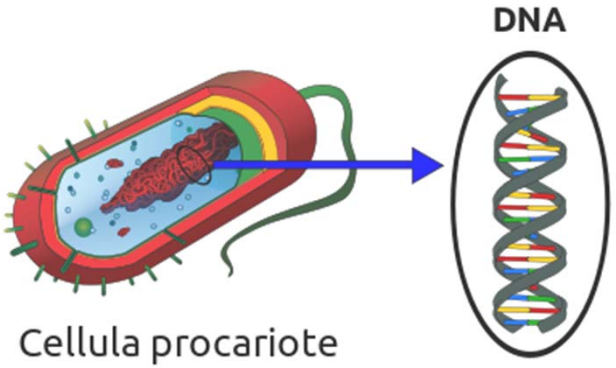
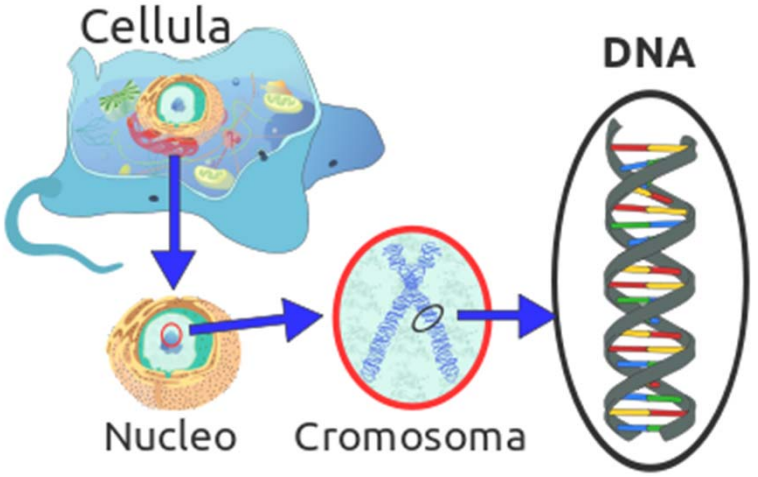
Carpire il segreto della vita con l'informatica

Giosuè Lo Bosco
Dipartimento di Matematica e Informatica, Università di Palermo, ITALY.

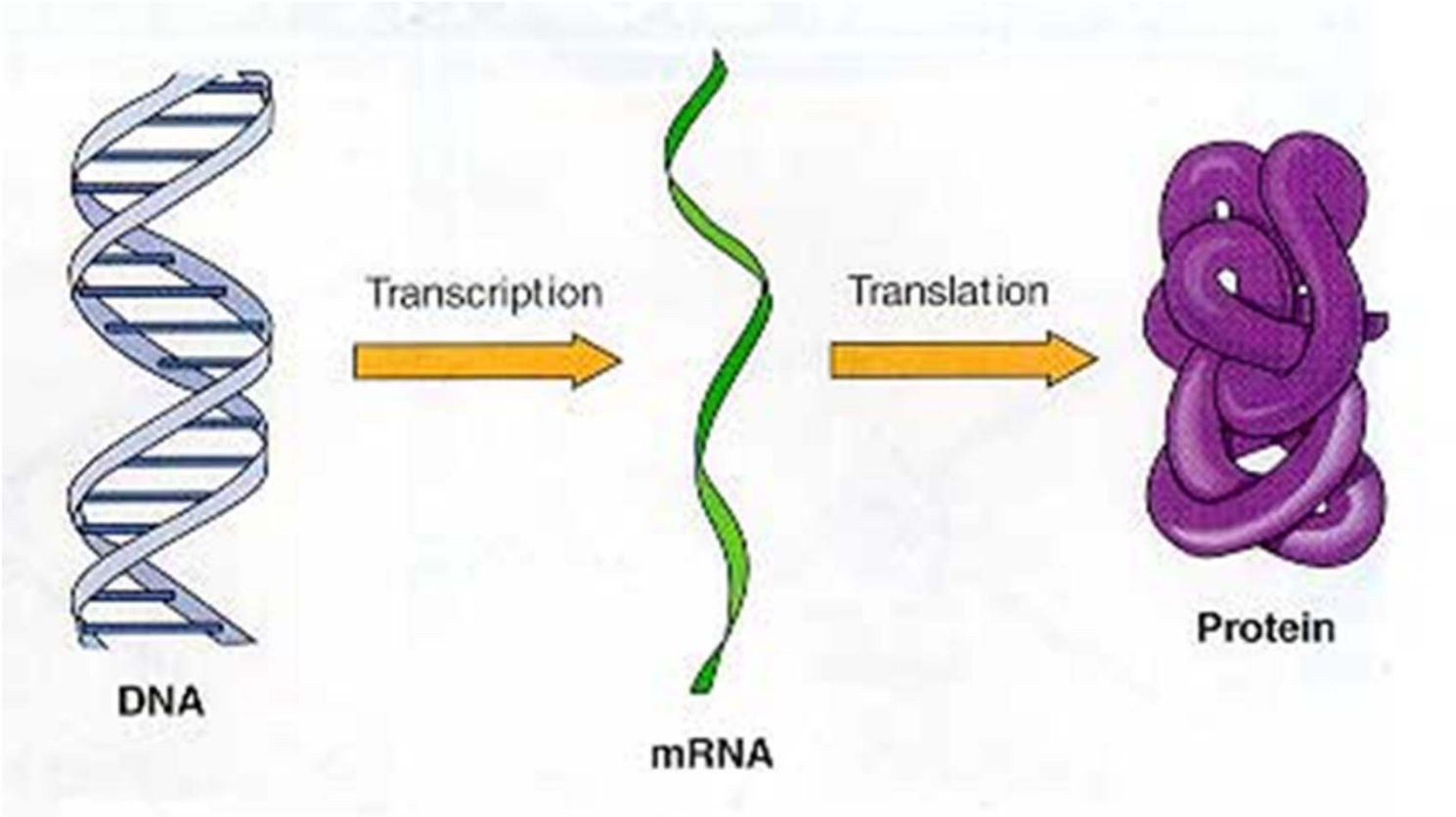
Lezioni Lincee Palermo, 26
Febbraio 2015

Alla base della vita degli organismi c'è il DNA





Dogma Centrale della Biologia Molecolare



Geni: porzioni di dna localizzate in precise posizioni all'interno della sequenza, che contengono tutte le informazioni necessarie per la produzione di una proteina.

La *genetica* è la branca della biologia, che studia i geni, l'ereditarietà e la variabilità genetica.

Informatica e genetica (bioinformatica 1.0)

Lezioni Lincee Palermo, 26 Febbraio 2015

Rappresentazione “informatica” del DNA

Nucleotidi

Alfabeto: {A,T,C,G}

Rappresentazione del DNA come *sequenza*

….atcgtagtctgatgctgatctagttcgtatctatcgtacac….

$3*10^{12}$ nucleotidi nel DNA umano !

Sequenziamento:

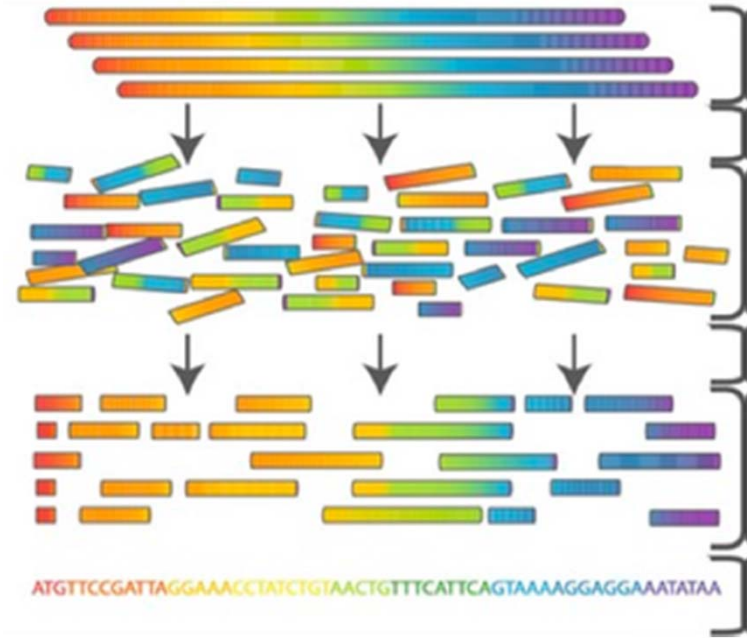
Il sequenziamento del DNA è un processo che serve a mettere in fila le basi (Adenina, Citosina, Guanina e Timina) che costituiscono il frammento di DNA in analisi, in modo da poterlo leggere propriamente ed analizzare.

https://www.youtube.com/watch?v=6ldtdWjDwes

Progetto Genoma:

Iniziato nel 1990, lo scopo di questo progetto era sequenziare tutto il genoma umano.

Nel 2000 la prima bozza, nel 2003 genoma completo.



Genomi clonati

Taglio di segmenti a dimensione variabile

Segmenti non ordinati

Assemblaggio

Problema informatico

Risultato dell’assemblaggio

Combinazione

Il DNA viene letto da una “macchina” e produce proteine

Abbiamo informazioni su come questa macchina agisce, quando inizia e come viene codificata una proteina a partire da un gene.

Dal punto di vista informatico

Input di tipo stringa (lunga $3*10^{12}$ caratteri):…..atcgtagtctgatgctgatctagttcgtat….

Software che legge la stringa, e produce un risultato (proteina).

Abbiamo la sequenza di DNA, e se volessimo trovare I geni ?

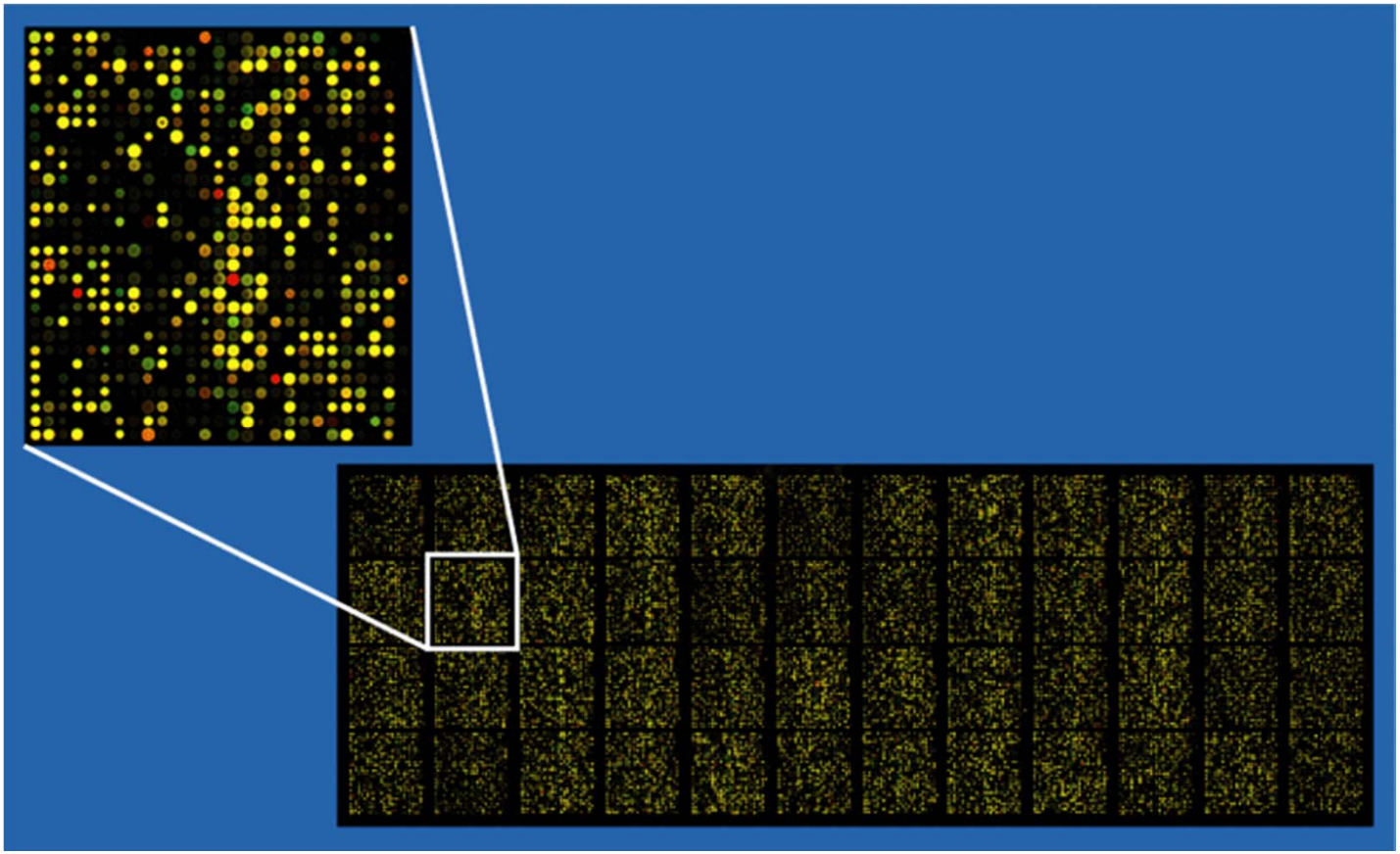
Input di tipo stringa (lunga $3*10^{12}$ caratteri):…..atcgtagtctgatgctgatctagttcgtat….

Gene (stringa di circa 8000 caratteri):…. aatcgtagtctgatgc….

Come facciamo a trovare le regioni nel DNA che sono uguali al gene ?

Tipico problema informatico detto *Pattern Matching* sul quale molte soluzioni sono state fornite (anche indipendentemente dal problema biologico)

Microarray e bioteconologie



Lezioni Lincee Palermo, 26 Febbraio 2015

Bioinformatica:

Descrivere dal punto di vista numerico e statistico i fenomeni biologici fornendo ai risultati biochimici e biologici un corredo di strumenti analitici e numerici

Discipline coinvolte: informatica, matematica applicata, statistica.

La bioinformatica principalmente si occupa di:

1) fornire modelli statistici validi per l'interpretazione dei dati provenienti da esperimenti di biologia molecolare e biochimica al fine di identificare tendenze e leggi numeriche

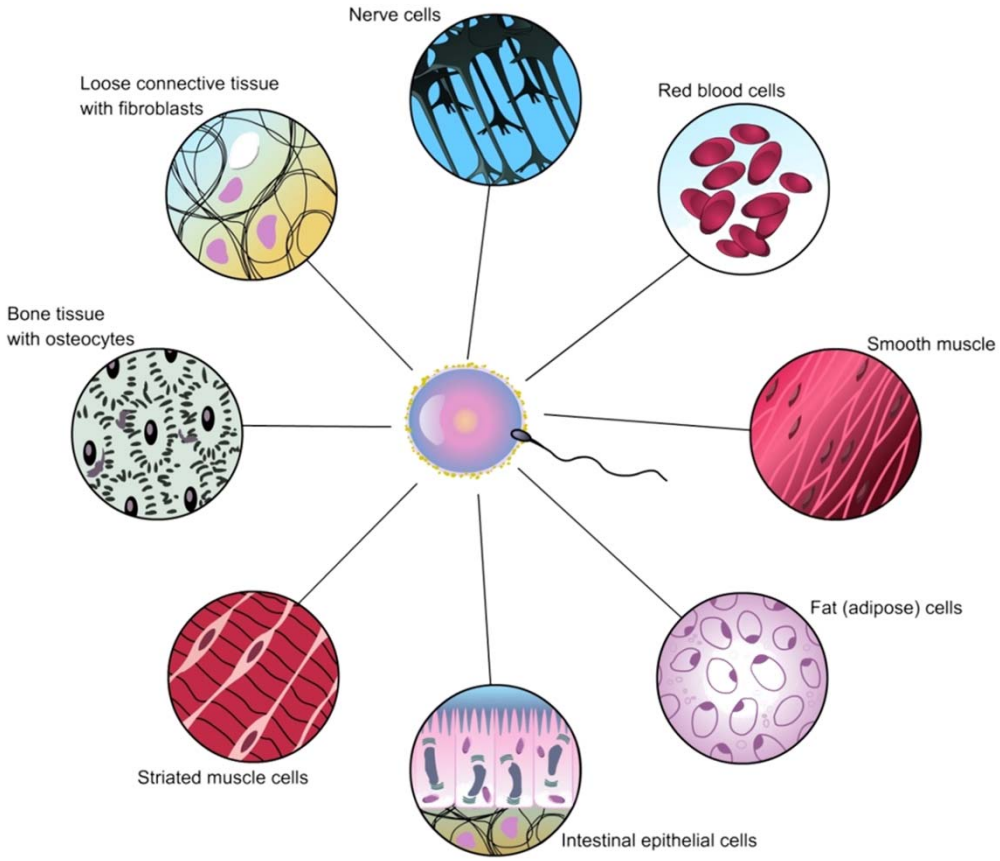
2) generare nuovi modelli e strumenti matematici per riuscire a comprendere i sistemi biologici

3) organizzare le conoscenze acquisite a livello globale in basi di dati al fine di rendere tali dati accessibili a tutti, e ottimizzare gli algoritmi di ricerca dei dati stessi per migliorarne l'accessibilità.

Informatica ed epigenetica (bioinformatica 2.0)

Lezioni Lincee Palermo, 26 Febbraio 2015

Oservazione: Quasi tutte le cellule del nostro corpo condividono lo stesso DNA, ma l'effetto finale è differente



Probabilmente, esistono altri meccanismi che generano il *fenotipo*

Dalla *genomica* all'*epigenomica*

Nuova branca : *Epigenomica*

"E' la branca della genomica che studia tutte le modificazioni ereditabili che variano l'espressione genica pur non alterando la sequenza del DNA»

Da un punto di vista “informatico”

L’epigenetica può essere vista come la logica di controllo del “software”

AAA-AGT-GCA-ATT-GCT-GGT-ATC-AAT-CT

GGT-AAT-GGT-GTA-TTA-AAC-TTA-AAA-AT

ATT-AGA-GAT-ATT-TTA-ATT-CAA-AGA-GA

AAA-CAT-AAT-CAT-TTT-GCT-GTG-AAA-CA

GAA-AAA-TGG-TTC-CAA-TTA-CCT-TCT-TC

TTT-GAG-AAA-TTA-GAA-ATT-TAT-AAA-AA

AAA-GAA-GAT-TGC-ATT-TAT-TAT-CCC-AT

AAG-TAT-GAT-AAA-TAT-AAT-TCT-ATT-GT

TAT-TTT-GAT-TTC-TAT-AAA-ATC-ATA-AGT

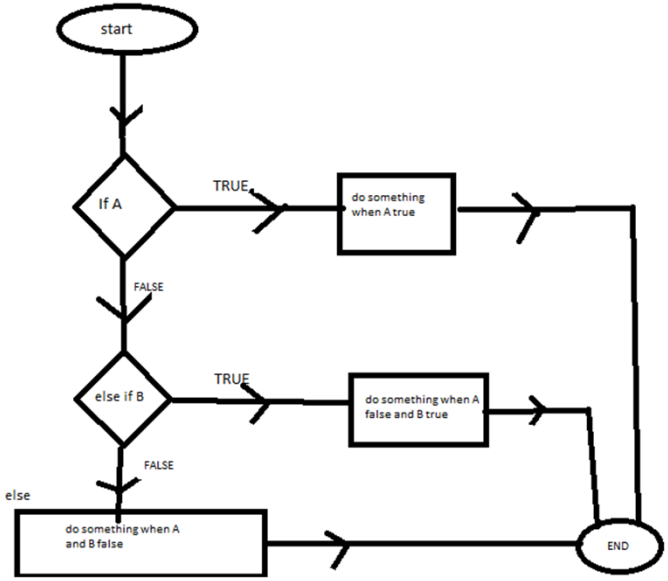
TCA-TTT-AAT-TTA-CGT-GGA-ACA-CTA-AA

TTT-ATA-CCT-GTA-GCT-AAT-TTA-CCA-AC

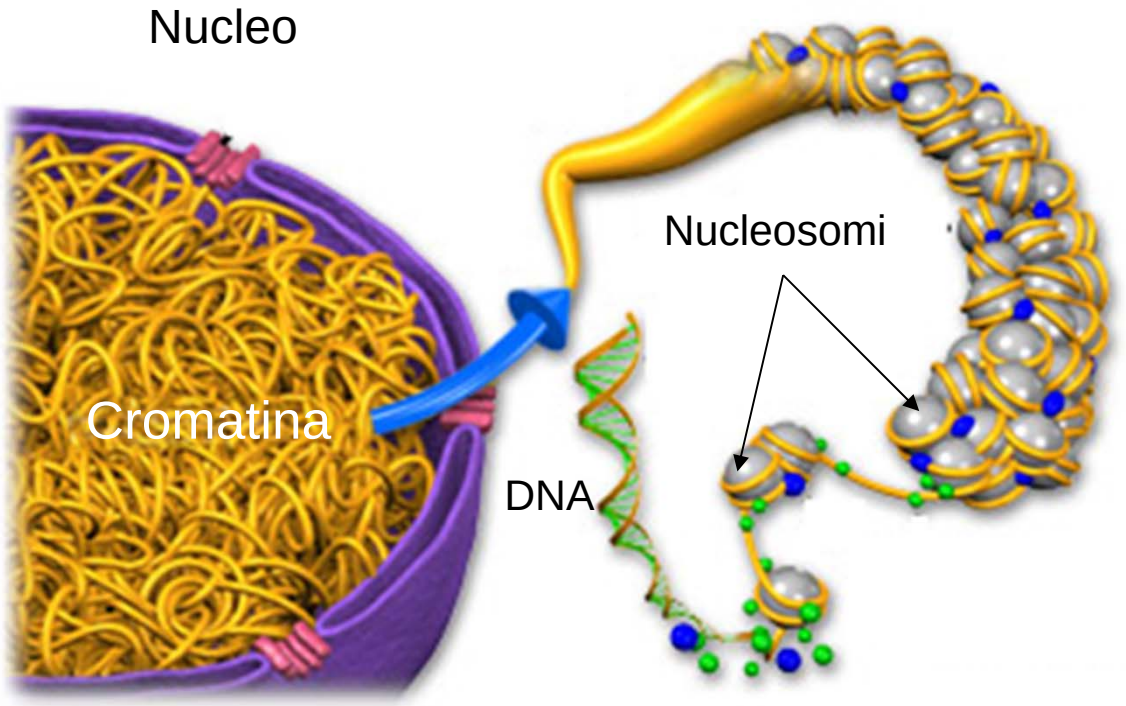
AGT-AAA-AAC-ACG-GTT-GAA-TTA-TAT-TT

CAT-AAT-GCT-TCT-ACT-ATT-ATT-GAA-CC

GTT-CCT-GCC-ACA-ATA-ATT-TGT-TTA-GA

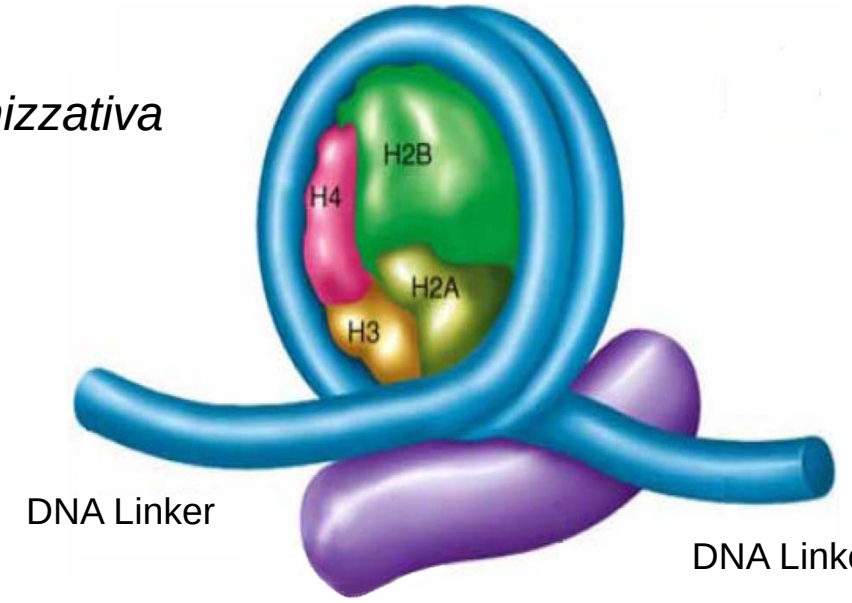


Organizzazione del DNA



Nulceosoma

Il nucleosoma è l'unità organizzativa fondamentale della cromatina.

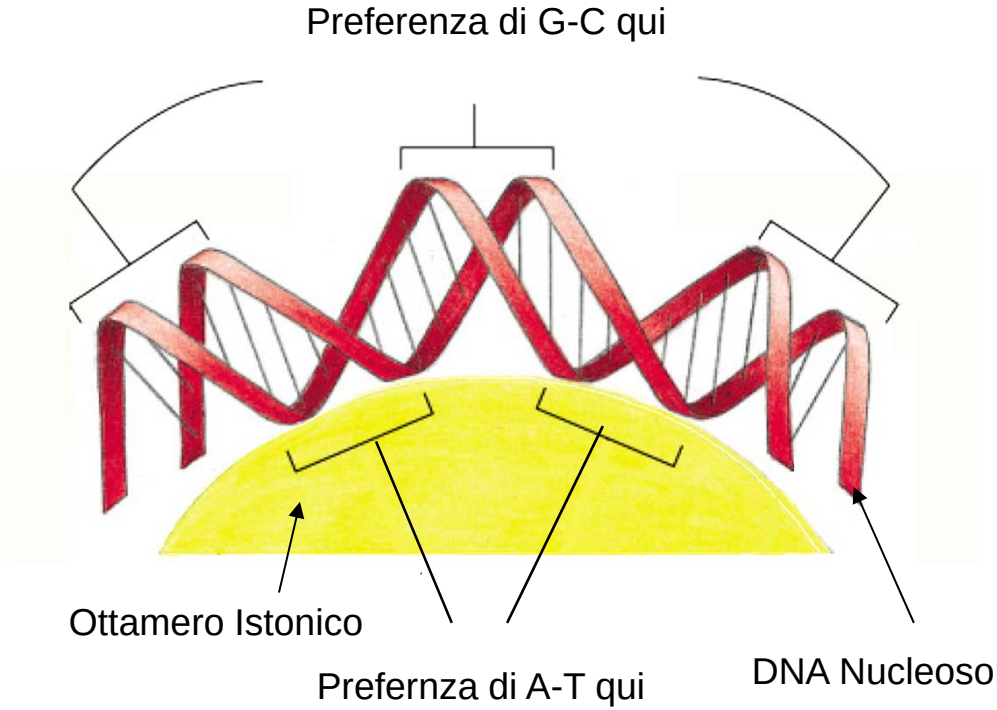


Nucleosoma: *Ottamero Istonico (due molecole di H2A, H2B, H3 e H4) dove ~147 bp di DNA sono saldamente avvolti.*

75-90% del DNA è avvolto sui nucleosomi.

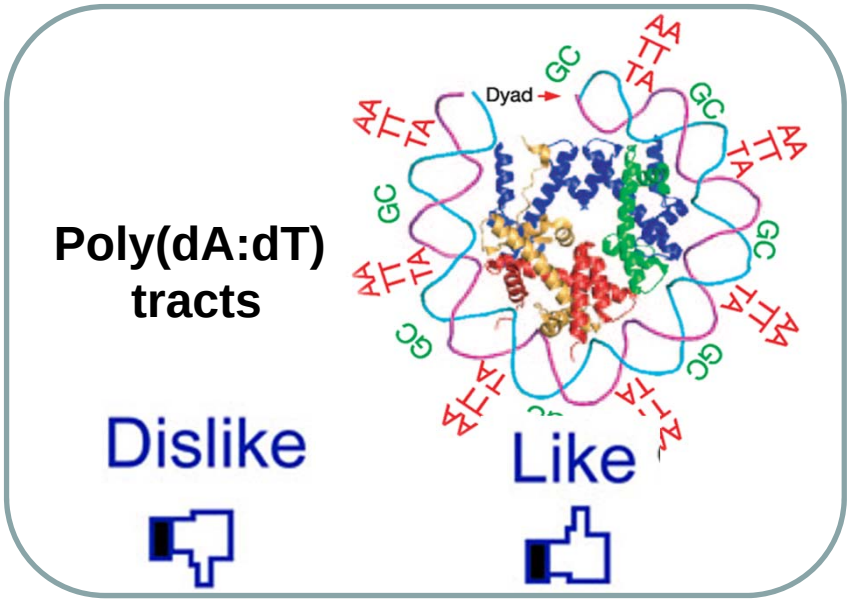
Specificità di sequenza

Affinità di legame dell'ottamero istonico



Sappiamo pure quali sequenze piacciono e quali non piacciono ad i nucleosomi

Specificità della sequenza e nucleosomi



Perché studiare la cromatina (nucleosomi)

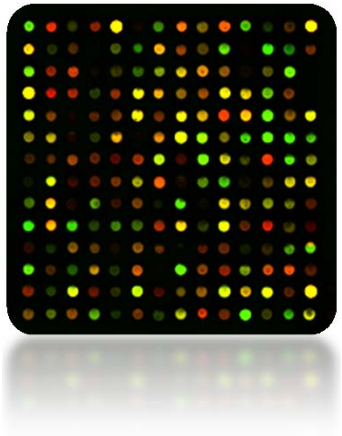
La struttura della cromatina è dinamica ed è coinvolta in molti processi inclusa l'espressione genica.

Il DNA è avvolto nella cromatina e solo una piccola porzione del genoma è accessibile in ogni tipo di cellula.

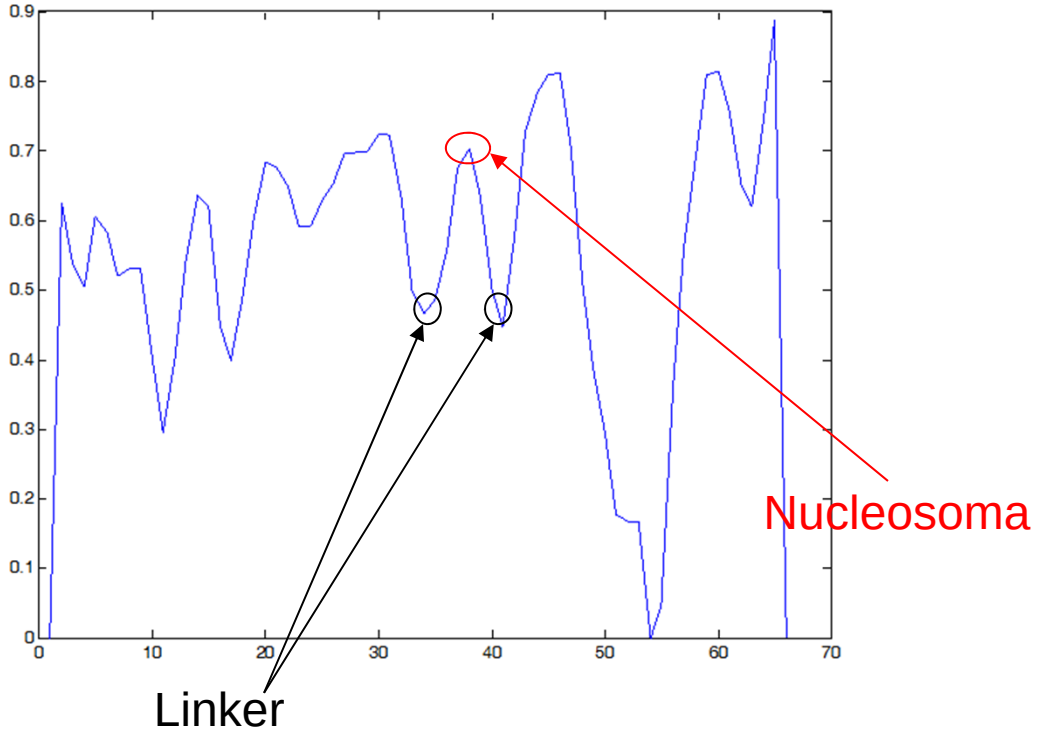
Le alterazioni nella struttura della cromatina sono coinvolte in diverse malattie, tra cui il cancro.

Tipi di input, stringhe e segnali

```
CATTAAGCCCGTATTTCAATT
AAACTCTTCATATT
ATGAATTAATGCAGTTCAA
GCTTAGTCTGCCTTAT
TATAAGCAATAGCTTCTA
```

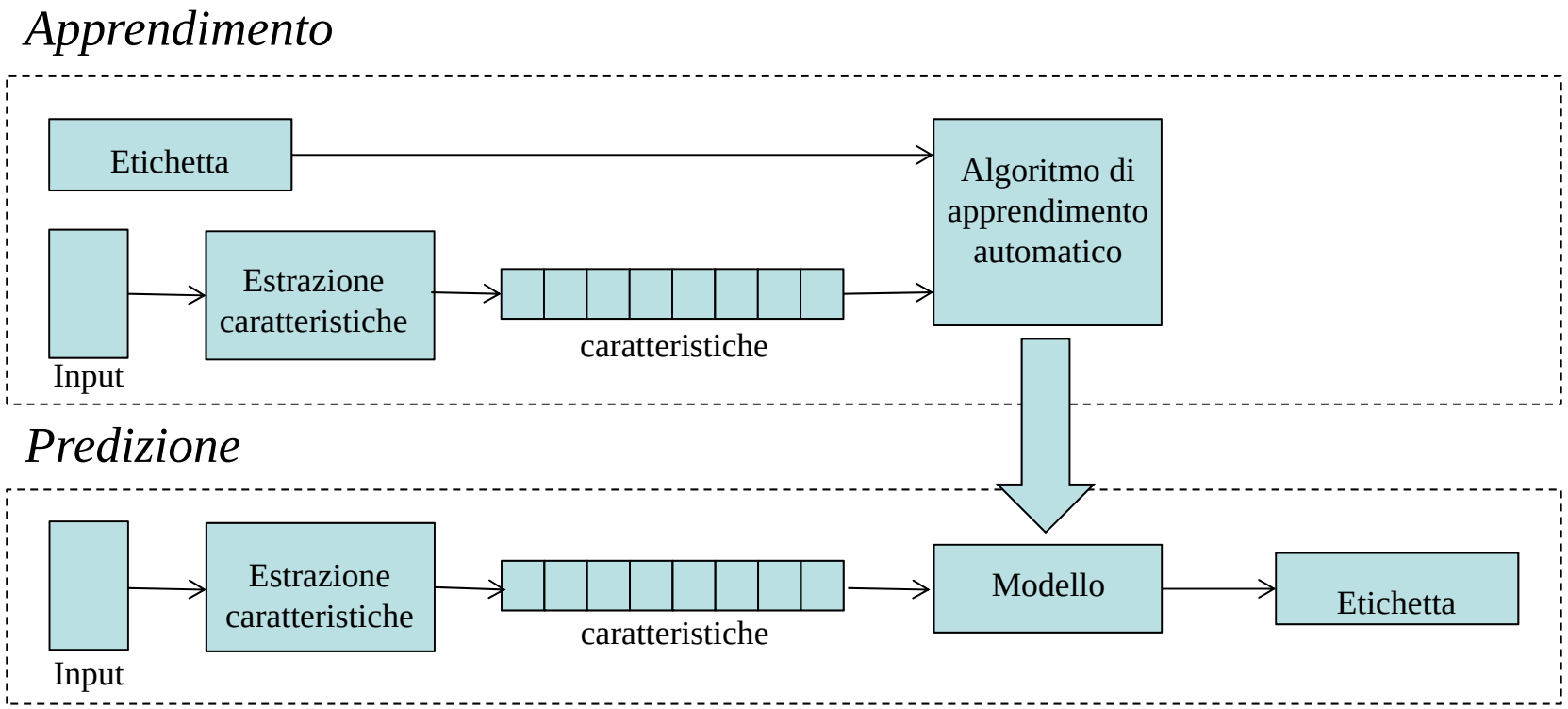


Presenza di AA, AT, TA



Identificazione di picchi e valli

Ma questo è un problema di apprendimento automatico!



Approccio 1, input stringhe:

Fase di estrazione delle caratteristiche: occorre associare ad un input “letterale” un insieme di numeri che rappresenteranno l’input all’algoritmo di apprendimento automatico

Es: finestra di dimensione, K=2

$s = AACTCGTC$ $W = AA\ AC\ AT\ AG\ CA\ CC\ CT\ CG\ TA\ TC\ TT\ TG\ GA\ GC\ GT\ GG$

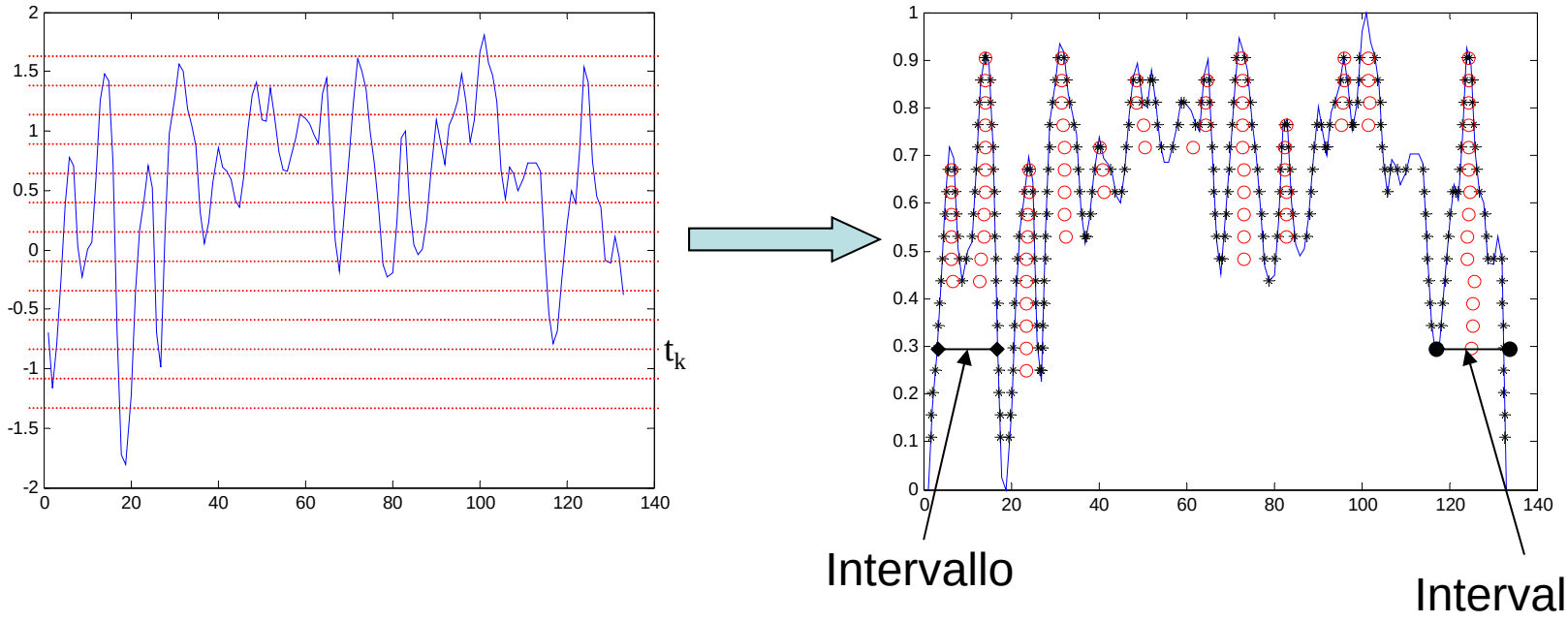
$$x_s = [1\ \ 1\ \ 0\ \ 0\ \ 0\ \ 0\ \ 1\ \ 1\ \ 0\ \ 2\ \ 0\ \ 0\ \ 0\ \ 0\ \ 1\ \ 0]$$

$$x_s = \left[\frac{1}{7}\ \ \frac{1}{7}\ \ 0\ \ 0\ \ 0\ \ 0\ \ \frac{1}{7}\ \ \frac{1}{7}\ \ 0\ \ \frac{2}{7}\ \ 0\ \ 0\ \ 0\ \ 0\ \ \frac{1}{7}\ \ 0\right]$$

Nota: l’estrazione delle caratteristiche secondo questo metodo, è motivato dalle considerazioni fatte in precedenza sulla presenza di sequenze di dimensione 2 !

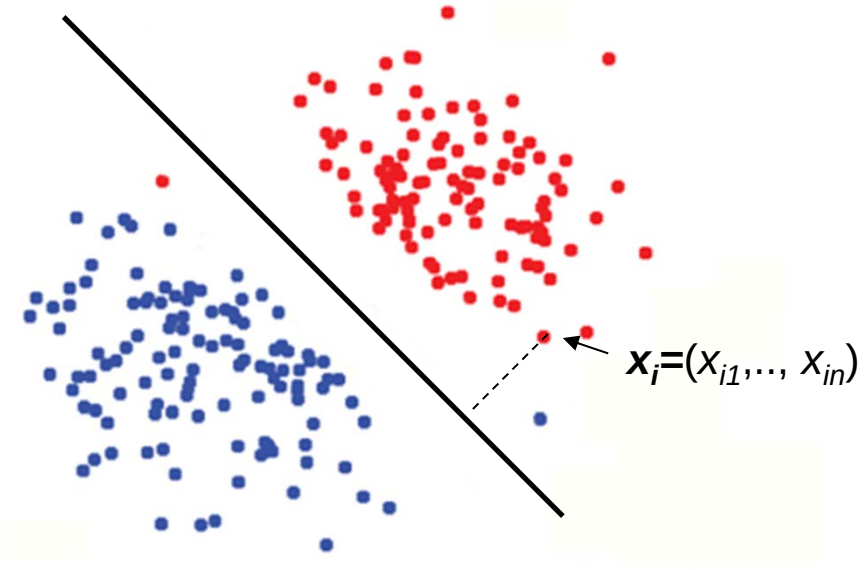
Approccio 2, segnali provenienti da un microarray:

Fase di estrazione delle caratteristiche: occorre associare ad un input “segnale” un insieme di numeri che rappresenteranno l’input all’algoritmo di apprendimento automatico



Nota: l’estrazione degli intervalli, ci consente di riconoscere la *forma a campana* che rappresenta la presenza del nucleosoma.

Un idea semplice su un algoritmo di apprendimento automatico



Bioinformatica

La bioinformatica ha a che fare con le scienze della vita, ovvero con qualcosa che riguarda anche la nostra salute

Oggi giorno esistono delle piattaforme biotecnologiche in grado di fornire impressionanti quantità di dati, la cui collezione ed analisi coinvolge necessariamente l'infomratica.

Basta riflettere sul fatto, che l'attività di ricerca in tale ambito si svolge in modo "interdisciplinare" ovvero i gruppi che lavorano sulle scoperte scientifiche sono composti da Biologi, Fisici, Matematici, Informatici.