# Αλγόριθμοι Εύρεσης Ομοιοτήτων Ακολουθιών
## Μέρος ΙΙ: Ευριστικές μέθοδοι αναζήτησης σε βάσεις δεδομένων


**Βασίλης Προμπονάς, PhD**
**Ερευνητικό Εργαστήριο Βιοπληροφορικής**

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

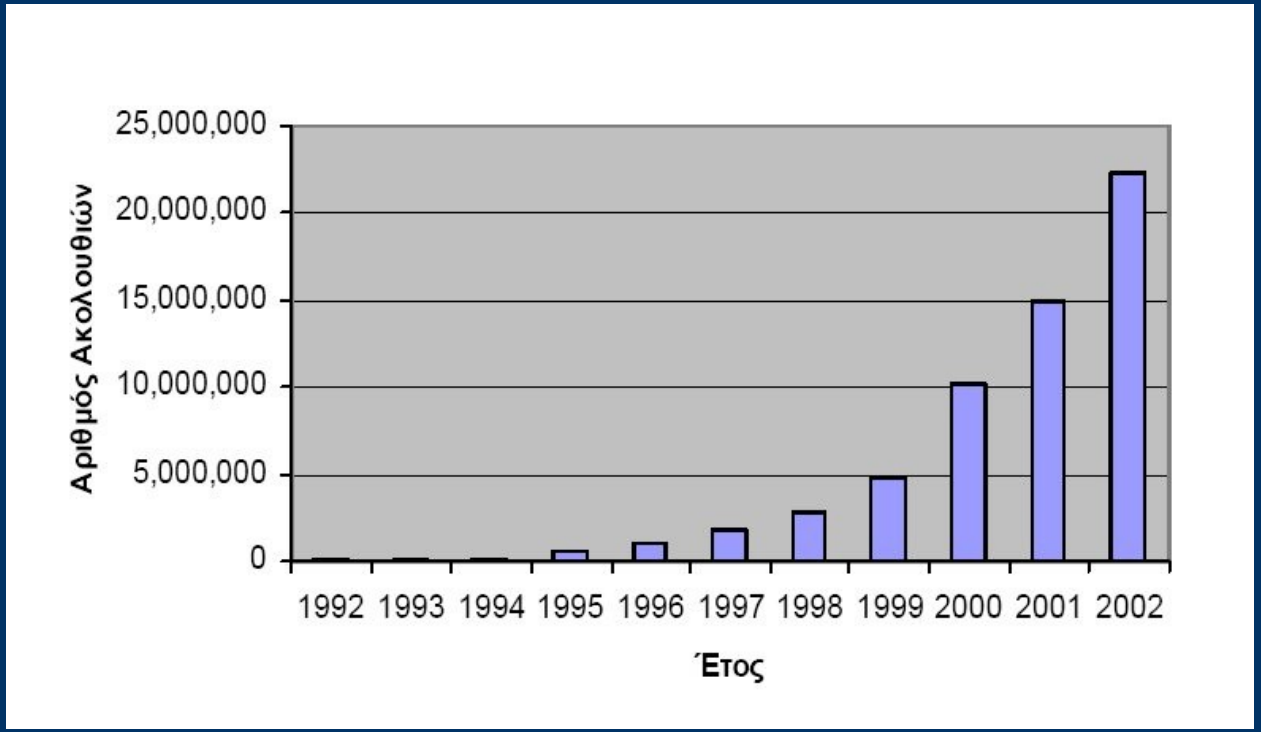
τηλ: 00357-22892879 (εσωτ. 2879)
email: vprobon@ucy.ac.cy, vprobon@biol.uoa.gr

# *Σύνοψη*

- Πού είναι το πρόβλημα??
  - Ρυθμός καταχώρησης δεδομένων ακολουθιών
  - Αποδοτικότητα αλγορίθμων ΔΠ
- Πως θα "κόψουμε δρόμο"??
  - Κόψε τις γωνιές!
  - γρήγορΑ!!
  - **ΕΚΡΗΞΗ**!!!
- Συζήτηση
  - ...

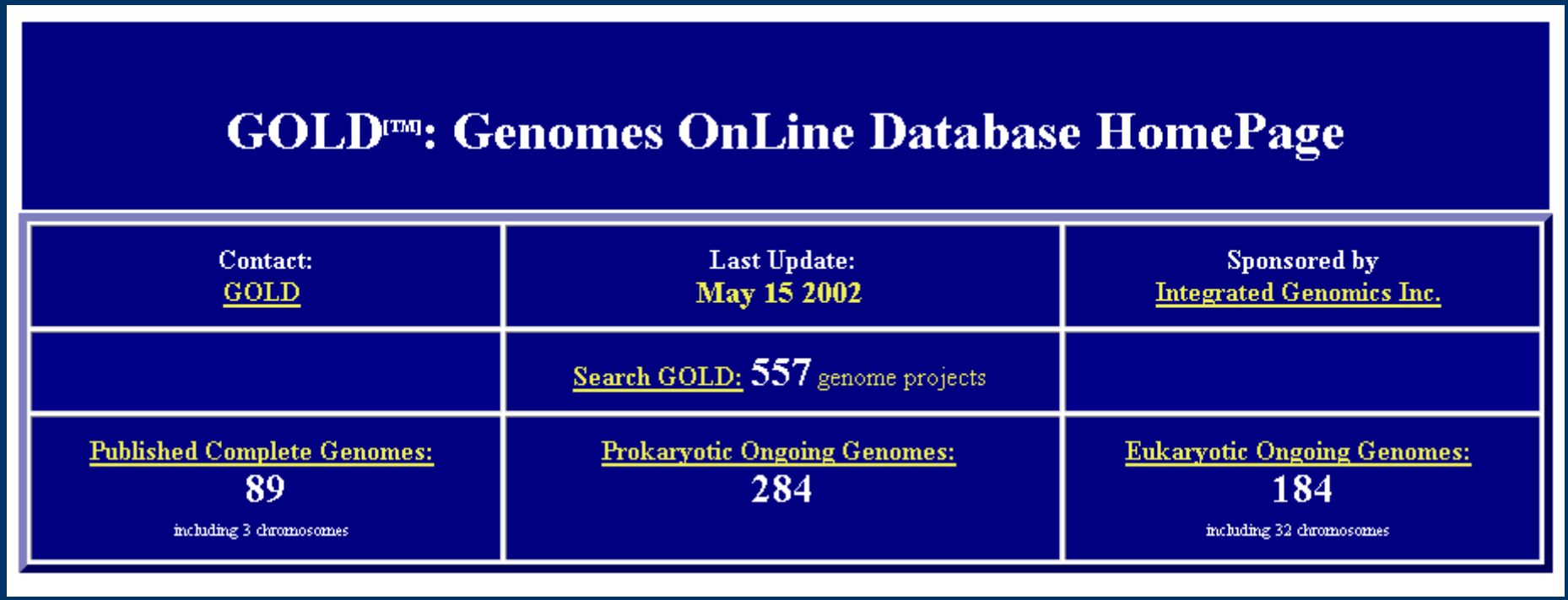
# Πού είναι το πρόβλημα??

- Υπάρχει πρόβλημα?
  - Διαρκής συσσώρευση νέων δεδομένων ακολουθιών
  - Ακολουθεί το νόμο του Moore



# *Γιατί όμως??*

**Μάϊος 2002**



GOLD™: Genomes OnLine Database HomePage

| Contact: GOLD | Last Update: May 15 2002 | Sponsored by Integrated Genomics Inc. |
|---|---|---|
| | Search GOLD: 557 genome projects | |
| Published Complete Genomes: 89 including 3 chromosomes | Prokaryotic Ongoing Genomes: 284 | Eukaryotic Ongoing Genomes: 184 including 32 chromosomes |

**ΣΗΜΕΡΑ**

**http://www.genomesonline.org/**

# *... μα και τι έγινε ...?*

- Έχουμε σχετικά αποδοτικούς αλγορίθμους ΔΠ
  - ***Εγγυημένη*** βέλτιστη στοίχιση
  - $2^{2N} => N^2$ [O(MN)]
  - Ναι, αλλά όταν το N μεγαλώνει πολύ γρήγορα ...

# Ευριστικές Λύσεις
## *(κοινώς: Μαφίες, Κομπίνες ή Παγαποντιές ...)*

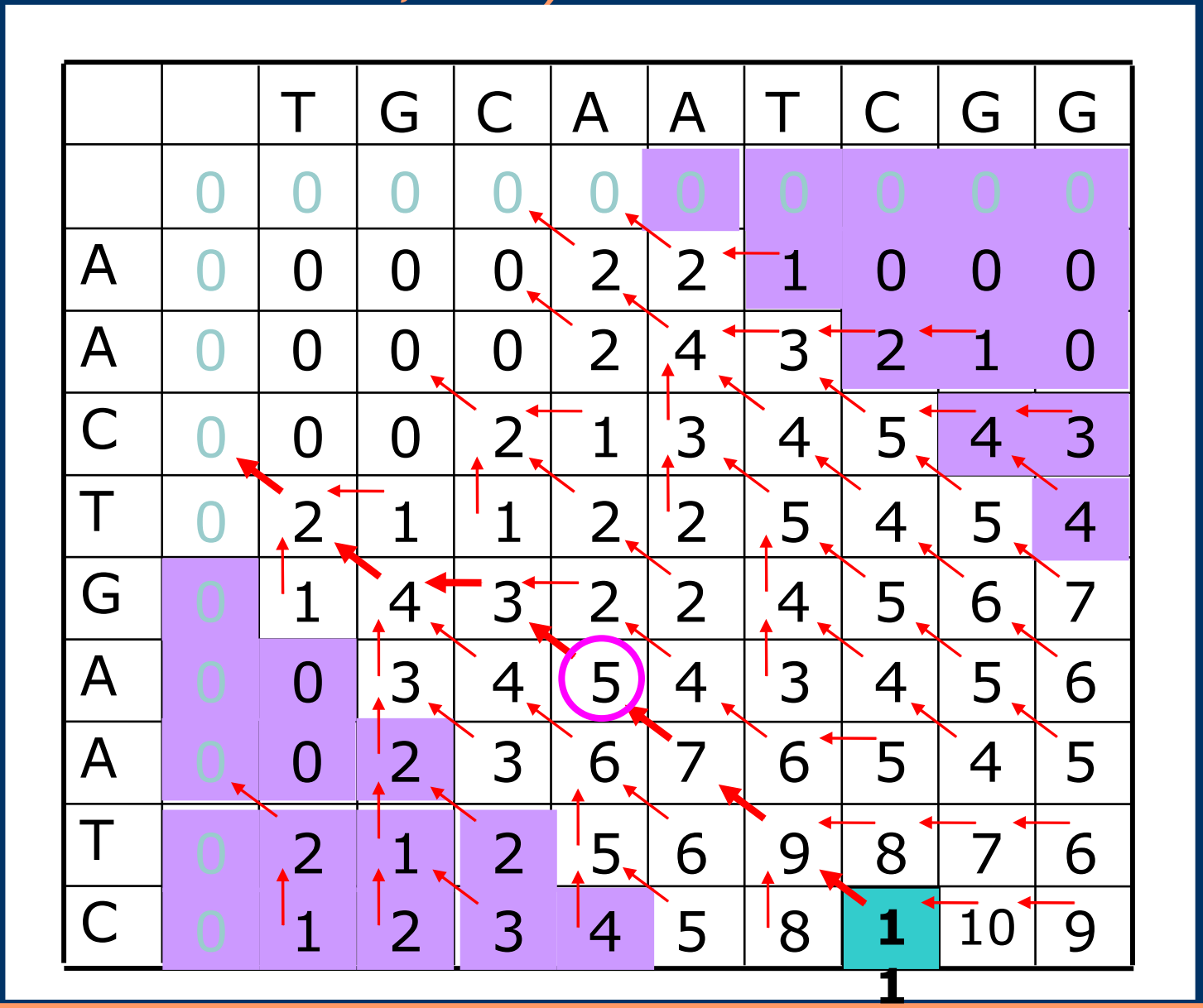
- Μείωση του χρόνου εκτέλεσης
  - Πώς???
- Εύρεση παραδεκτών (όχι εγγυημένα βέλτιστων) λύσεων
  - Πόσο καλές?
- Ισορροπία μεταξύ ***ΑΠΟΔΟΣΗΣ*** (χρόνος) και ***ΕΥΑΙΣΘΗΣΙΑΣ*** (ποιότητα)

# Ευριστικές Λύσεις (1.Κοπή Γονιών)
*(Kruskal and Sankoff, 1983)*

- Μείωση των υπολογισμών στον Πίνακα ΔΠ
- Παραδοχή
  - "Η διαδρομή που αντιστοιχεί στη βέλτιστη στοίχιση βρίσκεται κοντά στην κύρια διαγώνιο του πίνακα ΔΠ"
  - Η παραδοχή αυτή αποκλείει στοιχίσεις (Ποιές??)
  - Πότε ισχύει αυτό???
- Με λίγα λόγια: **Κόβουμε τις γωνίες**

# Ευριστικές Λύσεις (1.Κοπή Γονιών)

*(Kruskal and Sankoff, 1983)*



| | | 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 | **1** | 10 | 9 |
| | | | | | | | | **1** | | |

# Ευριστικές Λύσεις (1.Κοπή Γονιών)
## (Kruskal and Sankoff, 1983): Η περίπτωση m=n

$$\forall a_i, b_j \tau\eta\varsigma \sigma\tau o\acute{\iota}\chi\iota\sigma\eta\varsigma |i-j| \leqslant K, K \in \mathbb{N}, 0 \leqslant K \leqslant n$$

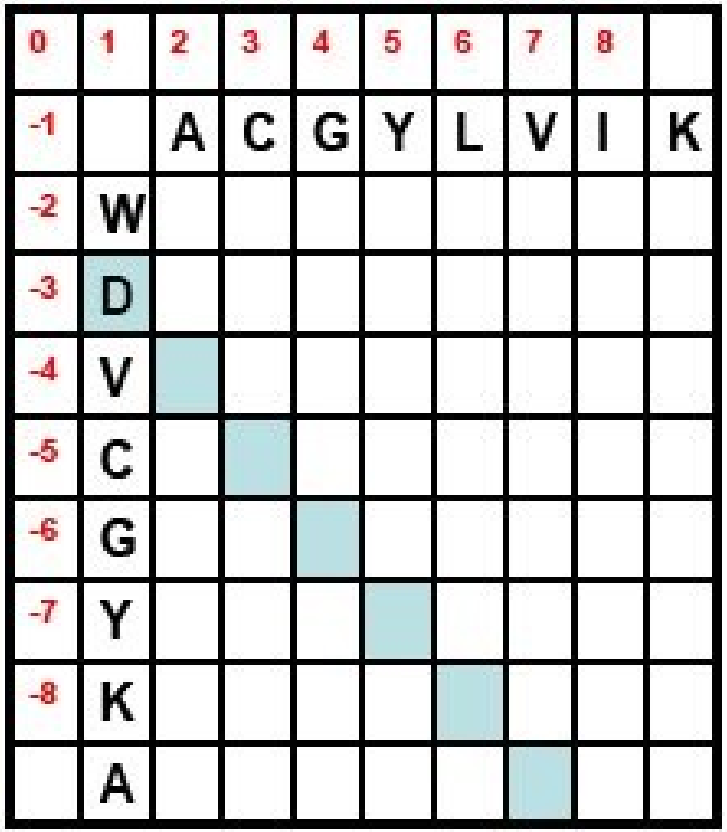
- Τί συμβαίνει όταν K=0, K=n ??
- Ποιό είναι το κέρδος για μια τυχαία τιμή του K?
- Πώς γενικεύω όταν m≠n?
- Μειονεκτήματα ...

# Ευριστικές Λύσεις (2.FASTA)
## (Pearson and Lipman, 1988; Pearson, 1990)

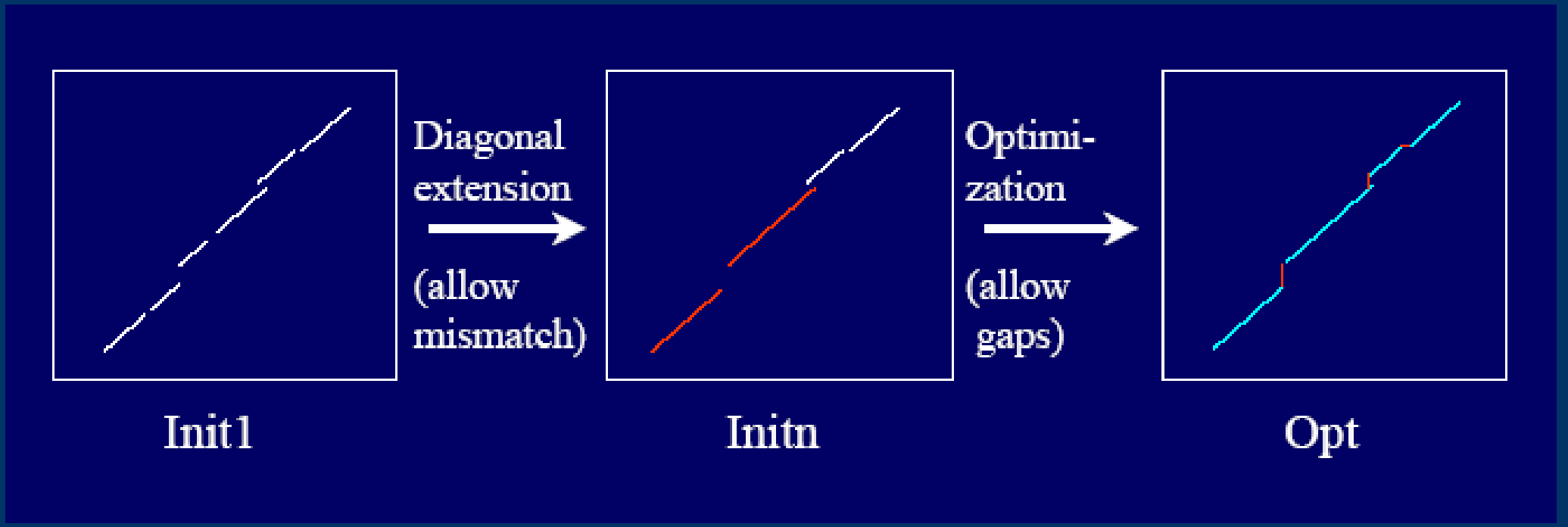
- ***Ευρετήρια*** (hash-tables) και ταυτόσημες ***k-πλέτες*** (k-tuples)
- Διαδικασία 4 βημάτων
  - Διαγώνιες Ταυτότητας
  - Top-10 rescoring (init1)
  - Ενοποίηση "γειτονικών" περιοχών (initn)
  - Band-alignment με πλήρη ΔΠ (opt)

# Ευριστικές Λύσεις (2.FASTA)
## (Pearson and Lipman, 1988; Pearson, 1990)

| K-tuple (K=1) | position in SEQ1 | SEQ2 | offset SEQ1-SEQ2 |
|---|---|---|---|
| A | 1 | 8 | -7 |
| C | 2 | 4 | -2 |
| D | - | 2 | X |
| G | 3 | 5 | -2 |
| I | 7 | - | X |
| K | 8 | 7 | 1 |
| L | 5 | - | X |
| V | 3 | 6 | -3 |
| W | - | 1 | X |
| Y | 4 | 6 | -2 |



# Ευριστικές Λύσεις (2.FASTA)
*(Pearson and Lipman, 1988; Pearson, 1990)*



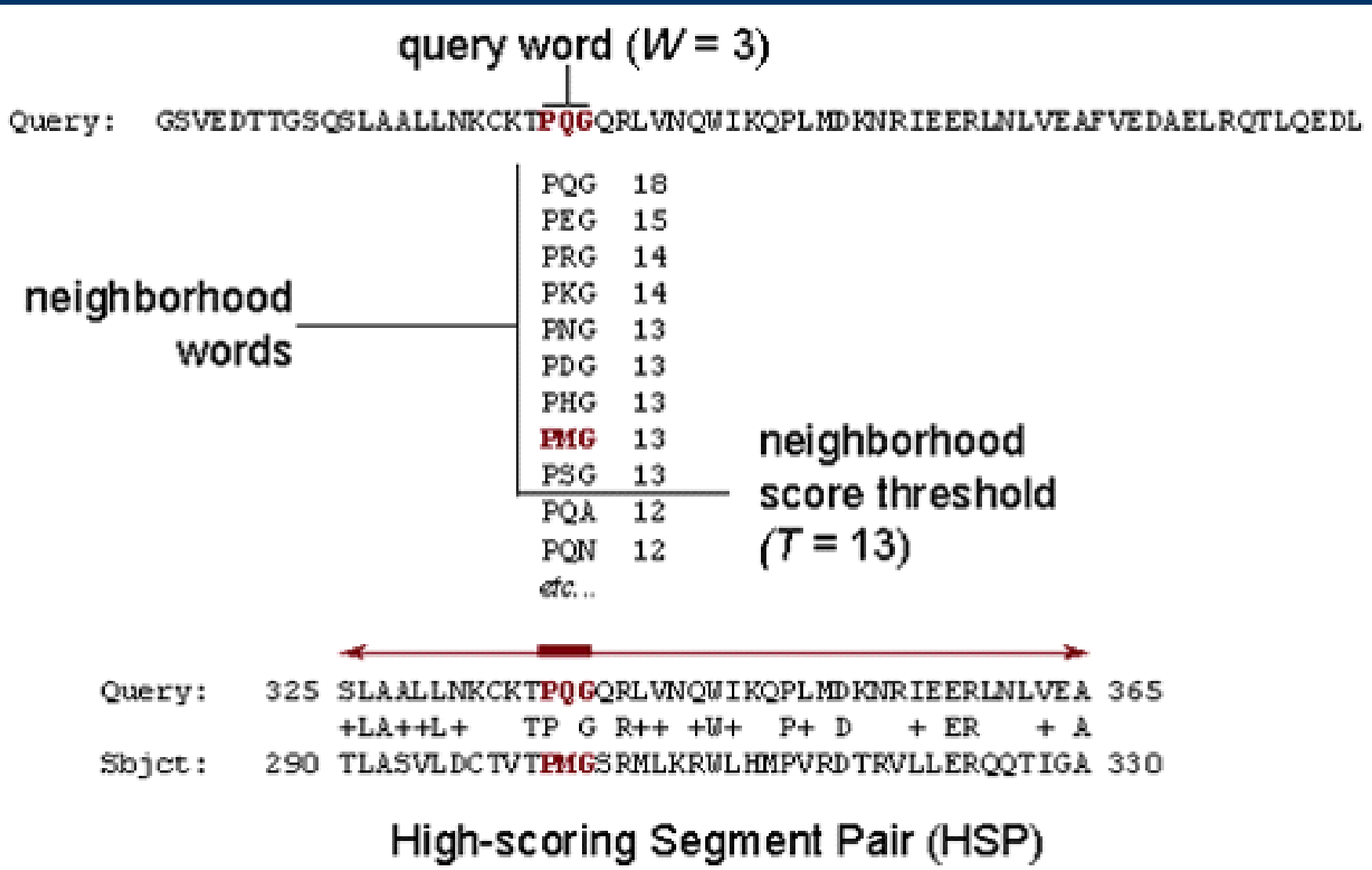
# Ευριστικές Λύσεις (3.BLAST)
*(Altschul, et al, 1990; Altschul, et al, 1997)*

- **B**asic **L**ocal **A**lignment **S**earch **T**ool
- Κατάτμηση των Ακολουθιών σε ***λέξεις*** (‘words’) και κατασκευή ευρετηρίων με λέξεις ‘υψηλής συνάφειας’

- Αναζήτηση ζευγών τμημάτων με υψηλό score (High-scoring Segment Pairs)
  - Τμήματα που στοιχίζονται (με ή χωρίς κενά ??)
  - Έχουν τοπικά μέγιστο score (δεν είναι δυνατόν να βελτιωθεί με επέκταση ή ‘κόψιμο’)
  - score > S (τιμή κατωφλίου)

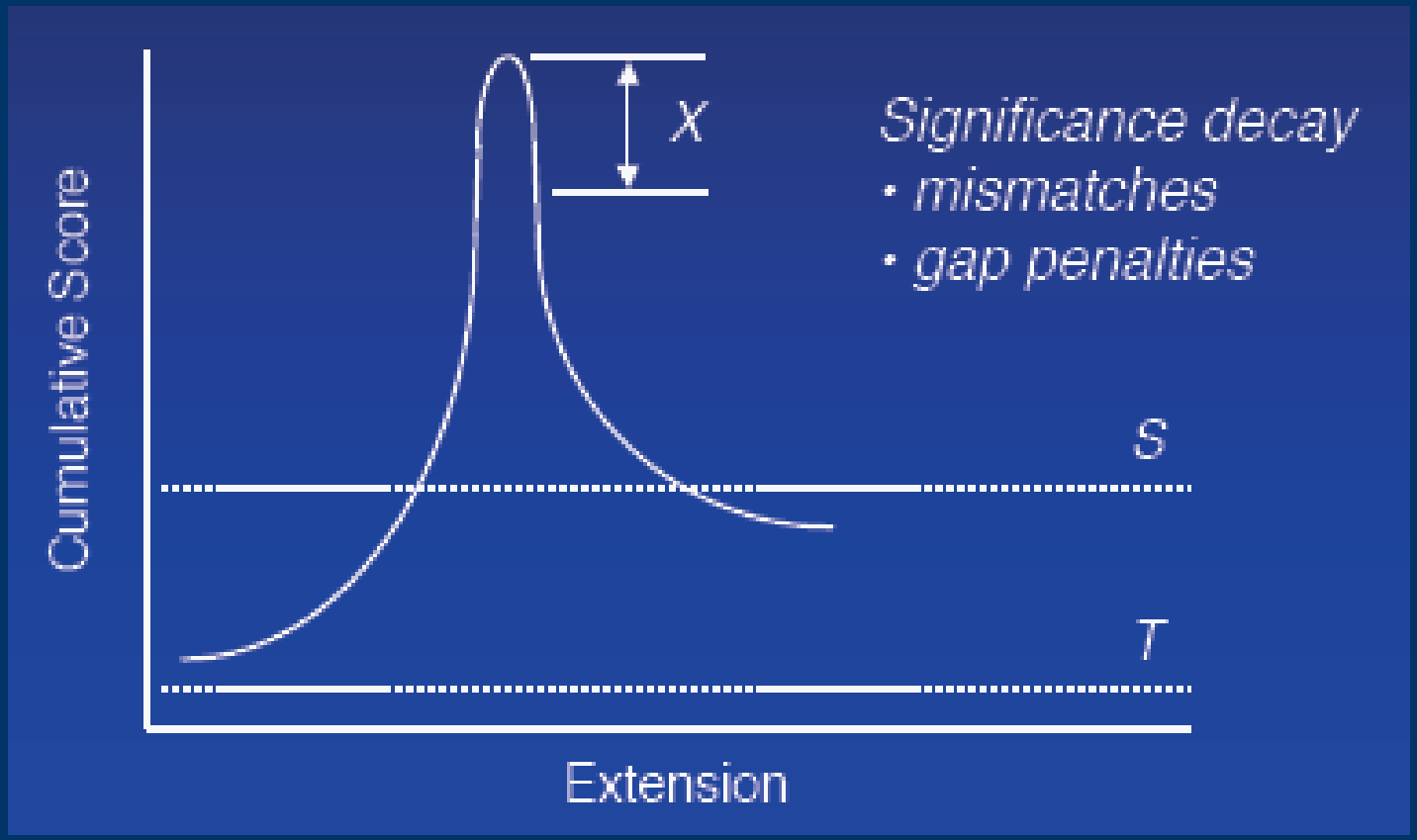
- Υλοποίηση που βασίζεται σε «Μηχανές Πεπερασμένων Καταστάσεων»

# Ευριστικές Λύσεις (3.BLAST)
## (Altschul, et al, 1990; Altschul, et al, 1997)



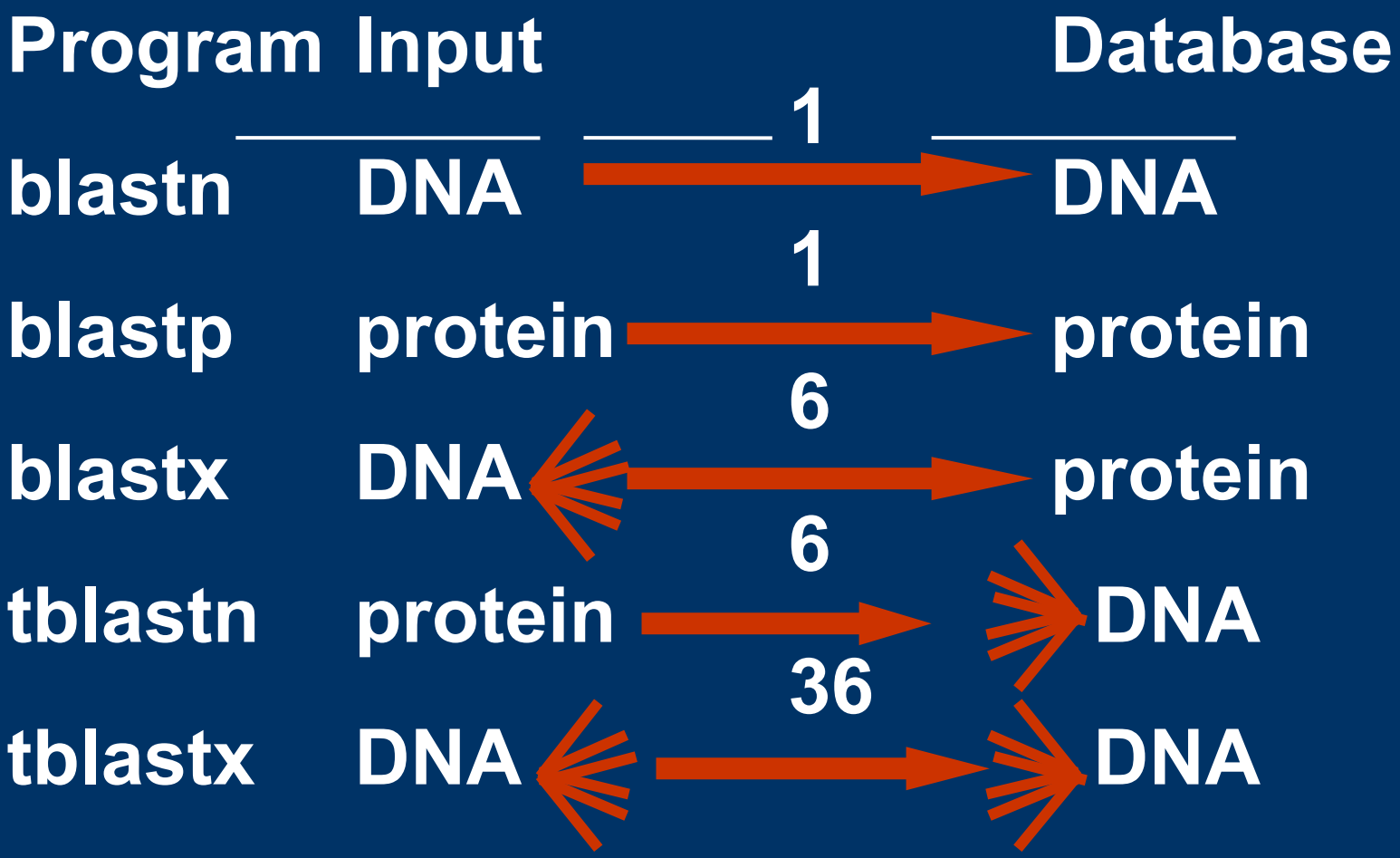
# Ευριστικές Λύσεις (3.BLAST)
## (Altschul, et al, 1990; Altschul, et al, 1997)

```
Query:   325  SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA  365
              +LA++L     TP+G R++ +W+ +P+ D    + ER    + A
Sbjct:   290  TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA  330
```



# Ευριστικές Λύσεις (3.BLAST)

*(Altschul, et al, 1990; Altschul, et al, 1997)*




From: Jonathan Pevsner, 2003

# Ευριστικές Λύσεις (3.BLAST)
*(Altschul, et al, 1990; Altschul, et al, 1997)*

- Μήκος λέξης W – Ταχύτητα αναζήτησης
- Ταχύτητα αναζήτησης – Ευαισθησία
- Παράμετροι, T, S (??)

- ... και ένα μικρό **live DEMO**
  **http://www.ncbi.nlm.nih.gov/BLAST/**

# *Συζήτηση*

- ...