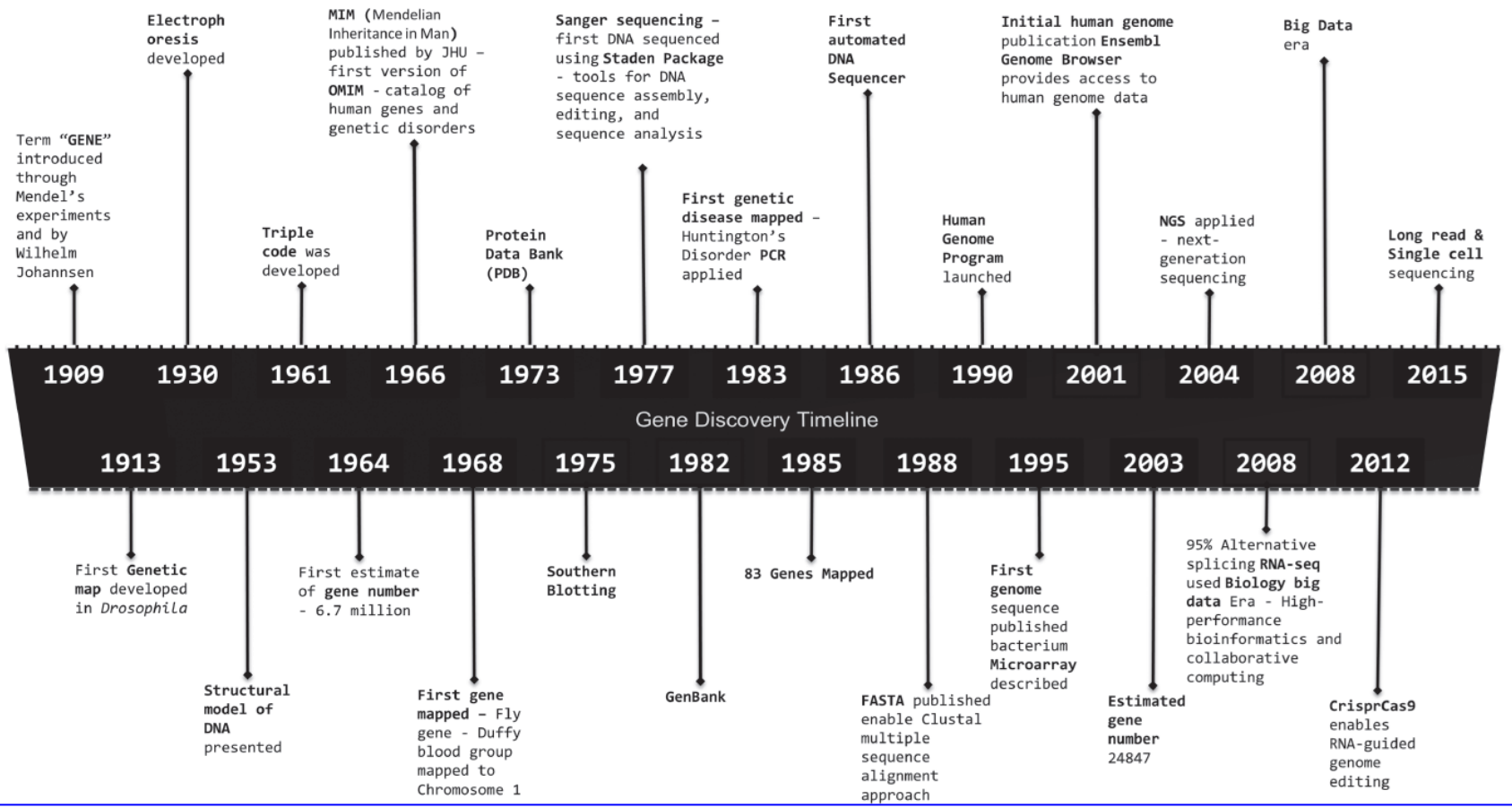
The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®.


SRA database growth
20,890,661,604,612,706 total bases
8,322,631,978,557,471 open access bases
Size, terabases
0
10
100
1000
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
Total bases
Open access bases
10/25/2018 06:07am
SRA

Workflow delle analisi bioinformatiche NGS per la ricerca biomedica



Principali scoperte e invenzioni nel campo della genomica



Saman Zeeshan, Ruoyun Xiong, Bruce T Liang, Zeeshan Ahmed, 100 Years of evolving gene–disease complexities and scientific debutants, *Briefings in Bioinformatics*, https://doi.org/10.1093/bib/bbz038

I passi per arrivare alla sequenza completa del nostro genoma

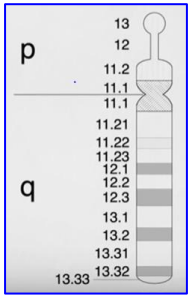
https://www.youtube.com/watch?v=MNFUf8dqk68

- Nel 1990 comincia il progetto “Genoma Umano”

- Nel 1995 sequenziano il primo genoma batterico (Haemophilus Influenzae)

- Nel 1998 fu fondata la Celera Genomics per il sequenziamento del “Genoma umano”

CELERA
A PECorporation Business

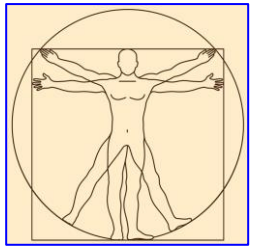


- Nel 1999 viene sequenziato interamente il primo cromosoma umano (Chr 22)

I passi per arrivare alla sequenza completa del nostro genoma

https://www.youtube.com/watch?v=MNFUf8dqk68

- Nel 2000 viene sequenziato il genoma completo di “Drosophila melanogaster”



- Nel 2001 viene sequenziato il primo draft del genoma umano

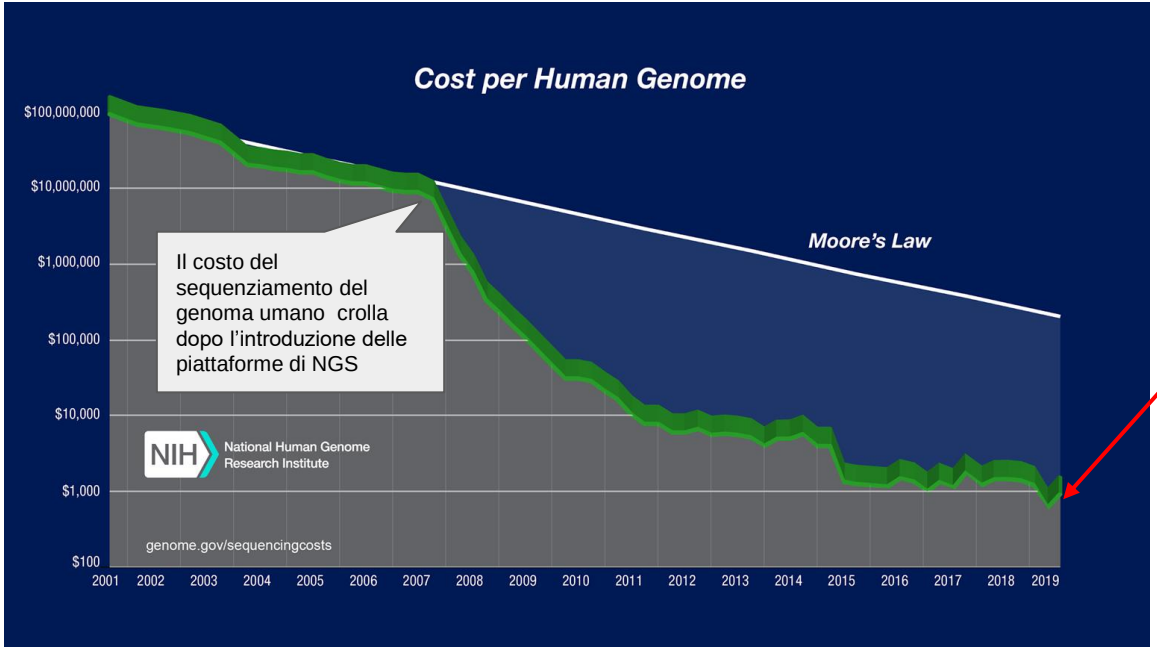
- Nel 2002 viene sequenziato il genoma completo del topo

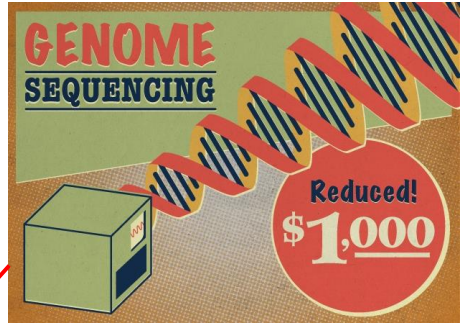
❑ Nel 2003 viene annunciata la sequenza completa del genoma umano

Il progetto genoma umano: 13 anni per arrivare alla sequenza completa del nostro genoma

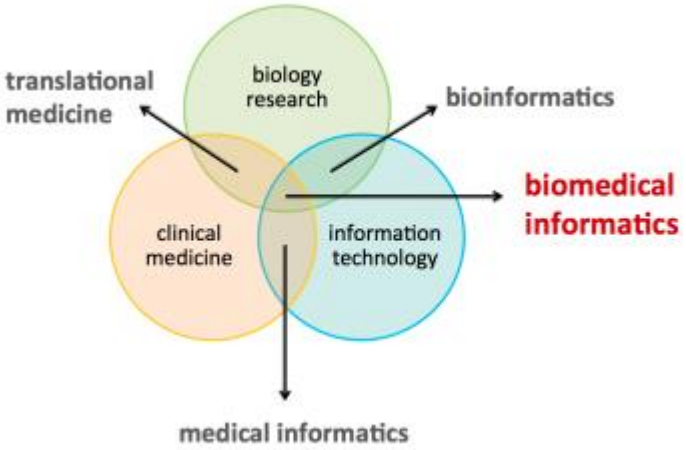
Costo del progetto:
circa **3 000 000 000 $**

La *bioinformatica, oggi*





La *bioinformatica, oggi*



Clinical bioinformatics

Tempo necessario per sequenziare un genoma umano oggi ≈ 24 ore

La riduzione dei tempi e dei costi rende possibile l'approccio bioinformatico anche nel mondo della medicina -> clinical bioinformatics

La *bioinformatica, oggi*

- E' una scienza "pratica", che supporta con innumerevoli applicazioni tutti gli aspetti delle "life science" ed in particolare quelli molecolari.
- Tende a dare risposte "rapide", di solito le analisi *in silico* sono più rapide di quelle *in vitro.*
- Utilizzare metodi informatici è l'unico modo per integrare e organizzare l'enorme mole di informazioni che le moderne tecniche sperimentali producono.
- Tutte le analisi al computer devono essere sempre validate in laboratorio sperimentale.

La *bioinformatica, oggi*

Ieri…

Un gene

- ✓ Dove si trova?
- ✓ Quanti trascritti produce?
- ✓ Per quali proteine codifica?
- ✓ In quali processi funzionali è coinvolto?
- ✓ Qual è la funzione delle proteine codificate? Subiscono modifiche post-traduzionali?
- ✓ Da quali fattori di trascrizione è regolato?
- ✓ Ci sono mutazioni che causano fenotipi particolari?

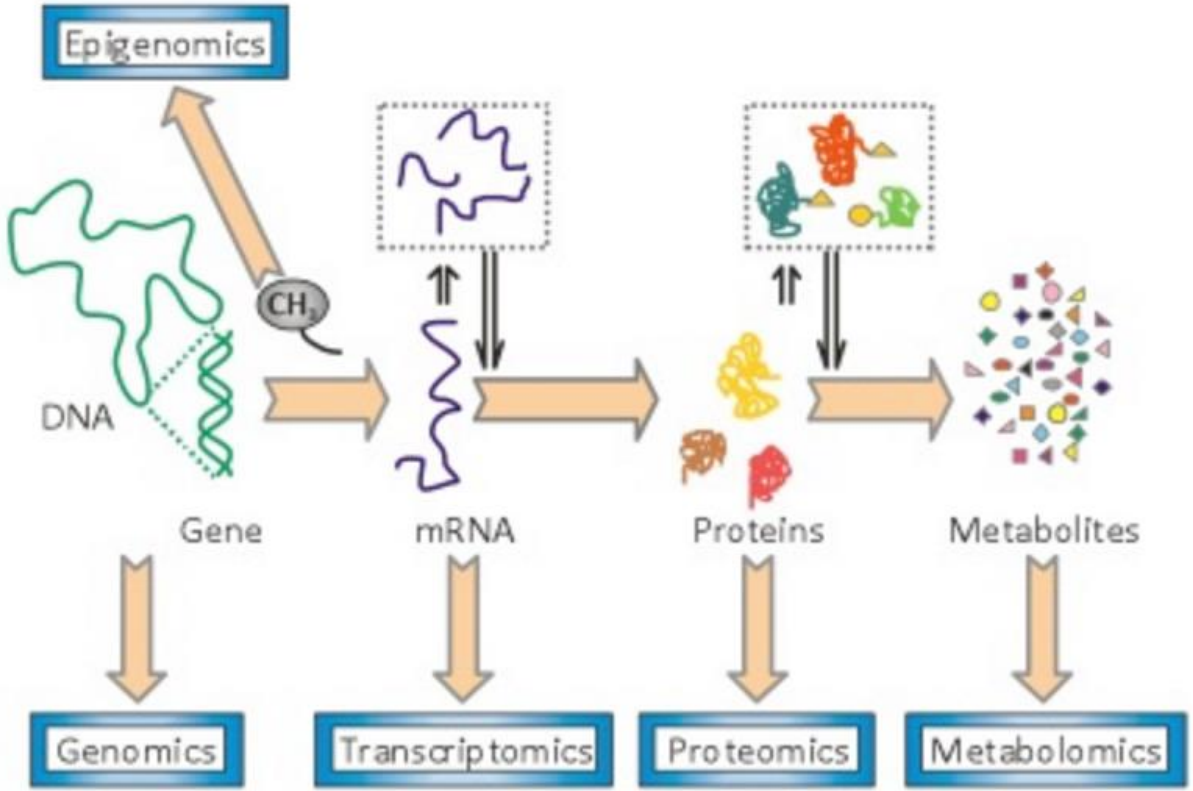
Oggi…

- ✓ Come sono organizzati TUTTI i geni di un genoma? (Genomica)
- ✓ Quali e quanti sono i trascritti che producono (Trascrittomica).
- ✓ E quali le proteine? (Proteomica)
- ✓ Qual è la funzione dei geni e delle proteine da essi codificate (genomica funzionale)
- ✓ Quali le interazioni tra questi e quelle? (Interattomica/System biology)
- ✓ Come sono evoluti i geni e i genomi (Genomica comparata)
- ✓ A quali varianti genomiche è associato un dato fenotipo? (Genomica)

La *bioinformatica, oggi*

- Siamo nell'era delle *–omiche post genomiche*:
  - Trascrittomica
  - Epigenomica
  - Proteomica
  - Metagenomica
  - Genomica comparata
  - Metabolomica
  - Etc..

La *bioinformatica, oggi*



La *bioinformatica, oggi*

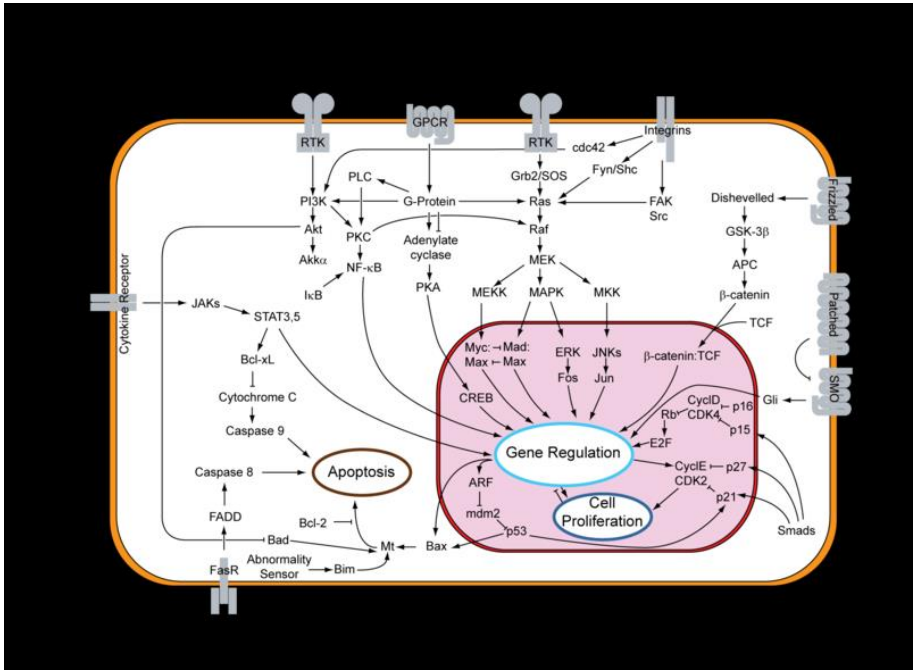
Ieri…

TP53

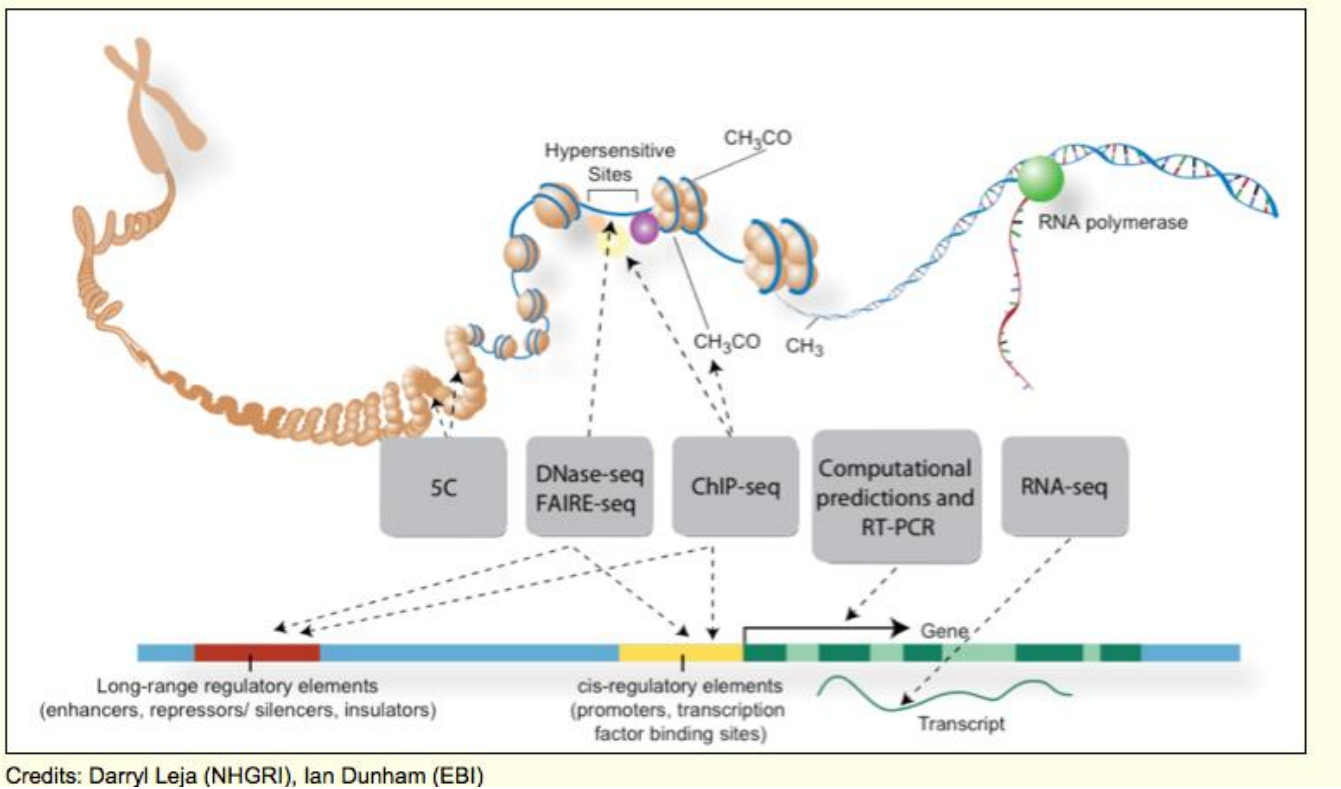
- Braccio corto del cromosoma 17
- Codifica per la proteina P53, un fattore di trascrizione.
- E' una proteina che ha un'importante funzione in caso di danno al DNA, attiva i pathway per la riparazione e blocca il ciclo cellulare.
- In caso il danno sia troppo esteso, induce l'apoptosi.
- E' un oncosoppressore.

Oggi…

TP53



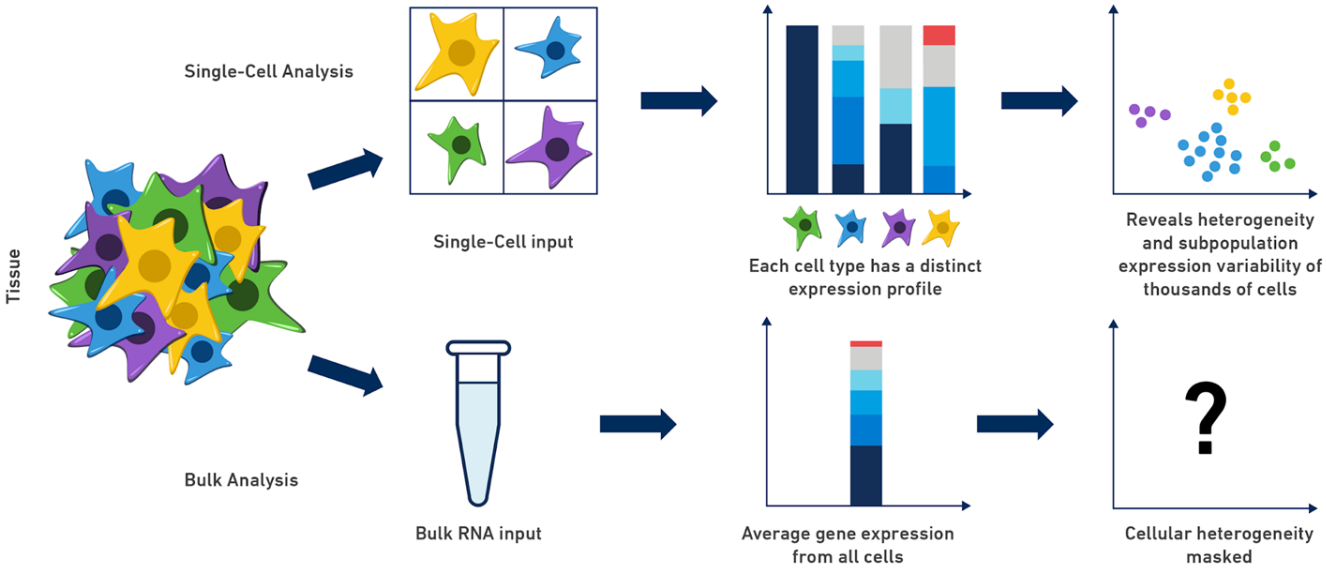
La *bioinformatica, oggi*



Credits: Darryl Leja (NHGRI), Ian Dunham (EBI)

Contemporaneamente su tutto il genoma, di ogni tipo cellulare.

La *bioinformatica, oggi*



Le tecniche più recenti consentono di vedere cosa accade all'interno della singola cellula (single-cell RNA-Seq, single-cell ChIP-Seq, etc..)!

La *bioinformatica, oggi*

La quantità di dati e informazioni prodotte rende sempre più necessaria la presenza di infrastrutture adeguate per la loro corretta gestione.

NCBI
National Center for Biotechnology Information

EMBL-EBI

elixir

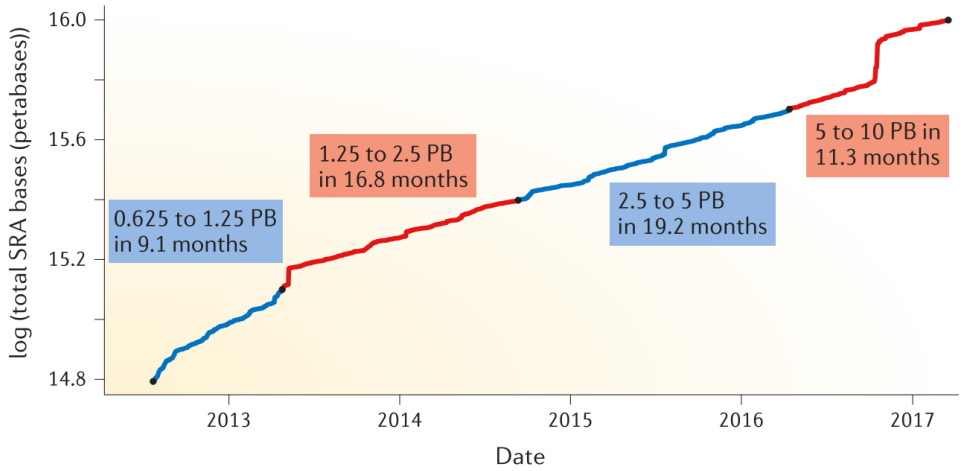
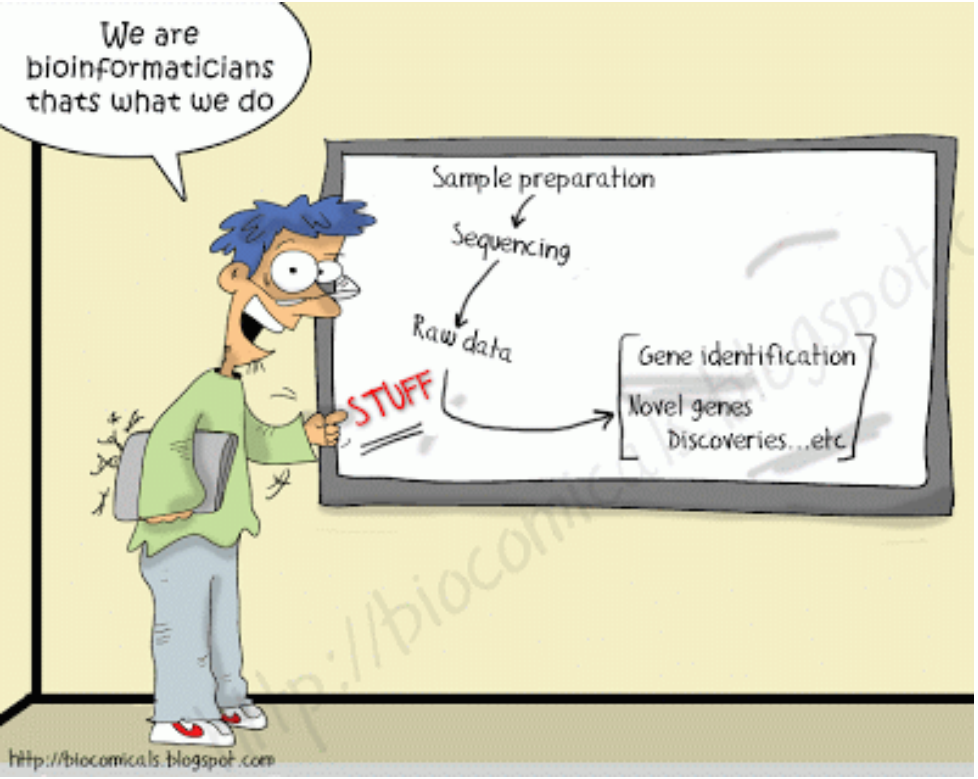


Figure 1 | **Increase in storage of next-generation sequencing data.** From July 2012 to March 2017, the amount of genomic data (total bases) in the Sequence Read Archive (SRA) doubled four times. The large jump in October 2016 is chiefly due to data from the Trans-Omics for Precision Medicine (TOPMed) project. As of October 2017, the SRA contains about 14 petabases (millions of billions of bases) of data. PB, petabases.

La *bioinformatica, oggi*

Name	Website	Description
1000 Genomes Project (1KGP)[102]	www.internationalgenome.org	This project includes whole-genome and exome sequencing data from 2,504 individuals across 26 populations
Cancer Cell Line Encyclopedia (CCLE)[115]	portals.broadinstitute.org/ccle	This resource includes data spanning 1,457 cancer cell lines
Encyclopedia of DNA Elements (ENCODE)[33]	www.encodeproject.org	The goal of this project is to identify functional elements of the human genome using a gamut of sequencing assays across cell lines and tissues
Genome Aggregation Database (gnomAD)[13]	gnomad.broadinstitute.org	This resource entails coverage and allele frequency information from over 120,000 exomes and 15,000 whole genomes
Genotype–Tissue Expression (GTEx) Portal[15,16]	gtexportal.org	This effort has to date performed RNA sequencing or genotyping of 714 individuals across 53 tissues
Global Alliance for Genomics and Health (GA4GH)[92]	genomicsandhealth.org	This consortium of over 400 institutions aims to standardize secure sharing of genomic and clinical data
International Cancer Genome Consortium (ICGC)[14]	icgc.org	This consortium spans 76 projects, including TCGA
Million Veterans Program (MVP)[19]	www.research.va.gov/mvp	This US programme aims to collect blood samples and health information from 1 million military veterans
Model Organism Encyclopedia of DNA Elements (modENCODE)[25,85]	www.modencode.org	The goal of this effort is to identify functional elements of the *Drosophila melanogaster* and *Caenorhabditis elegans* genomes using a gamut of sequencing assays
Precision Medicine Initiative (PMI)[18]	allofus.nih.gov	This US programme aims to collect genetic data from over 1 million individuals
The Cancer Genome Atlas (TCGA)[116]	cancergenome.nih.gov	This resource includes data from 11,350 individuals spanning 33 cancer types
Trans-Omics for Precision Medicine (TOPMed)[17]	https://www.nhlbiwgs.org	The goal of this programme is to build a commons with omics data and associated clinical outcomes data across populations for research on heart, lung, blood and sleep disorders

Il bioinformatico



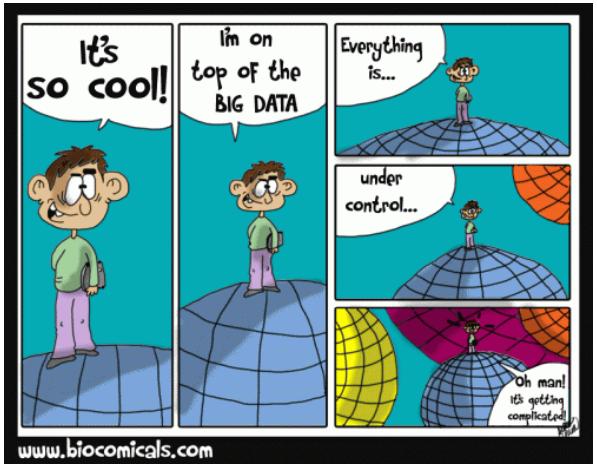
BIOINFORMATICS SCIENTIST

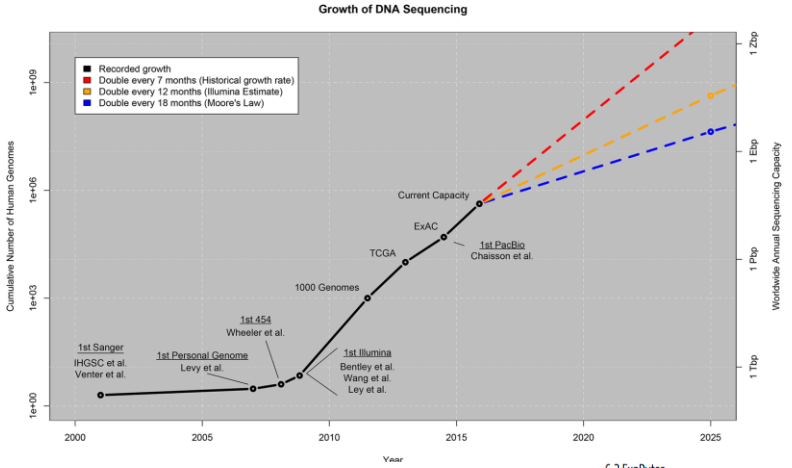
N. [BI-O-IN-FOR-MAT-ICS SCI-EN-TIST]

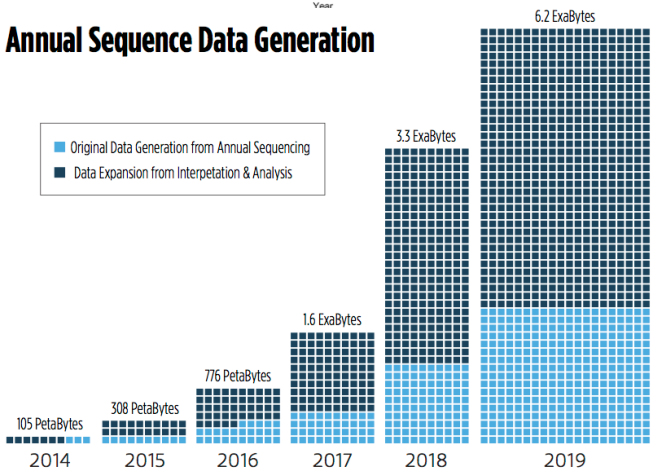
1. A person who solves problems you can't.
2. One who does precision guess work based on unreliable data provided by those of questionable knowledge.

see also WIZARD, MAGICIAN

La sfida: i big data



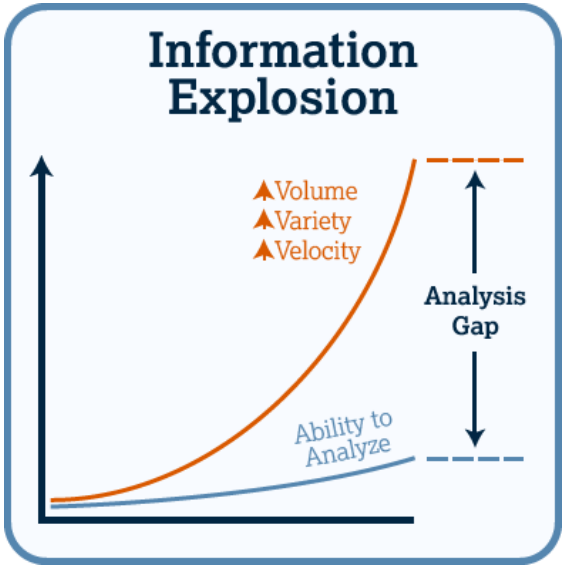




Perché è importante per voi?

- Per passare l'esame di questo corso , ma è il minore dei problemi ☺
- Perché fare ricerca oggi in campo biomolecolare senza avere la capacità di utilizzare o quantomeno comprendere gli strumenti di base della bioinformatica è quasi impossibile, e le cose stanno "peggiorando" molto velocemente. Essere "analfabeta bioinformatico" per un biologo è un po' come essere "analfabeta digitale" per un cittadino ☺ Si rimane tagliati fuori!
- Perché i semplici concetti e strumenti che vedremo in questo corso potranno esservi utili in molte occasioni in futuro. Non ragionate mai per compartimenti stagni!
- Ma soprattutto… perché può aprirvi molte opportunità di lavoro.

Perché è importante per voi?
(se la conosci, non la eviti)

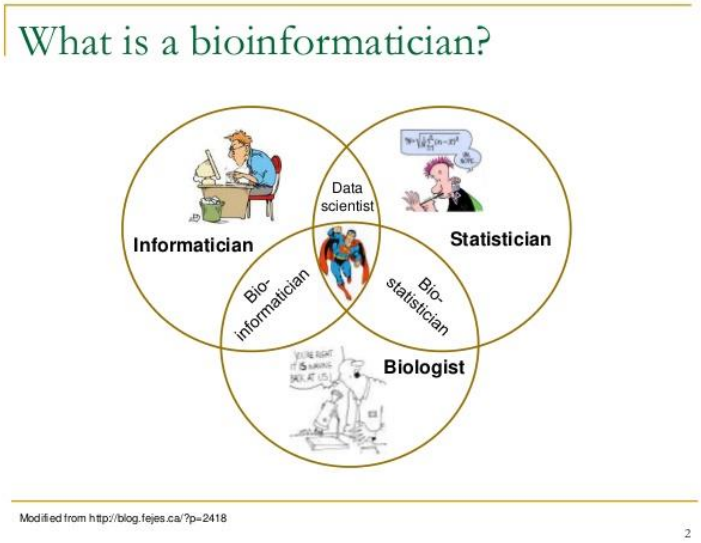


Questo "gap" rappresenta una miniera di opportunità per chi vorrà e saprà coglierle!

"Biology easily has 500 years of exciting problems to work on."
Donald Knuth (Stanford Professor & famous computer scientist)

Cosa serve per fare bionformatica

- ✓ E’ una materia inerentemente multidisciplinare. C’è chi ci arriva dalla biologia, altri dall’informatica. Molti altri ci sono arrivati dalla fisica o dalla statistica.



Avere solide basi di biologia molecolare è indispensabile, non potete sperare di analizzare ciò che non capite!

Informatica e statistica sono importanti ma si possono imparare “strada facendo”, man mano che si presentano i problemi che esse ci aiutano a risolvere.

I “tipi” di bionformatico: l’utilizzatore occasionale

- ✓ Di solito è un biologo “wet” che lavora nel campo della ricerca biomolecolare.
- ✓ Comprende i concetti base della bioinformatica e sa utilizzare gli strumenti di uso più “generale” come browser genomici, banche dati, programmi di allineamento o per disegnare i primer, etc…
- ✓ Sa utilizzare anche qualche programma più “specifico” che lo aiuta nel suo lavoro di tutti i giorni. Sa inoltre come maneggiare i tipi di dati prodotti dal suo laboratorio.
- ✓ Spesso è il “bioinformatico” di riferimento del laboratorio “wet” in cui lavora, tenuto in grande considerazione dai colleghi per la sua capacità di operare “magie” sui dati.

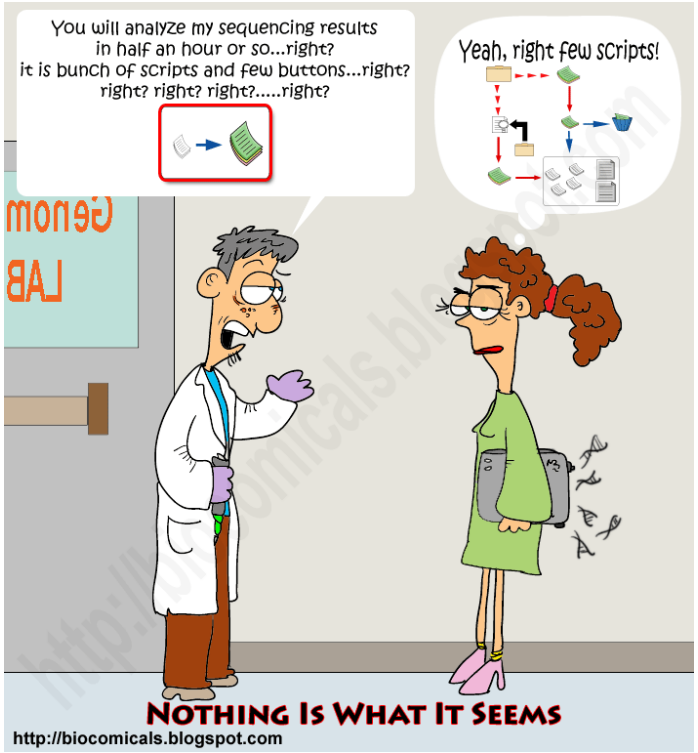
I “tipi” di bionformatico: l’utilizzatore professionale

- ✓ Ha fatto della bioinformatica la sua occupazione principale. Conosce bene le principali banche dati biologiche generali e specializzate.
- ✓ Conosce i principali strumenti bioinformatici da utilizzare in base ai dati da analizzare e ai risultati da ottenere e si tiene aggiornato quando ne compaiono di nuovi.
- ✓ Sa destreggiarsi con i principali “tipi di dati” biologici e sa adattarsi velocemente quando si trova davanti a qualcosa di nuovo e mai visto prima.
- ✓ Di solito lavora in grandi gruppi di ricerca o fa da bioinformatico di riferimento a più gruppi che si appoggiano a lui per le loro analisi. Riceve attestati di stima o di odio in base ai risultati delle analisi (che spesso dipendono dalla qualità dei dati di partenza).

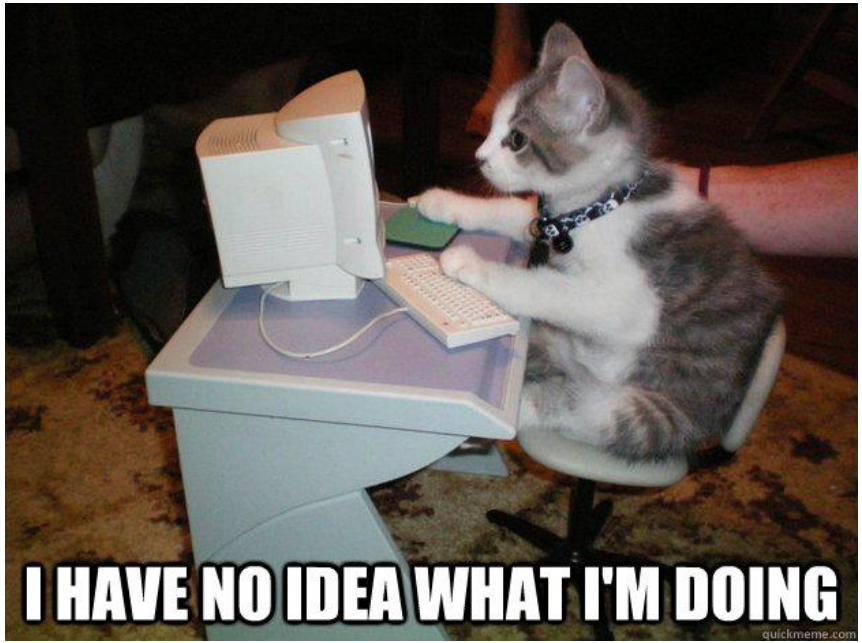
I “tipi” di bionformatico: lo sviluppatore

- ✓ E’ come l’utilizzatore professionale, ma non si accontenta degli strumenti preparati da altri.
- ✓ Pensa a nuove soluzioni per i problemi che gli vengono posti (stando attento a non reinventare la ruota!)
- ✓ E’ in grado di implementare nuove soluzioni utilizzando gli strumenti dell’informatica (programmazione) e della statistica.
- ✓ E’ in grado di capire le differenze tra i vari approcci alla soluzione di un problema e valutare quello che di volta in volta conviene seguire e perché.
- ✓ Suscita reazioni contrastanti tra gli altri biologi.

Ci sono svantaggi…



Sindrome del biologo ansioso: ci ho messo 6 mesi a produrre questi dati adesso hai 20 minuti per dirmi cosa dicono.



Rischio: i computer fanno sempre quello che gli si dice alla lettera. A volte è difficile accorgersi che quanto gli chiediamo di fare è "sbagliato".

… e vantaggi

Experimental Biologist
Computational Biologist
NO EATING
NO DRINKING
NO BREATHING
NO OPEN TOE SHOES
NO SHORTS
NO MINI SKIRTS
ALWAYS LAB COATS ON
LIFE IS BEAUTIFUL
www.biocomicals.com

Don't splash water on my laptop
Now, you're exaggerating
Bioinformaticians: They can work from anywhere
www.biocomicals.com

KEEP CALM AND CALL A BIOINFORMATICIAN

DON'T PANIC