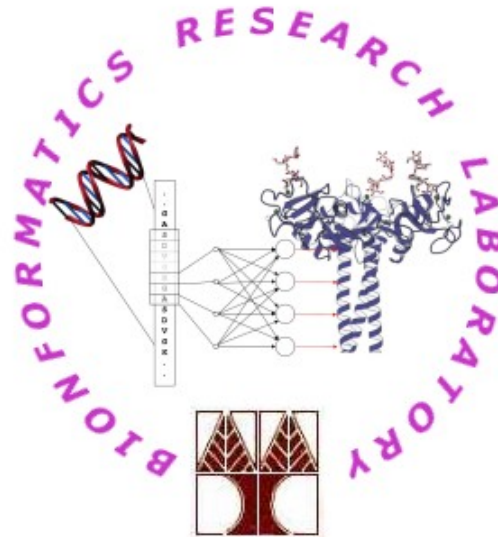


# Αλγόριθμοι Εύρεσης Ομοιοτήτων Ακολουθιών

## Μέρος II:

### Ευριστικές μέθοδοι αναζήτησης σε βάσεις δεδομένων



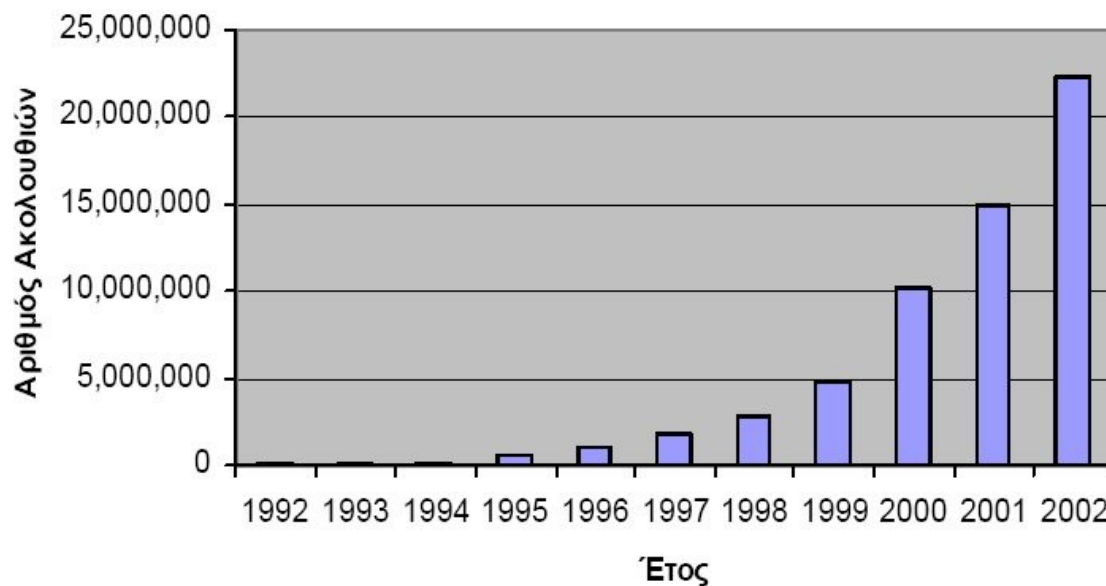
Vasilis Promponas  
Bioinformatics Research Laboratory  
Department of Biological Sciences  
University of Cyprus

# ΣΥΝΟΨΗ

- 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<sup>™</sup>: Genomes OnLine Database HomePage**

Contact: <u>GOLD</u>	Last Update: <b>May 15 2002</b>	Sponsored by <u>Integrated Genomics Inc.</u>
	<u>Search GOLD:</u> <b>557</b> genome projects	
<u>Published Complete Genomes:</u> <b>89</b> <small>including 3 chromosomes</small>	<u>Prokaryotic Ongoing Genomes:</u> <b>284</b>	<u>Eukaryotic Ongoing Genomes:</u> <b>184</b> <small>including 32 chromosomes</small>

**ΣΗΜΕΡΑ**

<http://www.genomesonline.org/>

## ... μα και τι έγινε ...?

- Έχουμε σχετικά αποδοτικούς αλγορίθμους ΔΠ
  - Εγγυημένη βέλτιστη στοίχιση
  - $2^{2N} \Rightarrow N^2$  [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 CPU/μνήμη) και ΕΥΑΙΣΘΗΣΙΑΣ (ποιότητα)

# Ευριστικές Λύσεις (1. Κοπή Γωνιών)

(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	<b>11</b>	10	9



# Ευριστικές Λύσεις (1. Κοπή Γονιών)

(Kruskal and Sankoff, 1983): Η περίπτωση  $m=n$

$$\forall a_i, b_j \text{ της στοίχισης } |i-j| \leq K, K \in \mathbb{N}, 0 \leq K \leq n$$

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

Να το δουλέψετε  
στο σπίτι,  
θα το συζητήσουμε  
στην επόμενη διάλεξη

# 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)

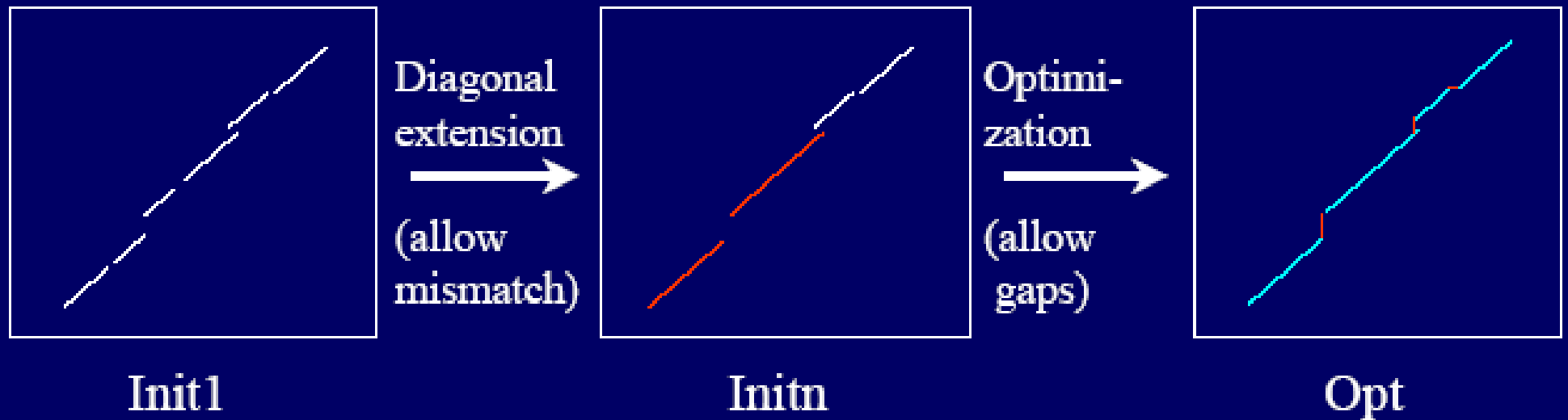
(Pearson and Lipman, 1988; Pearson, 1990)

K-tuple (K=1)	position in		offset
	SEQ1	SEQ2	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

0	1	2	3	4	5	6	7	8	
-1		A	C	G	Y	L	V	I	K
-2	W								
-3	D								
-4	V								
-5	C								
-6	G								
-7	Y								
-8	K								
	A								

# Ευριστικές Λύσεις (2.FASTA)

(Pearson and Lipman, 1988; Pearson, 1990)



# Ευριστικές Λύσεις (3.BLAST)

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

- Basic Local Alignment Search Tool
  - 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)

query word ( $w = 3$ )

Query: GSVEDTTGSQSLAALLNKCKT**PQG**QRLVNQWIKQPLMDKNRIEERLNLVEAFVEDAELRQTLQEDL

neighborhood  
words

- PQG 18
- PEG 15
- PRG 14
- PKG 14
- PNG 13
- PDG 13
- PHG 13
- PMG** 13
- PSG 13
- PQA 12
- PQN 12
- etc...

neighborhood  
score threshold  
( $T = 13$ )

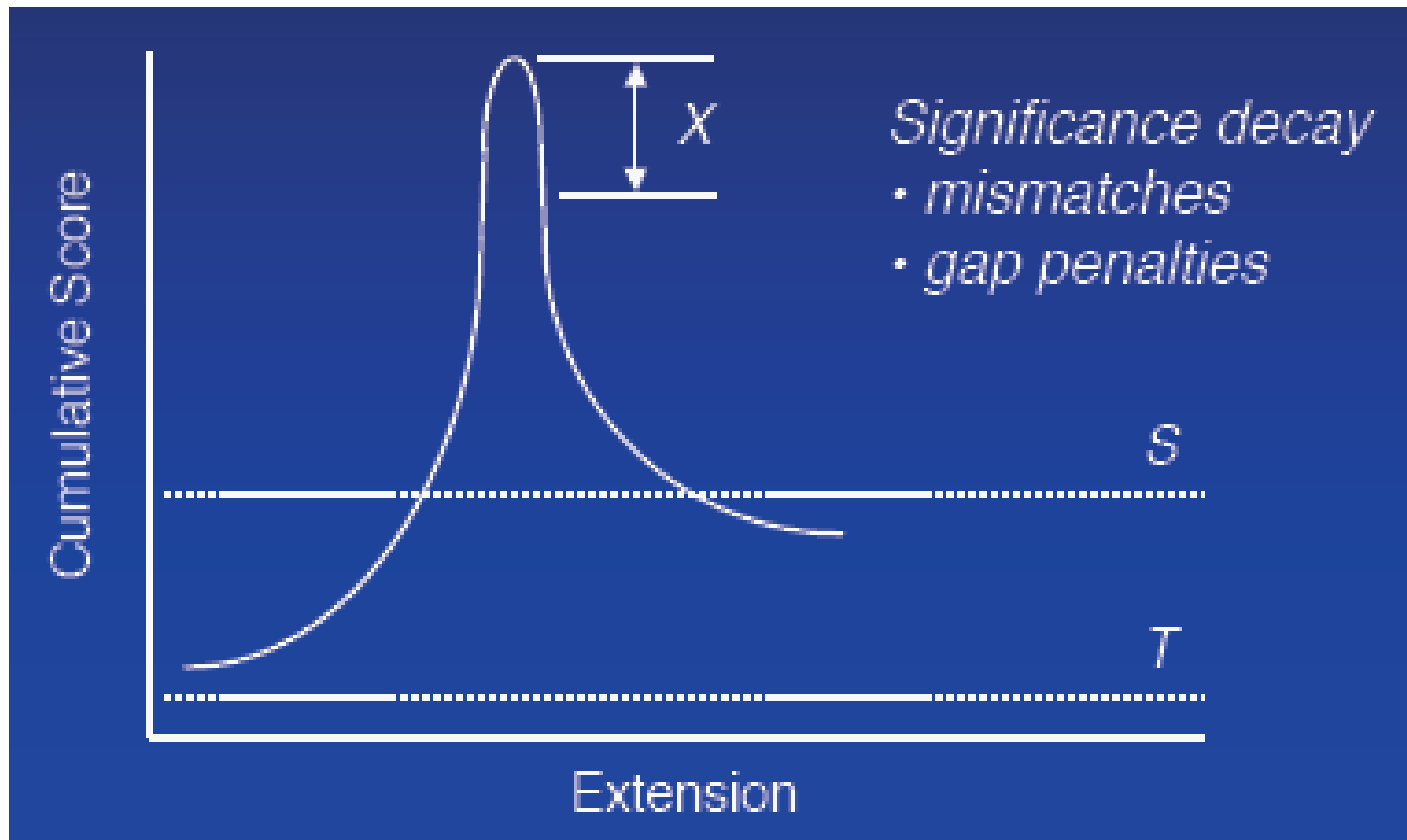
Query: 325 SLAALLNKCKT**PQG**QRLVNQWIKQPLMDKNRIEERLNLVEA 365  
 +LA++L+ TP G R++ +W+ P+ D + ER + A  
 Sbjct: 290 TLASVLDCTV**PMG**SRLKRWLHMPVRDTRVLLERQQTIGA 330

High-scoring Segment Pair (HSP)

# Ευριστικές Λύσεις (3.BLAST)

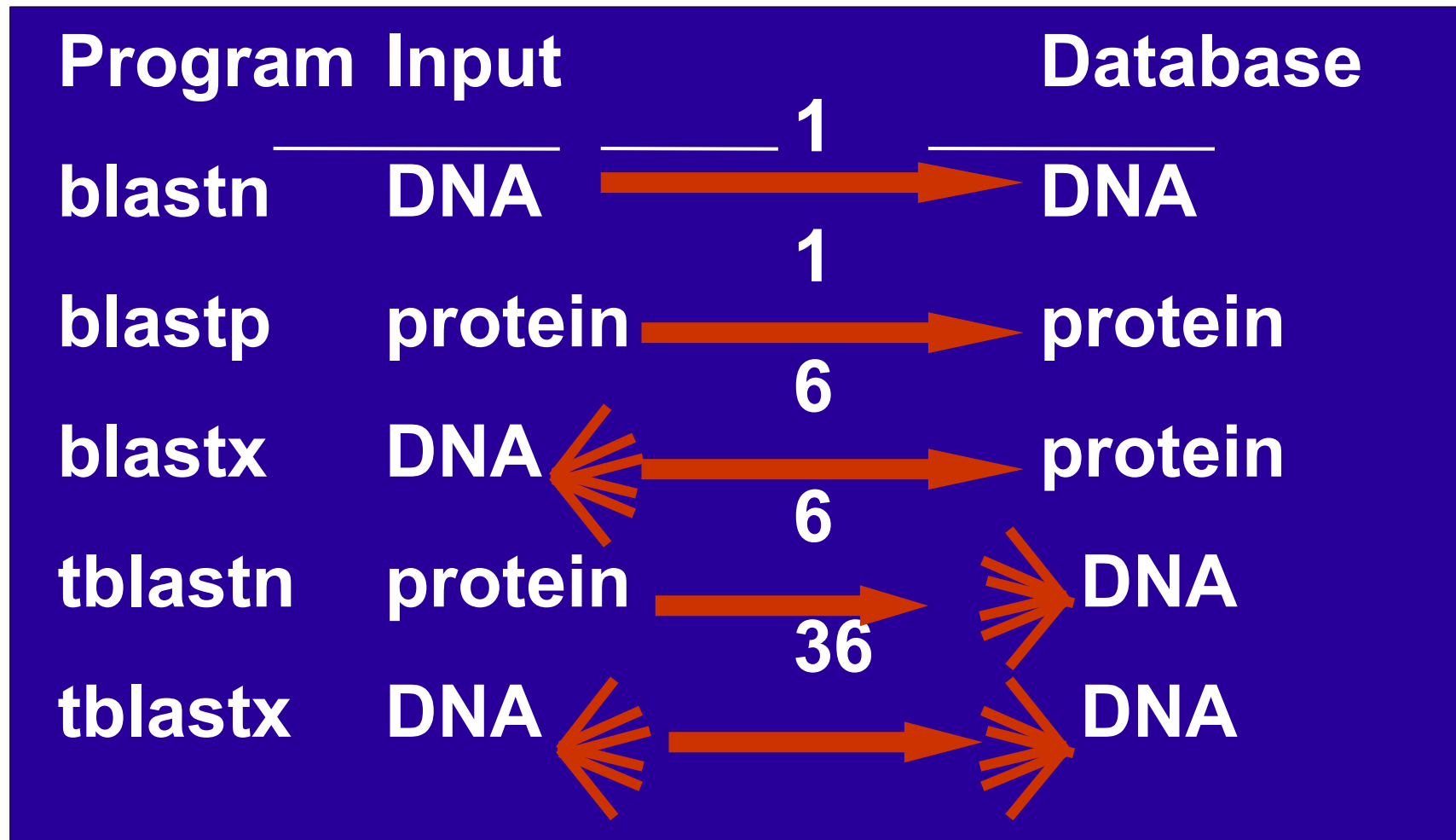
(Altschul, et al, 1990; Altschul, et al, 1997)

		←	▬	→	
Query:	325	SLAALLNKCKT	PQGQRLVNQWIKQPLMDKNRIEERLNLVEA	365	
		+LA++L	TP+G R++ +W+ +P+ D + ER + A		
sbjct:	290	TLASVLDCTVT	PMGSRMLKRWLHMPVRDTRVLLERQQTIGA	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 ...

## Διαδικτυακές πηγές

**GOLD** <http://genomesonline.org>

**NCBI BLAST server** <http://www.ncbi.nlm.nih.gov/BLAST/>

**Επιπλέον διδακτικό υλικό:**

<http://troodos.biol.ucy.ac.cy/>