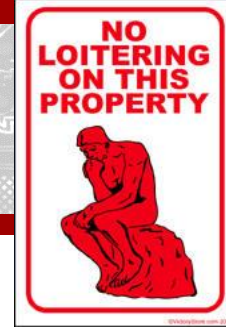


Dati per la vita

*Dall'analisi del genoma alla medicina
personalizzata
con l'aiuto dell'informatica*

Giancarlo Mauri

Università di Milano-Bicocca



“Ogni tentativo di impiegare metodi matematici per lo studio delle questioni biologiche deve essere considerato profondamente irrazionale e contrario allo spirito della biologia.”

“Se l'analisi matematica dovesse mai giocare un ruolo importante in biologia - un'aberrazione che è per fortuna quasi impossibile - avrebbe inizio una degenerazione rapida e diffusa di quella scienza.”

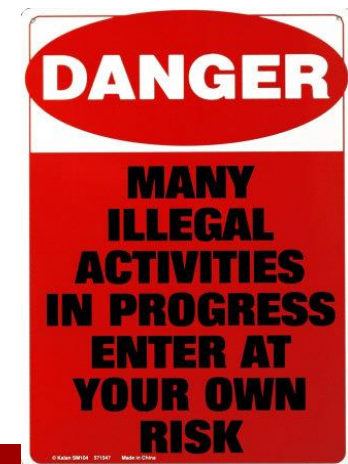
Auguste Comte, Philosophie Positive, 1830



Lezioni Lincee



Milano-Bicocca



Due secoli dopo ...

La biologia è cambiata:

Da scienza empirica qualitativa

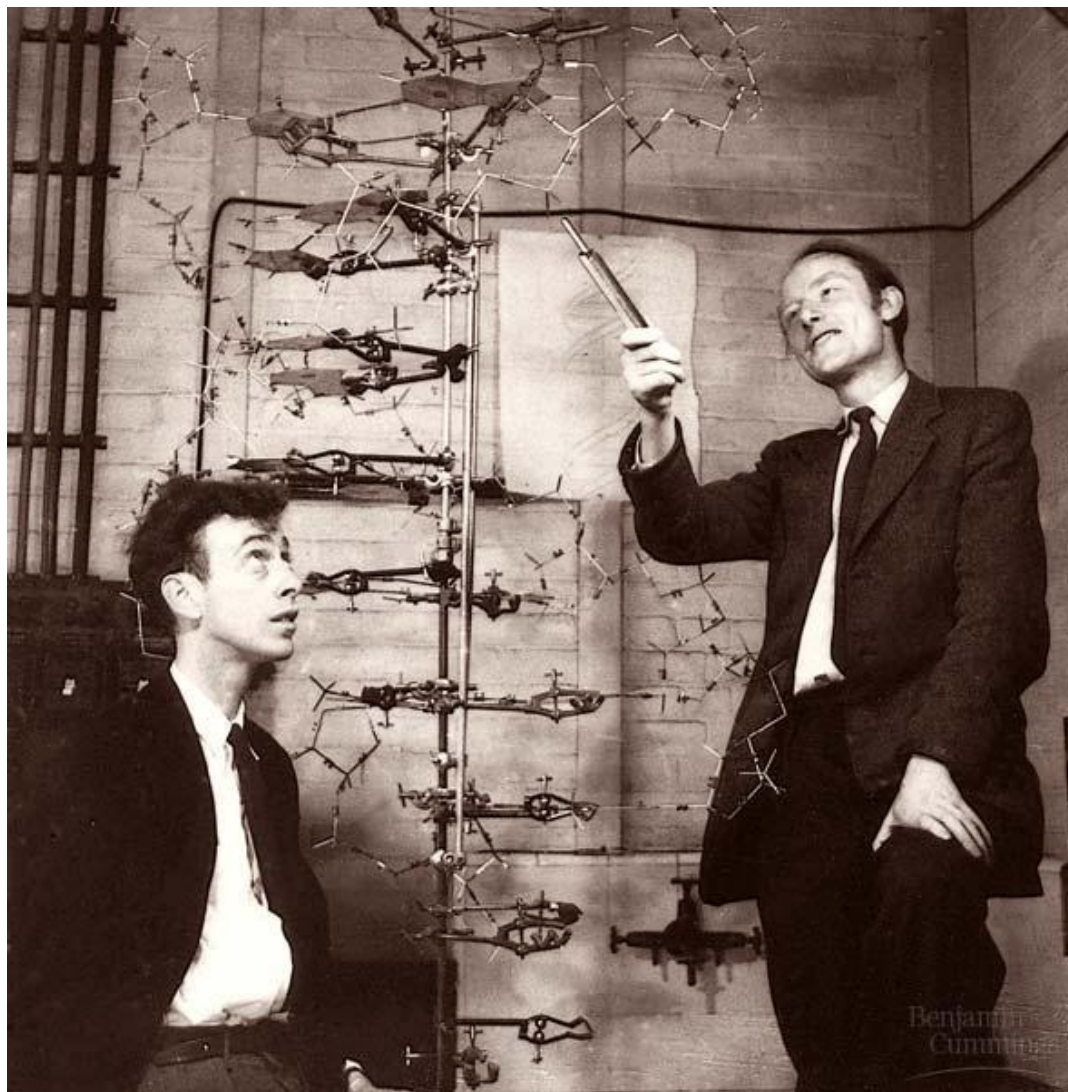


A scienza esatta quantitativa



Le servono concetti, metodi e strumenti matematici e computazionali

1953: Watson and Crick scoprono la struttura del DNA



Le basi biologiche: il DNA

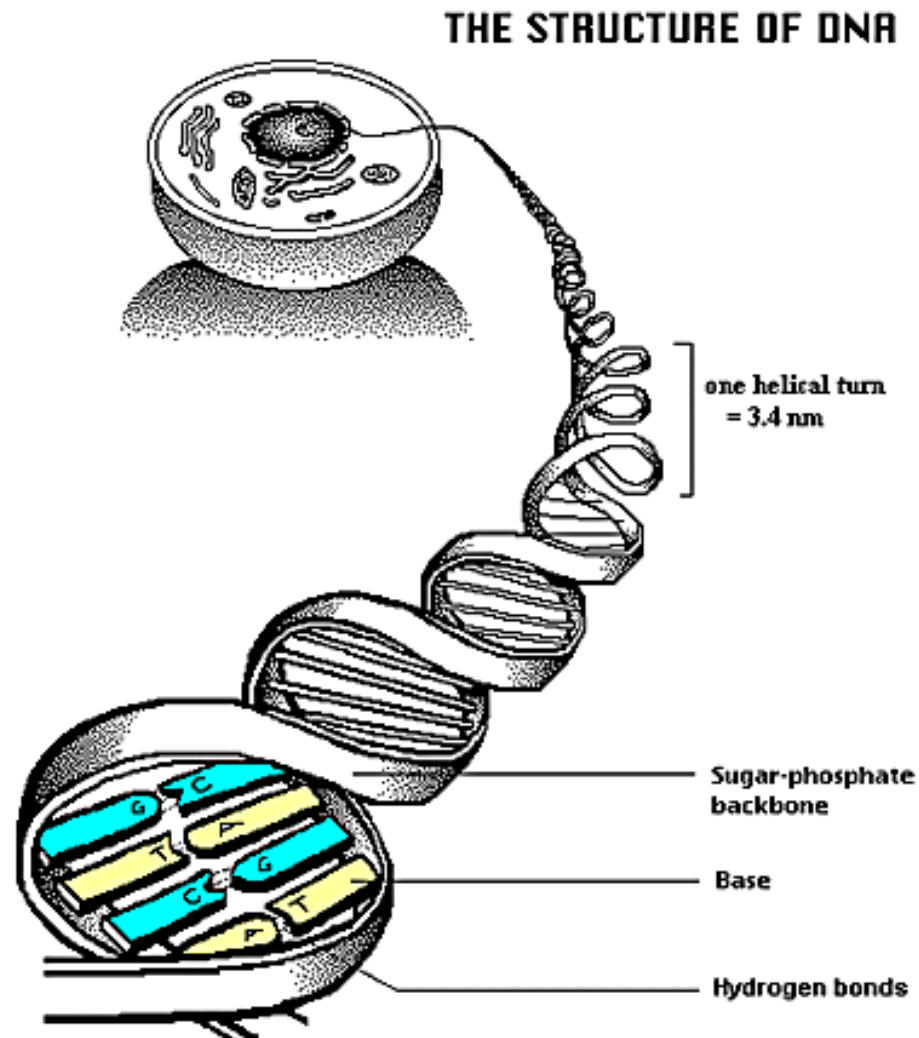
- ◆ Un organismo è un dispositivo naturale autoriproducentesi che manipola materia, energia ed *informazione*
- ◆ L'archivio delle informazioni è il *genoma*
- ◆ Il genoma è una molecola di **DNA**, una doppia elica contenente un messaggio scritto usando un alfabeto di quattro lettere (**A, C, G, T**), i *nucleotidi*, a due a due complementari

Le basi biologiche: il genoma

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

■ Batterio (<i>E. Coli</i>)		4.8 Mbp
■ Lievito		14.4 Mbp
■ Verme (<i>C. elegans</i>)		100.0 Mbp
■ Moscerino (<i>Drosophila</i>)		165.0 Mbp
■ Uomo (<i>Homo sapiens</i>)		3.3 Gbp

Le basi biologiche: il DNA



Il DNA e le proteine

- ◆ Il DNA governa la costruzione e il funzionamento dell'organismo attraverso la produzione di proteine
 - sono le principali responsabili della struttura e delle attività dell' organismo
 - costituite da venti diversi aminoacidi
 - Lunghezza media: 200-300 aminoacidi
- ◆ Ogni proteina è costruita a partire da un gene
- ◆ I geni costituiscono una piccola parte del genoma (< 10%) e devono essere individuati

Il Progetto Genoma Umano

◆ Obiettivi

- Determinare la sequenza del DNA umano
- Identificare i circa 30.000 geni
- Immagazzinare le informazioni in banche dati
- Sviluppare strumenti per l'analisi

Il Progetto Genoma Umano

1986

progetto pilota (5,3 M\$)

1990

inizio del progetto quindicennale

1995

il primo obiettivo è raggiunto con un anno di anticipo

Il Progetto Genoma Umano

1998

- ◆ Nasce **Celera Genomics**: promette di completare il sequenziamento in 3 anni
- ◆ **HGP** anticipa di due anni il termine

1999

- ◆ **HGP** anticipa dal 2001 al 2000 le previsioni per la prima bozza
- ◆ 1 dicembre: annunciato il sequenziamento del primo cromosoma umano intero

Il Progetto Genoma Umano

2000

- ◆ Aprile : Celera Genomics annuncia di aver completato il sequenziamento; scoppia la "guerra del DNA"
- ◆ Giugno: annuncio congiunto HGP - Celera

2001

- ◆ 12 febbraio: pubblicazione di un "draft" del genoma

2003

- ◆ Aprile: Annuncio del completamento del Progetto Genoma Umano

Perché servono i computer e gli algoritmi

◆ Generare i dati: l'assemblaggio di frammenti

- Alcune copie del genoma sono spezzettate in frammenti corti (200 ÷ 700 bp)
- I frammenti vengono sequenziati
- Attraverso le sovrapposizioni di frammenti si cerca di ricostruire la sequenza:

è come un puzzle da 50 milioni di pezzi

L'assemblaggio di frammenti

GCCTGAGAAATGGCTACCACATC

CCACATCCACGGAGAGG

L'assemblaggio di frammenti

GCCTGAGAAATGGCTACACATC
CCACATCACGGAGAGG

GCCTGAGAAATGGCTACCACATCCACGGAGAGG

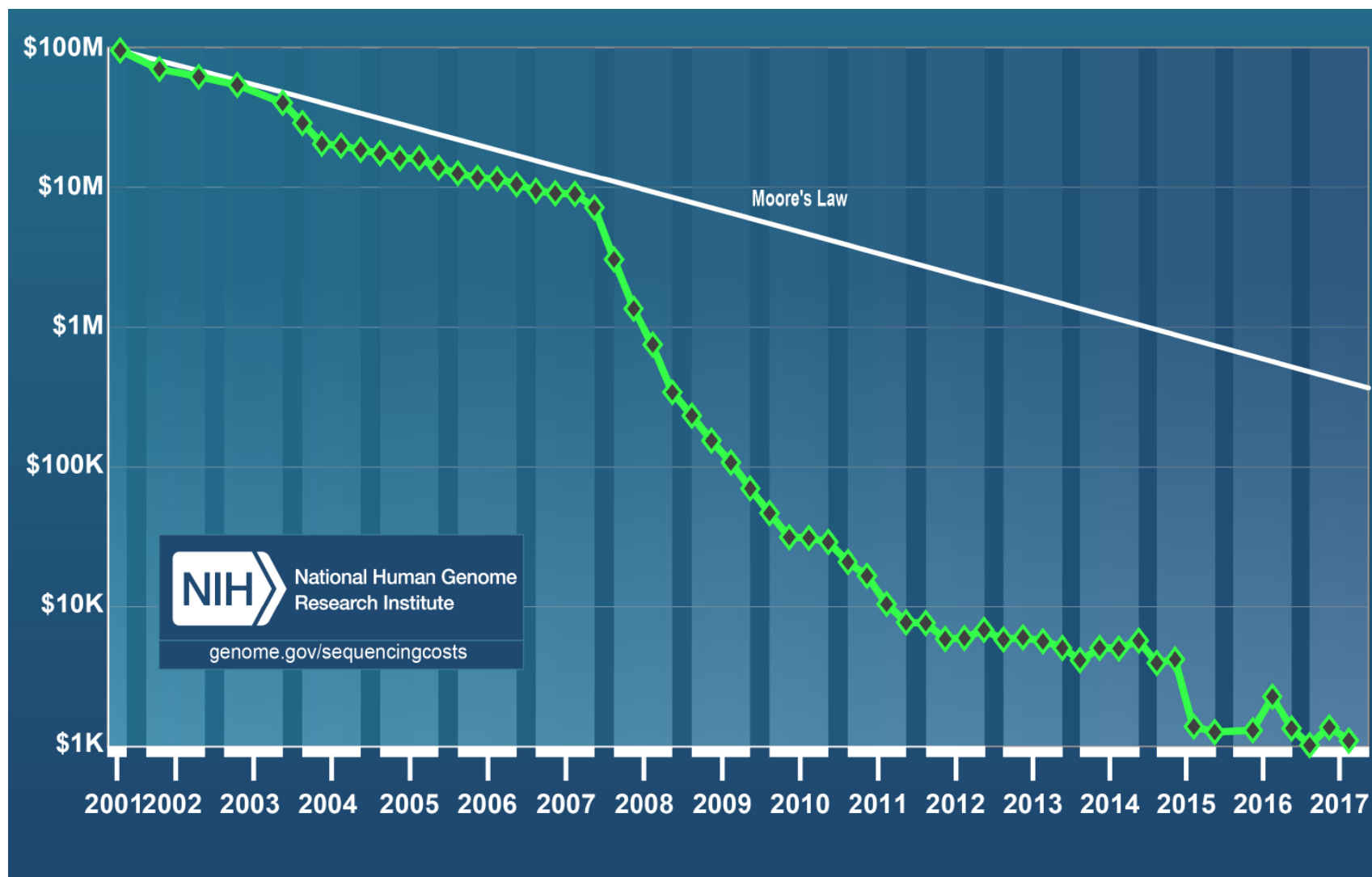
TTCCGGAGAGGGAGCCTGAG

L'assemblaggio di frammenti

GCCTGAG AATGGCTACCACATCCACGGAGAGG
TTCCGGAGAGGG GCCTGAG

TTCCGGAGAGGGAGCCTGAGAAATGGCTACCACATCCACGGAGAGG

Costo per genoma



Il diluvio dei dati

- ◆ Le tecnologie NGS producono miliardi di dati nucleotidici per esperimento, a un costo relativamente basso.
- ◆ I sequenziatori moderni sono in grado di analizzare l'equivalente di un genoma umano ogni 14 minuti al costo di \$5,000.
- ◆ Questo tasso è 400 volte maggiore rispetto all'anno 2000, quando la bozza del genoma umano fu pubblicata per la prima volta. Si prevede che ulteriori sviluppi si tradurranno in sequenziatori tra 1000 e 1.000.000 di volte più produttivi nei prossimi dieci anni.

◆ Dati di interesse scientifico generale

- Dati di sequenza
- Dati di espressione genica
- Proteomica
- Reti regolatorie
- Fenotipi
- ...



◆ Dati di singoli individui

Il diluvio dei dati

- ❖ **Sequenze**
- ❖ **Mutazioni**
- ❖ **Strutture**
- ❖ **Espressione genica**
- ❖ **Annotazioni**
- ❖ **Reti regolatorie**
- ❖ **Immagini**
- ❖ **Dati medici**

GenBank

187.066.846 sequences
199.823.644.287 bases
Aug 2015

UniProtKB/TrEMBL

50.825.784 sequences
16.880.602.444 amino acids
Sep 2015

Gene Expression Data

1.579.212 samples
Up to 50.000.000.000 data points
Sep 2015

Annotations

ENSEMBL database

20 species
more than 1.500 tables
more than 800M records
more than 1.500B data points

La sfida computazionale



Perché servono i computer e gli algoritmi

◆ Estrarre conoscenza dai dati

- Identificare i geni sulle sequenze di DNA
- Capire meccanismi di trascrizione del DNA
 - Reti regolatorie
- Trovare segnali sconosciuti
- Capire le proteine
 - Struttura
 - Funzione
- Comprendere i processi cellulari
 - Systems biology

Perché servono i computer

◆ Usare la conoscenza

- Individuare l'informazione sui geni che indica la presenza di una patologia (diagnosi)
- Riparare i geni
- Progettare farmaci con un reale impatto sulla malattia senza alterare il delicato equilibrio dell'organismo
- Verso la medicina personalizzata

I Computer sono velocissimi.

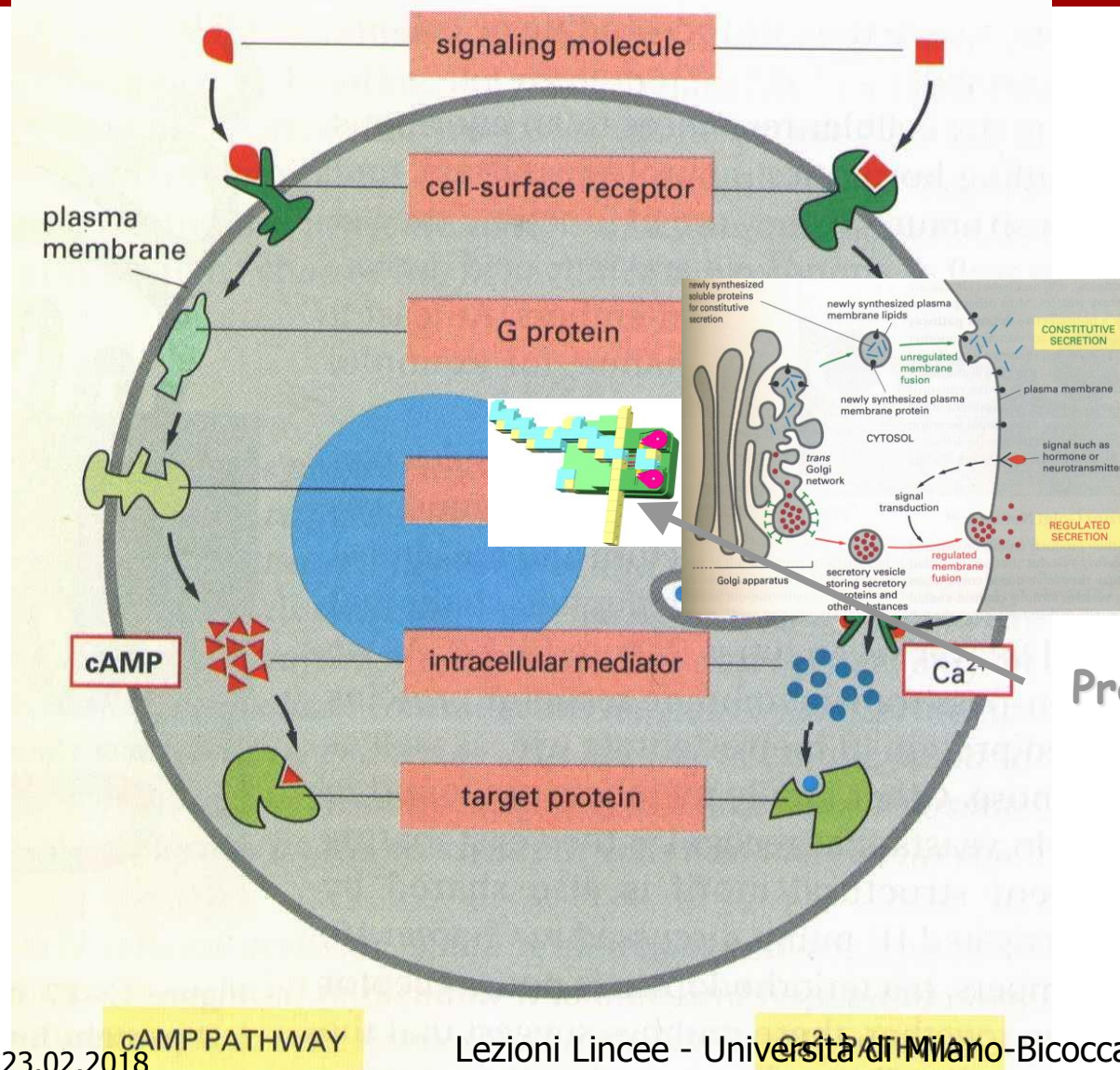
Ma ...

La soluzione di certi problemi
richiede tantissimo tempo!

Servono
Algoritmi
Efficienti

Nasce la Bioinformatica

La medicina nel “:2050Dottore in una cellula”



Programmable Computer