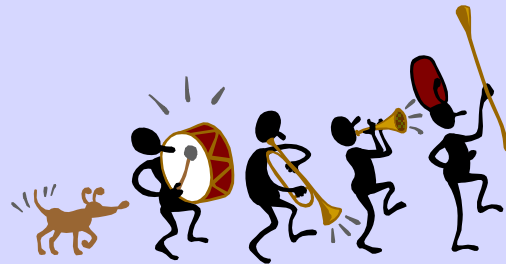


Dall'Informatica alla Biologia e ritorno



Paola Bonizzoni

Università degli Studi di Milano-Bicocca

Informatica e biologia



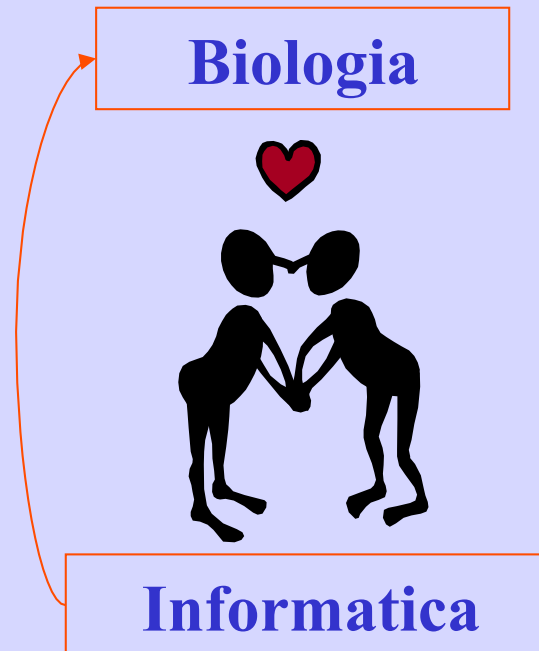
Bioinformatica:

Ricostruzione
evoluzione

Analisi di sequenze

Folding di Proteine

Simulazione di
processi biologici



L'informazione ed il DNA

- In un organismo i processi biologici elaborano *informazione*
- L'archivio delle informazioni è il *genoma*, costituito da un insieme di *cromosomi*
- Un cromosoma è una molecola di **DNA**, una doppia elica contenente un messaggio con alfabeto (A, C, G, T)

IL DNA



IL genoma

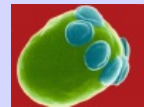
- Il genoma è formato da milioni di *basi* anche in organismi semplici:

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

Batterio (*E. Coli*)

4.8 Mbp

Lievito



14.4 Mbp

Verme (*C. elegans*)



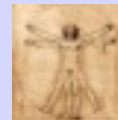
100.0 Mbp

Moscerino (*Drosophila*)



165.0 Mbp

Uomo (*Homo sapiens*)



3.3 Gbp

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

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 delle sequenze

Affrontare i problemi etici, legali e sociali collegati

Perché servono i computer: 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

- **Esempio**

ACCGT

CGTGC

TTAC

TACCGT

ACCGT

CGTGC

TTAC

TACCGT

TTACCGTGC (consensus)

Il DNA e le proteine

- Il DNA governa la costruzione e il funzionamento dell'organismo attraverso la produzione di *proteine*

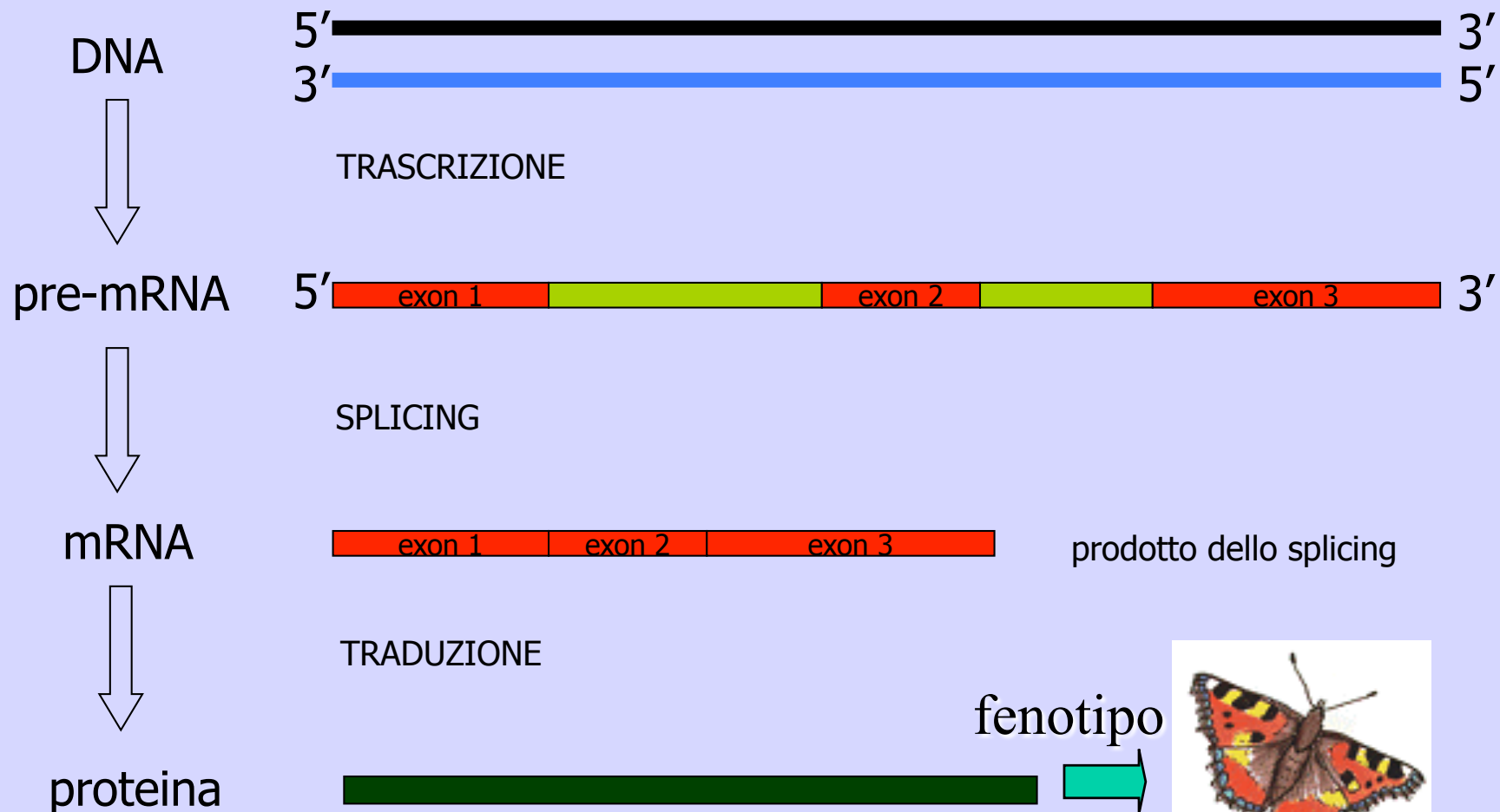
sono le principali responsabili della struttura e delle attività dell'organismo

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 DNA e l'espressione genica

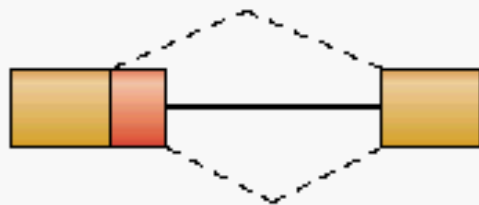


Lo splicing alternativo fa sì che un gene “produca” più proteine

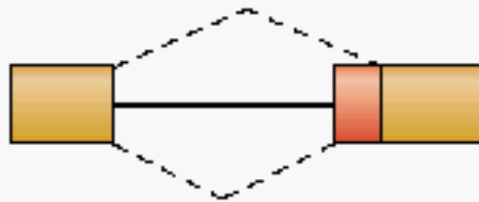
(a) Retained intron



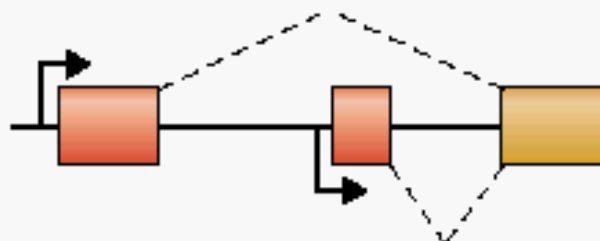
(b) Competing 5' splice sites



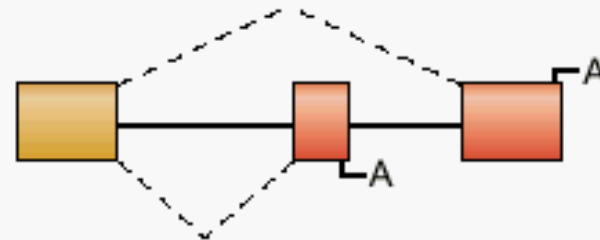
(c) Competing 3' splice sites



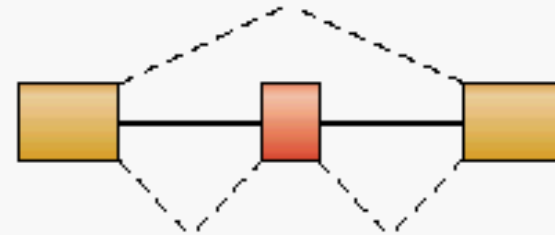
(d) Multiple promoters



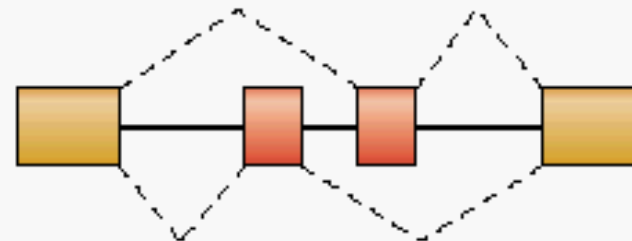
(e) Multiple poly(A) sites



(f) Cassette exons



(g) Mutually exclusive exons



Perché servono i computer - La “post-genomica”

- Estrarre un significato dai dati o sequenze
 - Identificare i geni sulle sequenze di DNA
 - Capire meccanismi di trascrizione del DNA
 - Trovare segnali sconosciuti
 - Capire l'espressione genica
 - Capire l'evoluzione
 - Capire le proteine

Quali “programmi” esegue il DNA?

Perché servono i computer

- **Usare la conoscenza**

- Individuare l'informazione sui geni che indica la presenza di una patologia (diagnosi)
- Riparazione o sostituzione dei geni
- Progettazione di farmaci con un reale impatto sulla malattia senza alterare il delicato equilibrio dell'organismo

Esempi: algoritmi

Calcolo mutazioni geniche tra genomi:

S1 = ACCGGT

S2 = AC-GGA

C cancellata, T sostituita con A

Input: due sequenze S1, S2 (genomi)

Output: minimo numero di mutazioni che trasformano S1 in S2

Esempi: algoritmi

Calcolo distanza tra genomi:

$G1 = 1, 2, 3, 4, 5, 6, 7, 8, 9$ $Geni = interi$

$G2 = 1, 2, \boxed{5, 4, 3}, 6, 7, \boxed{9, 8}$

Mutazioni = operazioni (es. inversioni)

Es. Inversione di 3,4,5  5,4,3

Input: due genomi $G1$ e $G2$

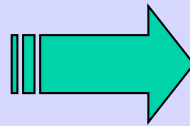
Output: minimo numero di inversioni che trasformano $G1$ in $G2$

Esempi: algoritmi

La **distanza evolutiva** tra due specie è il minimo numero di mutazioni che sono necessarie per trasformare un genoma nell'altro.



~ 180 mutazioni



Esempio: progetto ASPIC

Input: un gene

Output: trascritti (proteine del gene)



The screenshot shows the ASPic homepage. At the top left is the ASPic logo, which includes a stylized sunburst icon and the text "ASPic" in large red letters, followed by "Alternative Splicing Prediction" in smaller blue and red text. Below this is a description: "A web tool to detect the exon-intron structure of a gene by comparing its genomic sequence to the related cluster of ESTs". To the right of the ASPic logo are two logos: the "Laboratory of Bioinformatics and Comparative Genomics" logo, which is a circular diagram with various colored lines, and the "CASPUR" logo, which consists of a blue stylized 'C' and a red grid-like 'C'.

Below the logos is a navigation bar with the following links: Home, About Us, General Info, Architecture, Web Services, Run ASPic (highlighted in red), DB, and Help.

The main content area is titled "Aspic Homepage" in bold black text. Below the title is a paragraph of text:

ASPic (Alternative Splicing Prediction) is a web-based tool to detect the exon-intron structure of a gene by comparing its genomic sequence to the related cluster of ESTs. ASPic predicts constitutive and alternative splice sites through a novel methodology that uses a combined analysis of all EST alignments to make them most compatible to a common exon-intron structure of the gene considered. The project aims to investigate new computational methods for alternative splicing prediction and then to develop computational tools based on these procedures for gene variant detection.

ASPIC output

Detailed Intron view

In this page we will show all the introns found for your ASPic Id.
If you want a particular file about this process you'll try to look in the Download section.

Intron	Relative Start	Relative End	Absolute Start	Absolute End	Length	#ESTs	Donor &	Mismatch (15 bp)	Mismatch (15 bp)
Intron 1 [10355]-[41690]									
1	10,355	41,690	127,865						
2	41,811	42,776	127,864						
3	42,978	46,120	127,861						
4	46,199	47,195	127,860						
5	47,445	51,815	127,855						
6	47,445	68,639	127,838						
7	51,867	69,690	127,837						
8	52,077	68,639	127,838						

DONOR SCORE: 87

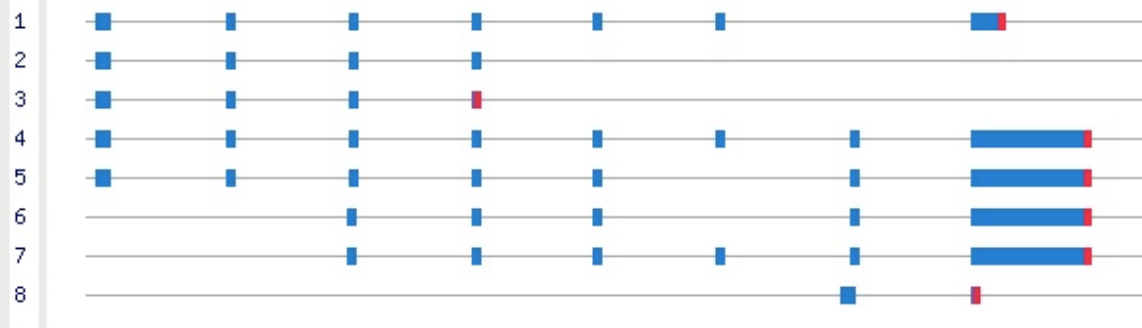
ACCEPTOR SCORE: 94

GGGTGAGT

ATTTTTCATGCAGA

TTTTCCCCTCCCCGGGTGAGTGGCGACGGCCGAGG----TATGCAATTTTTCATGCAGATATGGAATGTGCAG
 TTTTCCCCTCCCCCGG 38 39 ATATGGAATGTGCAG
 TTTTCCCCTCCCCCGG 82 83 ATATGGAATGTGCAG
 TTTTCCCCTCCCCCGG 82 83 ATATGGAATGTGCAG
 TTTTCCCCTCCCCCGG 82 83 ATATGGAATGTGCAG
 TTTTCCCCTCCCCCGG 82 83 ATATGGAATGTGCAG
 TTTTCCCCTCCCCCGG 139 140 ATATGGAATGTGCAG
 TTTTCCCCTCCCCCGG 179 180 ATATGGAATGTGCAG
 TTTTCCCCTCCCCCGG 199 200 ATATGGAATGTGCAG
 TTTTCCCCTCCCCCGG 201 202 ATATGGAATGTGCAG
 TTTTCCCCTCCCCCGG 275 276 ATATGGAATGTGCAG
 TTTTCCCCTCCCCCGG 333 334 ATATGGAATGTGCAG
 TTTTCCCCTCCCCCGG 416 417 ATATGGAATGTGCAG
 TTTTCCCCTCCCCCGG 330 331 ATATGGAATGTGCAG
 356 ATATGGAATGTGCAG
 347 ATATGGAATGTGCAG
 353 ATATGGAATGTGCAG
 355 ATATGGAATGTGCAG
 361 ATATGGAATGTGCAG
 388 ATATGGAATGTGCAG
 397 ATATGGAATGTGCAG

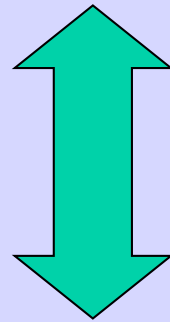
BX484339
 BT019452
 BT019453
 CV029098
 CR542254
 CR979152
 BE146366
 DR005472
 J04101
 BQ356500
 BM452565
 BI825043
 BM142037
 BE799215
 CK819846
 BC017314
 NM_005238
 X14798
 BI911387
 BP365215



Bonizzoni P, Rizzi R, Pesole G. ASPIC: a novel method to predict the exon-intron structure of a gene that is optimally compatible to a set of transcript sequences. BMC Bioinformatics. 2005 Oct 5;6(1):244

DNA e computer

Processi biologici



Processi di calcolo

DNA Computing... o calcolo naturale?



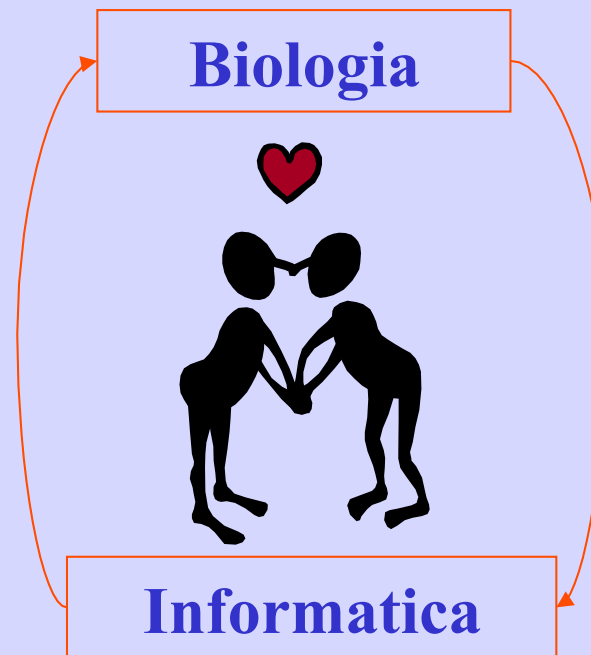
Bioinformatica:

Ricostruzione
evoluzione

Analisi di sequenze

Folding di Proteine

Simulazione di
processi biologici



Bio-inspired Computing:

DNA computing

Algoritmi genetici

Ant computing

Reti neurali

I limiti delle tecniche di computazione tradizionali

- Complessità “computazionale”
 - Algoritmi non praticabili (output ottenuto dopo secoli...):
 - problemi di ottimizzazione: numero elevato di soluzioni da controllare

Esempi problemi di ottimizzazione:

Riempire un container, o una nave

Trovare il percorso più corto per un commesso viaggiatore

Trovare la distribuzione migliore delle ambulanze in città

Realizzare la rete del metano con meno tubi possibile

Il calcolo “naturale”

- **Modelli di calcolo ispirati da meccanismi biologici**
 - Reti neurali
 - Algoritmi genetici
- **Modelli di calcolo basati su tecniche biologiche**
 - DNA computing
- **Modelli di calcolo basati sui principi quantistici**
 - Quantum computing

Verso il DNA computing

"In 1959, Richard Feynmann gave a visionary describing the possibility of building computer that were *sub-microscopic*. Despite remarkable progress in computer miniaturization, this goal is far to be achieved.

HERE THE POSSIBILITY OF COMPUTING
DIRECTLY WITH MOLECULES IS EXPLORED"

...

Science 1994



1959

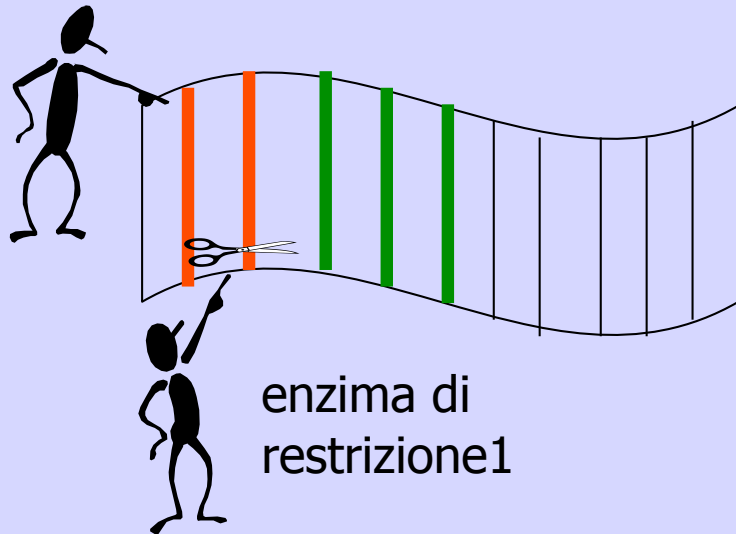
Richard Feynman immagina computer
“submicroscopici”

Computer quantistici

Computer Molecolari

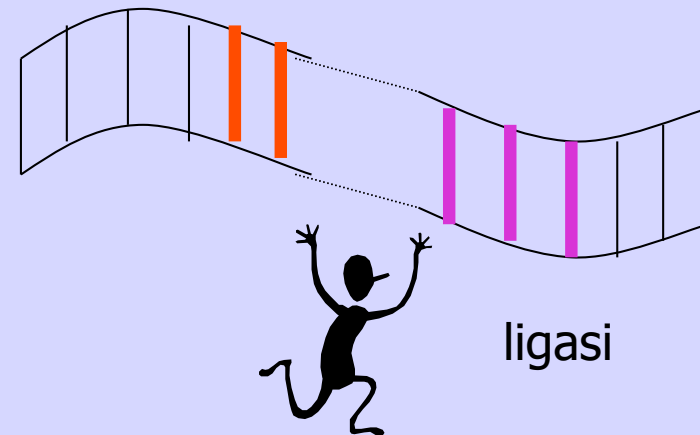
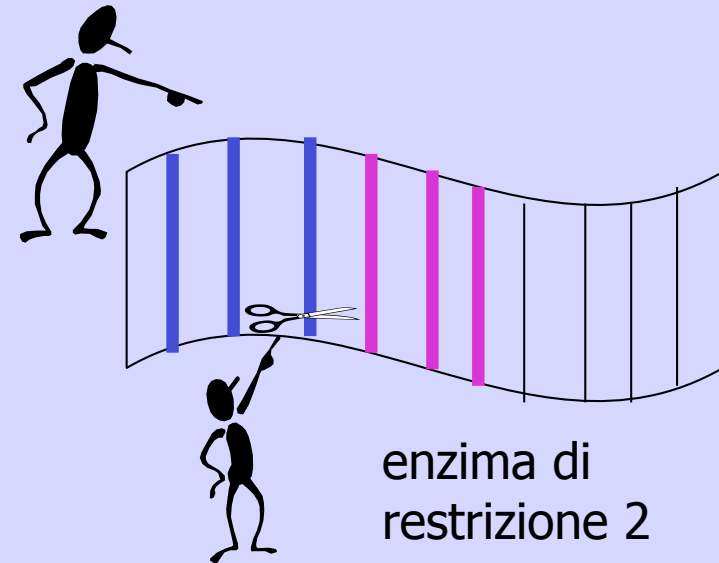


Verso il DNA computing



1987

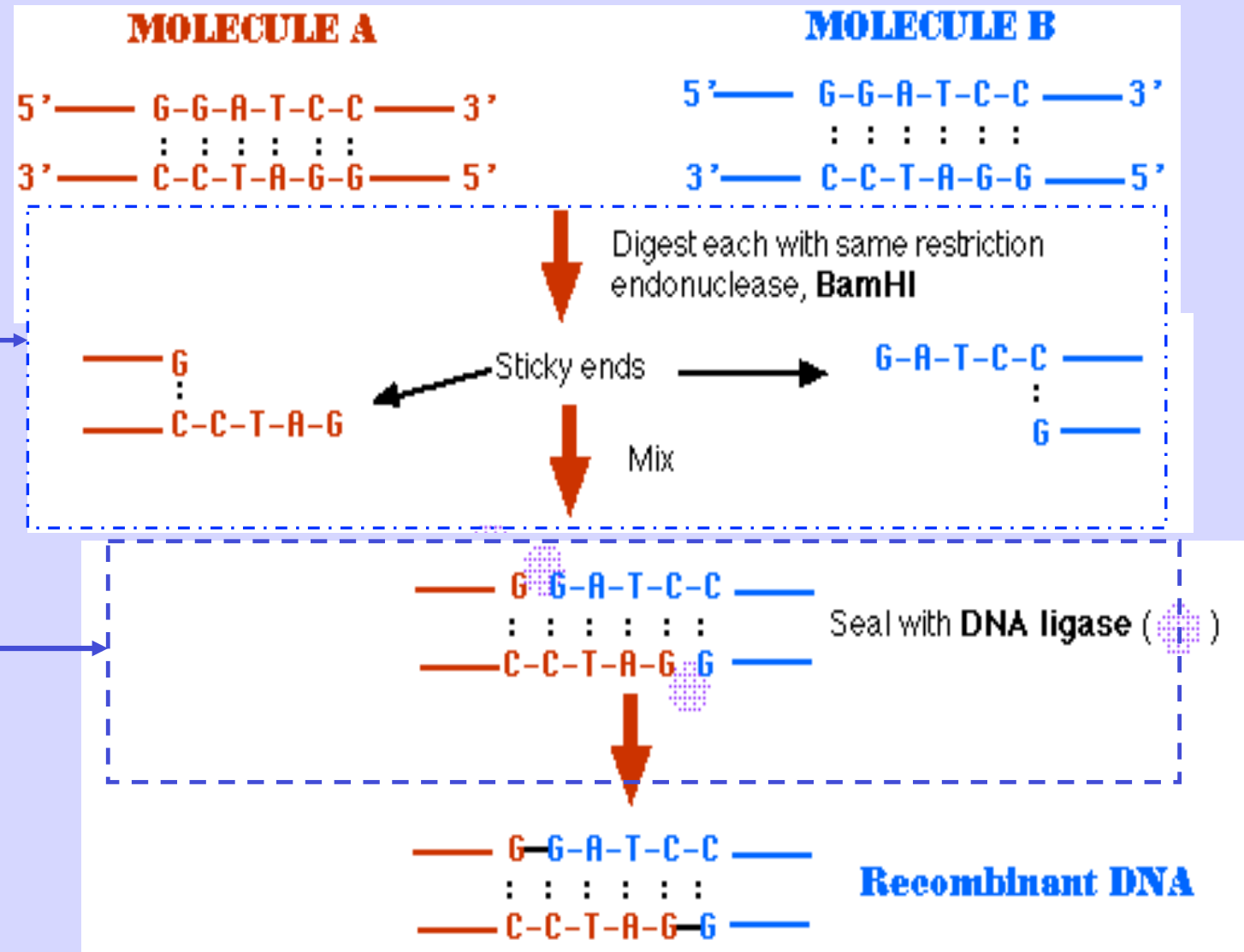
**Tom Head definisce i
sistemi di splicing,
modelli formali
dell'azione degli enzimi
di restrizione**



Verso il DNA computing

Cut:
Enzimi di
restrizione

Paste:
enzima
ligase



Verso il DNA computing

Tom Head 1987 (Bull. of Math. Biology)

Sistemi di calcolo basati sull'operazione di splicing

Input: insieme di molecole (**stringhe**)

Programma: insieme di enzimi (**regole di splicing**)

Output: un linguaggio o insieme di stringhe
(**risultato della computazione**)



Modelli di calcolo
equivalenti alle Macchine di
Turing !

Verso il DNA computing

1994

**Leonard Adleman risolve
HPP con tecniche
biologiche**

❑ La matematica nelle cellule!

❑ comportamento del DNA

come

Macchina di Turing



**Soluzione di
problemi NP Completi**

La metodologia tipica



ma...

1 secondo per eseguire il calcolo

600000 secondi per ottenere il risultato



Che vantaggi potrebbe darci il calcolo molecolare?



Parallelismo: 10^{20} operazioni in parallelo

Velocità: 10^{20} op/sec (vs 10^{12} op/sec)

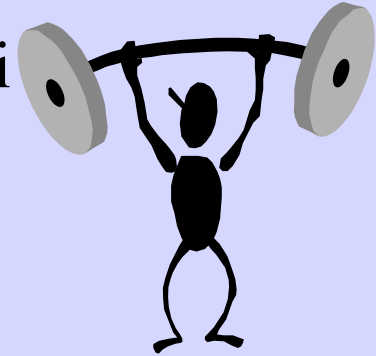
Efficienza Energetica: $2 \cdot 10^{19}$ op/Joule (vs 10^9 op/Joule)

Densità di informazione: 1 bit/nm³ (vs 1 bit x 10^{12} nm³)

L'altra faccia della luna ...

□ Limiti computazionali

- Istanze “grandi” (500 nodi) restano intrattabili
- Complessità in “peso”



DNA per risolvere HPP su 200 vertici

=

Più della massa della Terra

<< ... new computational complexity measure:

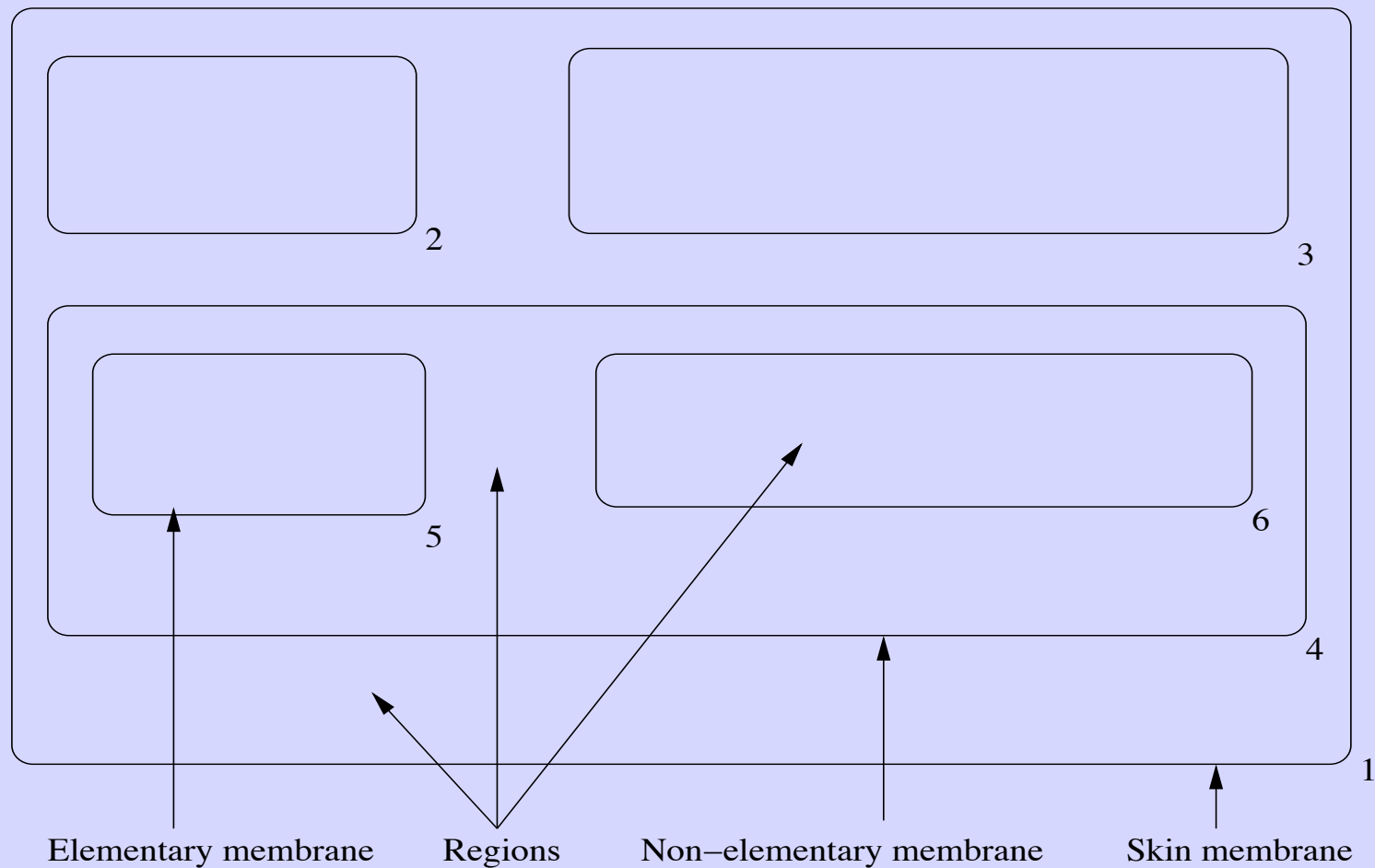
the weight of the computation >>

[Hartmanis, Bull. of EATCS 1996]

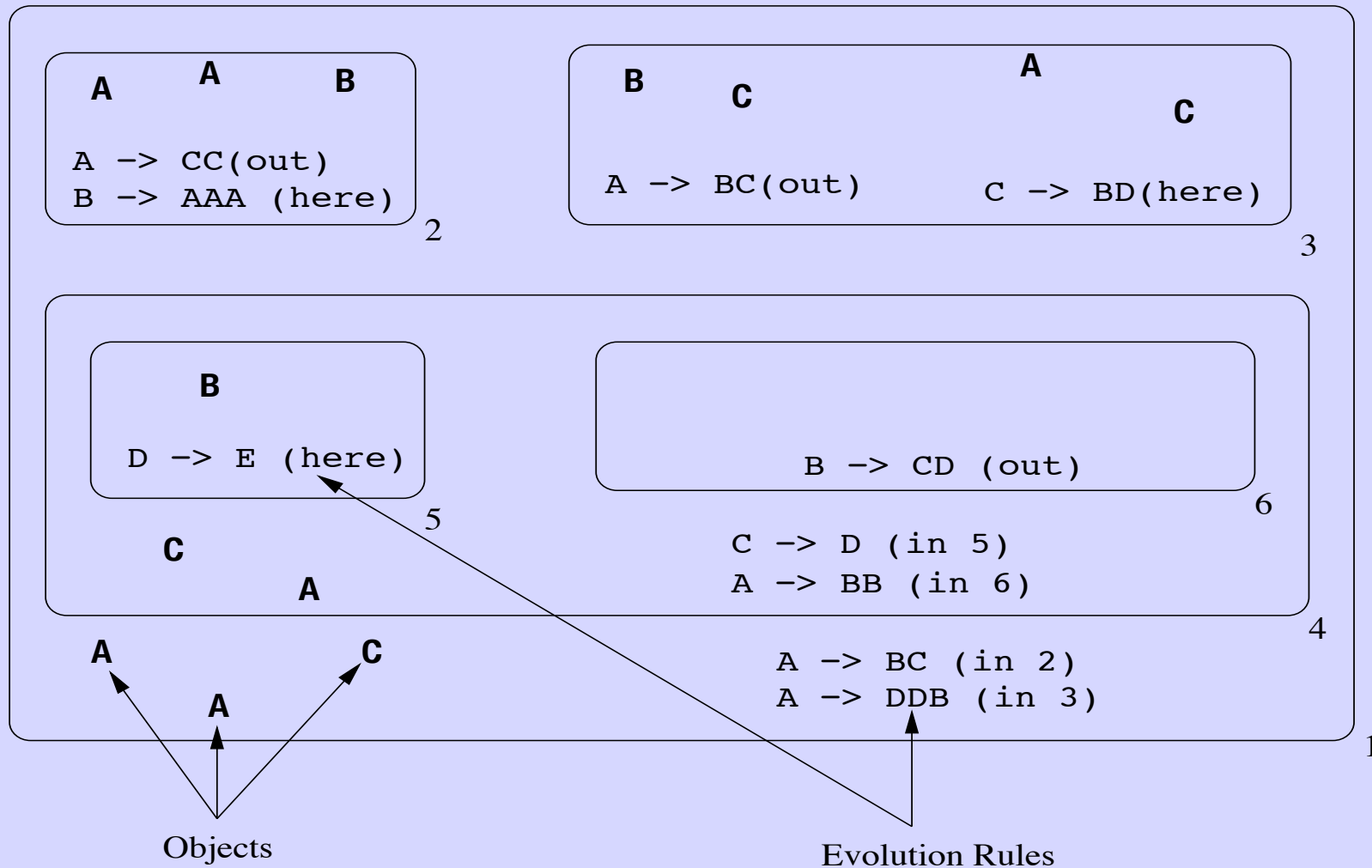
La teoria: “dry” DNA computing

- **Potenza computazionale**
 - Equivalente alle macchine di Turing
- **Modelli formali:**
 - Sistemi di accoppiamento (splicing)
 - Sistemi a inserimento/cancellazione
 - Sistemi a membrane
 - Sistemi autoassemblanti
 - ...

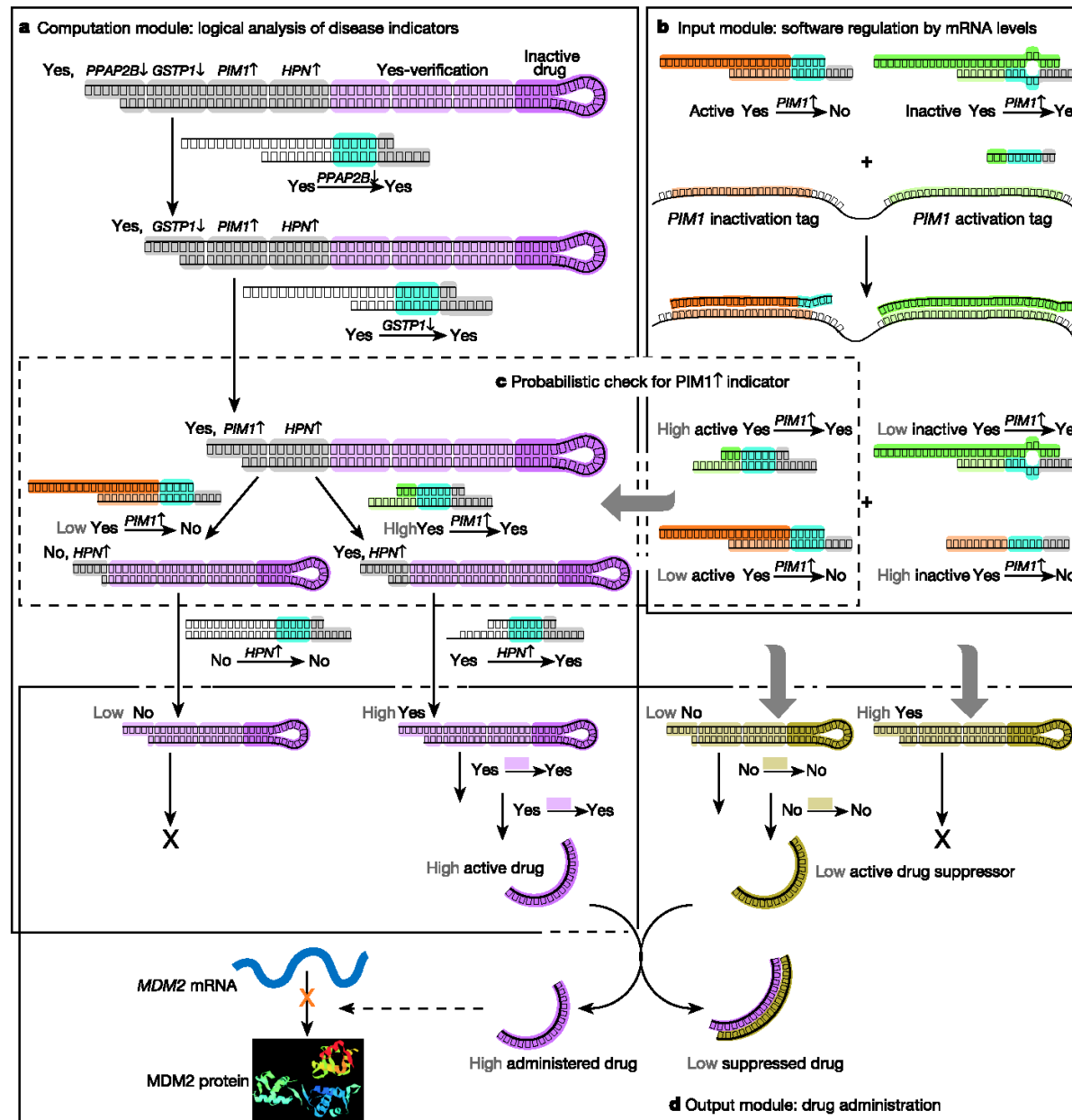
Sistemi a Membrane: la struttura



Oggetti e Regole di evoluzione



L'automa diagnostico di Shapiro



Conclusioni

