Bioinformatica: DNA e Algoritmi

Alberto Policriti

Dpt. of Mathematics and Informatics, University of Udine.

Applied Genomics Institute
IGA
ISTITUTO DI GENOMICA APPLICATA

Di cosa parleremo

In generale

- Definiamo i termini: DNA & Algoritmi
- Tecnologie ($\Rightarrow$ Problemi) $\Rightarrow$ *Sequenziamento e Assemblaggio*
- Bioinformatica: Passato e Futuro

Di cosa parleremo

In generale

- Definiamo i termini: DNA & Algoritmi
- Tecnologie ($\Rightarrow$ Problemi) $\Rightarrow$ *Sequenziamento e Assemblaggio*
- Bioinformatica: Passato e Futuro

Di cosa parleremo

In generale

- Definiamo i termini: DNA & Algoritmi
- Tecnologie ($\Rightarrow$ Problemi) $\Rightarrow$ *Sequenziamento e Assemblaggio*
- Bioinformatica: Passato e Futuro

Di cosa parleremo

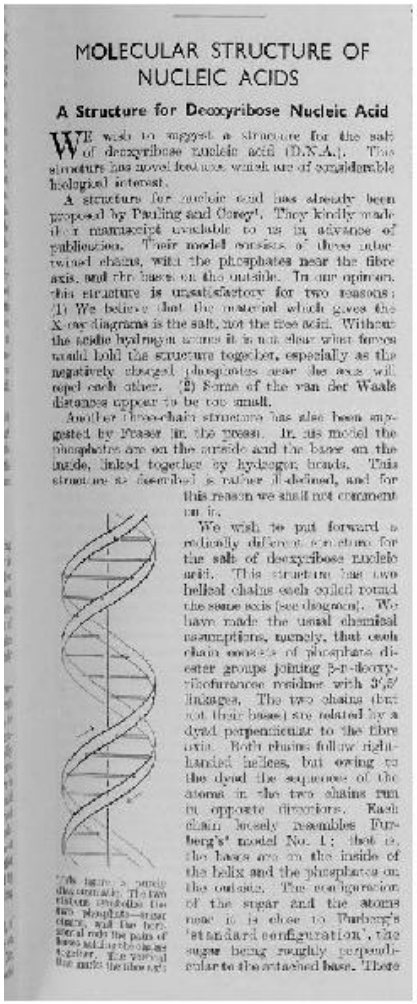
In generale

- Definiamo i termini: DNA & Algoritmi
- Tecnologie ($\Rightarrow$ Problemi) $\Rightarrow$ *Sequenziamento e Assemblaggio*
- Bioinformatica: Passato e Futuro

Le aree

- Algoritmica su stringhe
  - Matematica del discreto
  - String matching esatto e approssimato
- Systems biology
  - Matematica del continuo
  - Automi e biologia
- Computare usando il DNA

Crick e Watson: 1953



MOLECULAR STRUCTURE OF NUCLEIC ACIDS

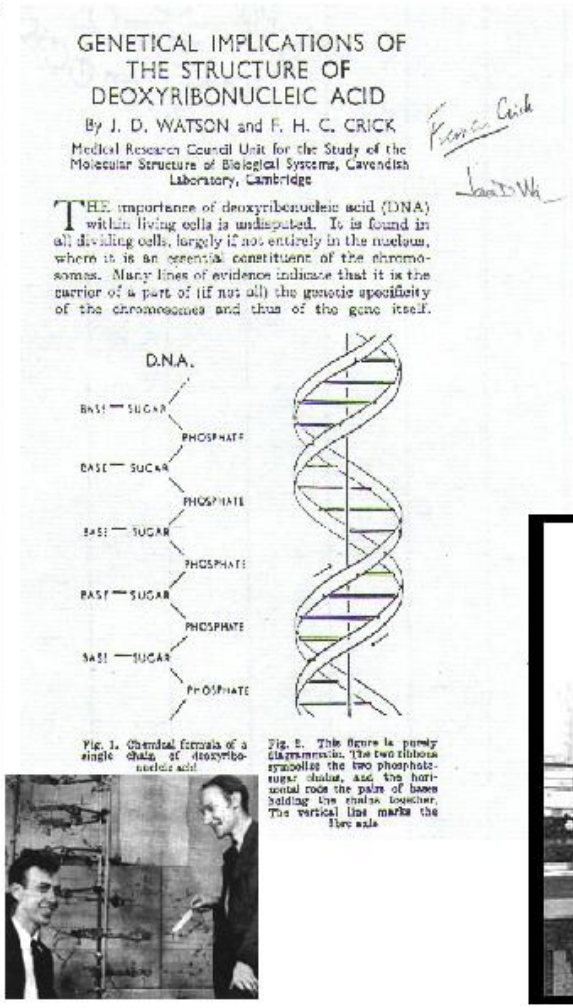
A Structure for Deoxyribose Nucleic Acid

We wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey[1]. They kindly made their manuscript available to us in advance of publication. Their model consists of three interwined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate diester groups joining β-D-deoxyribofuranose residues with 3',5' linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed helices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions. Each chain loosely resembles Furberg's[2] model No. 1; that is, the bases are on the inside of the helix and the phosphates on the outside. The configuration of the sugar and the atoms near it is close to Furberg's 'standard configuration', the sugar being roughly perpendicular to the attached base. There



GENETICAL IMPLICATIONS OF THE STRUCTURE OF DEOXYRIBONUCLEIC ACID

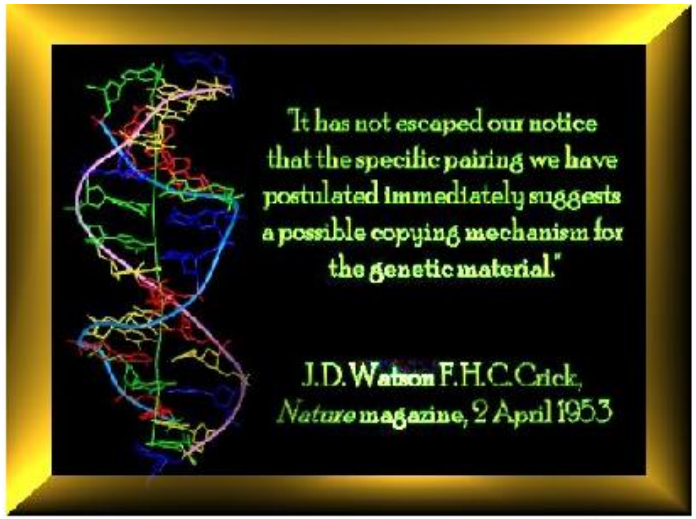
By J. D. WATSON and F. H. C. CRICK

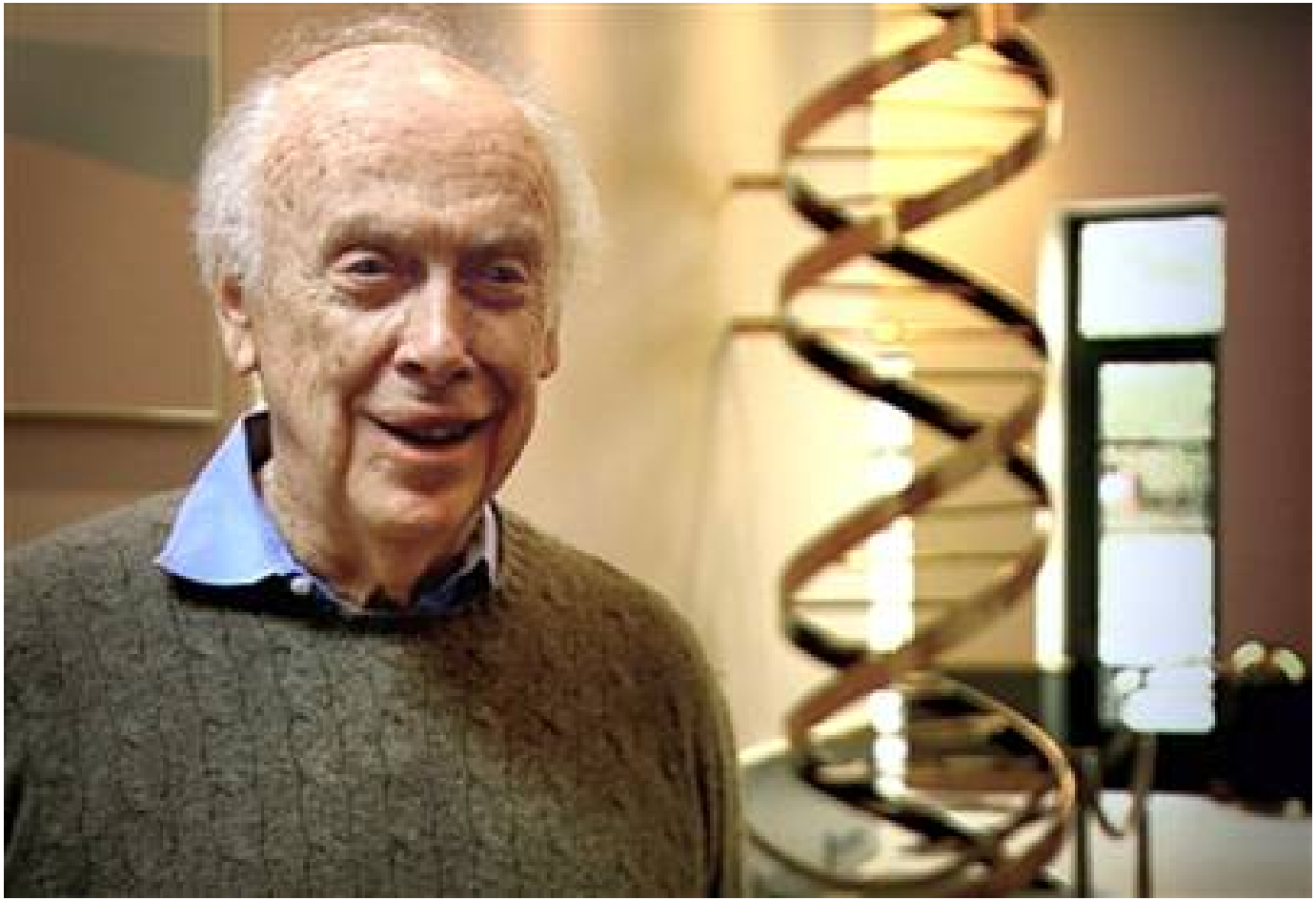
Medical Research Council Unit for the Study of the Molecular Structure of Biological Systems, Cavendish Laboratory, Cambridge

THE importance of deoxyribonucleic acid (DNA) within living cells is undisputed. It is found in all dividing cells, largely if not entirely in the nucleus, where it is an essential constituent of the chromosomes. Many lines of evidence indicate that it is the carrier of a part of (if not all) the genetic specificity of the chromosomes and thus of the gene itself.

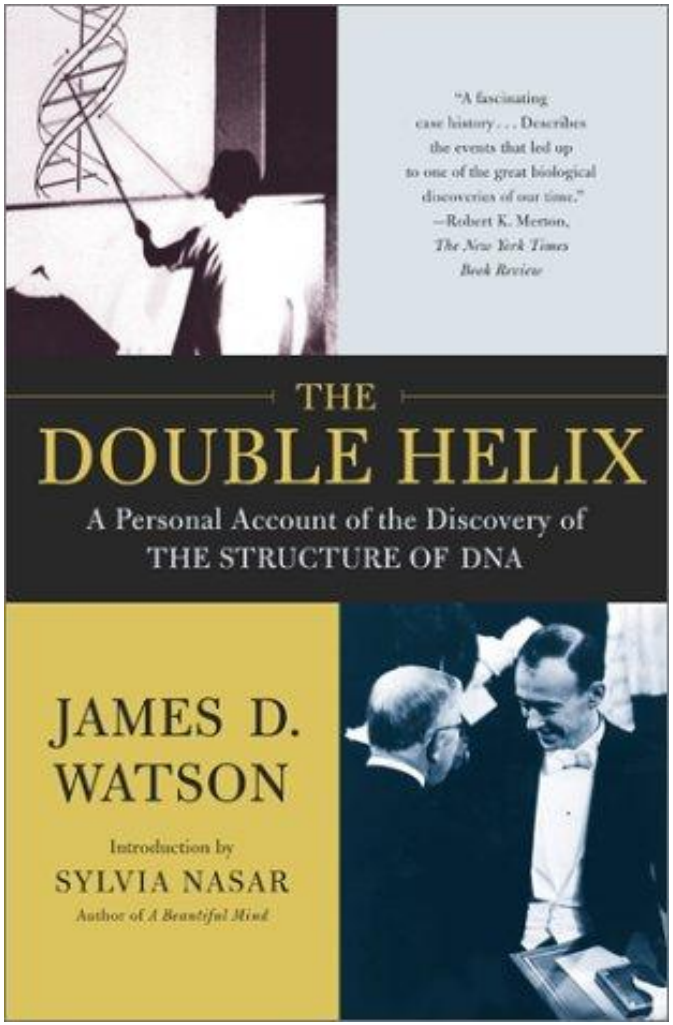
Fig. 1. Chemical formula of a single chain of deoxyribonucleic acid

Fig. 2. This figure is purely diagrammatic. The two ribbons symbolize the two phosphate-sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

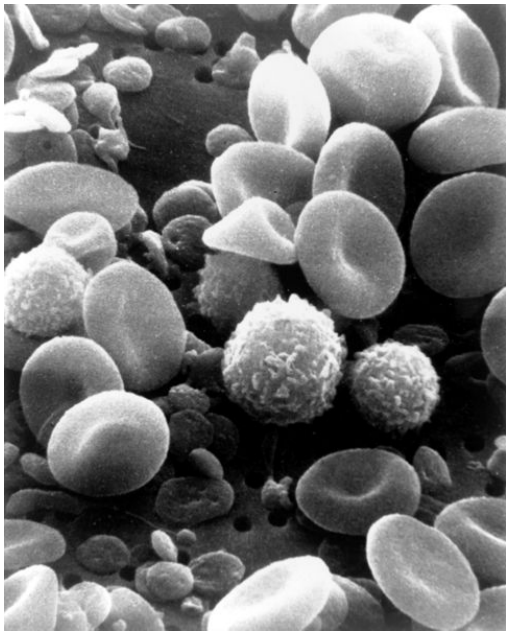


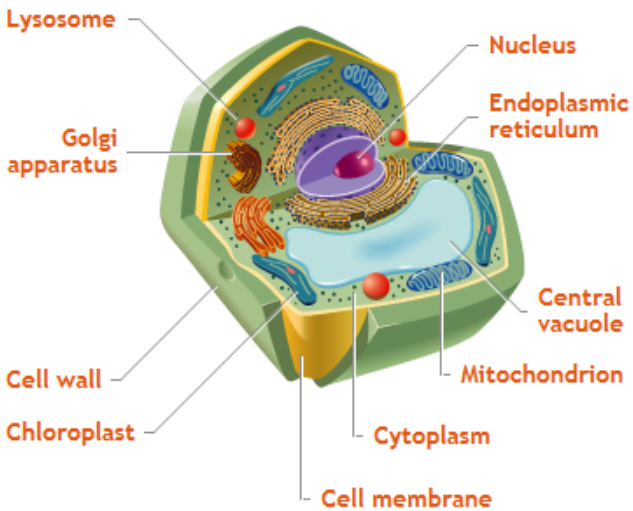


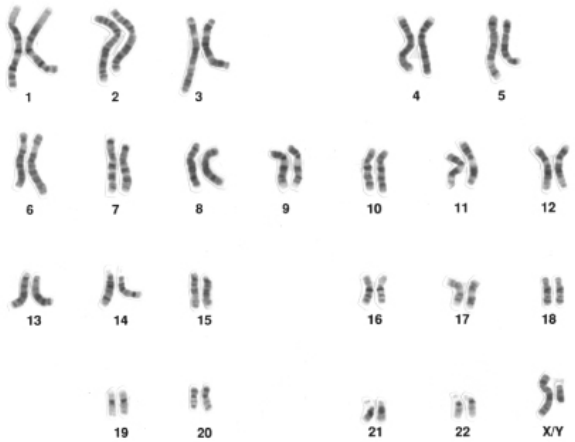
Lettura interessante



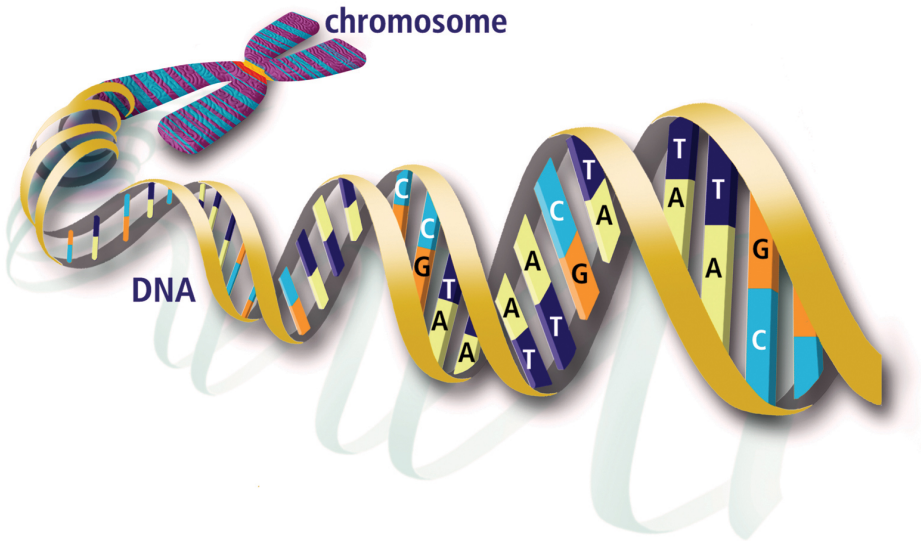
Da cosa partiamo



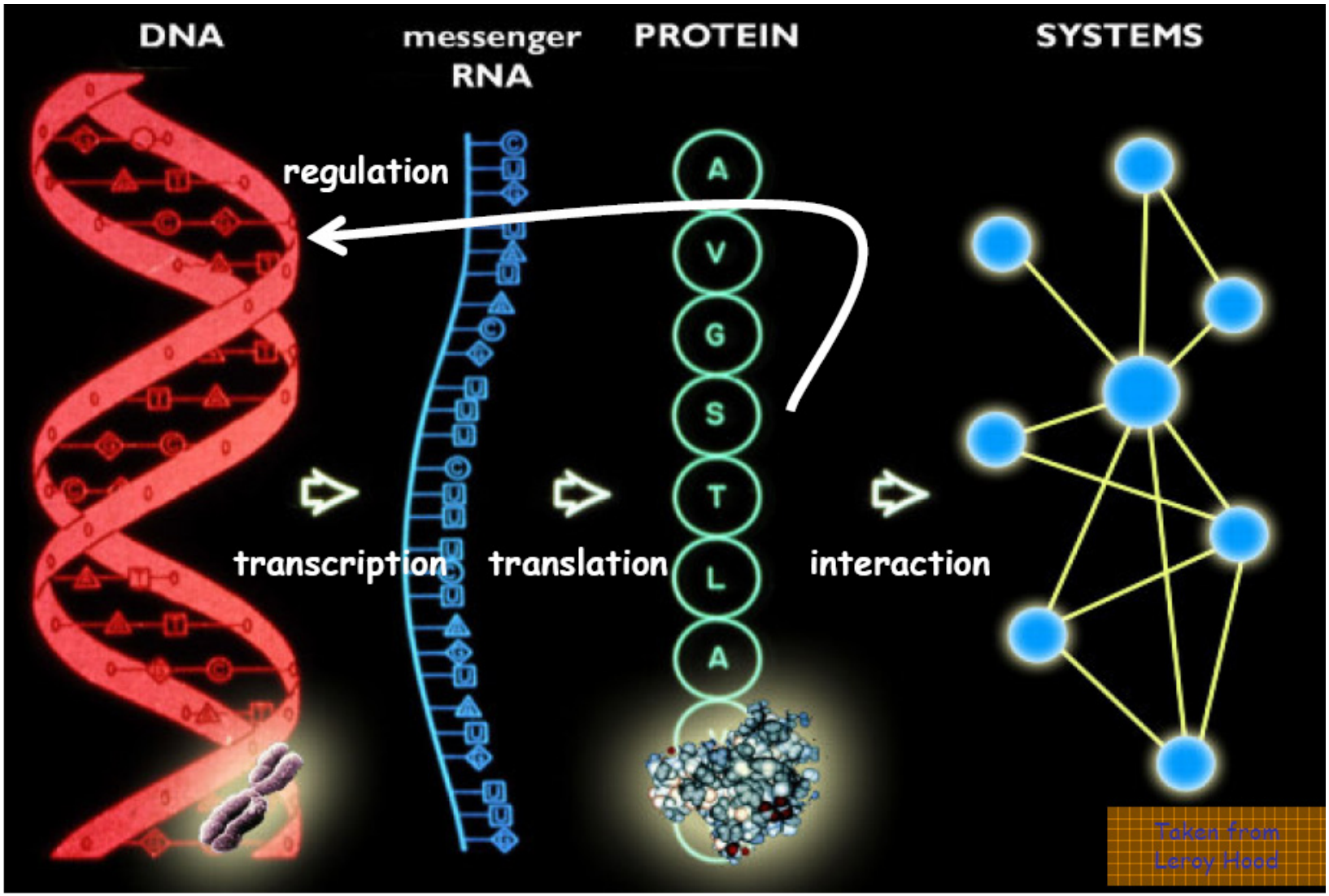




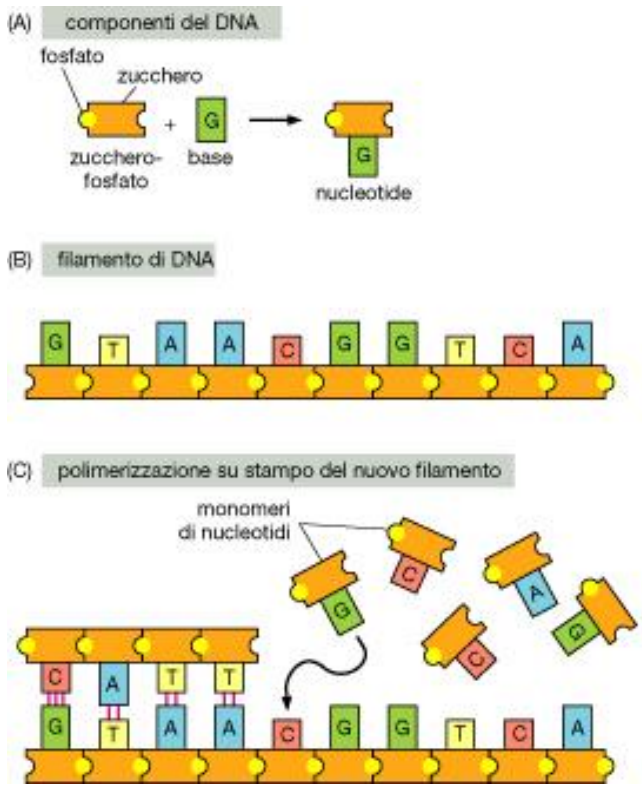
Da cosa partiamo

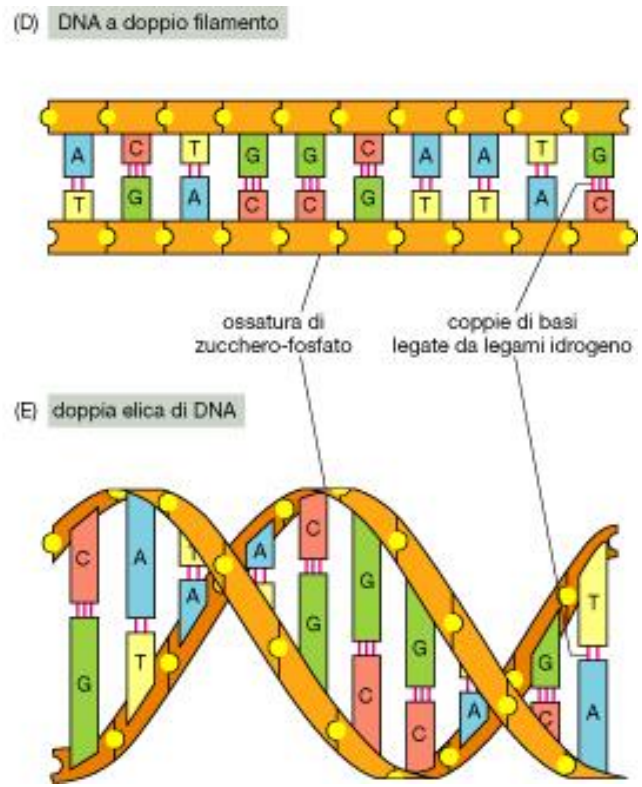


Da cosa partiamo

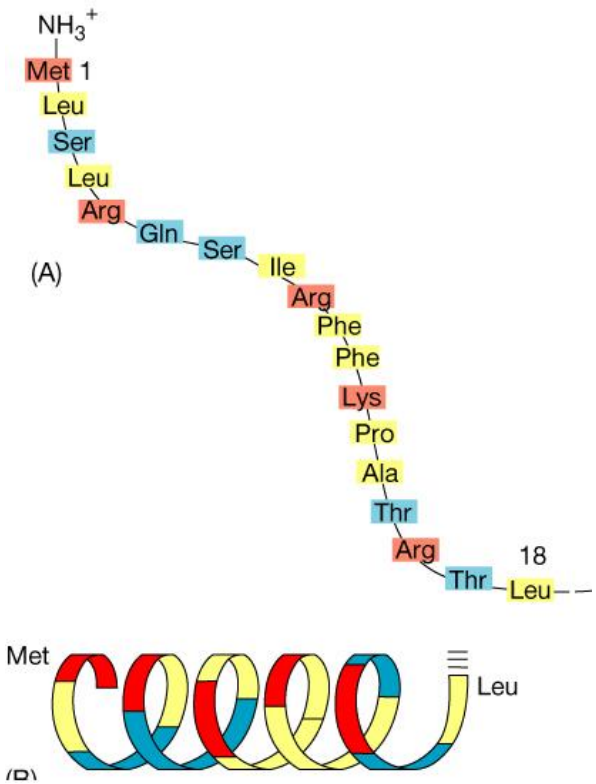


Memorizzazione dell'informazione

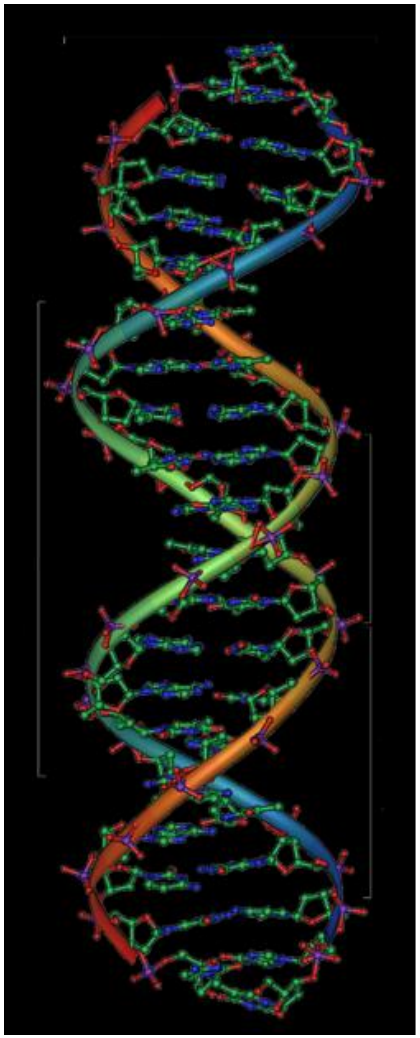




Memorizzazione dell'informazione



DNA



Le caratteristiche

- un "*turn*" di DNA: 3.4nm 10.5 bases
- una cellula: 2m di DNA
- un uomo: 10^{13} celule
- .... un mucchio di DNA!

Estremamente delicato da manipolare

La filosofia è scritta in questo grandissimo libro che continuamente ci sta aperto innanzi a gli occhi (io dico l'universo), ma non si può intendere se prima non s'impara a intender la lingua, e conoscer i caratteri, ne' quali è scritto. Egli è scritto in lingua matematica, e i caratteri son triangoli [A-T], cerchi [C-G], ed altre figure geometriche [Met, Lys, Leu, ...], senza i quali mezi è impossibile a intenderne umanamente parola; senza questi è un aggirarsi vanamente per un oscuro laberinto.

Galileo Galilei

Il nostro obiettivo

TCAGGGCCAGCAGCAAAGTTTCCAGCTGATCATCCTGATGGTGCGCGGTGAGTAAGACACCACCTTCCGG
GAGGTTTTGGCGAAAAACGTCGTAGCGGGCCTGACGGGCAAGATCCTCAAGACTTTGCCGGGCCTTTTTC
TCAATTTCAACACGTTTTGCTATGAATGGTACTGTCAGTACATCGCAAACCTGCGCGCAATGGGTCAGCC'
ATTGGTCAGCCTTAGGATTTAAACCATGATGCACATGCACTGCCTGCAGGGAAAAGCCGTCTCGCTGAGA
TATATAACGGGCCAGCAATTCAAGTAACACGCGGGAATCGAGGCCGCCGCTGAATCCAACCAGTAAGTGC
CCGGGCCGGATCAGCTGATCCAGCACAGAAAAGACACTTGCTTCCAATTCATGATGACTCACAACAGATA
CTTCATCGATAAAGTACAATATGCGGATATTAGCAGATCCTGTTTGTCATTACGTGAACCGGGATAATAA
GCCATATTGCTTGCGGCTCGTTTACACGATAACGATGATCGGATGGAAGAAATAAAAAAAGCCCCTGAGC
GGGGCTTAAAGAAAAGGAAAGCATTAATACTTAACGGGTTTCCAGCGGAGCGAAGCTCTTCACAAGATCA
TCAATGGCTTTCATCTGTTGCAGGAATGGCTCCAGCTTATCCAGTGGTAATGCACTAGGACCATCACAAC
GTGCATGTTCCGGATCAGGATGCGACTCAATAAACAAGCCAGCCAGACCAACCGCCATACCAGCGCGTGC
CAGTTCAGTCACTTGCTCACGACGACCACTTGAAGCCGCACCCAGCGGGTCACGACACTGCAGGGAATGC
GTTACGTCAAAAATCACCGGGCTGCCCTGGCTGGCATCTTTCATCACGCCGAAGCCCAGCATGTCGACCA
CCAGATTGTCATAACCGAAATTCACACCGCGTTCACACAGAATAATACGGTCGTTACCGCATTCGGCGAA
TTTCTCAACGATGTTCTTAACCTGTCCCGGGCTCAGGAACTGCGGTTTTTTCACGTTAATCACGTTACCT
GTTTTAGCCAGAGCTTCCACCAGATCAGTCTGACGCGCCAAGAAGGCCGGTAATTGCAGCACATCCACG
CTTCACCGATCGGTTTGGCTTGCCATGTTTCATGCACATCTGTGATCAGGCTGACACCGAAGGTTTTCTT
CAGTTCTTCAAAAATACGCAGACCTTCTTCCATACCCGGACCACGGTAGGAATGAACTGAGGAGCGGTTG
GCTTTATCCCAGCTGGCTTTAAATACCAAAGGGATGCCCAGTTTTTCAGTTACGGTGACGTAGGTTTCAC
AGATCGACATCGCCAGATCCCGTGACTCCAGTACATTCATGCCACCAAACAGTACGAATGGCTGATCATT
AGCAACATTAATTCCGTTAAACTGAACGACTTTCTGTTTCATTCATCACTCCATCAGTGAAAAATTATGG
CGTCATGATCCAGCATCGCCAGCTGTACTTTCAGTACGGCTGCAACCGGATCCTGAGGACATTGTTCAAT
AAAATAGGTGTAATCATGAGCTGCCAGCTGCGGACATTCCAGCTGTTCATATACCAGCCCGCGATCACGG
ATTTCATAGGGGTCATCGGGGTTCATCGCCAGCAAAACCTCGCTGCAGCGTAACGCCTCCGGCAAACAAC
GACTTTGTAACATACAACCTTTAATCACGCTCAATAAACGTGACATTATCTGACGCTGATCTGTCTCTTT
TGTATATTTACTATCCATCCGGGTTAAGTCACCGAGTGTGGCTCGCAGCATTAATGACTGCTGCTCATAA
TCCCACTCTTCACCGGTAAACGGATCGATTAATACAGGCTTACTTTCCGGGAAAAGCAGCAAGAAGTTAC
CAGGGAAGCAAACACCACGCACCGGCAATTCTATCGAATGAGCCAGATGCATTAACACGATACCCAGCGT
GAGCGGCAAACCAGTGCGCTGCATCAGCACCTGATCCAGCAGGCAGTTTTCCGGTGCATAGTAATTCTGC
CAGTCACCGGAAAATTGCAATTCATGATAGAACGCATGCAACAGTTGTTCACGACGCCCTTTATCATCAG
GGGCCGTAATATACGGCCGGAGCTCTGTACACAAAGCGCGTAGCCGCTGCAATCCGTCGATCAATACCGG
TTTATCTTCTACATCTTCTCACCCCATCAATCCCACCATCCCAATATCCACCCTTCAAAATCACTATCC

Termini

Sequenziamento

- estrazione e frammentazione del DNA
- *sequenziamento* dei frammenti e generazione delle reads

wet-lab activity

Assemblaggio

- *stoccaggio* delle reads (+ altre informazioni)
- *assemblaggio* delle reads (corte) in contigs (lunghe)

dry-lab activity

La combinatorica del problema

Vincoli

- L'attività di *sequenziamento* produce **reads** di lunghezza limitata
- Il **costo** di un progetto è proporzionale al numero di reads prodotte

La combinatorica del problema

Vincoli

- L'attività di *sequenziamento* produce **reads** di lunghezza limitata
- Il **costo** di un progetto è proporzionale al numero di reads prodotte

L'input degli algoritmi di *assemblaggio*

La combinatorica del problema

Vincoli

- L'attività di *sequenziamento* produce **reads** di lunghezza limitata
- Il **costo** di un progetto è proporzionale al numero di reads prodotte

L'input degli algoritmi di *assemblaggio*

GACTTTGTAACATACAACCTTTAATCACGCTCAATATGACATTATCTGACGCTGATCTGTCTC...

La combinatorica del problema

Vincoli

- L'attività di *sequenziamento* produce **reads** di lunghezza limitata
- Il **costo** di un progetto è proporzionale al numero di reads prodotte

L'input degli algoritmi di *assemblaggio*

GACTTTGTAACATACAACCTTTAATCACGCTCAATATGACATTATCTGACGCTGATCTGTCTC...

GACTTTGTA ACATACAAC CTTTAATCACG CTCAATATGACAT TATCTGACGCTGA TCTGTCTC...

La combinatorica del problema

Vincoli

- L'attività di *sequenziamento* produce **reads** di lunghezza limitata
- Il **costo** di un progetto è proporzionale al numero di reads prodotte

L'input degli algoritmi di *assemblaggio*

GACTTTGTAACATACAACCTTTAATCACGCTCAATATGACATTATCTGACGCTGATCTGTCTC...

CTTTAATCACG
ACATACAAC
CTCAATATGACAT
TATCTGACGCTGA
TCTGTCTC
GACTTTGTA
...

La combinatorica del problema

Vincoli

- L'attività di *sequenziamento* produce **reads** di lunghezza limitata
- Il **costo** di un progetto è proporzionale al numero di reads prodotte

L'input degli algoritmi di *assemblaggio*

GACTTTGTAACATACAACCTTTAATCACGCTCAATATGACATTATCTGACGCTGATCTGTCTC...

CTTTAATCACG
ACATACAAC
CTCAATATGACAT
TATCTGACGCTGA
TCTGTCTC
GACTTTGTA
...

?

La combinatorica del problema

Vincoli

- L'attività di *sequenziamento* produce **reads** di lunghezza limitata
- Il **costo** di un progetto è proporzionale al numero di reads prodotte

La soluzione: *tante* copie

La combinatorica del problema

Vincoli

- L'attività di *sequenziamento* produce **reads** di lunghezza limitata
- Il **costo** di un progetto è proporzionale al numero di reads prodotte

La soluzione: *tante* copie

GACTTTGTA

La combinatorica del problema

Vincoli

- L'attività di *sequenziamento* produce **reads** di lunghezza limitata
- Il **costo** di un progetto è proporzionale al numero di reads prodotte

La soluzione: *tante* copie

GACTTTGTA
TTTGTAACATAC

La combinatorica del problema

Vincoli

- L'attività di *sequenziamento* produce **reads** di lunghezza limitata
- Il **costo** di un progetto è proporzionale al numero di reads prodotte

La soluzione: *tante* copie

GACTTTGTAACATACAAC
TTTGTAACATAC

La combinatorica del problema

Vincoli

- L'attività di *sequenziamento* produce **reads** di lunghezza limitata
- Il **costo** di un progetto è proporzionale al numero di reads prodotte

La soluzione: *tante* copie

GACTTTGTAACATACAAC
TTTGTAACATACAACCTTTAA

La combinatorica del problema

Vincoli

- L'attività di *sequenziamento* produce **reads** di lunghezza limitata
- Il **costo** di un progetto è proporzionale al numero di reads prodotte

La soluzione: *tante* copie

GACTTTGTAACATACAACCTTTAATCACG
TTTGTAACATACAACCTTTAA

La combinatorica del problema

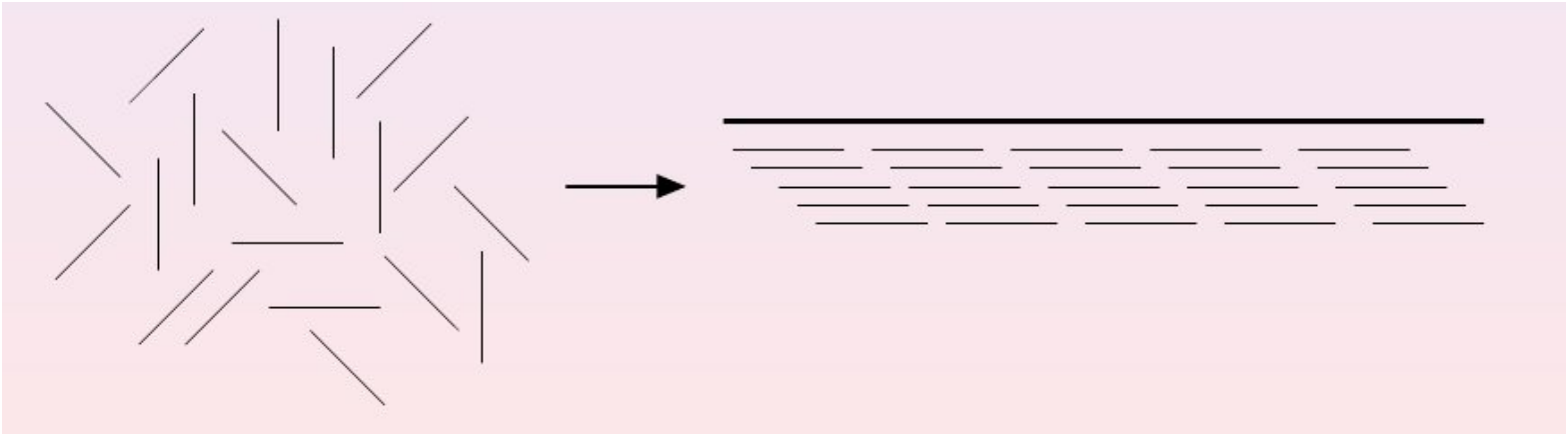
Vincoli

- L'attività di *sequenziamento* produce **reads** di lunghezza limitata
- Il **costo** di un progetto è proporzionale al numero di reads prodotte

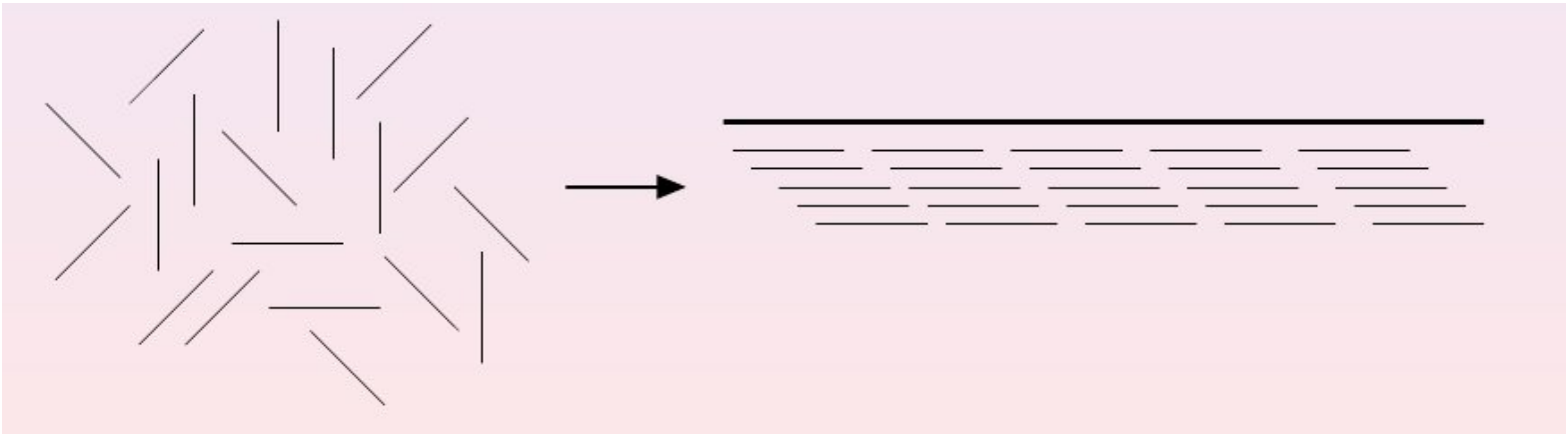
La soluzione: *tante* copie

GACTTTGTAACATACAACCTTTAATCACG . . .
TTTGTAACATACAACCTTTAATCACGCTCAATA . . .

Ripuliamo "matematicamente" il problema



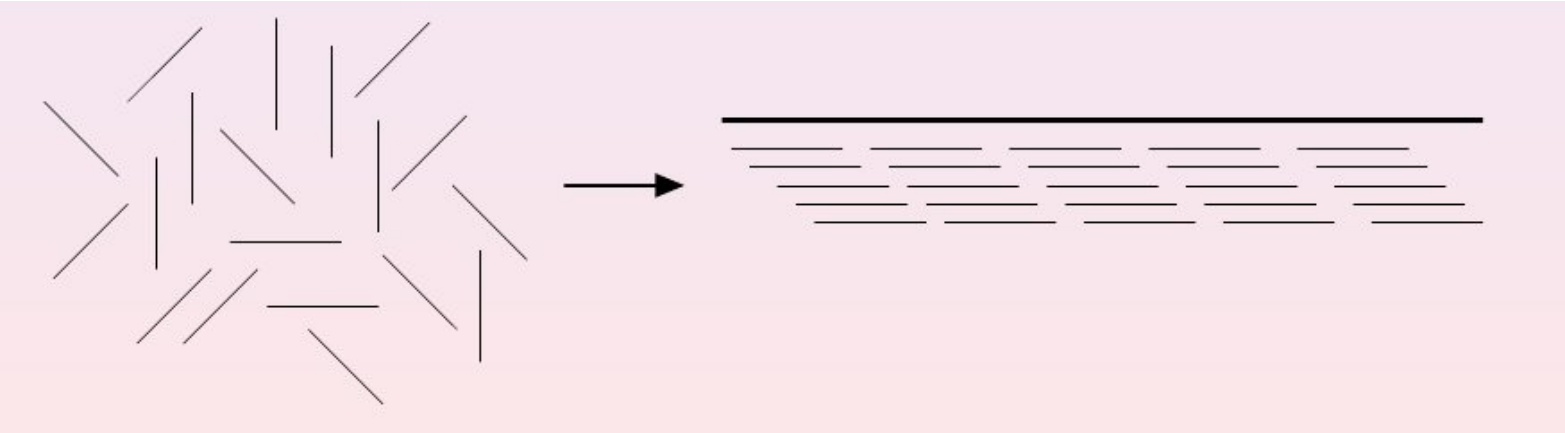
Ripuliamo "matematicamente" il problema



Definizione (Shortest Superstring Problem)

Dato un insieme di stringhe $\{s_1, s_2, ..., s_n\}$ trova la più corta super-stringa T tale che tutte le s_i siano sotto-stringhe di T .

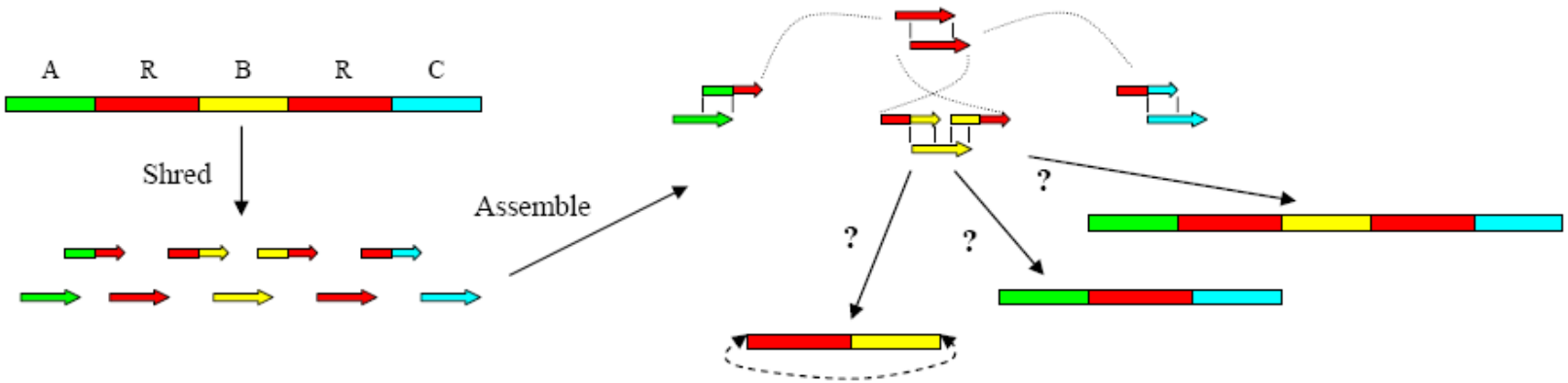
Ripuliamo "matematicamente" il problema



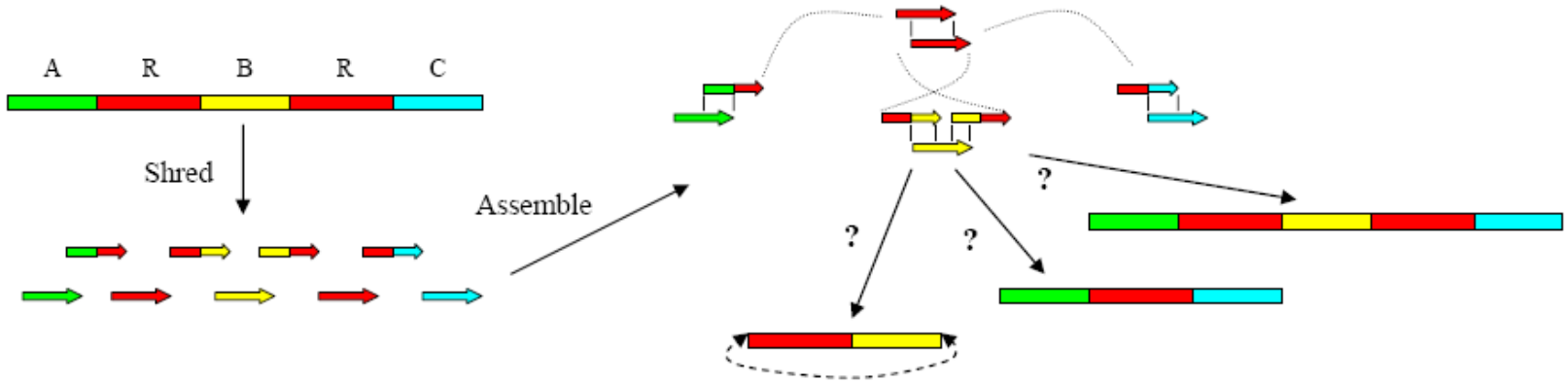
Due problemi:

1. $\mathcal{NP}$-completenezza [Gallant et al. 1980]
2. Il problema dell'assemblaggio è formulato in modo scorretto!

Elementi ripetuti

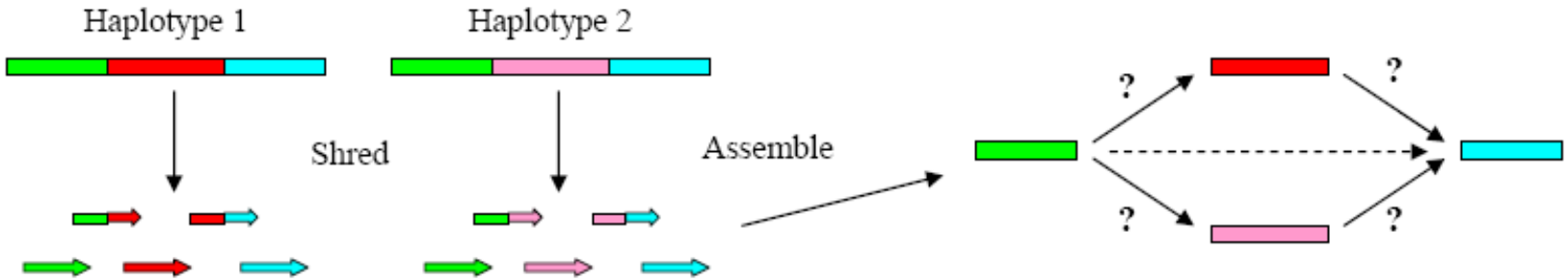


Elementi ripetuti



...

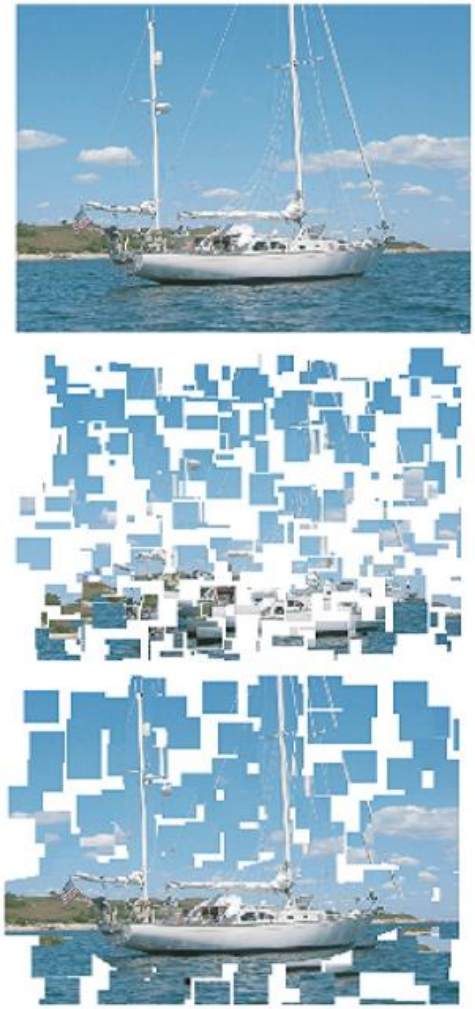
e, in aggiunta,



Morale

Assemblaggi di "qualità"

- Alte "coperture" (tanti *genomi equivalenti)*
- Tante reads
- reads lunghe
- Buon sw (*assembler*)
- Buon hw (parallelo?)
- ...



L'*allineamento di sequenze*

Input

...GTTGATTAGCTTATCCCAAAGCAAGGCACTGAAAATGCTAG
GTGATGTAGCTTAACCCAAGCAAGGCACTAAAAATGCCTAG ...

L'*allineamento di sequenze*

Input

...GTTGATTAGCTTATCCCAAAGCAAGGCACTGAAAATGCTAG
GTGATGTAGCTTAACCCAAGCAAGGCACTAAAAATGCCTAG ...

Output

...GTTGAT_TAGCTTATCCCAAAGCAAGGCACTGAAAATG_CTAG
GT_GATGTAGCTTAACCCAA_GCAAGGCACTAAAAATGCCTAG...

L'*allineamento di sequenze*

Input

...GTTGATTAGCTTATCCCAAAGCAAGGCACTGAAAATGCTAG
GTGATGTAGCTTAACCCAAGCAAGGCACTAAAAATGCCTAG ...

	G	T	T	G	A	T	T	A	G	C	T	T	A
G	0	1	2	3	4	5	6	7	8	9	10	11	12
T	1	0	1	2	3								
G	2	1	1	1	2								
A	3	2	2	2	1								
T	4	3	2										
G	5	4	3										
T	6	5	4										
A	7	6	5										

Output

...GTTGAT_TAGCTTATCCCAAAGCAAGGCACTGAAAATG_CTAG
GT_GATGTAGCTTAACCCAA_GCAAGGCACTAAAAATGCCTAG...

Un altro problema algoritmico (semplice?)

Input

T testo su un alfabeto Σ e P pattern su Σ.

Un altro problema algoritmico (semplice?)

Input

T testo su un alfabeto Σ e P pattern su Σ.

Output

Tutte le occorrenze di P in T

Un altro problema algoritmico (semplice?)

Input

T testo su un alfabeto Σ e P pattern su Σ.

... è naturale aspettarsi $|P| \cdot |T|$ operazioni di confronto ma si puó fare molto meglio!

Output

Tutte le occorrenze di P in T

Pitagora: "Tutto è numero"

Idea

Reads (corte) *sono* Numeri (grandi)

Fondamentali

Definizione

- $\Sigma_{dna} = \{a, c, g, t\}$ *alfabeto;*
- P *pattern* $|P| = m$*;*
- T *testo and* $|T| = n$*;*

T_s *è la stringa di* m *caratteri* $T[s, \ldots, s + m - 1]$.

Stringhe in Σ_{dna} *sono* numberi in base 4

$$\mathsf{P} \rightsquigarrow p \qquad \mathsf{T}_s \rightsquigarrow t_s$$

Fondamentali

Definizione

- $\Sigma_{dna} = \{a, c, g, t\}$ *alfabeto;*
- P *pattern* $|P| = m$*;*
- T *testo and* $|T| = n$*;*

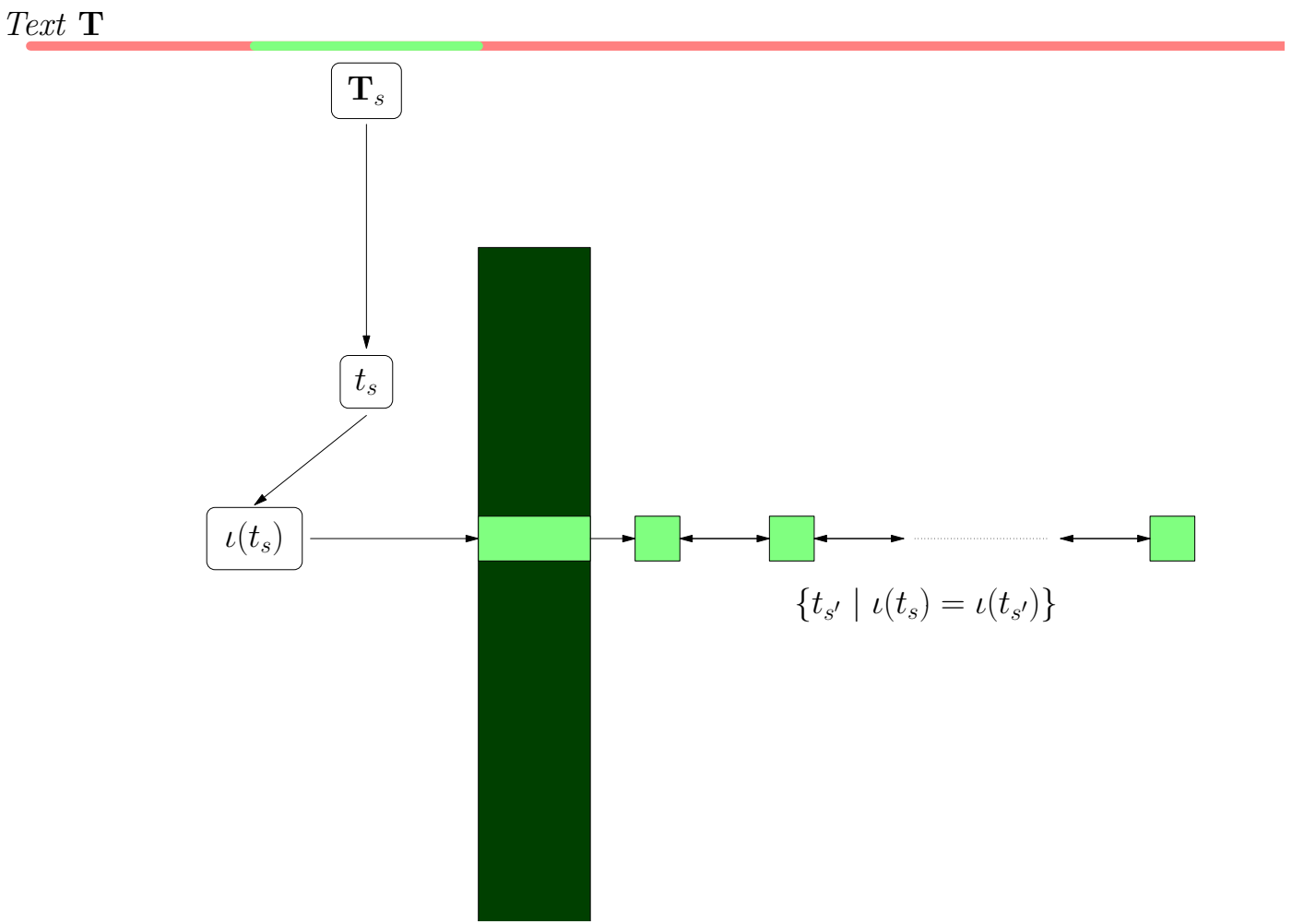
T_s *è la stringa di* m *caratteri* $T[s, \ldots, s + m - 1]$.

Stringhe in Σ_{dna} *sono* numberi in base 4

$$\mathsf{P} \rightsquigarrow p \qquad \mathsf{T}_s \rightsquigarrow t_s$$

$$a \equiv 0 \quad c \equiv 1 \quad g \equiv 2 \quad t \equiv 3$$

Indice

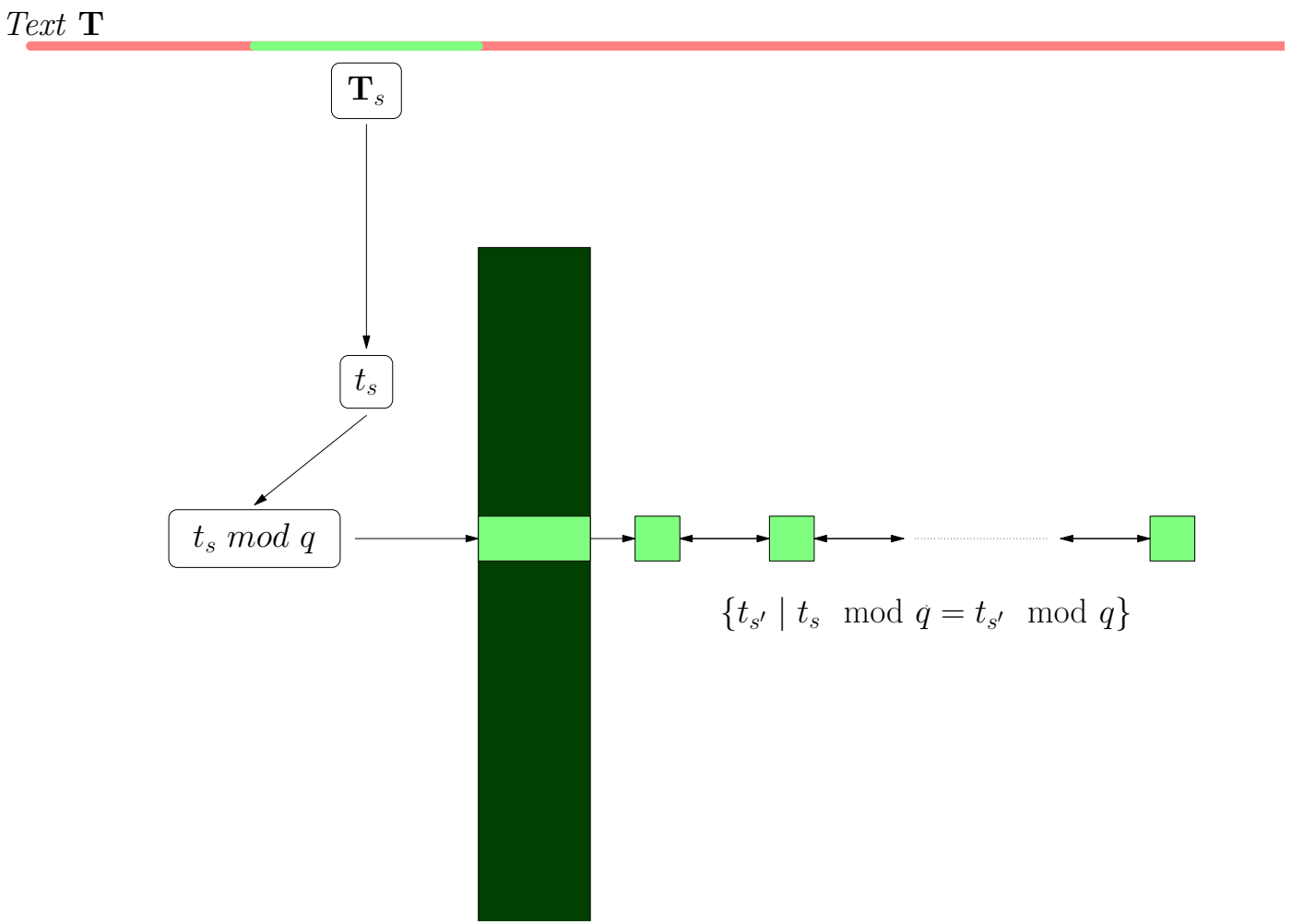


Indice

Problema

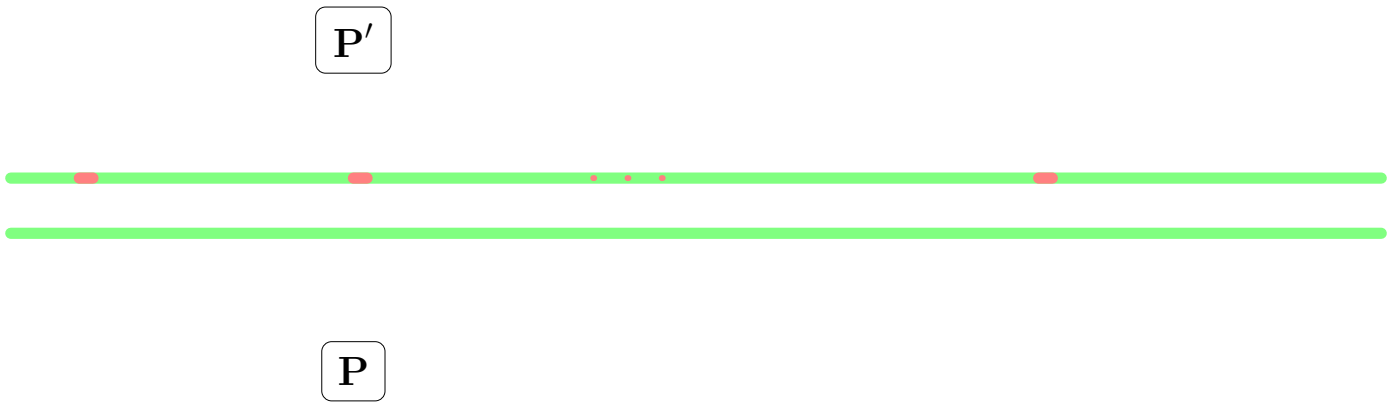
I numeri diventano è GRANDI

Indice



... e gli errori?

... e gli errori?



... e gli errori?

... e gli errori?

$\mathbf{P}' - \mathbf{P}$

$$|\{\mathbf{P}' \mid d_H(\mathbf{P}, \mathbf{P}') \leq d\}| = \sum_{d' \leq d} \binom{m}{d'} (\Sigma - 1)^{d'}$$

... e gli errori?

$\mathbf{P}' - \mathbf{P}$

$$|\{\mathbf{P}' \mid d_H(\mathbf{P}, \mathbf{P}') \leq d\}| = \sum_{d' \leq d} \binom{m}{d'} (\Sigma - 1)^{d'}$$

$$p' - p \in \mathcal{Z}(d, m)$$

... e gli errori?

$\mathbf{P}' - \mathbf{P}$

$$|\{\mathbf{P}' \mid d_H(\mathbf{P}, \mathbf{P}') \leq d\}| = \sum_{d' \leq d} \binom{m}{d'} (\Sigma - 1)^{d'}$$

$$p' - p \in \mathcal{Z}(d, m)$$

$$(p' - p \mod q) \in ???$$

rNA

randomized Numerical Aligner

- *opportuno* q^*
  ($\Rightarrow$ "piccolo" $\mathcal{Z}$)
- diamo *struttura* a $\mathcal{Z}$
  ($\Rightarrow$ implementiamo un test di appartenenza veloce)
- euristiche

rNA

randomized Numerical Aligner

- *opportuno* q^*
  ($\Rightarrow$ "piccolo" $\mathcal{Z}$)
- diamo *struttura* a $\mathcal{Z}$
  ($\Rightarrow$ implementiamo un test di appartenenza veloce)
- euristiche

Complessità

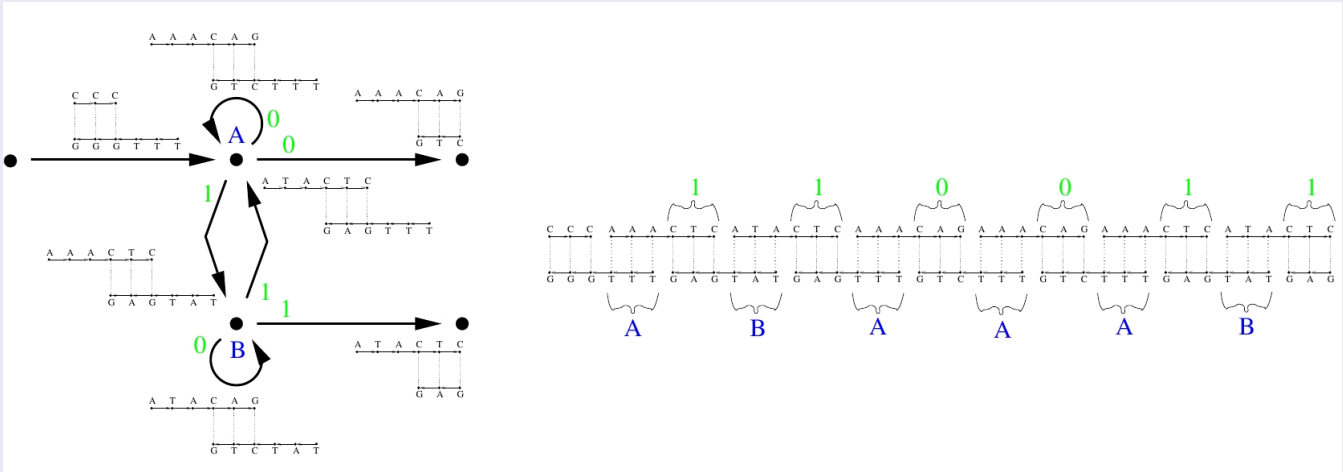
Proporzionale, in media, alla dimensione del pattern.

Una notte del 1993 L. Adleman stava leggendo "The molecular Biology of the Gene"...

... si sedette sul letto e disse a sua moglie:"Dio mio, queste cose possono calcolare "

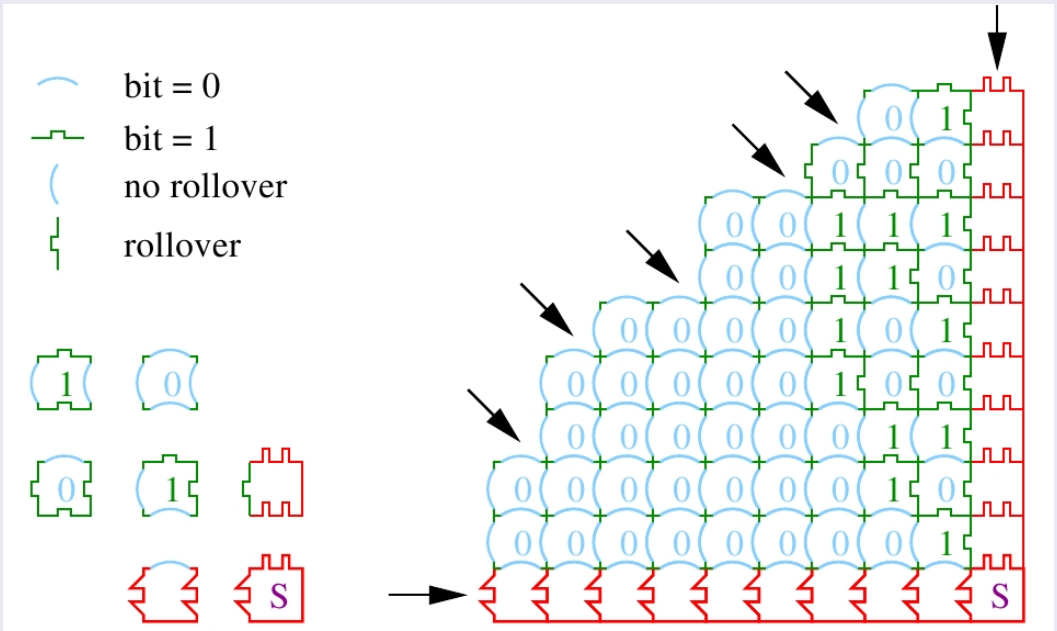
Una notte del 1993 L. Adleman stava leggendo "The molecular Biology of the Gene"...

... si sedette sul letto e disse a sua moglie:"Dio mio, queste cose possono calcolare "



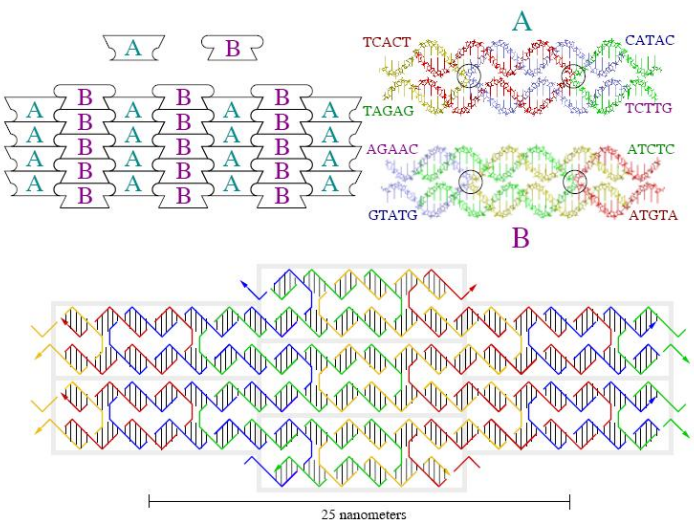
DNA self-assembly

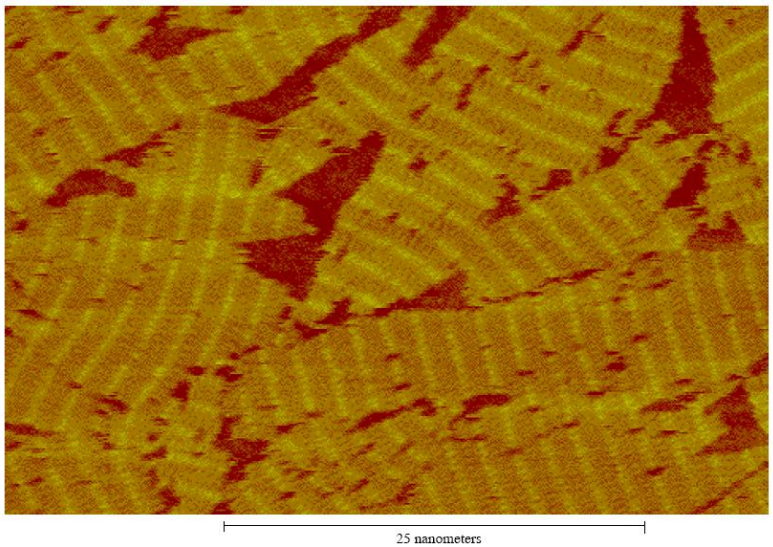
Si sa da tempo che è possibile computare "*in 2D*"



DNA self-assembly

Oggi si possono sintetizzare molecole di DNA che eseguono tali computazioni!





DNA self-assembly

Per saperne di più

DNA Computing by Self-Assembly

Erik Winfree
Departments of Computer Science and Computation & Neural Systems
California Institute of Technology
Pasadena, California

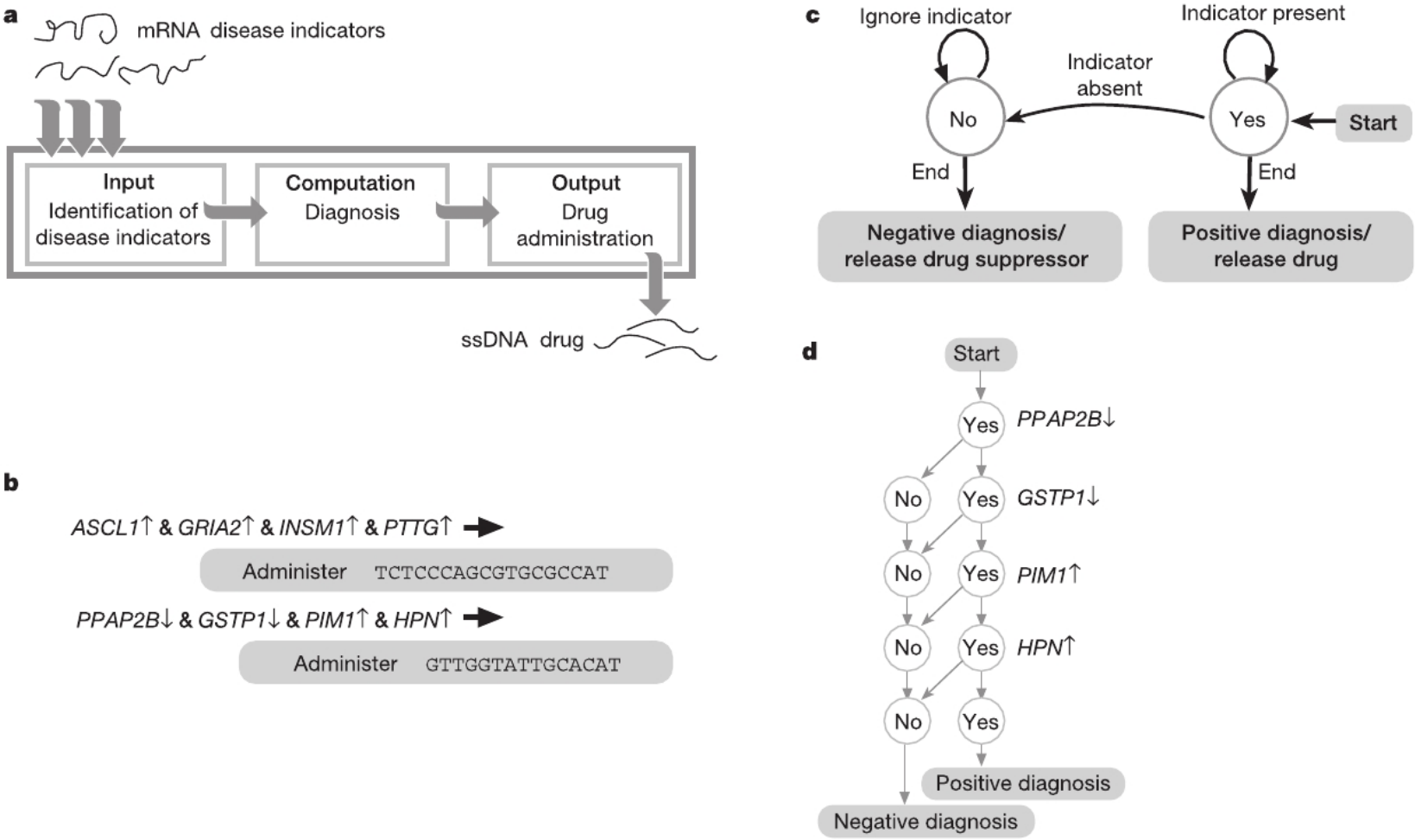
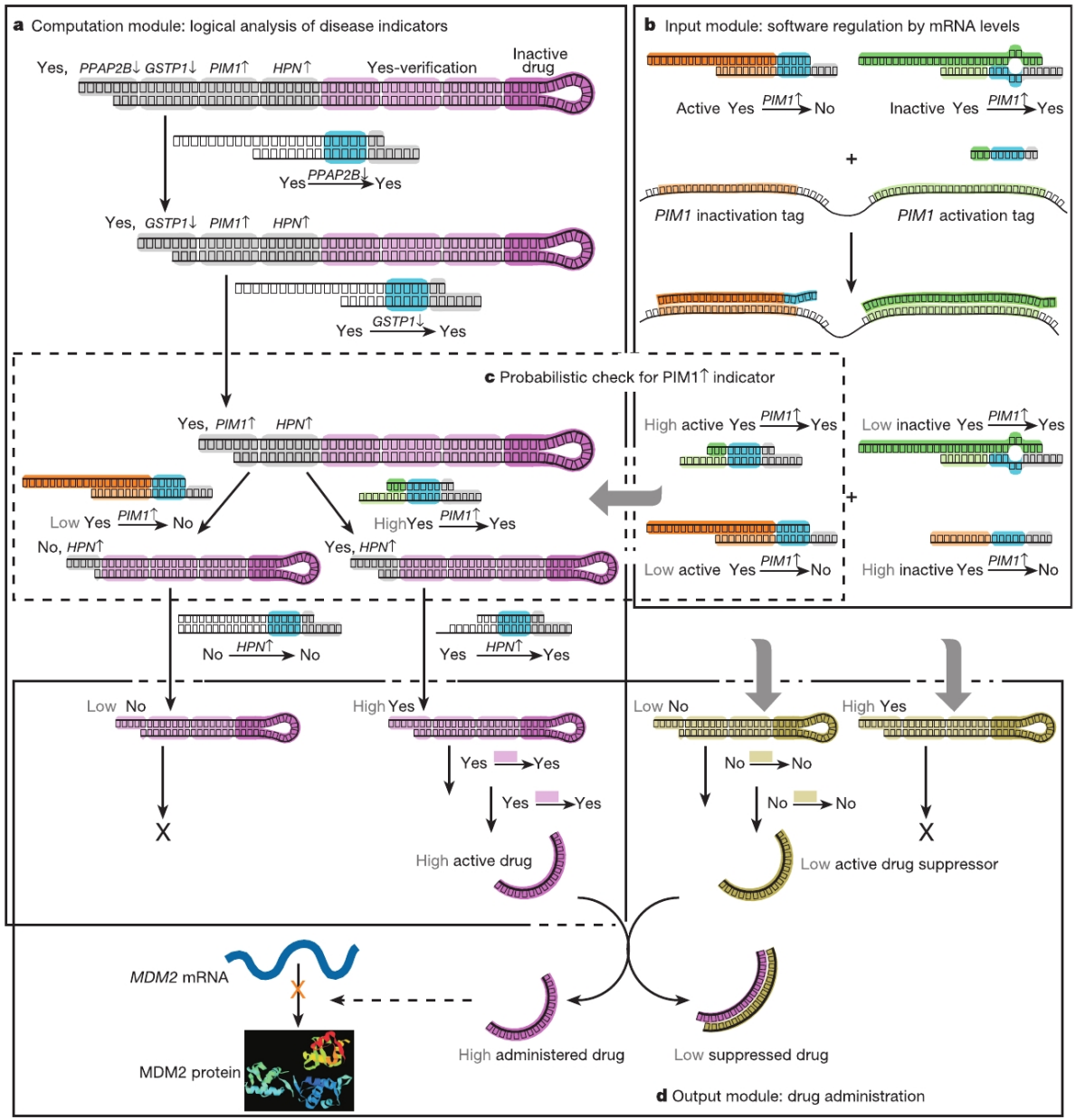


Figure 1 Logical design and logical operation of the molecular computer. **a**, Function and modular organization of the molecular computer. **b**, Example diagnostic rules for simplified models of SCLC[19] and prostate cancer[20], indicating overexpression (↑) or underexpression (↓) of a disease-related gene. The first rule states that if genes *ASCL1*, *GRIA2*, *INSM1* and *PTTG1* are overexpressed then administer the ssDNA molecule TCTCCCAGCGTGCGCCAT (oblimersen), purported to be an antisense therapy drug for SCLC[27]. The second rule states that if the genes *PPAP2B* and *GSTP1* are underexpressed and the genes *PIM1* and *HPN* are overexpressed then administer the ssDNA molecule GTTGGTATTGCACAT, purported to be a drug for prostate cancer[25]. **c**, Transition diagram of the diagnostic automaton. **d**, The computation that diagnoses prostate cancer.

NATURE | doi:10.1038/nature02551 | www.nature.com/nature


a Computation module: logical analysis of disease indicators
Yes, PPAP2B↓ GSTP1↓ PIM1↑ HPN↑
Yes-verification
Inactive drug
Yes PPAP2B↓ Yes
Yes, GSTP1↓ PIM1↑ HPN↑
Yes GSTP1↓ Yes
b Input module: software regulation by mRNA levels
Active Yes PIM1↑ No
Inactive Yes PIM1↑ Yes
PIM1 inactivation tag
PIM1 activation tag
c Probabilistic check for PIM1↑ indicator
Yes, PIM1↑ HPN↑
Low Yes PIM1↑ No
HighYes PIM1↑ Yes
No, HPN↑
Yes, HPN↑
No HPN↑ No
Yes HPN↑ Yes
High active Yes PIM1↑ Yes
Low inactive Yes PIM1↑ Yes
Low active Yes PIM1↑ No
High inactive Yes PIM1↑ No
Low No
High Yes
X
Yes Yes
Yes Yes
High active drug
No No
No No
Low active drug suppressor
MDM2 mRNA
MDM2 protein
High administered drug
Low suppressed drug
d Output module: drug administration

Domanda interessante

qual è l'hardware e quale il software?

Per saperne di più

letters to nature

An autonomous molecular computer for logical control of gene expression

Yaakov Benenson[1,2], Binyamin Gil[2], Uri Ben-Dor[1], Rivka Adar[2] & Ehud Shapiro[1,2]

[1]*Department of Computer Science and Applied Mathematics and* [2]*Department of Biological Chemistry, Weizmann Institute of Science, Rehovot 76100, Israel*

Early biomolecular computer research focused on laboratory-scale, human-operated computers for complex computational problems[1–7]. Recently, simple molecular-scale autonomous programmable computers were demonstrated[8–15] allowing both input and output information to be in molecular form. Such computers, using biological molecules as input data and biologically active molecules as outputs, could produce a system for 'logical' control of biological processes. Here we describe an autonomous biomolecular computer that, at least *in vitro*, logically analyses the levels of messenger RNA species, and in response produces a molecule capable of affecting levels of gene expression. The computer operates at a concentration of close to a trillion computers per microlitre and consists of three programmable modules: a computation module, that is, a stochastic molecular automaton[12–17]; an input module, by which specific mRNA levels or point mutations regulate software molecule concentrations, and hence automaton transition probabilities; and an output module, capable of controlled release of a short single-stranded DNA molecule. This approach might be applied *in vivo* to biochemical sensing, genetic engineering and even medical diagnosis and treatment. As a proof of principle we programmed the computer to identify and analyse mRNA of disease-related genes[18–22] associated with models of small-cell lung cancer and prostate cancer, and to produce a single-stranded DNA molecule modelled after an anticancer drug.

Taking our cue from the terminology of medical treatment, we consider that our molecular computer performs *in vitro* a computational version[23,24] of 'diagnosis'—the identification of a combination of mRNA molecules at specific levels, which in our example is a highly simplified model of cancer—and 'therapy'—production of a biologically active molecule, which in our case is a drug-like single-stranded (ss)DNA with known anticancer activity (Fig. 1a). The computer operation is governed by a 'diagnostic rule' that encodes medical knowledge in simplified form (Fig. 1b). The left-hand side of the rule consists of a list of molecular indicators for a specific disease, and its right-hand side indicates a molecule to be released, which could be a drug for that disease. For example, the diagnostic rule for prostate cancer states[20] that if the genes *PPAP2B* and *GSTP1* are underexpressed and the genes *PIM1* and hepsin (*HPN*) are overexpressed then administer the ssDNA molecule GTTGGTATTGGACATG, which inhibits[25] the synthesis of the protein MDM2 by binding to its mRNA. The computer design is flexible in that any sufficiently long RNA molecule can function as a molecular indicator and any short ssDNA molecule, up to at least 21 nucleotides, can be administered.

The computation module is a molecular automaton[12–15] (Supplementary Fig. S1) that processes such a rule as depicted in Fig. 1c. The automaton has two states: positive (Yes) and negative (No). The computation starts in the positive state and if it ends in that state we call the result 'positive diagnosis', otherwise it is called 'negative diagnosis'. To facilitate rule processing by the automaton, the left-hand side of the diagnostic rule is represented as a string of symbolic indicators, or symbols for short, one for each molecular

Conclusioni

Campo stimolante

- Tanti bei problemi computazionali
- Tecnologie sofisticate
- Forte impatto sulle scienze della Vita

Conclusioni

È probabilmente vero in linea di massima che nella storia del pensiero umano gli sviluppi piú fruttuosi avvengono frequentemente in quei punti di interferenza fra due diverse linee di pensiero.

W. Heisenberg

Definizione di Algoritmo

Informale (Markov, 1960)

Processo di calcolo che soddisfa tre requisiti principali:

1. Precisione prescrittiva. *Definitezza*
2. Applicabilità a dati iniziali variabili entro limiti prefissati. *Generalità*
3. Processo orientato ad ottenere un risultato, calcolabile in base ai dati iniziali. *Effettività*

Formale (Gödel, Church, Turing, ...)

Funzioni ricosrive,M lambda calcolo. macchine di Turing, ...