Αλγόριθμοι Εύρεσης Ομοιοτήτων Ακολουθιών
Μέρος Ι: Στοιχίσεις ακολουθιών κατά ζεύγη

Βασίλης Προμπονάς, PhD
Ερευνητικό Εργαστήριο Βιοπληροφορικής

Τμήμα Βιολογικών Επιστημών
Νέα Παν/πολη, Γραφείο Β161
Πανεπιστήμιο Κύπρου
Ταχ.Κιβ. 20537
1678, Λευκωσία
ΚΥΠΡΟΣ

τηλ: 00357-22892879 (εσωτ. 2879)
email: vprobon@ucy.ac.cy, vprobon@biol.uoa.gr

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”
- Συζήτηση

Αλγόριθμοι

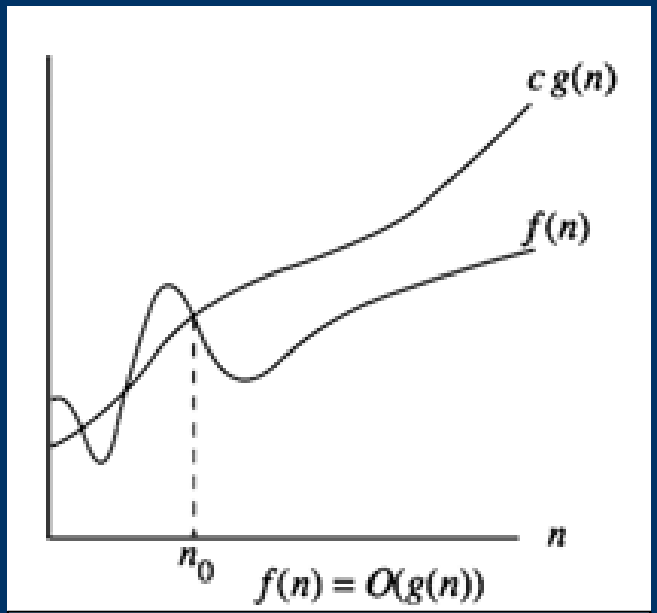
- Ορισμός: Αλγόριθμος είναι μια καλά προσδιορισμένη διαδικασία για την επίλυση μιας ***κλάσης*** προβλημάτων:
  - Συγκεκριμένα δεδομένα εισόδου
  - Πεπερασμένο πλήθος βημάτων
  - Επίλυση Προβλήματος
- Χαρακτηριστικά Αλγορίθμων
  - Ορθότητα
  - Αποδοτικότητα

Αλγόριθμοι II

- 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

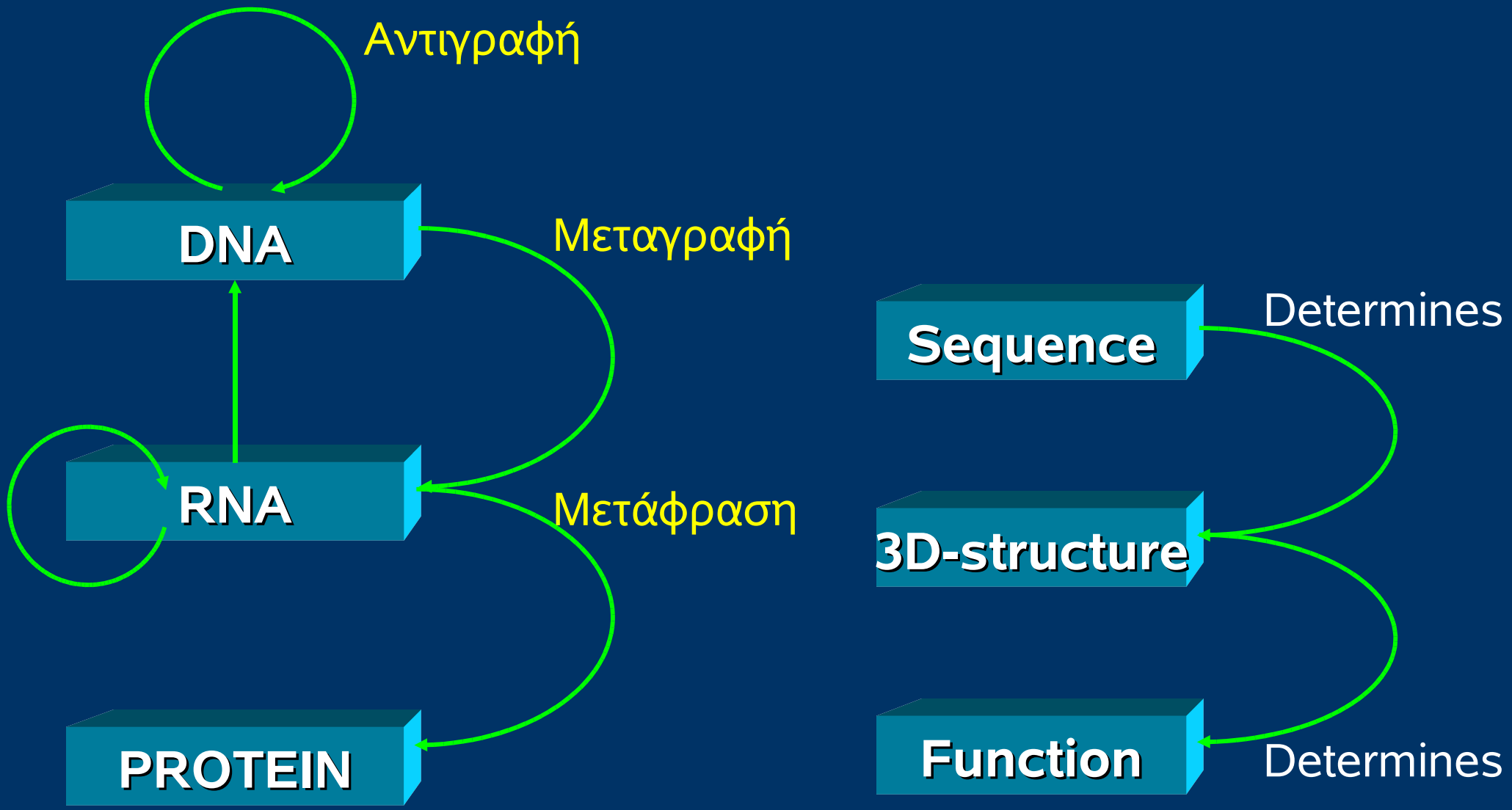


από Cormen *et. al*

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

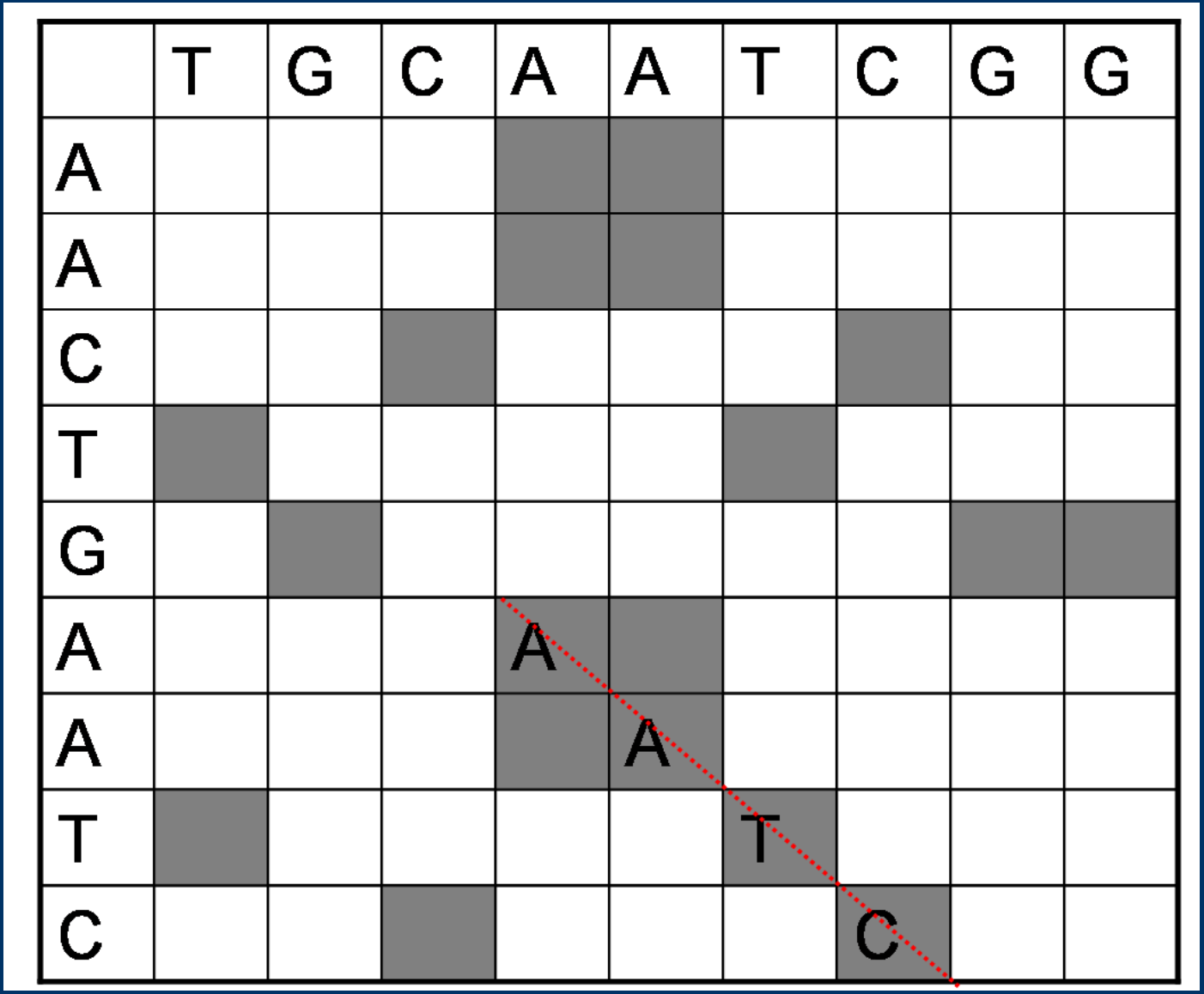
```
S1  HFCGGSLINEQWVVSAGHC

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

Ολική Στοίχιση

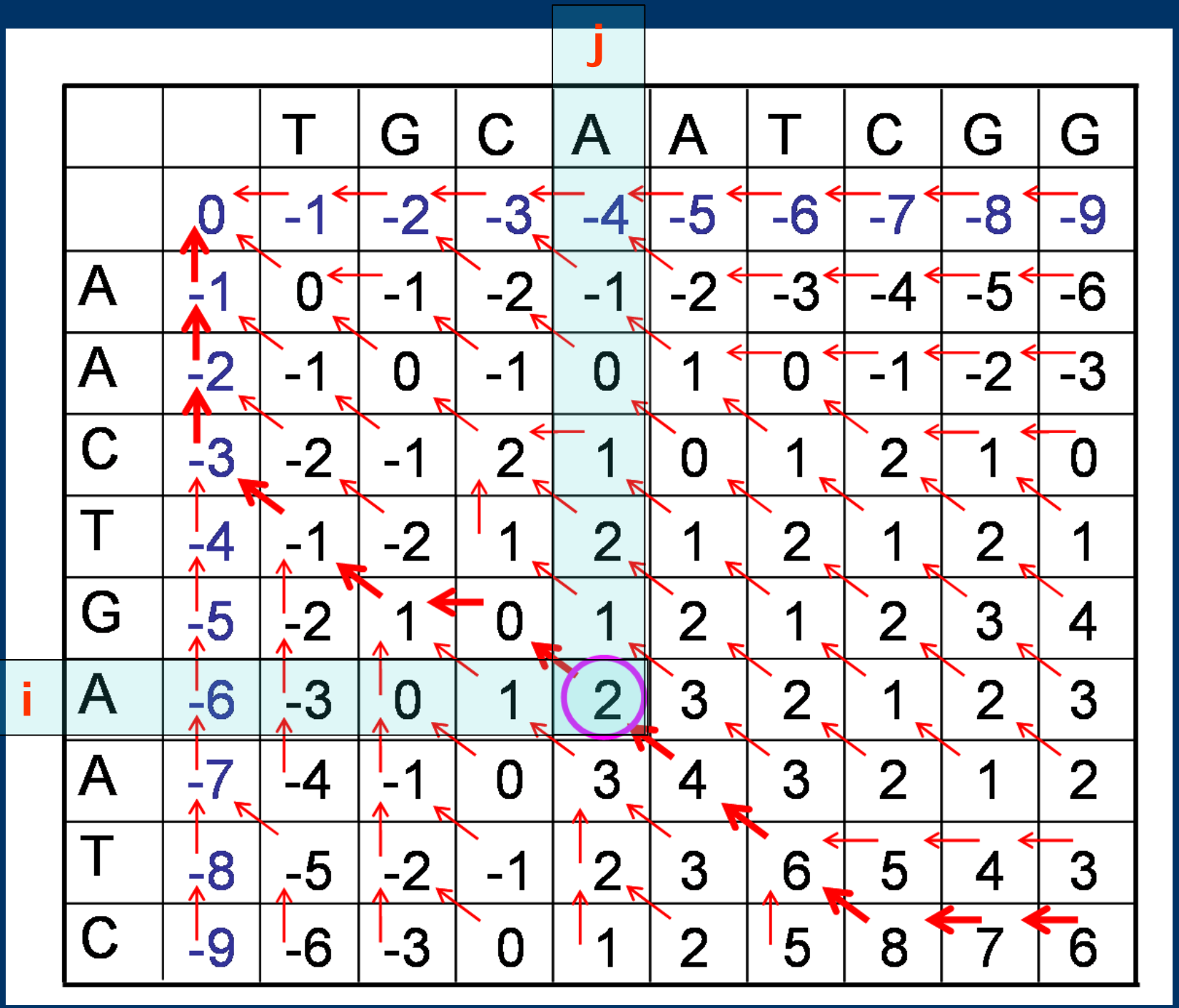
J. Mol. Biol. (1970) 48, 443–453

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

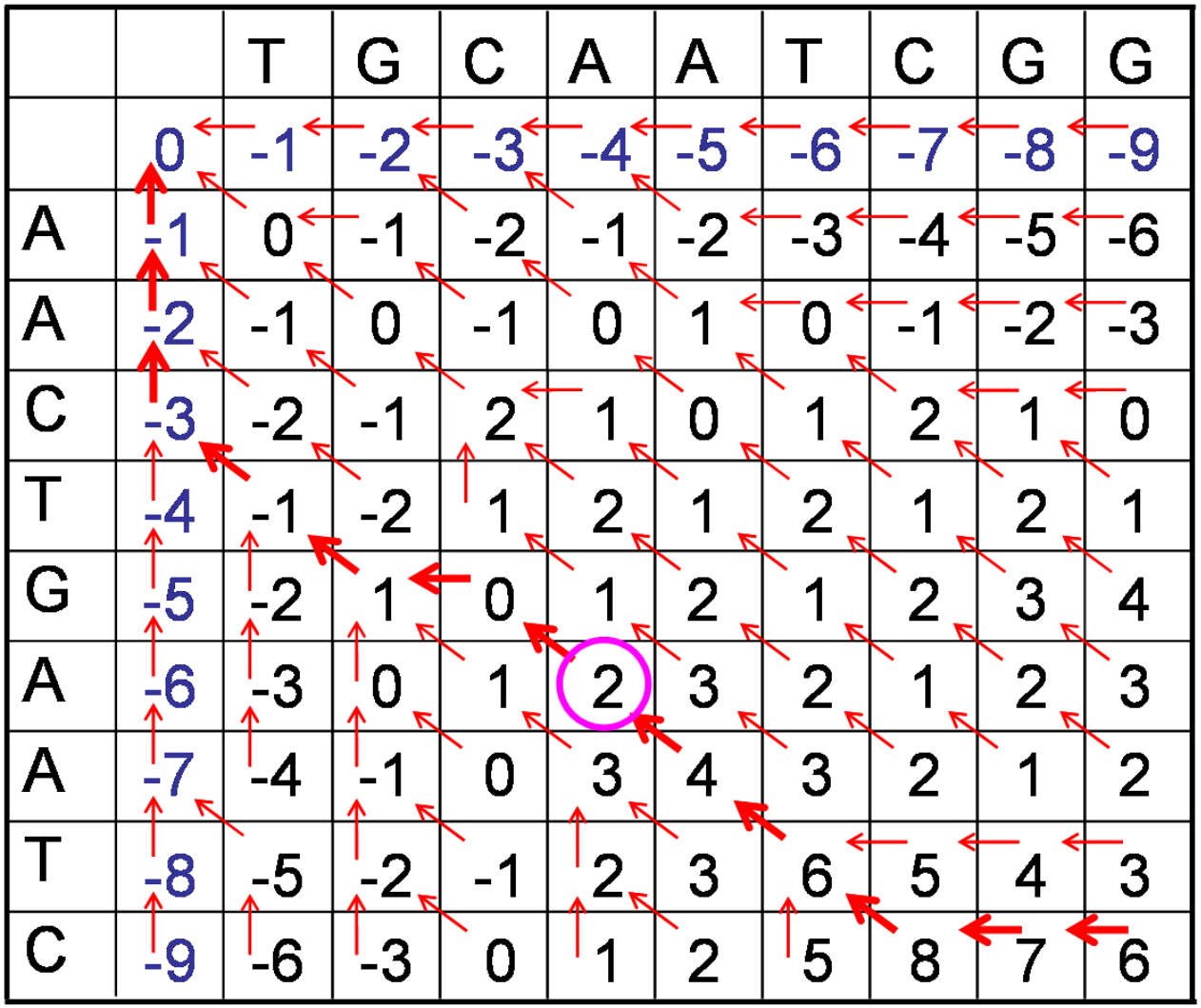
Saul B. Needleman and Christian D. Wunsch

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 -

$$F_{i,j} = \max \begin{cases} F_{i-1,j-1} + S_{x_i,y_j} \\ F_{i-1,j} - g \\ F_{i,j-1} - g \end{cases}$$


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

Τοπική Στοίχιση

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

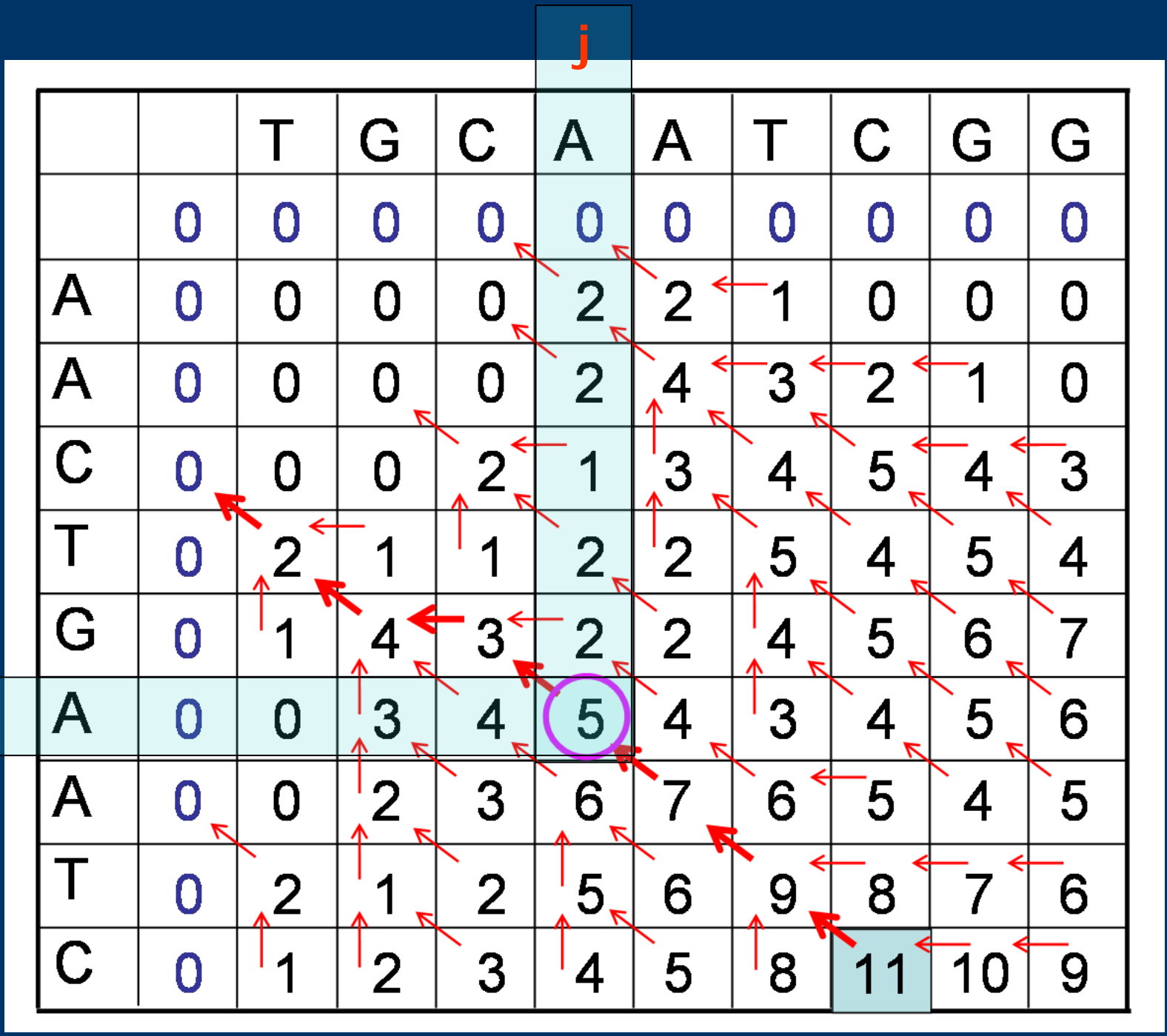
T. F. Smith

M. S. Waterman

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

- ...