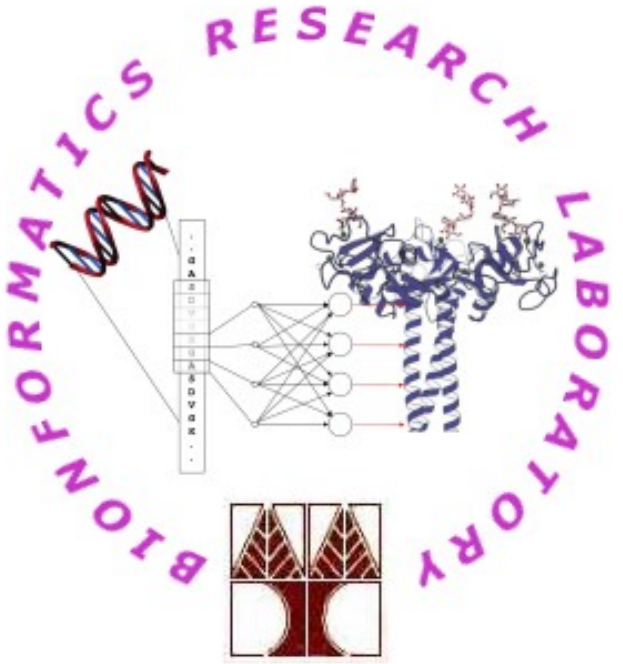
# Αλληλουχίες βιολογικών μακρομορίων
## Δομή, λειτουργία, εξέλιξη




Vasilis Promponas
Bioinformatics Research Laboratory
Department of Biological Sciences
University of Cyprus

# ΣΥΝΟΨΗ

- Εισαγωγή
- Ακολουθίες: Οργάνωση  Δεδομένων
- Υπολογιστική Ανάλυση Ακολουθιών
- Στοιχεία Μοριακής Εξέλιξης
  - Ομοιότητα vs Ομολογία
- Σύγκριση ακολουθιών κατά ζεύγη
- Συζήτηση ...

# Ακολουθίες: Οργάνωση Δεδομένων

- GeneBank/EMBL/DDBJ
- UNIPROT
  - SwissProt-PIR-TREMBL

**PRIMARY**

- PROSITE
- PFAM

**SECONDARY**

# SwissProt (http://www.expasy.ch/sprot)

```
ID   INS_PIG         STANDARD;      PRT;   108 AA.
AC   P01315; Q9TSJ5;
…
DE   INSULIN PRECURSOR.
GN   INS.
OS   Sus scrofa (Pig).
…
CC   -!- FUNCTION: INSULIN DECREASES BLOOD GLUCOSE CONCENTRATION. IT
CC       INCREASES CELL PERMEABILITY TO MONOSACCHARIDES, AMINO ACIDS AND
…
DR   EMBL; AF064555; AAC77920.1; ALT_INIT. [EMBL / GenBank / DDBJ]
…
KW   Insulin family; Hormone; Glucose metabolism; Signal; 3D-structure.
FT   SIGNAL        1     24
FT   CHAIN        25     54       INSULIN B CHAIN.
…
SQ   SEQUENCE   108 AA;  11671 MW;  CB4491B429858EBE CRC64;
     MALWTRLLPL LALLALWAPA PAQAFVNQHL CGSHLVEALY LVCGERGFFY TPKARREAEN
     PQAGAVELGG GLGGLQALAL EGPPQKRGIV EQCCTSICSL YQLENYCN
//
```

# PIR (http://pir.georgetown.edu)

```
>P1;IPPG
insulin precursor - pig
C;Species: Sus scrofa domestica (domestic pig)
…
C;Accession: A01583; A94572; S16492; A60835; B60835
…
C;Keywords: hormone; pancreas
F;1-30/Domain: insulin chain B #status experimental
F;1-30,64-84/Product: insulin #status experimental
F;33-63/Domain: connecting peptide #status experimental
F;64-84/Domain: insulin chain A #status experimental
F;7-70,19-83,69-74/Disulfide bonds: #status experimental
>P1;IPPG

FVNQHLCGSH LVEALYLVCG ERGFFYTPKA RREAENPQAG AVELGGGLGG LQALALEGPP
QKRGIVEQCC TSICSLYQLE NYCN*
```

# Prosite (http://www.expasy.ch/prosite)

```
ID   INSULIN; PATTERN.
AC   PS00262;
…
DE   Insulin family signature.
PA   C-C-{P}-x(2)-C-[STDNEKPI]-x(3)-[LIVMFS]-x(3)-C.
…
DO   PDOC00235;
//
```

## … and Documentation …

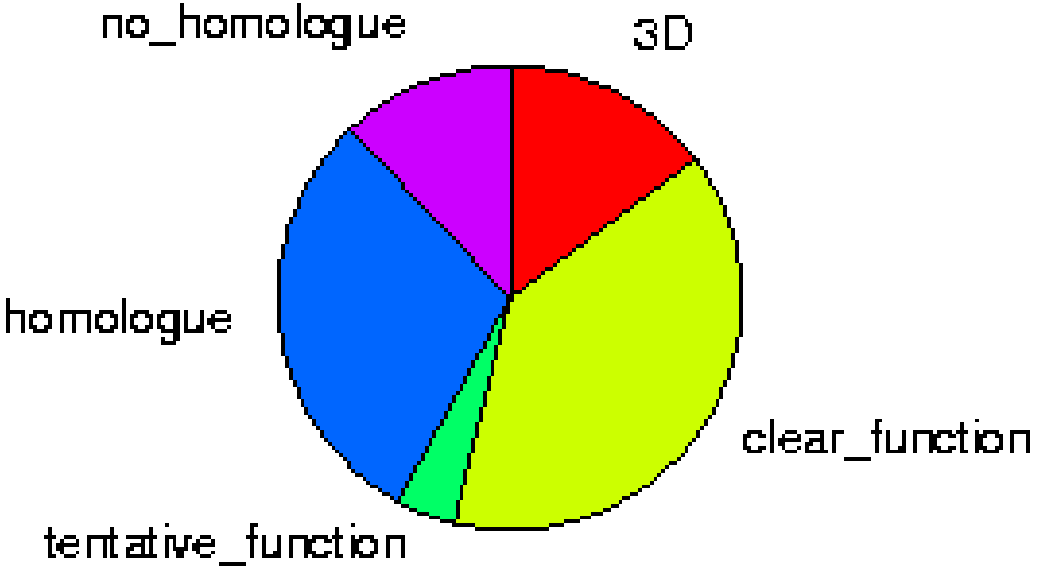
```
{PDOC00235}
{PS00262; INSULIN}
{BEGIN}
****************************
* Insulin family signature *
****************************
The insulin family of proteins [1] groups a number of active peptides which are
evolutionary related. This family currently consists of:
…
{END}
```

# Υπολογιστική Ανάλυση Ακολουθιών

- Μέθοδοι Βασισμένοι στην Ανίχνευση Ομοιότητας
- Εμπειρικές Μέθοδοι
- Τεχνικές Μηχανικής Μάθησης
- Αυτοματοποιημένα ή «με το χέρι??»

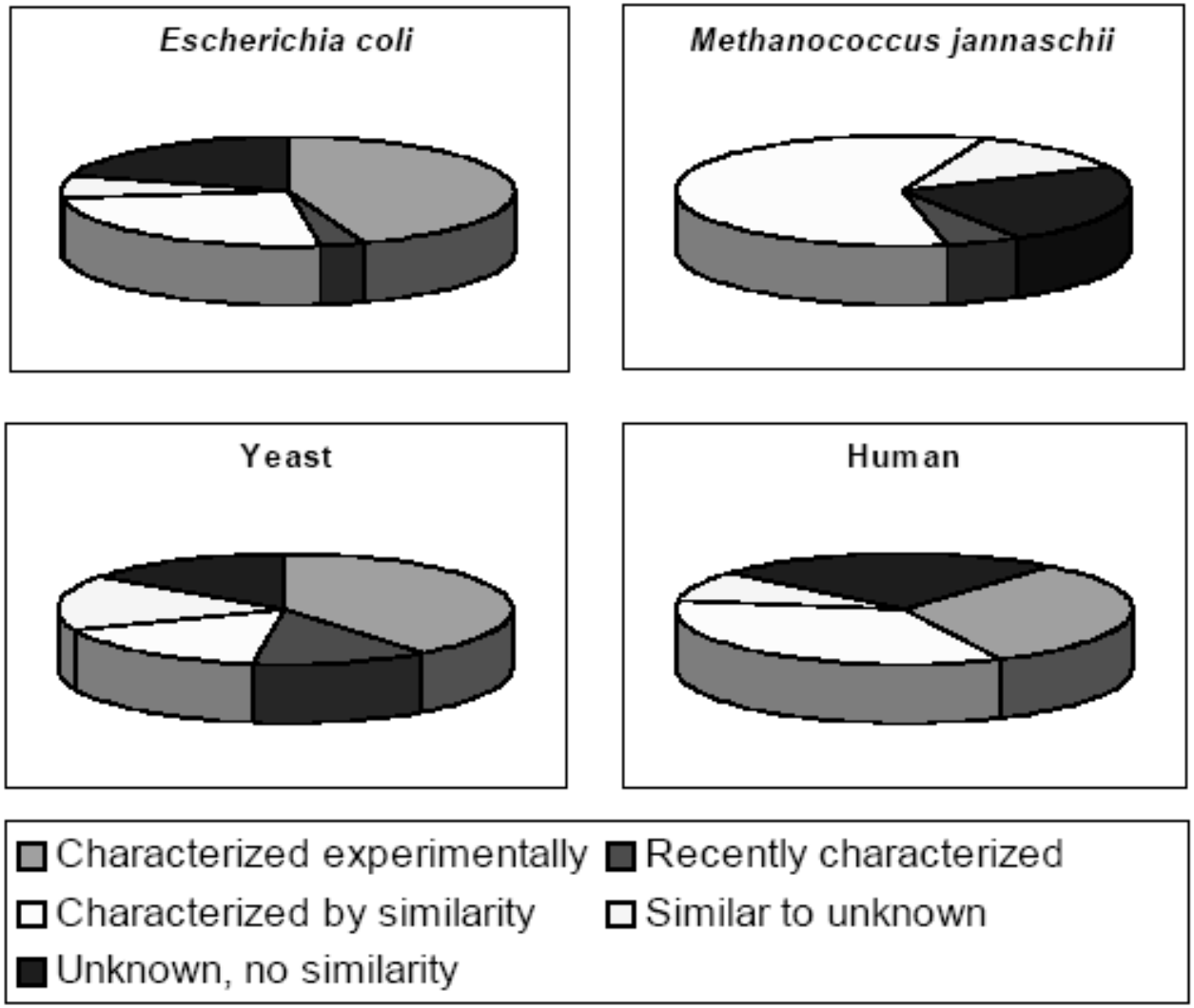
# Αυτοματοποιημένα ??

Αυτοματοποιημένος Σχολιασμός βασισμένος σε ομοιότητες (σύστημα GeneQuiz, Μάϊος 2000) για τα ORFs του γονιδιώματος του Αρχαίου *Methanococcus jannaschii.*



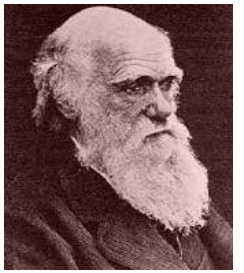

http://jura.ebi.ac.uk:8765/ext-genequiz//genomes/mj0005/index.html

# Αυτοματοποιημένα ?? (2)




Από Koonin EV and Galperin M, (2003) "Sequence – Evolution – Function: Computational Approaches in Comparative Genomics"

# Στοιχεία Μοριακής Εξέλιξης

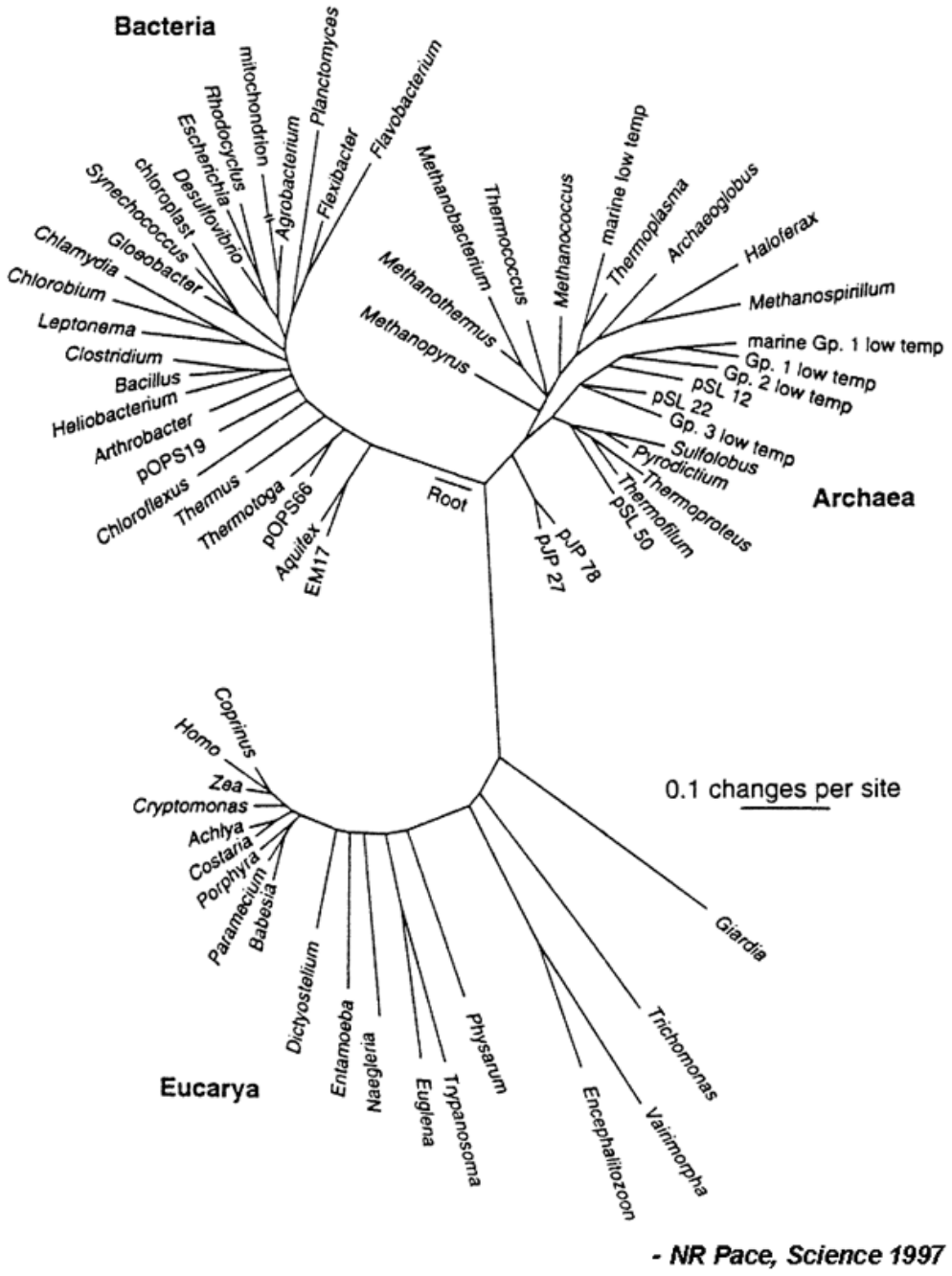


- Η Εξελικτική θεωρία αποτελεί θεμέλειο λίθο της σύγχρονης Βιολογίας
- Έρευνα σε εντελώς διαφορετικούς τομείς (π.χ. ανατομία, γονιδιωματική) επωφελείται από την κατανόηση των αλλαγών των οργανισμών στην πορεία του χρόνου
- Είναι δυνατόν να κατανοήσουμε καλύτερα τη συσχέτιση μεταξύ μορίων μελετώντας τις αλλαγές που υπέστησαν στην πορεία του χρόνου

*Nothing in biology makes sense except in the light of evolution.*

- *Theodosius Dobzhansky, 1973*

# The Tree of Life




*- NR Pace, Science 1997*

# Στοιχεία Μοριακής Εξέλιξης (2)

- Μοριακή?
  - Ποιά μόρια ...
- Εξέλιξη
  - Ποιοί μηχανισμοί/διαδικασίες ...
  - Μικρή vs Μεγάλη κλίμακα
    - Σημειακές μεταλλάξεις (Συνώνυμες, Σιωπηλές, Indels, ...)
    - Indels
    - Ανασυνδυασμός (crossing over – gene conversion)
    - Αναστροφές

# Το μοντέλο 1-παραμέτρου των Jukes-Cantor (1969)

- ΠΑΡΑΔΟΧΕΣ
  - Οι διαφορετικές σημειακές μεταλλάξεις (A->C, A->G, A->T, C->A...) είναι **ΙΣΟΠΙΘΑΝΕΣ (a)**
  - Οι πιθανότητες μετάλλαξης σε κάθε θέση είναι **ΑΝΕΞΑΡΤΗΤΕΣ**

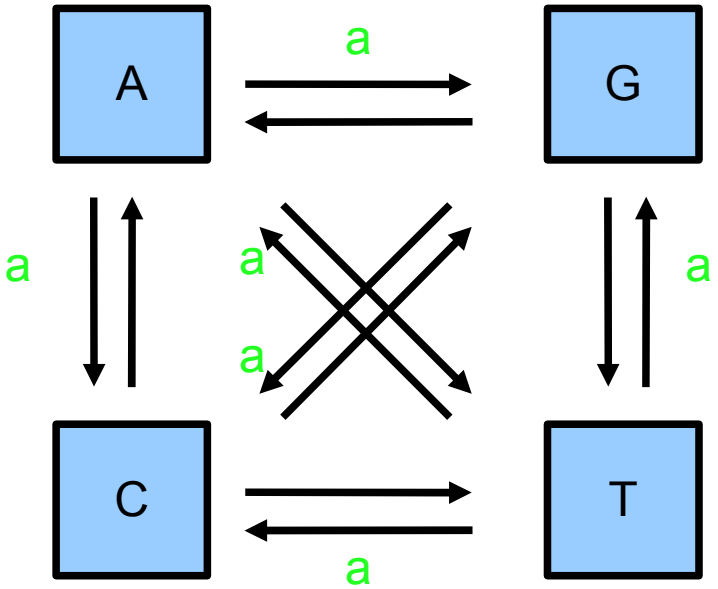
| | To: | A | G | C | T |
|---|---|---|---|---|---|
| From: | A | 1-3a | a | a | a |
| | G | a | 1-3a | a | a |
| | C | a | a | 1-3a | a |
| | T | a | a | a | 1-3a |

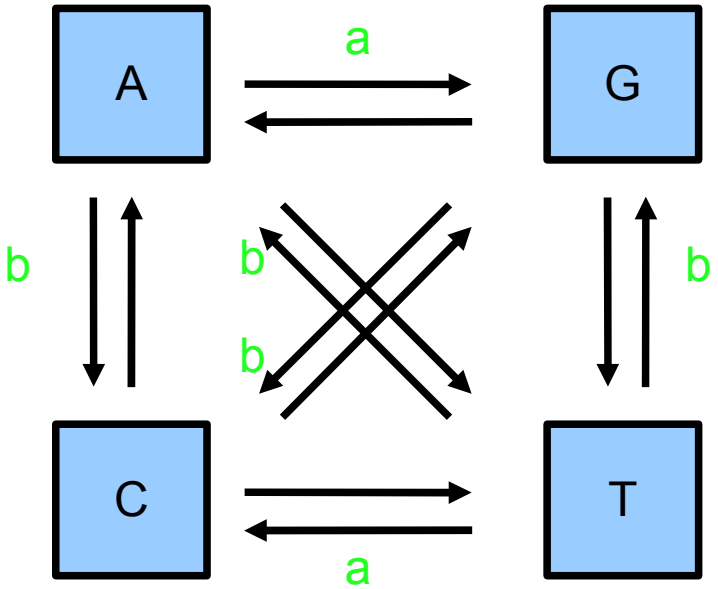
$$P_x(t+1)=(1-3a)P_x(t)+a(1-P_x(t))$$

# Το μοντέλο 2-παραμέτρων του Kimura (1980)

- ΠΑΡΑΔΟΧΕΣ
  - Οι διαφορετικές σημειακές μεταλλάξεις (A->C, A->G, A->T, C->A...) δεν είναι **ΙΣΟΠΙΘΑΝΕΣ**
    - **Οι μεταπτώσεις (A<->G, C<->T) είναι συχνότερες από τις μεταστροφές**
  - Οι πιθανότητες μετάλλαξης σε κάθε θέση είναι **ΑΝΕΞΑΡΤΗΤΕΣ**

# JC vs Kimura



# Δέντρα και Εξελικτική Ιστορία (Ι)

- Περιγραφή εξελικτικών σχέσεων ΜΕΤΑΞΥ ειδών

rodents

birds

snakes

crocodiles

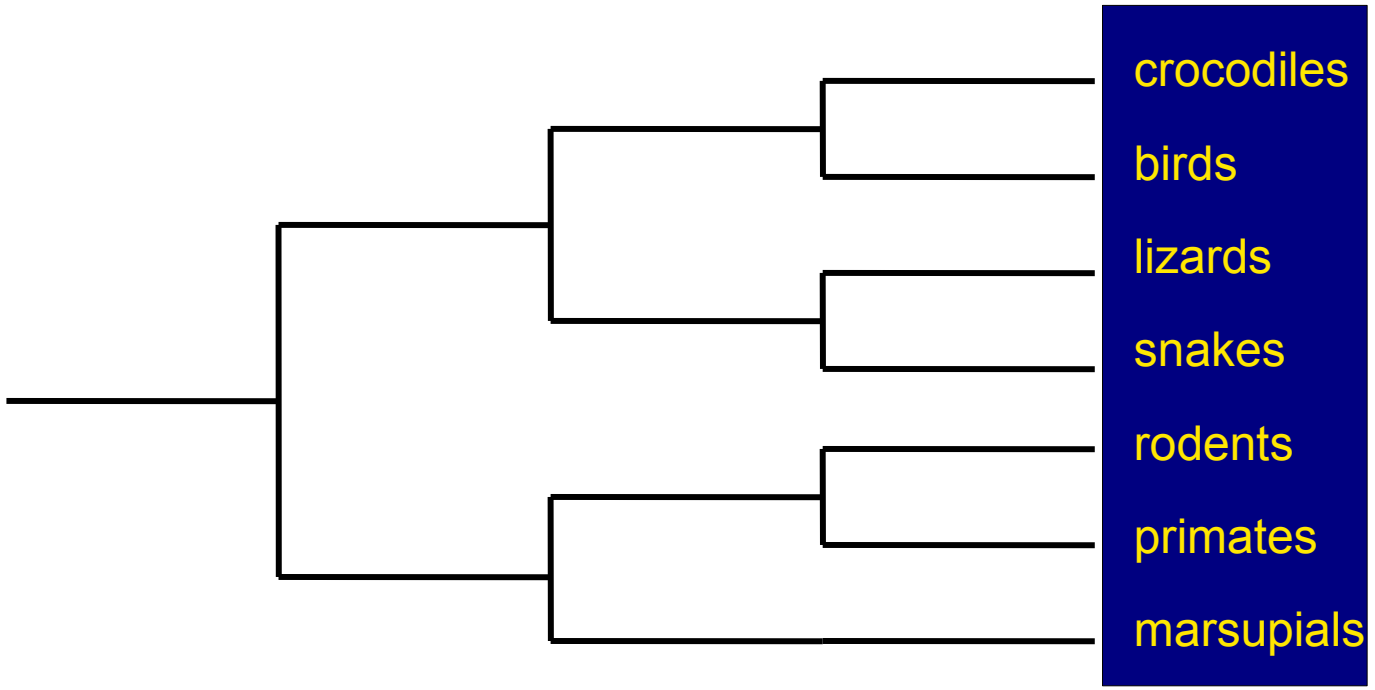
primates

marsupials

lizards

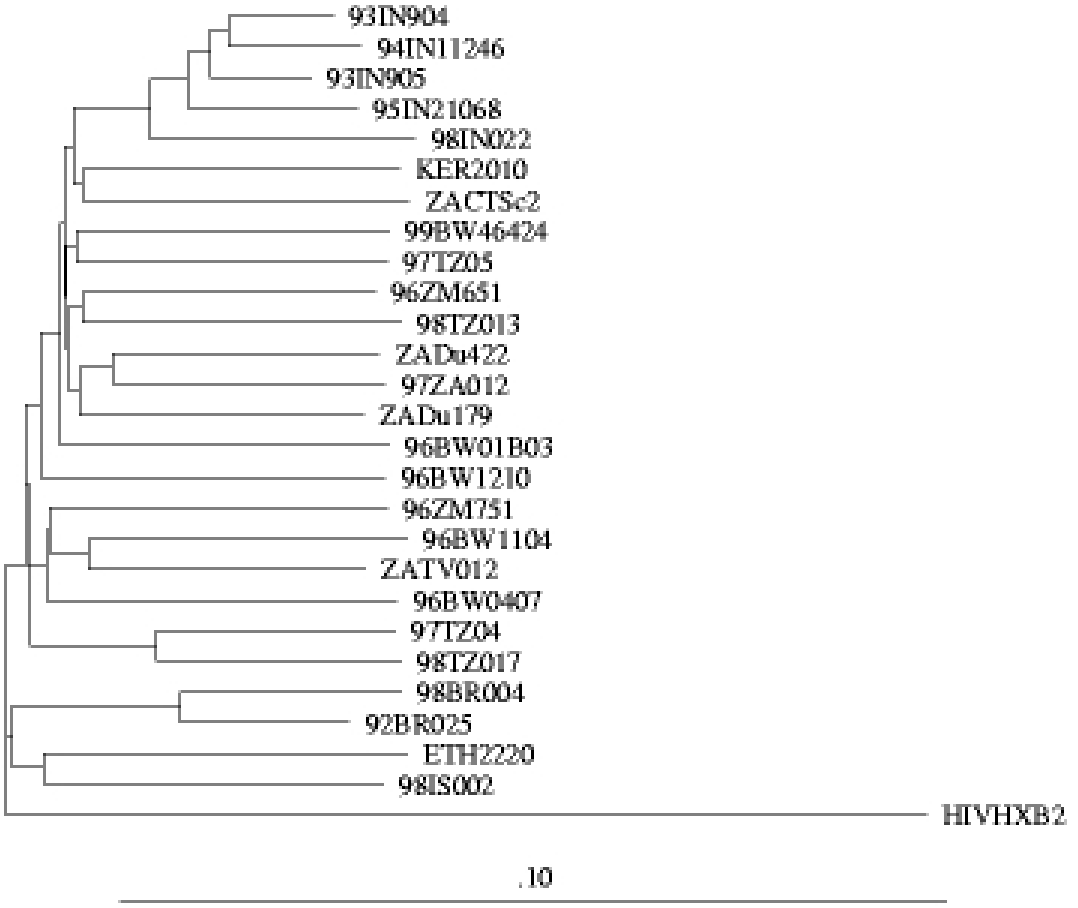
# Δέντρα και Εξελικτική Ιστορία (Ι)

- Περιγραφή εξελικτικών σχέσεων ΜΕΤΑΞΥ ειδών

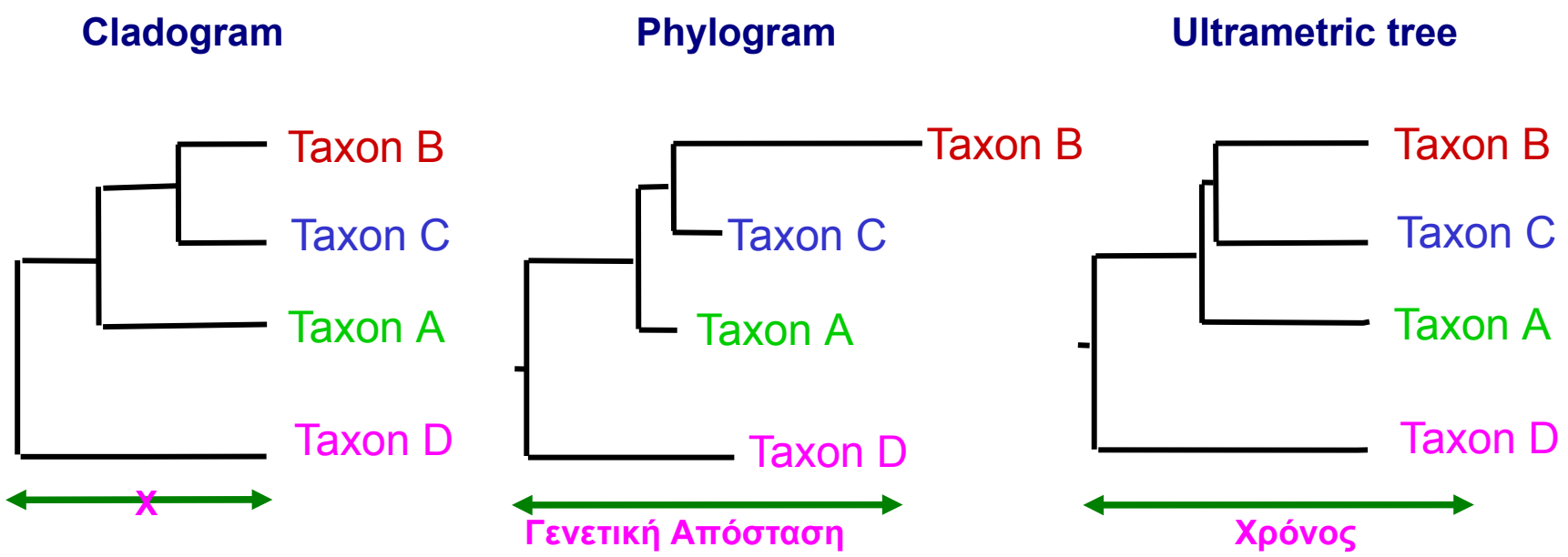


# Δέντρα και Εξελικτική Ιστορία (II)

- Περιγραφή εξελικτικών σχέσεων ΕΝΤΟΣ ειδών

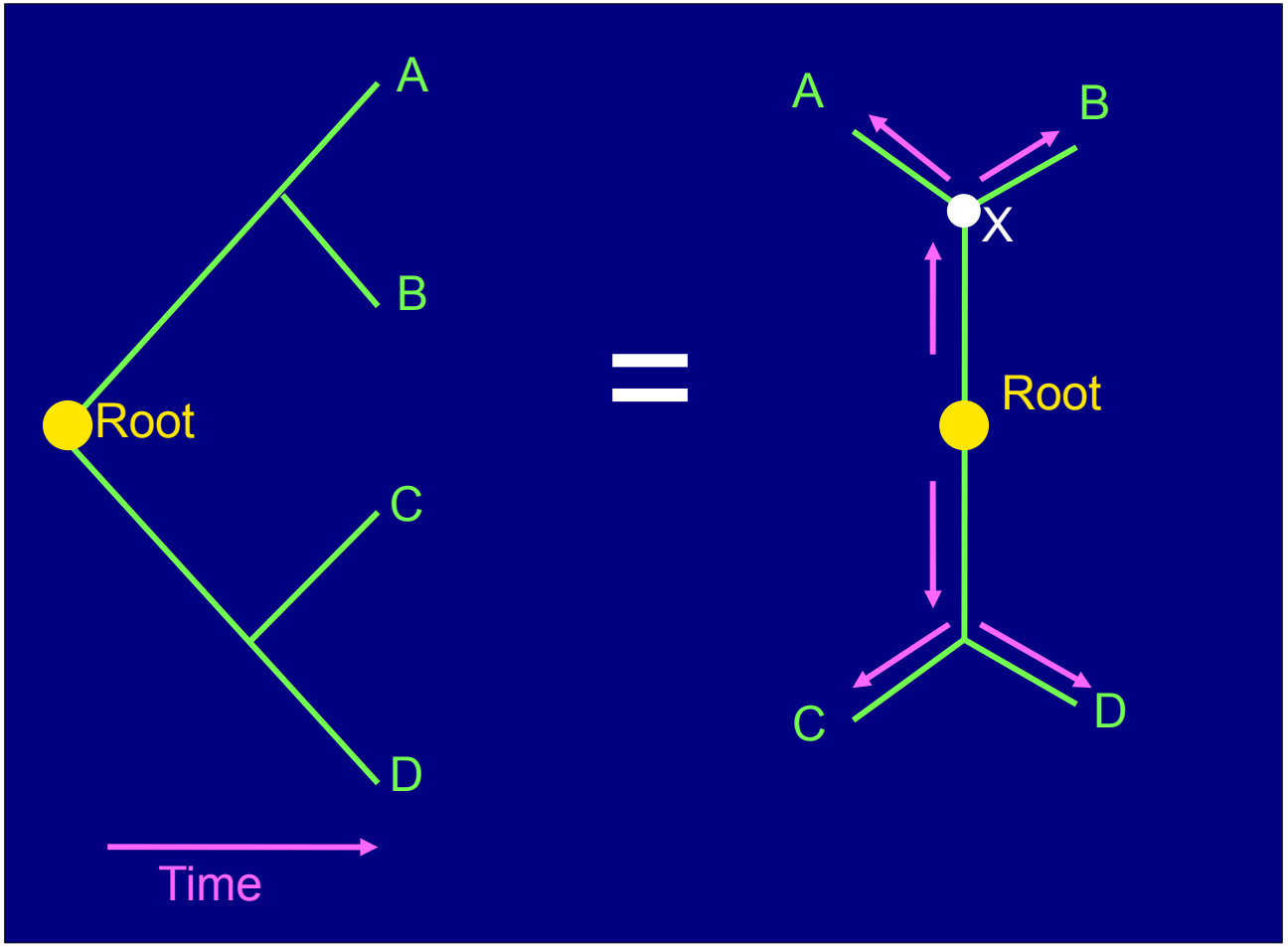


# Τύποι Δέντρων

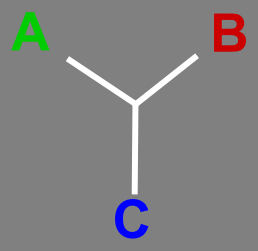


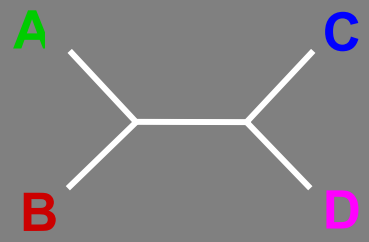
# ... κι άλλοι Τύποι Δέντρων ...

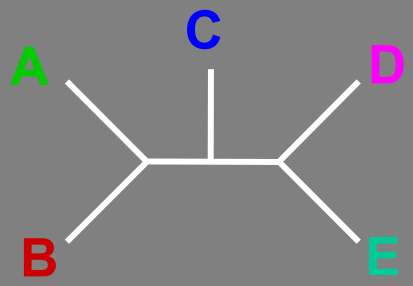
# ... κι άλλοι Τύποι Δέντρων ...

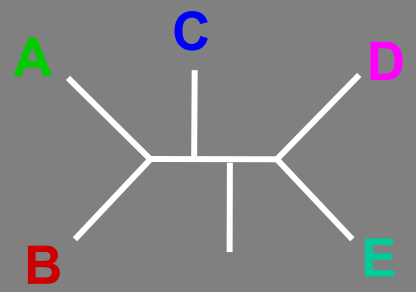


# Δέντρα χωρίς ρίζα



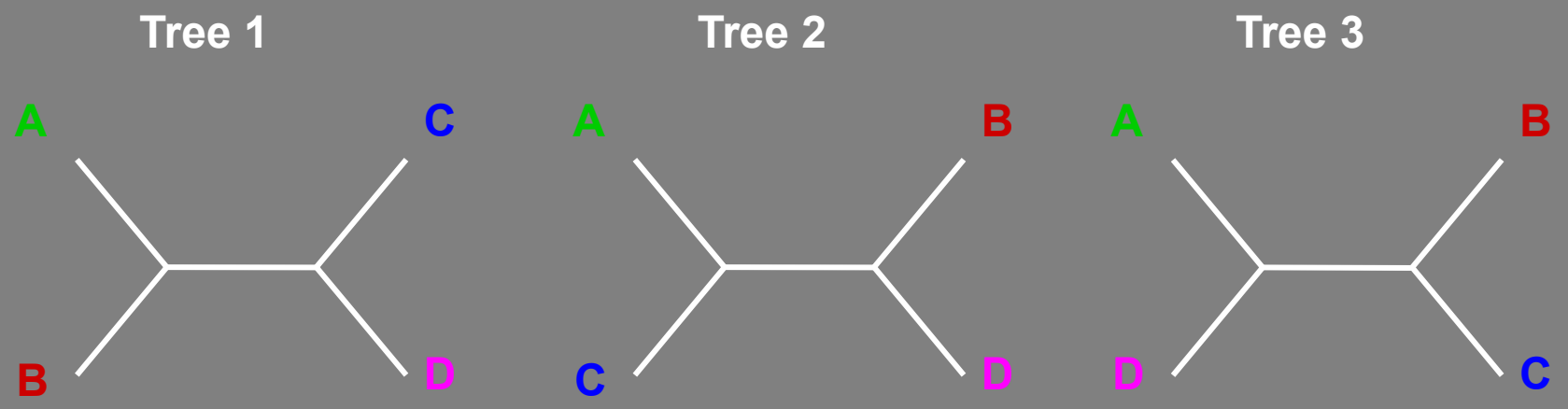



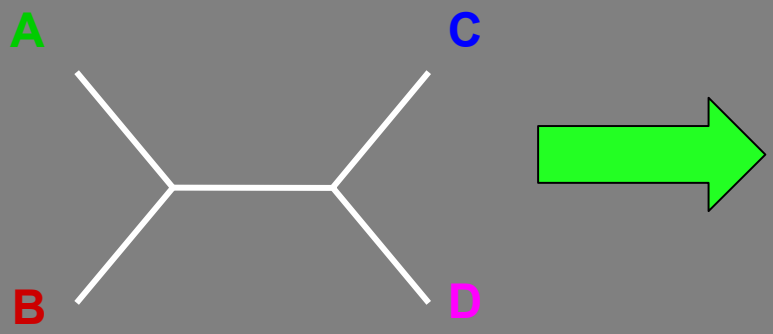

| # Taxa (N) | # Unrooted trees |
|---|---|
| 3 | 1 |
| 4 | 3 |
| 5 | 15 |
| 6 | 105 |
| 7 | 945 |
| 8 | 10,935 |
| 9 | 135,135 |
| 10 | 2,027,025 |
| . | . |
| . | . |
| . | . |
| . | . |
| 30 | ~3.58 x $10^{36}$ |

$$\frac{(2N - 5)!}{2^{N-3}(N - 3)!} = \text{\# unrooted trees for N taxa}$$

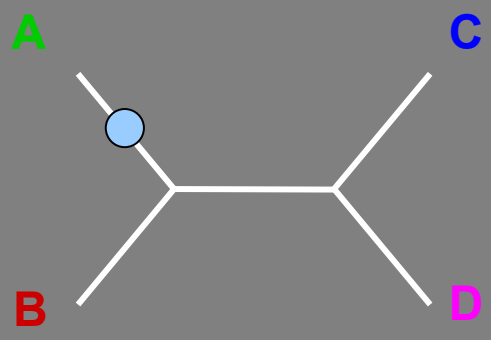
# Δέντρα χωρίς ρίζα (II)

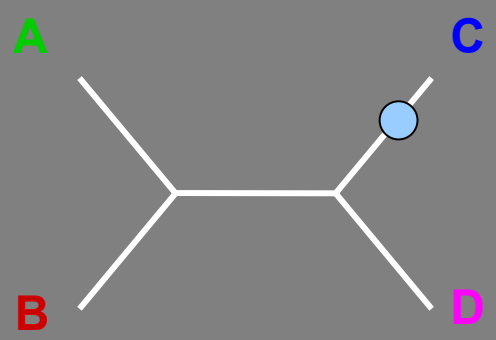


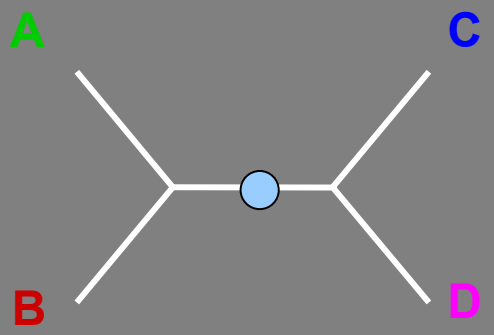
# Που “κολλάει” η ρίζα ??

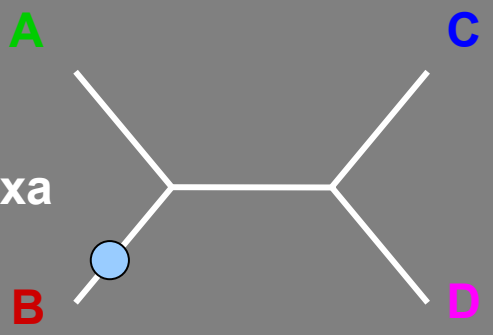


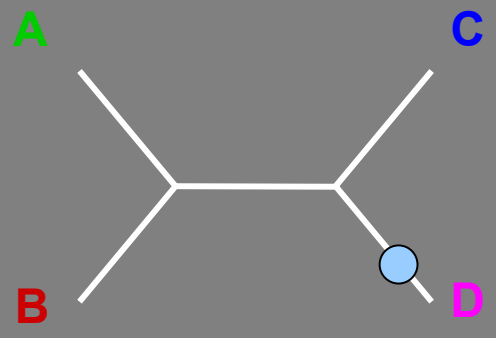


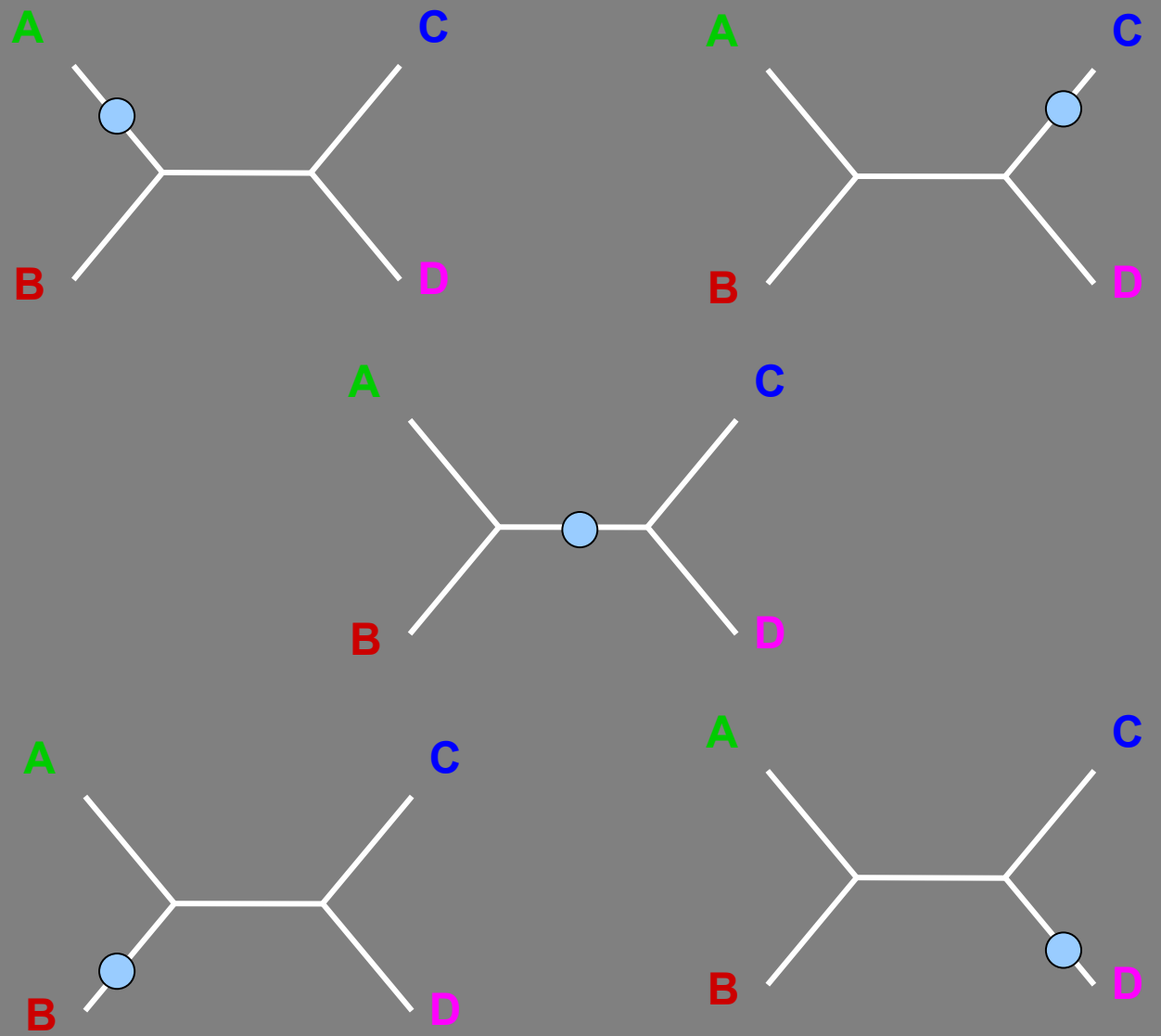




$$\frac{(2N-3)!}{2^{N-2}(N-2)!} = \text{\# rooted trees for N taxa}$$

# … ε, και … ?!

- Τα 5 αυτά δέντρα υποδεικνύουν ***ΔΙΑΦΟΡΕΤΙΚΕΣ*** εξελικτικές σχέσεις!!!



# Μα είναι πολλά τα Δέντρα ???

| Number of Taxa | Number of unrooted trees | Number of rooted trees |
|---|---|---|
| 3 | 1 | 3 |
| 4 | 3 | 15 |
| 5 | 15 | 105 |
| 6 | 105 | 945 |
| 7 | 945 | 10395 |
| 8 | 10395 | 135135 |
| 9 | 135135 | 2027025 |
| 10 | 2027025 | 34459425 |
| 20 | 2.22E+020 | 8.20E+021 |

# Πώς φτιάχνονται τα Δέντρα ?

?

# Ομολογία

(... και άλλες όμορφες ιστορίες ... )

- ΟΜΟΛΟΓΙΑ: η σχέση δύο χαρακτήρων οι οποίοι κατάγονται (συνήθως με απόκλιση) από ένα ΚΟΙΝΟ ΠΡΟΓΟΝΙΚΟ χαρακτήρα

- Χαρακτήρες:
  - Γενετικοί
  - Μορφολογικοί/Δομικοί
  - Συμπεριφορά

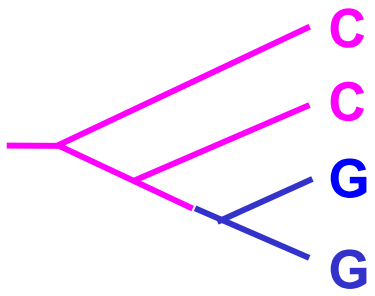
**ΠΟΙΟΤΙΚΟ ΚΡΙΤΗΡΙΟ (ΝΑΙ/ΟΧΙ)**

# DR Scooter strikes again!!

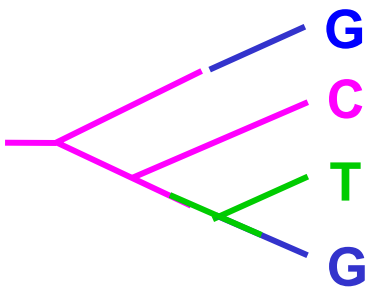
**ΚΥΡΙΑ ΜΟΥ ΕΙΣΤΕ ΠΟΛΥ ΕΓΚΥΟΣ !!!**
**Καλό, ε??**

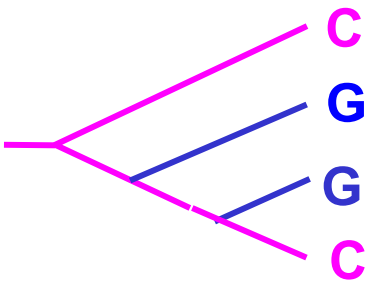
# Ομοιότητα και Ομολογία (I)

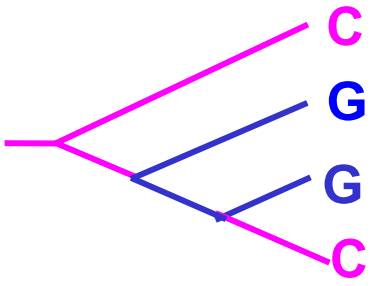
**ΟΜΟΛΟΓΙΑ (HOMOLOGY)**




**ΟΜΟΠΛΑΣΙΑ (HOMOPLASY)**



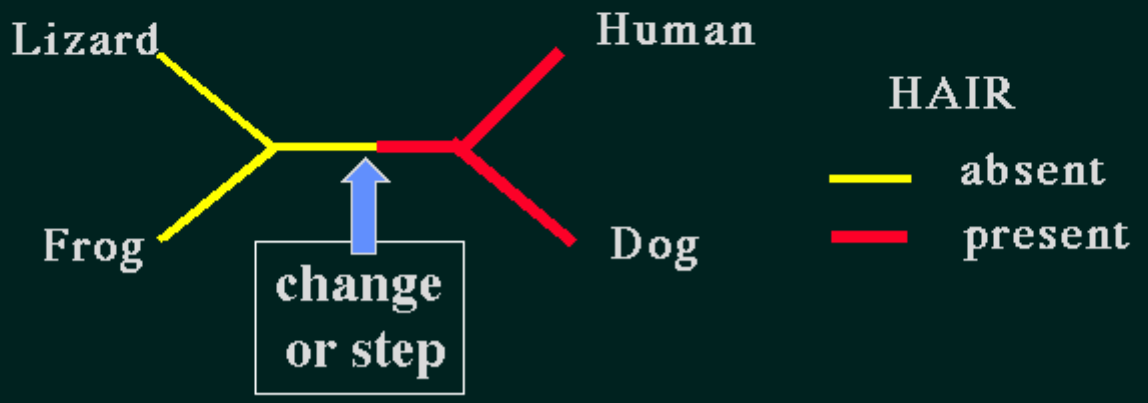



# Ομοιότητα και Ομολογία (II)

## Unique and unreversed characters - Hair

- Because hair evolved only once and is unreversed it is homologous and provides unambiguous evidence for the clade Mammalia



# Ομοιότητα και Ομολογία (III)

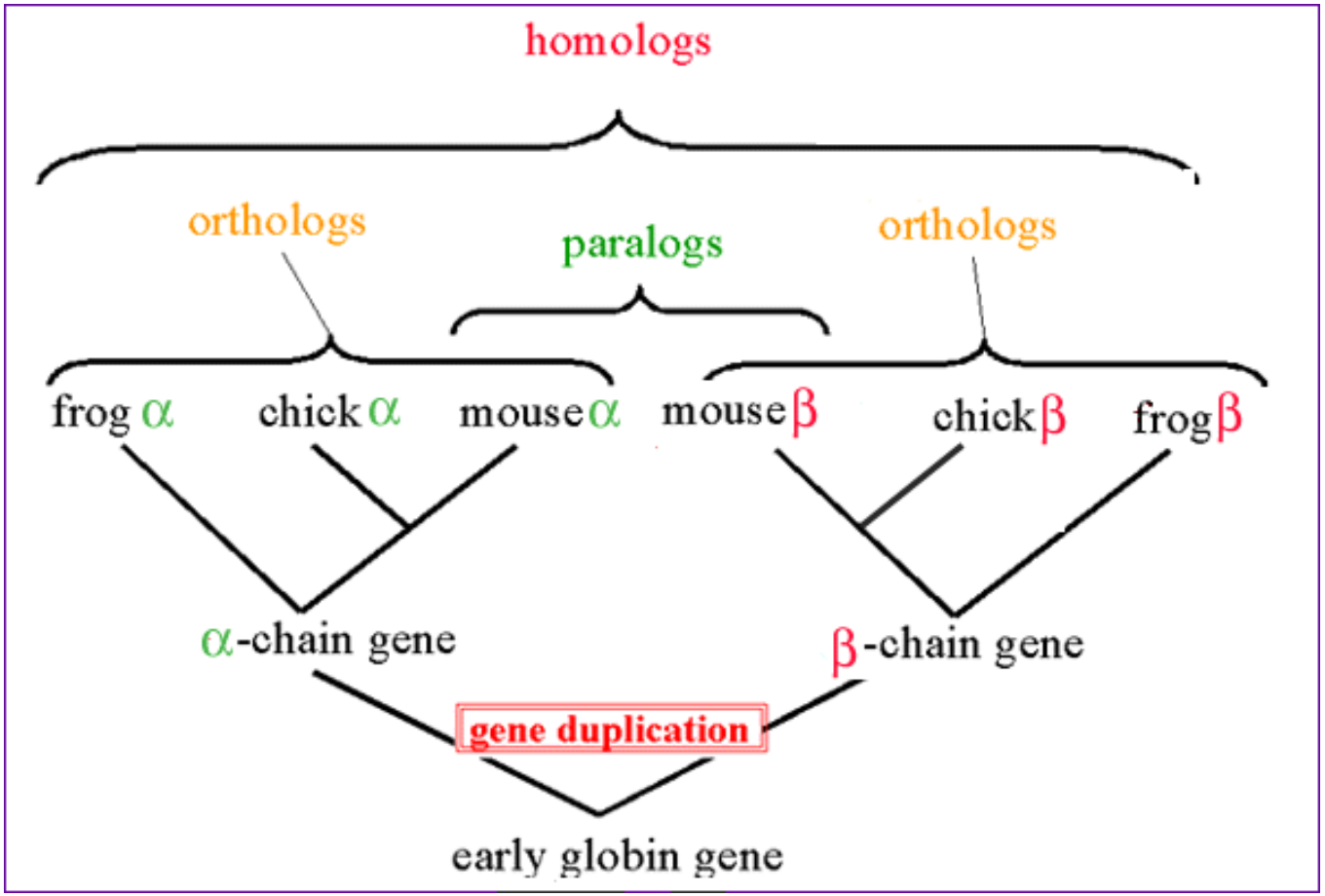
## Homoplasy - independent evolution - Tails

- Loss of tails evolved independently in humans and frogs - there are two steps on the true tree

# Σχέσεις Ομολογίας

- Ορθολογία
  - Ομόλογοι χαρακτήρες προϊόντα ***ΕΙΔΟΓΕΝΕΣΗΣ***
- Παραλογία
  - Ομόλογοι χαρακτήρες προϊόντα ***ΓΟΝΙΔΙΑΚΟΥ ΔΙΠΛΑΣΙΑΣΜΟΥ***
- Ξενολογία
  - Ομόλογοι χαρακτήρες προϊόντα ***ΟΡΙΖΟΝΤΙΑΣ ΜΕΤΑΦΟΡΑΣ***
- [Συνολογία]
  - Ομόλογοι χαρακτήρες προϊόντα π.χ. ενδοσυμβίωσης

# Σχέσεις Ομολογίας (II)




Πηγή: http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Orthology.html

# Αλγόριθμοι Εύρεσης Ομοιοτήτων Ακολουθιών

## Μέρος Ι: Στοιχίσεις ακολουθιών κατά ζεύγη

# Σύνοψη

- Αλγόριθμοι
  - Βασικοί Ορισμοί
  - Στοιχεία Ανάλυσης Αλγοριθμικής Πολυπλοκότητας

- Σύγκριση ή Στοίχιση Ακολουθιών?
  - Dot Matrix Plots
  - Μέθοδοι Δυναμικού Προγραμματισμού
    - Τοπικές  Στοιχίσεις
    - Ολικές Στοιχίσεις
    - “Πιάσε μια απ'όλα”

# Αλγόριθμοι

- Ορισμός: Αλγόριθμος είναι μια καλά προσδιορισμένη διαδικασία για την επίλυση μιας ***κλάσης*** προβλημάτων:
  - Συγκεκριμένα δεδομένα εισόδου
  - Πεπερασμένο πλήθος βημάτων
  - Επίλυση Προβλήματος

- Χαρακτηριστικά Αλγορίθμων
  - Ορθότητα
  - Αποδοτικότητα

# Αλγόριθμοι ΙΙ

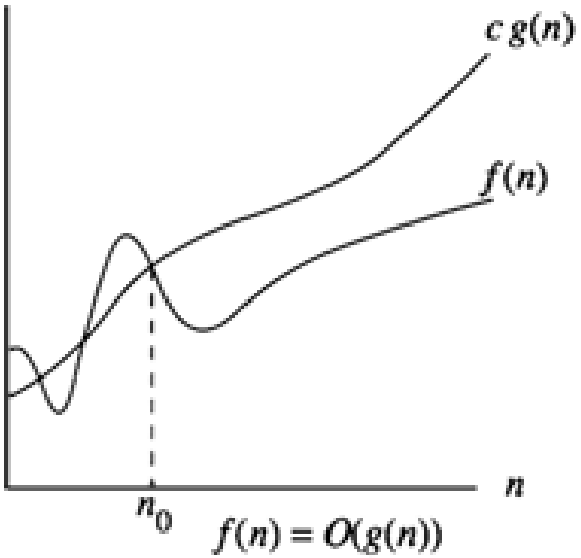
- Μας ενδιαφέρουν οι ***ΟΡΘΟΙ*** αλγόριθμοι [πάντα???]
- Αξιολόγηση της Αποδοτικότητας
  - ***Πρακτική Εφαρμογή***
  - ***Ασυμπτωτική Συμπεριφορά*** συναρτήσει του μεγέθους των δεδομένων εισόδου

# Ασυμπτωτική Συμπεριφορά

- Αναφέρεται σε οποιαδήποτε συνάρτηση
- Ιδιαίτερο ενδιαφέρον:
  - Χαρακτηριστικά εκτέλεσης (Χρόνος, Μνήμη, κλπ) ως συνάρτηση του "μεγέθους" ***n*** του προβλήματος

- Ασυμπτωτικά <=> Μεγάλο ***n***

- O(***g(n)***), Ω(***g(n)***), Θ(***g(n)***)
  - f(***n***)=O(***n***) linear
  - f(***n***)=O(***n***$^2$) quadratic
  - f(***n***)=O(log(***n***)) logarithmic
  - f(***n***)=O(c$^n$) exponential

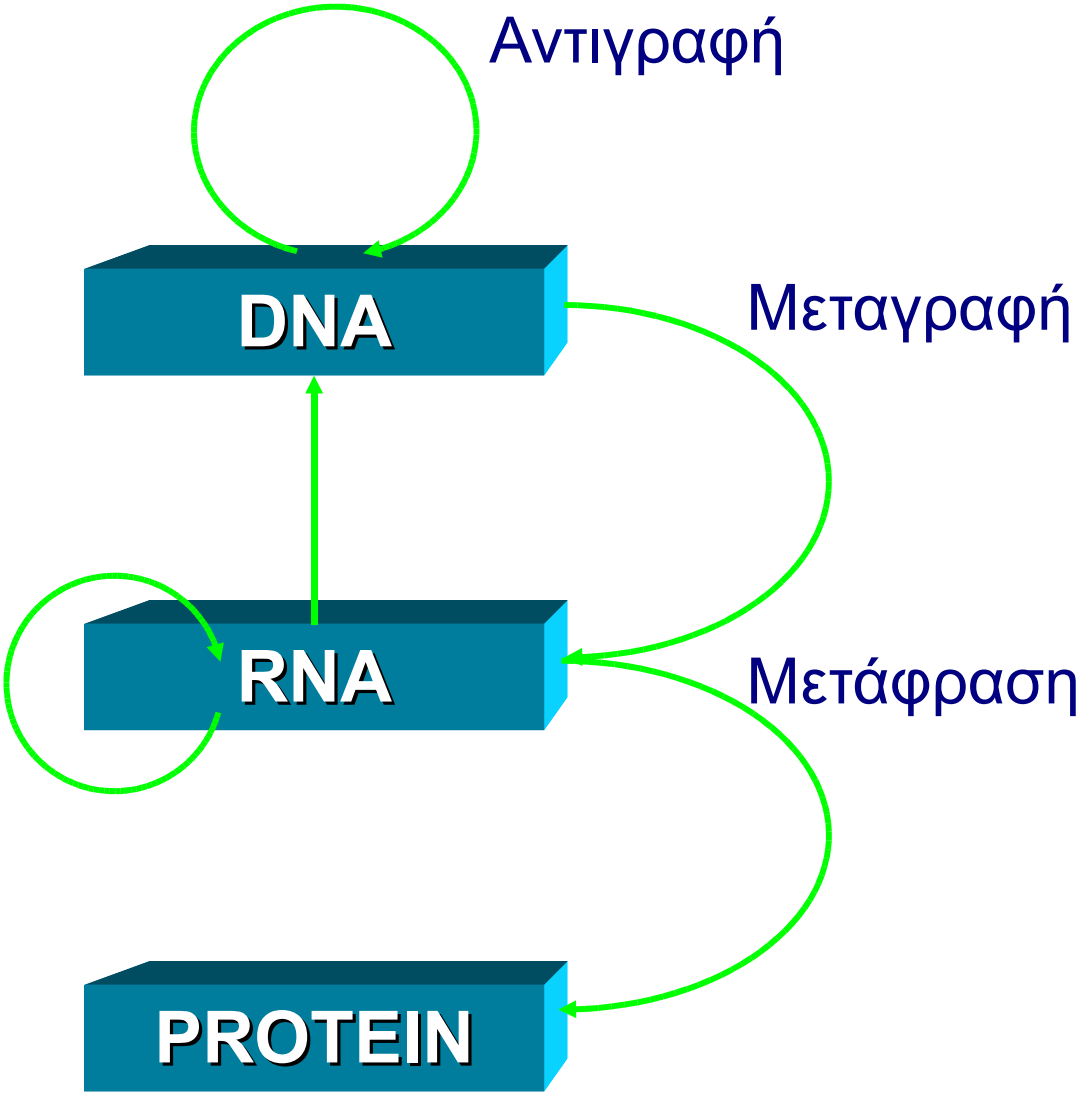


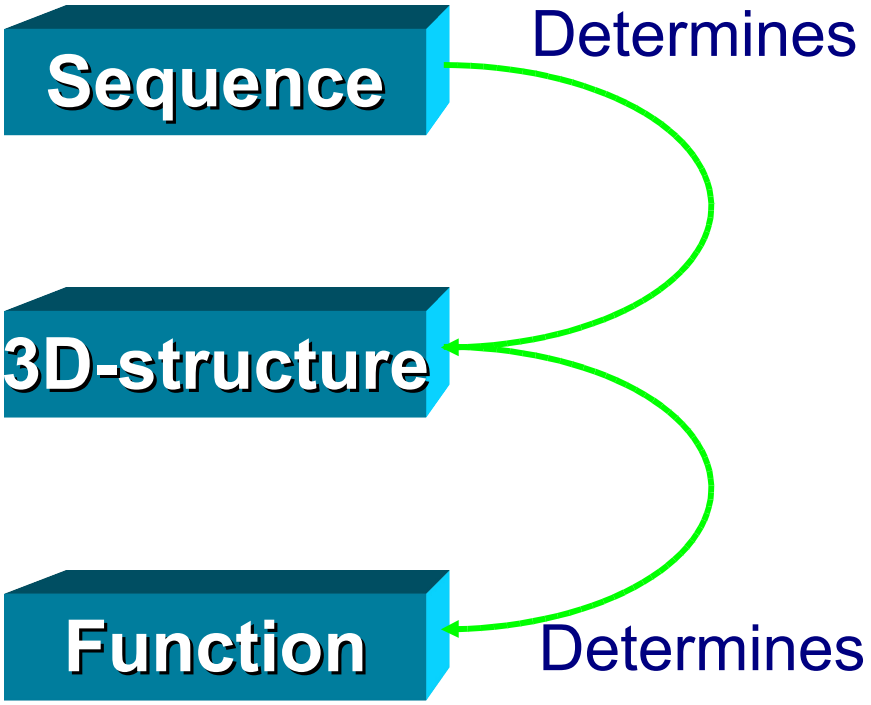

από Cormen *et. al*

# Σύγκριση; Για ποιο λόγο; Πώς;

- Ήταν πάντα … '**trendy**'
- Ο τρόπος με τον οποίο θα εφαρμοσθεί εξαρτάται από:
  - Τύπο/Πλήθος δεδομένων
  - Ερώτημα (?)
- Στηριζόμαστε στο γεγονός ότι:
  - ΟΜΟΙΟΤΗΤΑ ΑΚΟΛΟΥΘΙΩΝ => ???

# ΑΚΟΛΟΥΘΙΑ == ΠΛΗΡΟΦΟΡΙΑ



# ΟΜΟΙΟΤΗΤΑ => ??

- Δομική/Λειτουργική Συσχέτιση
- Εξελικτική Σχέση
- Εντοπισμός 'κρίσιμων' καταλοίπων
- Εύρεση χαρακτηριστικών μοτίβων

# Σύγκριση Δύο Ακολουθιών (pairwise alignment)

- Τύποι Σύγκρισης
- Μέτρο Σύγκρισης
- Αντικειμενικότητα
- Σημαντικότητα

# Σύγκριση Δύο Ακολουθιών (pairwise alignment)

- Τύποι Σύγκρισης
  - Τοπική, Ολική
  - Πρωτεΐνη/DNA/RNA
  - Τί ιδιότητες έχουν οι ακολουθίες μου??
- Μέτρο Σύγκρισης
- Αντικειμενικότητα
- Σημαντικότητα

# Σύγκριση Δύο Ακολουθιών (pairwise alignment)

- Τύποι Σύγκρισης
- Μέτρο Σύγκρισης
  - Χρειαζόμαστε ένα μοντέλο => ΠΙΝΑΚΕΣ ΑΝΤΙΚΑΤΑΣΤΑΣΗΣ
    - Εξελικτική Σχέση
      - Αντικαταστάσεις (substitutions)
      - Προσθήκες (insertions)
      - Εξαλείψεις (deletions)
    - Δομική Αντιστοιχία
    - Φυσικοχημικές Ιδιότητες
- Αντικειμενικότητα
- Σημαντικότητα

# Στοίχιση Ακολουθιών Κατά Ζεύγη (Pairwise alignment)

- Τύποι Σύγκρισης
- Μέτρο Σύγκρισης
- Αντικειμενικότητα
  - Ποσοτικά vs Ποιοτικά Κριτήρια
  - Αυτοματοποίηση (??)
- Σημαντικότητα

# Στοίχιση Ακολουθιών Κατά Ζεύγη (Pairwise alignment)

- Τύποι Σύγκρισης
- Μέτρο Σύγκρισης
- Αντικειμενικότητα
- Σημαντικότητα
  - **... so what ???**

# Στοίχιση Ακολουθιών Κατά Ζεύγη (Pairwise alignment)

```
S1  HFCGGSLINEQWVVSAGHC

S2  HFCGASIYNENYATAGHC
```

**Τμήματα ακολουθιών Θρυψίνης**
**S1: Ποντικός**
**S2: Αστακός**

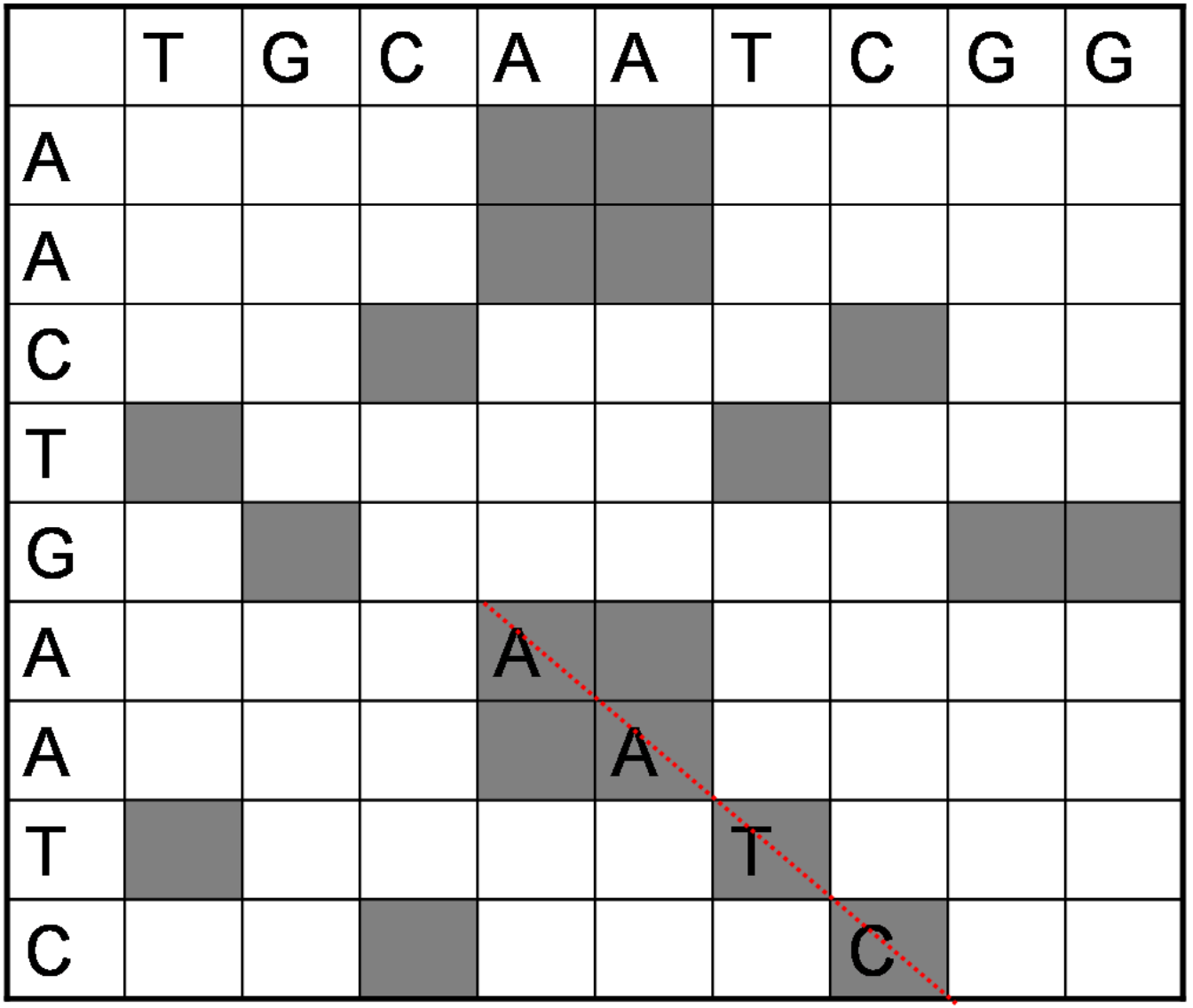
```
S1  HFCGGSLINEQWVVSAGHC

S2  HFCGASIYNENYA-TAGHC
```

**Με το ΧΕΡΙ!!!**
**ή το μάτι ..**

```
          S-S
     ┌─────────┐
S1  HFCGGSLINEQWVVSAGHC
    HFCG S  NE     AGHC
S2  HFCGASIYNENYA-TAGHC
```

# Πίνακες Διαγραμμάτων Σημείων (Dot Matrix Plots)



# Πίνακες Διαγραμμάτων Σημείων (Dot Matrix Plots)

- **Πλεονεκτήματα**
  - Οπτικοποίηση
  - Εύκολη (σχετικά) κατασκευή
  - Μικρές (σχετικά) Υπολογιστικές Απαιτήσεις
- **Μειονεκτήματα**
  - Αντικειμενικότητα
  - Σημαντικότητα
- ***ΣΗΜΑΝΤΙΚΟ!!!*** ***Στοίχιση == Διαδρομή***

# Μέθοδοι Δυναμικού Προγραμματισμού

- Προγραμματισμού (;)
- Αναζήτηση των Βέλτιστων Λύσεων μέσα από ΜΕΓΑΛΑ σύνολα λύσεων

$$\binom{2N}{N} = \frac{(2N)!}{(N!)^2} \simeq \frac{2^{2N}}{\sqrt{2\pi N}}$$

- Αντιμετώπιση με στρατηγική από «Κάτω προς τα επάνω»
  - Διαίρει και βασίλευε (;)
  - Αλγοριθμική πολυπλοκότητα ~ $N^2$

J. Mol. Biol. (1970) 48, 443–453

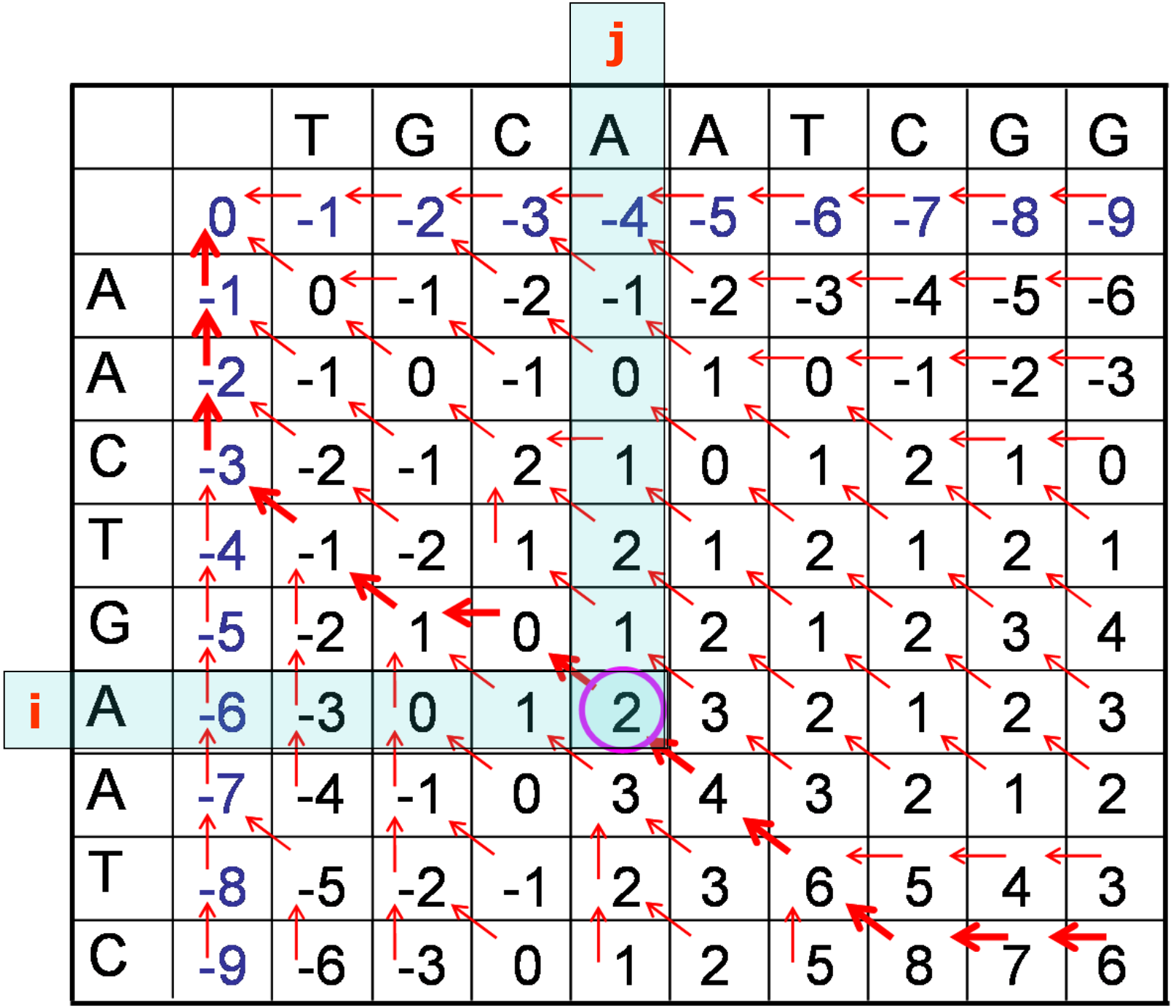
# Ολική Στοίχιση

A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins

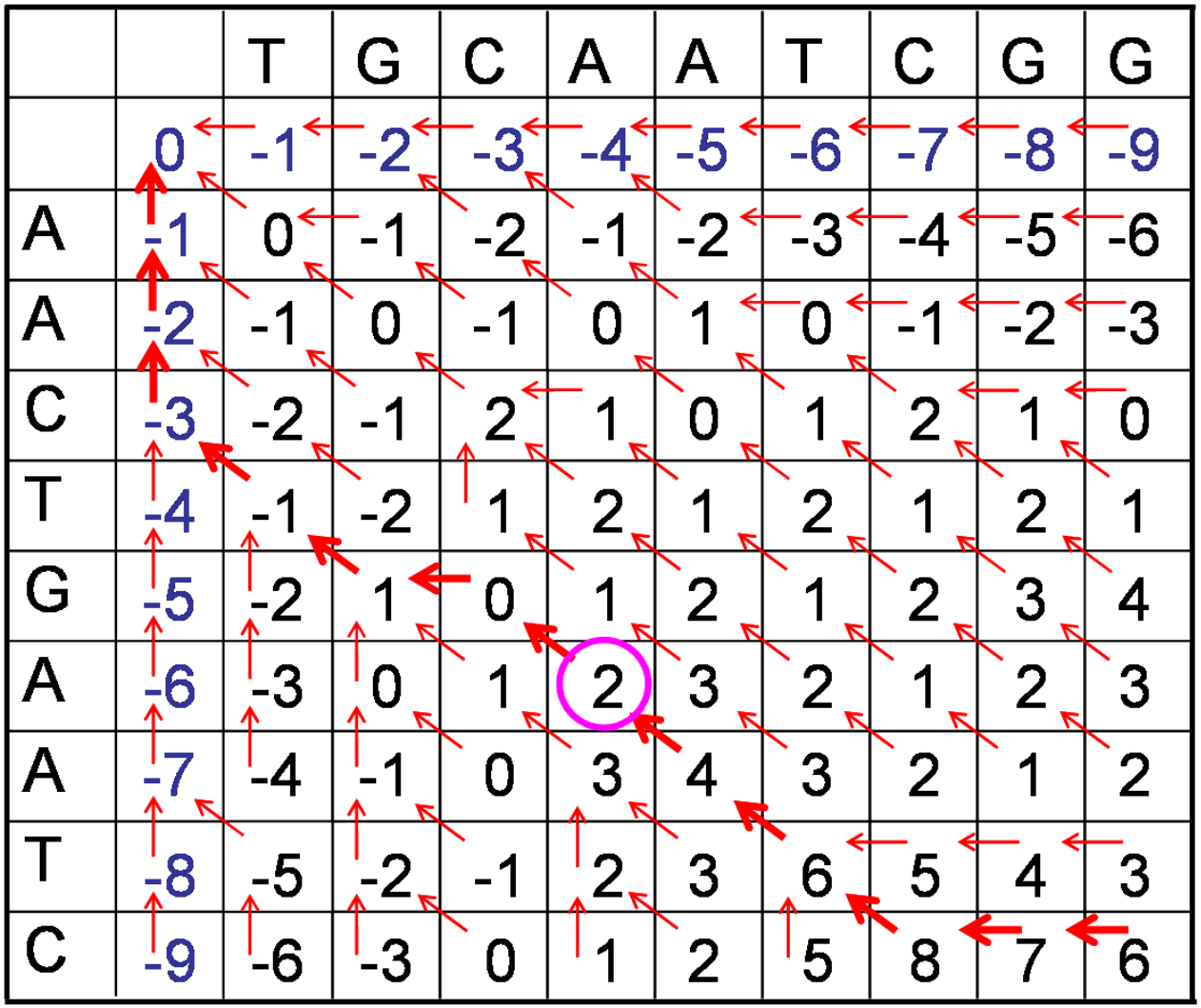
Saul B. Needleman and Christian D. Wunsch

| | | |
|---|---|---|
| T G C A $x_i$ | T G C $x_i$ - | T G C A $x_i$ |
| T A C A $y_i$ | T G C A $y_i$ | T G C $y_i$ - |

$$\mathbf{F}_{i,j} = \max \begin{cases} \mathbf{F}_{i-1,j-1} + \mathbf{S}_{x_i,y_j} \\ \mathbf{F}_{i-1,j} - g \\ \mathbf{F}_{i,j-1} - g \end{cases}$$


j
i
T G C A A T C G G
0 -1 -2 -3 -4 -5 -6 -7 -8 -9
A -1 0 -1 -2 -1 -2 -3 -4 -5 -6
A -2 -1 0 -1 0 1 0 -1 -2 -3
C -3 -2 -1 2 1 0 1 2 1 0
T -4 -1 -2 1 2 1 2 1 2 1
G -5 -2 1 0 1 2 1 2 3 4
A -6 -3 0 1 2 3 2 1 2 3
A -7 -4 -1 0 3 4 3 2 1 2
T -8 -5 -2 -1 2 3 6 5 4 3
C -9 -6 -3 0 1 2 5 8 7 6

```
---TGCAATCGG
AACTG-AATC--
```



| | | T | G | C | A | A | T | C | G | G |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | -1 | -2 | -3 | -4 | -5 | -6 | -7 | -8 | -9 |
| A | -1 | 0 | -1 | -2 | -1 | -2 | -3 | -4 | -5 | -6 |
| A | -2 | -1 | 0 | -1 | 0 | 1 | 0 | -1 | -2 | -3 |
| C | -3 | -2 | -1 | 2 | 1 | 0 | 1 | 2 | 1 | 0 |
| T | -4 | -1 | -2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 |
| G | -5 | -2 | 1 | 0 | 1 | 2 | 1 | 2 | 3 | 4 |
| A | -6 | -3 | 0 | 1 | 2 | 3 | 2 | 1 | 2 | 3 |
| A | -7 | -4 | -1 | 0 | 3 | 4 | 3 | 2 | 1 | 2 |
| T | -8 | -5 | -2 | -1 | 2 | 3 | 6 | 5 | 4 | 3 |
| C | -9 | -6 | -3 | 0 | 1 | 2 | 5 | 8 | 7 | 6 |

*J. Mol. Biol.* (1981), **147**, 195–197

T. F. SMITH

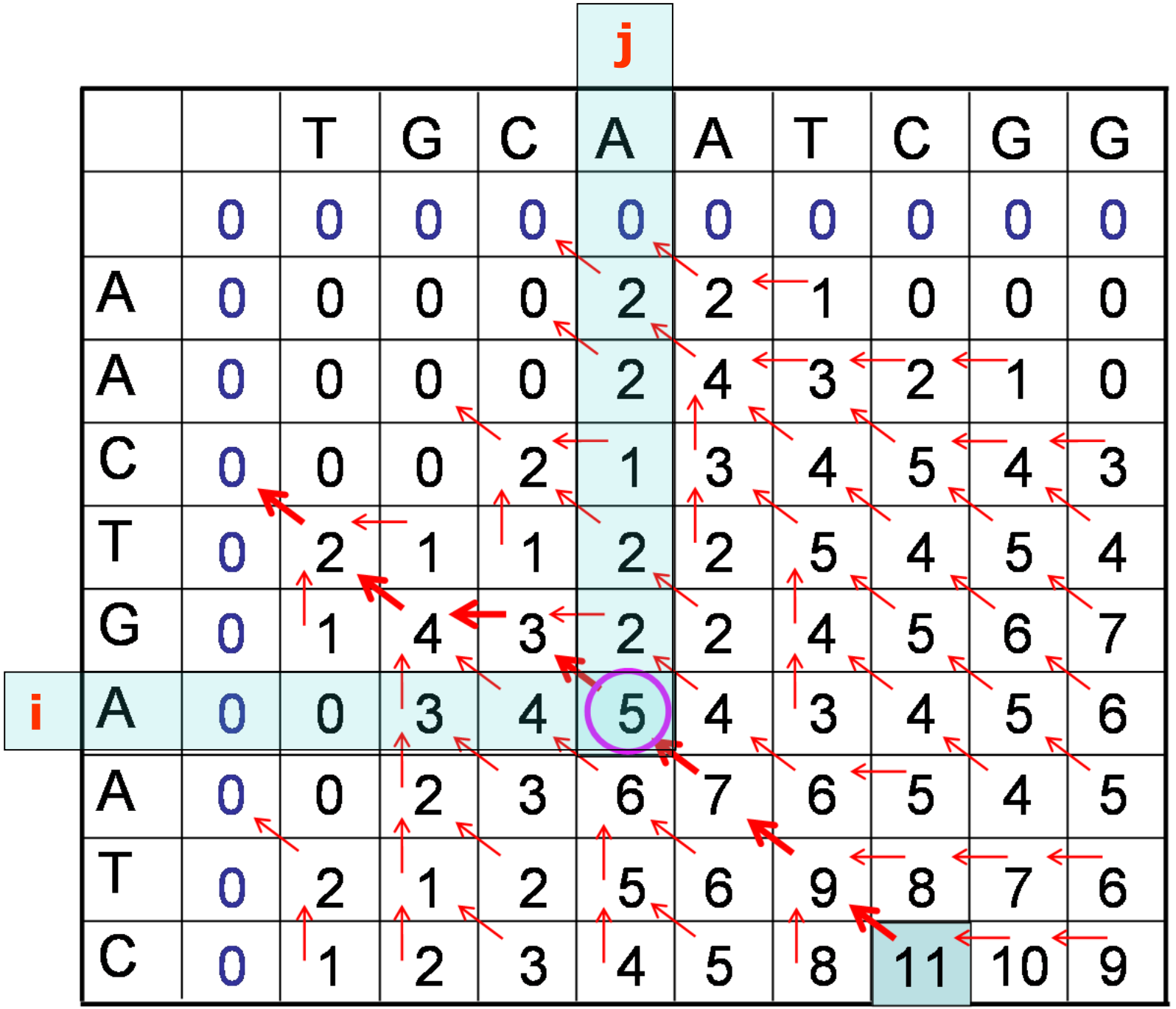
M. S. WATERMAN

# Τοπική Στοίχιση

**Identification of Common Molecular Subsequences**

| | | |
|---|---|---|
| T G C A $x_i$ | T G C $x_i$ - | T G C A $x_i$ |
| T A C A $y_i$ | T G C A $y_i$ | T G C $y_i$ - |

$$\mathbf{F}_{i,j} = \max \begin{cases} \mathbf{F}_{i-1,j-1} + \mathbf{S}_{x_i,y_j} \\ \mathbf{F}_{i-1,j} - g \\ \mathbf{F}_{i,j-1} - g \\ 0 \end{cases}$$


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

**TGCAATC**

**TG-AATC**

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

# Συζήτηση

- ...

# Συζήτηση ...