

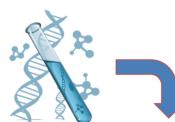
Seconda parte

(Alcuni) Problemi, Algoritmi e strutture dati per la Bioinformatica

1

Viaggio al centro del DNA

Prof.ssa P. Bonizzoni - R. Rizzi (Univ. Milano-Bicocca, AlgoLab)



+ algoritmi

60 billion bp per day
(2012)



+ algoritmi

Terapia
medica

sono stringhe...

2

Viaggio al centro del DNA

Prof.ssa P. Bonizzoni - R. Rizzi (Univ. Milano-Bicocca, AlgoLab)

genoma è una sequenza?

grafo - Pan-genome

sono grafi

+ algoritmi ... → Terapia medica

3

Breve reminder

- Genoma: collezione di cromosomi presenti nell'individuo
- Cromosomi: all'interno della cellula (nel nucleo)* [uomo è organismo diploide: 46 cromosomi, ogni cromosoma in coppia (23 padre, 23 madre), quindi 2 chr1, 2 chr2, ... Nel topo 40, nel cane 78]
 - Ognuno è rappresentato da filamenti lineari di DNA "attorcigliati". In particolare, due si uniscono nella parte centrale (centromero).
- DNA: contiene il nostro patrimonio genetico.

4

* Precisazione

Il genoma umano è composto da due genomi:

- il genoma nucleare (di cui parleremo)
- il genoma mitocontriale (mtDNA)

E' una molecola di DNA **circolare** di circa 16000 nucleotidi, presente in copie numerose nei mitocondri (si trovano nel citoplasma), gli organelli che generano energia (ATP)
 - Contiene informazioni per la sintesi di alcune proteine

La sua ereditarietà è materna

Il 97% della sequenza di nucleotidi che lo compongono è codificante
 Contiene in tutto 37 geni

5

* Precisazione

"Date le sue dimensioni ridotte, a volte viene un po' 'trascurato,' e data la sua complessità, a volte viene analizzato in maniera impropria," afferma il dottor Legati. E il professor Ghezzi aggiunge: *"E' però indubbio, e sempre più evidente, il ruolo di alterazioni genetiche nel DNA mitocondriale non solo nelle cosiddette malattie mitocondriali primarie (encefalo e cardiomio-patie dovute a mutazioni nel DNA mitocondriale che causano difetti energetici soprattutto a cervello, cuore, muscolo) ma anche in molte patologie complesse, come le malattie neurodegenerative, e in forme tumorali".*

Per le sue caratteristiche strutturali, quindi, l'mtDNA è particolarmente adatto nei casi di campioni scarsi o degradati.

NEANDERTHAL DNA AND MODERN HUMANS
 Svante Pääbo Receives the 2022 Nobel Prize in Physiology or Medicine



6

GENOMA

Il genoma è la lunga molecola di DNA che regola la funzione e lo sviluppo di un organismo vivente ed è:

- ✓ situata nel nucleo cellulare
- ✓ avvolta a doppia elica (Watson-Crick, 1953)
- ✓ frammentata in cromosomi

Una regione di DNA genomico che ha una certa funzione prende il nome di locus

La sequenza primaria di un *locus* è chiamata sequenza genomica

7

GENOMA E GENE

- Il **genoma governa** la vita di un organismo attraverso la produzione di proteine
- Un gene è distribuito sul genoma (quindi è frammentato) ed è una regione (*locus*) del genoma che esprime una proteina.
- Ogni proteina è costruita a partire da un gene (non vale il viceversa: «scoperta» derivata dal sequenziamento)

esistono anche geni che “non riescono” ad esprimere una proteina (non-coding): si interrompe la trascrizione...

8

GENOMA E GENE

Un gene è una regione (*locus*) del genoma che esprime una proteina

Entrambe le catene del genoma (forward e reverse) contengono geni

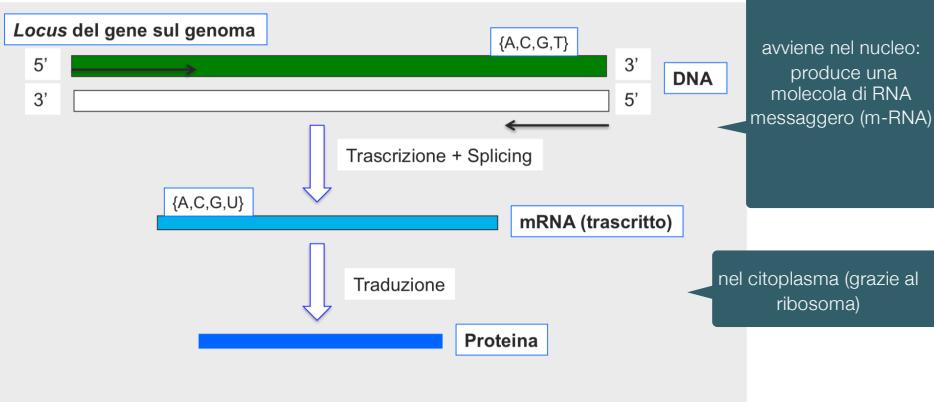
Il genoma umano contiene circa 20000 geni codificanti proteine



i geni verde scuro “trascrivono” da sinistra a destra,
gli altri da destra a sinistra

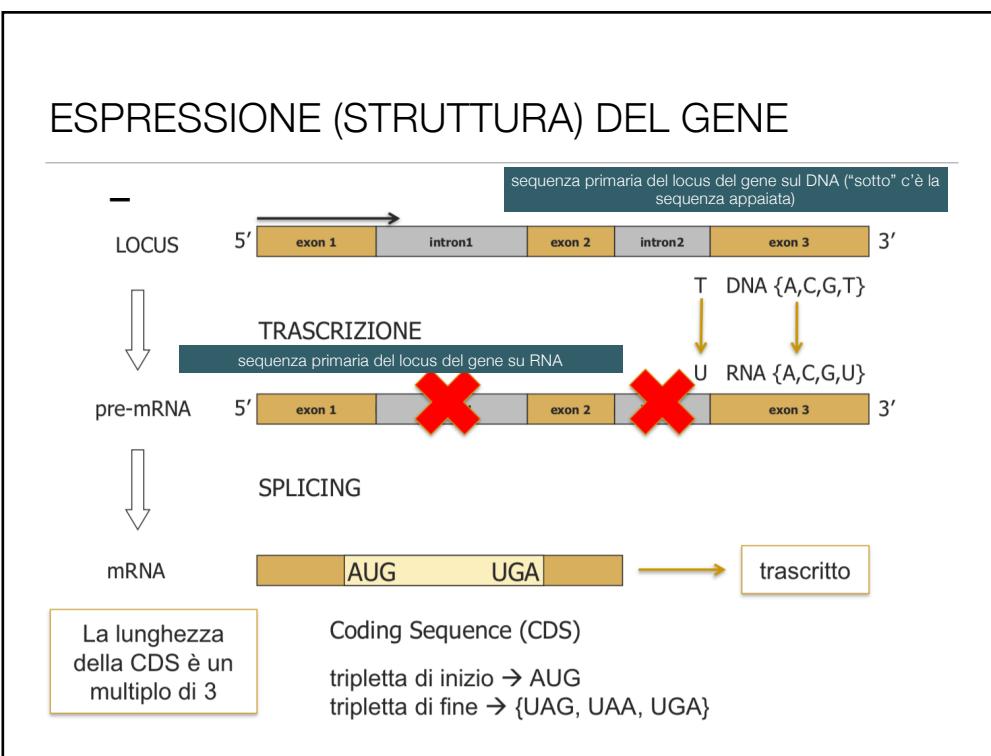
9

Dal GENE alla PROTEINA

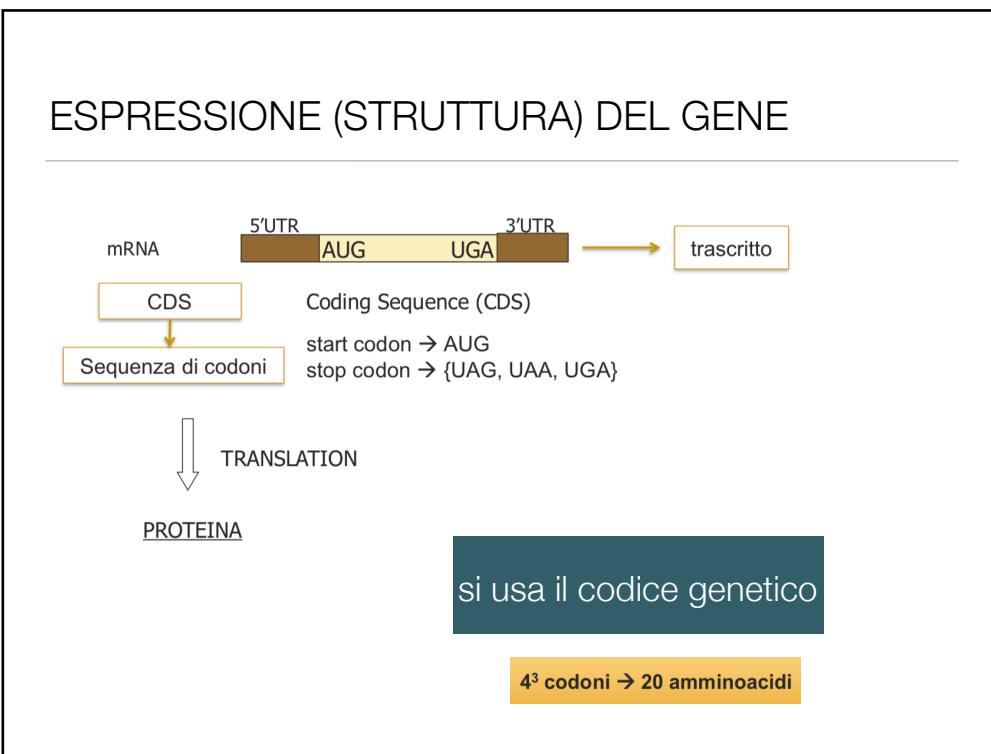


...ma se tutte le cellule hanno nel nucleo lo stesso genoma, perchè le cellule del fegato sono diverse da quelle del polmone?
Perchè cambia il profilo di espressione [non tutti i geni sono espressi in tutte le cellule, e per lo splicing alternativo vengono prodotte proteine diverse]

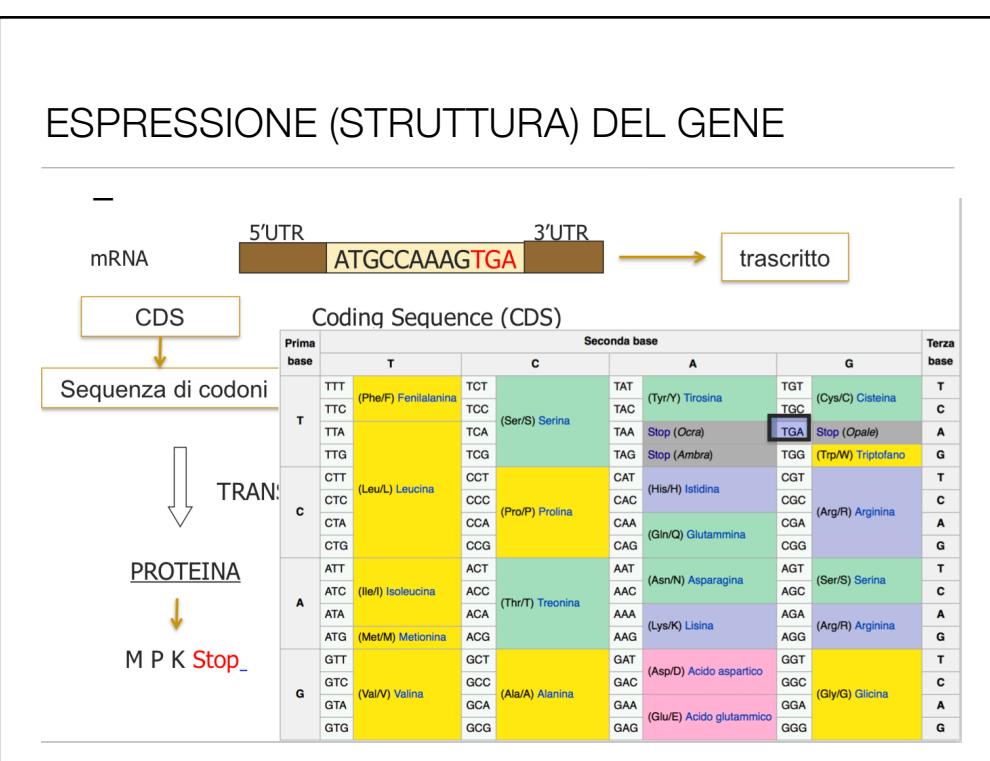
10



11



12



13

1. Confronto tra sequenze genomiche

- Per studiare aspetti evolutivi, somiglianze tra specie (anche per le sperimentazioni)
- Per capire cambiamenti nucleotidici o di riarrangiamento (ad es. ferro e ferritina)
- Vaccini, antibiotici

Cambiamenti nucleotidici:

Sappiamo che tutte le nostre cellule hanno lo stesso DNA
ma

le cellule tumorali non hanno il DNA di partenza, perdono la capacità di

produrre proteine

(con la radioterapia si colpiscono solo quelle cellule per «farle ammalare»
Non tendono a ripararsi e quindi muoiono)

14

1. Confronto tra sequenze genomiche

- Tra due sequenze: allineamento globale, locale, semiglobale
- Tra più sequenze: allineamento
- Tool di allineamento
- (Spliced alignment: riallineare un trascritto - mRNA messaggero prodotto dal gene - al genoma da cui proviene, perchè è discontinuo)

tecniche alignment-free

15

2. (Sequenziamento e) Assemblaggio

Sequenziamento del DNA

- **Il dato di sequenziamento**
 - **Tecnologie di sequenziamento:**
 - SANGER sequencing
 - Next-Generation Sequencing (NGS)
 - **Qualità del dato di sequenziamento**
 - Formato Standard FASTQ
- **Fragment Assembly e approcci graph-based:**
 - Overlap Graph (OG) e de Bruijn Graph (dBG)

dato
(DNA) ottenuto il
lab e usato in silico
per assemblare un
frammento del genoma

16

Grafi di assemblaggio

- **Overlap Graph** (adatti per long read): collego due frammenti usando la “sovraposizione”
- **De Bruijn Graph** (adatti per short reads): collego due k-mer con la relazione suffisso-prefisso.

Ma se esistono già assemblatori, perchè studiare questi grafi? Perchè posso usarli per altro:

BWT/FM-index/Suffix-tree/Bloom Filters/...

17

Idee per l'esame (se scegliete questa parte)

- Analisi di un problema (approfondendo leggermente quanto spiegato + articoli scientifici specifici, oppure su un problema nuovo)
- Analisi di un tool (articolo + sw), eventualmente con spunti di ricerca

→ Presentazione (slides e una tesina di descrizione)

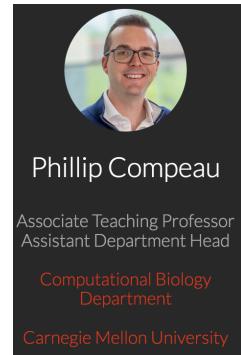
18

Materiale presentato...

[pdf+capitoli libro+articoli bio: sul sito]



<https://www.bioinformaticsalgorithms.org/>



[https://compeau.cbd.cmu.edu/teaching/
great-ideas-in-computational-biology/](https://compeau.cbd.cmu.edu/teaching/great-ideas-in-computational-biology/)

19

Materiale presentato...

<https://www.youtube.com/watch?v=yfXeKPt0nw4>



20

Outline delle nostre lezioni

Confronto tra sequenze

- ✓ Allineamento
- ✓ *Metodi alignment-free*
- ✓ Tool
- ✓ Alcuni spunti di lavoro...

Fragment assembly

- ✓ Strutture di indicizzazione (grafi)
- ✓ Tool
- ✓ Alcuni spunti di lavoro... (con BWT)

21

How Do We Compare Biological Sequences?

*Dynamic Programming and
Divide-and Conquer Algorithms*

Phillip Compeau and Pavel Pevzner.
Bioinformatics Algorithms: An Active Learning Approach
Cap. 6

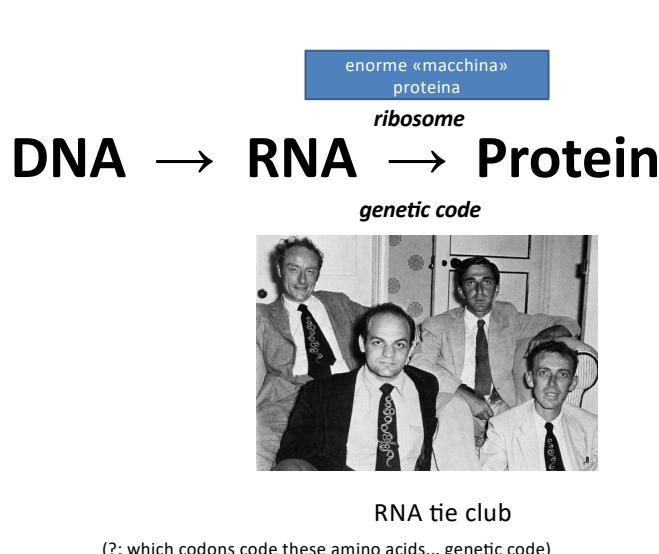
22

How Do We Compare Biological Sequences

- **From Sequence Comparison to Biological Insights**
- The Alignment Game and the Longest Common Subsequence
- The Manhattan Tourist Problem
- Dynamic Programming and Backtracking Pointers
- From Manhattan to the Alignment Graph
- From Global to Local Alignment
- Penalizing Insertions and Deletions in Sequence Alignment
- Space-Efficient Sequence Alignment
- Multiple Sequence Alignment

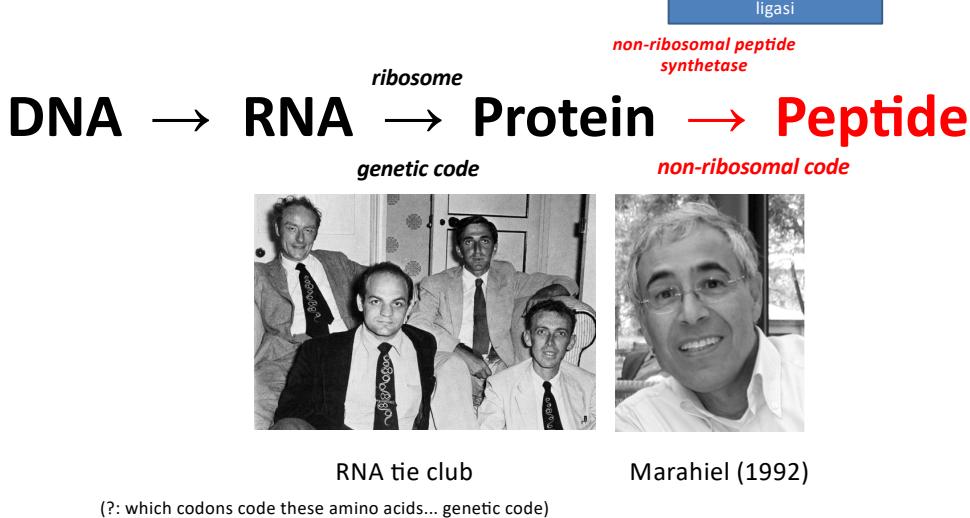
23

From Genetic Code to...



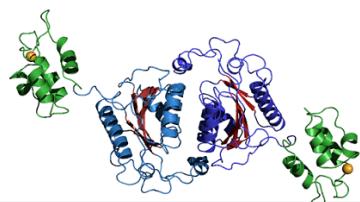
24

From Genetic Code to Non-Ribosomal Code



25

NRP Synthetase: A Giant Molecular Assembly Line



1939: NRPs in *Bacillus Brevis* → tyrocidine (NRPs encodes 10 amino acids)

STOP: What particular segment in NRPs is responsible of the coding? (i.e., what particular amino acid in NRP S. is responsible for synthetizing a specific codon in the tyrocidine?)

He knew that: for every synthesized amino acid there is a segment in the NRPs (roughly 600 amino acids long, called Adenylation domain)

26

NRP Synthetase: A Giant Molecular Assembly Line

Since tyrocidine is 10 amino acids long, there are 10 A-domain in NRPs

Question:
What is the NR segment code that decides what amino acid to synthesise?

Adenylation domains

A1 A2 A3 A4 A5 A6 A7 A8 A9 A10

NRP synthetase adds one amino acid at a time (via Adenylation domain)

27

Sintesi peptidica non-ribosomica (NRP)

- Dagli antibiotici impariamo che non c'è un solo modo per produrre (assemblare) amminoacidi in una proteina, ossia attraverso un ribosoma e usando l'RNA come un template.
- Usando la NRP: per ogni amminoacido che viene sintetizzato, c'è sempre un segmento nel composto (di circa 600 amminoacidi), chiamato Adenylation domain, che sintetizza un amminoacido alla volta. Per esempio, la tirocidina (primo antibiotico) è stata isolata da un batterio: NRP nel batterio codifica la tirocidina (10 amminoacidi). Ma quale particolare amminoacido o segmento è responsabile della sintesi del Valium, ad esempio?

28

Una cosa è certa...

- Ogni amminoacido viene sintetizzato dall'Adenylation domain, e siccome la tirocidina ne ha 10, devono esserci 10 Adenylation domain nel composto NRP.
- Quindi, quale è il codice non-ribosomico che decide quale amminoacido deve essere sintetizzato?

29

These Three A-domains Do Not Look Similar

even if they have the same function: add an amino acid to the peptide chain

```

YAFDLGYTCMFPVLLGGGELHIVQKETYTAPDEIAHYIKEHGITYIKLTPSLFHTIVNTASFAFDANFESLRILVLGGEKIIIPIDVIAFRKMYGHTEFINHYGPTTEATIGA
AFDVSAGDFARALLTGQLIVCPNEVKMDPASLYAIKKYDITIFEAUTPALVPLMEYIYEQKLDISQLQILIVGSDSCSMEDFKTLVSRGSTIRIVNSYGVTEACIDS
IAFDASSWEIYAPLLNGGTVVVCIDYYTTIDIKALEAVFKQHHIRGAMLPPALLKQCLVSAPTMISSEILFAAGDRLSSQDAILARRAVGSGVYNAYGPTENTVLS

```

30

These Three A-domains Do Not Look Similar

```

YAFDLGYTCMFPVLLGGGELHIVQKETYTAPDEIAHYIKEHGITYIKLTPSLFHIVNTASFAFDANFESLRLIVLGGEKIIPIDVIAFRKMYGHTEFINHYGPTREATIGA
AFDVSAGDFARALLTGGQLIVCPNEVKMDPASLYAIIKKYDITIFEATPALVIPLMEYIYEQKLDISQLQILIVGSDSCSMEDFKTLVSRGSTIRIVNSYGVTEACIDS
IAFDASSWEIYAPLNNGGTVVICIDYYTTIDIKALEAVFKQHHIRGAMLPPALLKQCLVSAPTMISSLEILFAAGDRLSSQDAILARRAVGSGVYNAYGPTENTVLS

```

just 3 conservative columns

31

Do they look similar now?

```

YAFDLGYTCMFPVLLGGGELHIVQKETYTAPDEIAHYIKEHGITYIKLTPSLFHIVNTASFAFDANFESLRLIVLGGEKIIPIDVIAFRKMYGHTEFINHYGPTREATIGA
-AFDVSAGDFARALLTGGQLIVCPNEVKMDPASLYAIIKKYDITIFEATPALVIPLMEYIYEQKLDISQLQILIVGSDSCSMEDFKTLVSRGSTIRIVNSYGVTEACIDS
IAFDASSWEIYAPLNNGGTVVICIDYYTTIDIKALEAVFKQHHIRGAMLPPALLKQCLVSAPTMISSLEILFAAGDRLSSQDAILARRAVGSGVYNAYGPTENTVLS

```

11 conservative columns

32

And now?

```

YAFDLGYTCMFPVLLGGGELHIVQKETYTAPDEIAHYIKEHGIITYIKLTPSLFHTIVNTASFAFDANFESLRLIVLGGEKIIPIDVIAFRKMYGHTE-FINHYGPTEATIGA
-AFDSAGDFARALLTGGQLIVCPNEVKMDPASLYAIIKKYDITIFEATPALVIPLMEYI-YEQKLDISQLQILIVGSDCSMEDFKTLVSRFGSTIRIVNSYGVTEACIDS
IAFDASSWEIYAPLNNGGTVVICIDYYTTIDIKALEAVFKQHHIRGAMLPPALLKQCLVSA---PTMISSLEILFAAGDRLSQDAILARRAVGSGV-Y-NAYGPTENTVLS

```

19 conservative columns encode the Conserved Core of A-domains!
They are similar.

33

Red Positions Encode Conservative Core of A-domains

```

YAFDLGYTCMFPVLLGGGELHIVQKETYTAPDEIAHYIKEHGIITYIKLTPSLFHTIVNTASFAFDANFESLRLIVLGGEKIIPIDVIAFRKMYGHTE-FINHYGPTEATIGA
-AFDSAGDFARALLTGGQLIVCPNEVKMDPASLYAIIKKYDITIFEATPALVIPLMEYI-YEQKLDISQLQILIVGSDCSMEDFKTLVSRFGSTIRIVNSYGVTEACIDS
IAFDASSWEIYAPLNNGGTVVICIDYYTTIDIKALEAVFKQHHIRGAMLPPALLKQCLVSA---PTMISSLEILFAAGDRLSQDAILARRAVGSGV-Y-NAYGPTENTVLS

```

Which positions are responsible for encoding **different** amino acids Asp, Orn, Val?

Asp: acido aspartico
Orn: ornitina
Val: valina

34

Blue Positions in A-domains Define Non-Ribosomal Code

8 amino acids located in different parts of the A-domains, code a single amino acid...

```

YAFDLGYTCMPPVLLGGELHIVQKETYTAPEIAHYIKEHGITYKLTPSLFHTIVNTASFAFDANFESIRLLIVLGGEKIIPIIDVIAFRKMYGHTE-FINHYGPTEATIGA
-AFDVSAGDFARALLTGGQLIVCPNEVKMDPASLYAIIKKYDITIFEATPALVIPLMEYI-YEQKLDISQIQILIVGSDSCSMEDFKTLVSRGSTIRIVNSGVTEACIDS
IAFDASSWBIYAPLLNGGTVVCIDYTTIDIKALEAVFKQHHIRGAMLPPALLKQCLVSA---PTMISSLEILFAGDRLSSQDAILARRAVGSGV-Y-NAYGPTENTVL

```

LTKVGHIG	Asp
VGEIGSID	Orn
AWMFAAVL	Val

35

Comparing Genes is a Fundamental Problem in Biology

Comparing Genes Problem:

- **Input:** Two genes.
- **Output:** How “similar” these genes are.

Goal: Convert this important biological question into a well-defined computational problem.

36

Try 1: Hamming Distance

Hamming Distance Problem:

- **Input:** Two strings.
- **Output:** The number of “mismatched” symbols in the two strings.

ATGCATGC
TGCATGCA Hamming distance = 8

STOP: What are the issues with this approach?

37

Try 1: Hamming Distance

Hamming Distance Problem:

- **Input:** Two strings.
- **Output:** The number of “mismatched” symbols in the two strings.



A **TGCATGC**
TGCATGCA Hamming distance = 8

Note: these strings have a long shared substring, it just doesn't line up perfectly.

38

Try 2: Longest Substring

Longest Shared Substring Problem:

- **Input:** Two strings.
- **Output:** The longest substring shared by both strings.

STOP: What are the weaknesses of using the length of a longest shared substring to represent the similarity between two strings?

39

Try 2: Longest Substring

Longest Shared Substring Problem:

- **Input:** Two strings.
- **Output:** The longest substring shared by both strings.

Consider the strings AAACAAACAAACAAACAAACAAA
and AAAGAAAGAAAGAAAGAAAGAAAGAAA
These strings are very similar, but they don't have a long shared substring in common.

40

Toward a Better Approach

STOP: What similarities do you see in these strings?

ATGCTTA
TGCATTAA

Key Point: we can find similarities if we “slide” the strings, letting symbols shift (but stay in same order).

A~~TGC~~-~~TTA~~-
-~~TGCATTAA~~A

41

Toward a More Accurate Problem

Symbol Matching Problem:

- **Input:** Two strings.
- **Output:** The greatest number of matched symbols in any “alignment” of the two strings.

A~~TGC~~-~~TTA~~-
-~~TGCATTAA~~A

42

Toward a More Accurate Problem

Symbol Matching Problem:

- **Input:** Two strings.
- **Output:** The greatest number of matched symbols in any “alignment” of the two strings.

Exercise: How many matches can you find if the strings are ATGTTATA and ATCGTCC? What algorithm did you use?

43

Matching Symbols as a Game

Growing alignment	Remaining symbols	Score
A T G T T A T A		
A T C G T C C		
A	T G T T A T A	+1
A	T C G T C C	
A T	G T T A T A	+1
A T	C G T C C	
A T -	G T T A T A	
A T C	G T C C	
A T - G	T T A T A	+1
A T C G	T C C	
A T - G T	T A T A	+1
A T C G T	C C	
A T - G T T	A T A	
A T C G T -	C C	
A T - G T T A	T A	
A T C G T - C	C	
A T - G T T A T	A	
A T C G T - C -	C	
A T - G T T A T A		
A T C G T - C - C		

44

From a Game to a Definition

Given two strings v and w , an **alignment** of v and w is a two-row matrix such that:

- the first row contains symbols of v
- the second row contains symbols of w
- each row may also contain **gap symbols** ("")
- no column has two gap symbols

$\text{A} \text{ T} - \text{G} \text{ T}$ $\text{A} \text{ T} \text{ C} \text{ G} \text{ T}$	T A T A $- \text{C} - \text{C}$	Matches
---	---	----------------

45

From a Game to a Definition

Given two strings v and w , an **alignment** of v and w is a two-row matrix such that:

- the first row contains symbols of v
- the second row contains symbols of w
- each row may also contain **gap symbols** ("")
- no column has two gap symbols

$\text{A} \text{ T} - \text{G} \text{ T} \text{ T} \text{A} \text{ T} \text{A}$ $\text{A} \text{ T} \text{ C} \text{ G} \text{ T} - \text{C} - \text{C}$	Mismatches
---	-------------------

46

From a Game to a Definition

Given two strings v and w , an **alignment** of v and w is a two-row matrix such that:

- the first row contains symbols of v
- the second row contains symbols of w
- each row may also contain **gap symbols** ("")
- no column has two gap symbols

A T - G T T A T A	Insertions
A T C G T - C - C	

47

From a Game to a Definition

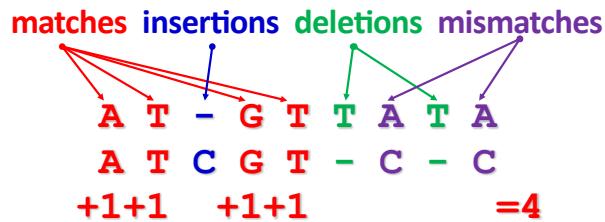
Given two strings v and w , an **alignment** of v and w is a two-row matrix such that:

- the first row contains symbols of v
- the second row contains symbols of w
- each row may also contain **gap symbols** ("")
- no column has two gap symbols

A T - G T T A T A	Deletions
A T C G T - C - C	

48

Recap



Alignment of two sequences is a two-row matrix:

1st row: symbols of the 1st sequence (in order) interspersed by “-”
 2nd row: symbols of the 2nd sequence (in order) interspersed by “-”