



Εισαγωγή στη Βιοπληροφορική

Μερικές διαφάνειες από διαφάνειες βιβλίου Neil C. Jones, Pavel A. Pevzner, Εισαγωγή στους Αλγορίθμους Βιοπληροφορικής

Και από το βιβλίο: Dan Gusfield Algorithms on Strings, Trees and Sequences, Cambridge University Press,

και από: Μπάγκος Παντελεήμων, Βιοπληροφορική, Έκδοση: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών

Γιατί χρειάζεται η βιοπληροφορική;

- Η βιοπληροφορική είναι ο συνδυασμός της βιολογίας και της πληροφορικής.
- Οι τεχνολογίες προσδιορισμού αλληλουχίας του DNA έχουν δημιουργήσει τεράστιες ποσότητες πληροφορίας, οι οποίες μπορούν να αναλυθούν αποτελεσματικά μόνο με υπολογιστές.
- Μέχρι τώρα, έχει προσδιοριστεί η αλληλουχία για >100000 είδη
 - Ο άνθρωπος, ο αρουραίος, ο χιμπατζής, η κότα, και πολλά άλλα.
- Καθώς αυξάνεται υπερβολικά ο όγκος και η πολυπλοκότητα των πληροφοριών, χρειάζονται περισσότερα υπολογιστικά εργαλεία για την ταξινόμηση των δεδομένων.
 - Η βιοπληροφορική έρχεται να σώσει την κατάσταση!!!

Τι είναι η βιοπληροφορική;

- Η βιοπληροφορική ορίζεται γενικά ως η ανάλυση, πρόβλεψη, και μοντελοποίηση των βιολογικών δεδομένων με τη βοήθεια των υπολογιστών



Βιο-πληροφορίες

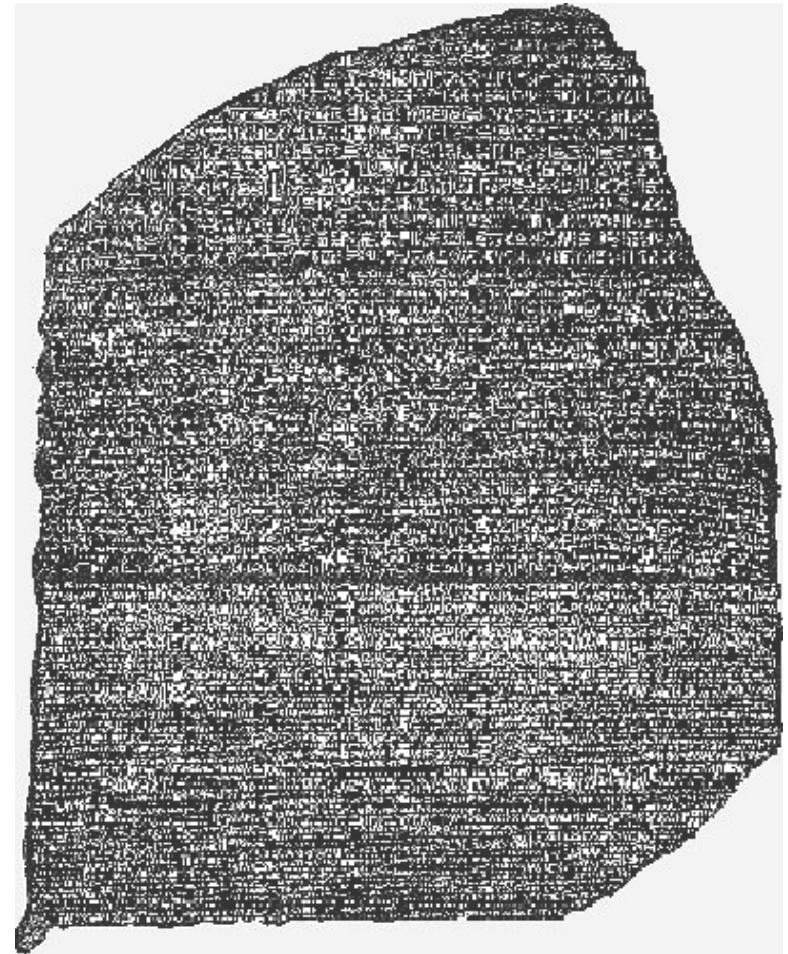
- Από τη στιγμή που ανακαλύψαμε τον τρόπο που το DNA λειτουργεί ως «εγχειρίδιο οδηγιών» της ζωής, η βιολογία έχει γίνει επιστήμη των πληροφοριών
- Τώρα που έχουμε προσδιορίσει την αλληλουχία πολλών διαφορετικών οργανισμών, είμαστε σε θέση να βρούμε το νόημα του DNA μέσω της *συγκριτικής γονιδιωματικής*, κατά τρόπο παρόμοιο με τη συγκριτική γλωσσολογία.
- Σιγά-σιγά, μαθαίνουμε το «συντακτικό» του DNA

Πληροφορίες αλληλουχιών

- Πολλές γραπτές γλώσσες αποτελούνται από διαδοχικά σύμβολα
- Ακριβώς όπως το ανθρώπινο κείμενο, οι γονιδιωματικές αλληλουχίες αναπαριστούν μια γλώσσα που γράφεται με τα σύμβολα A, T, C, G
- Πολλές τεχνικές αποκωδικοποίησης του DNA δεν διαφέρουν πολύ από τις αντίστοιχες τεχνικές αποκωδικοποίησης μιας αρχαίας γλώσσας

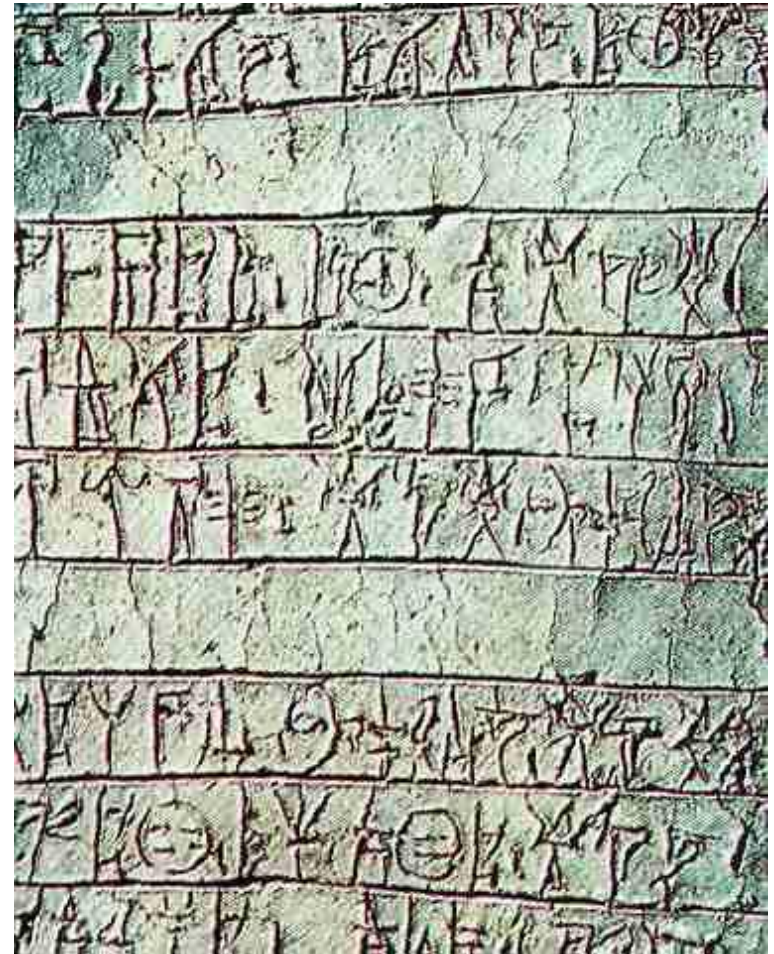
Η στήλη της Rosetta

- Με τη στήλη της Rosetta, οι γλωσσολόγοι μπόρεσαν να λύσουν τον κώδικα των αιγυπτιακών ιερογλυφικών
- Η εγχάρακτη επιγραφή στην ελληνική γλώσσα παρείχε ενδείξεις σχετικά με τη σημασία των ιερογλυφικών.
- Αυτό είναι ένα παράδειγμα της συγκριτικής γλωσσολογίας



Γραμμική Β

- Στις αρχές του εικοστού αιώνα, οι αρχαιολόγοι ανακάλυψαν πήλινες πινακίδες στο νησί της Κρήτης
- Η άγνωστη γλώσσα ονομάστηκε «Γραμμική Β»
- Η γραφή θεωρήθηκε ότι ανήκε σε μια αρχαία μινωική γλώσσα, και αποτέλεσε μυστήριο για 50 χρόνια



Τι είναι η Συγκριτική Γλωσσολογία;

- Η επιστήμη που μελετά τις σχέσεις μεταξύ διαφορετικών γλωσσών.
- Εξετάζει φωνητικές, μορφολογικές, συντακτικές και λεξιλογικές ομοιότητες.
- Χρησιμοποιείται για την ανασύνθεση πρωτογλωσσών και την κατανόηση της γλωσσικής εξέλιξης.

Η Στήλη της Ροζέττας και η Συγκριτική Γλωσσολογία

- Περιείχε το ίδιο κείμενο σε τρεις γλώσσες: Ιερογλυφικά, Δημοτική και Αρχαία Ελληνικά.
- Ο Σαμπολιόν χρησιμοποίησε τη συγκριτική μέθοδο για να αποκρυπτογραφήσει τα ιερογλυφικά.
- Η σύγκριση των ελληνικών λέξεων με τα αιγυπτιακά σύμβολα οδήγησε στην ανακάλυψη του φωνητικού τους συστήματος.

Συγκριτική Γλωσσολογία & Γονιδιωματική: Ομοιότητες

- Ανάλυση ομοιοτήτων και διαφορών μεταξύ γλωσσών και DNA.
- Ανακατασκευή προγονικών μορφών (πρωτογλώσσες & αρχαίοι γενετικοί πρόγονοι).
- Χρήση δέντρων εξέλιξης (Γλωσσικά & Φυλογενετικά Δέντρα).

Κοινές Αρχές & Εξελικτικές Διαδικασίες

- Κοινή καταγωγή: Οι γλώσσες και τα είδη προέρχονται από έναν κοινό πρόγονο.
- Μεταβολές με τον χρόνο: Φωνητικές αλλαγές στις γλώσσες, μεταλλάξεις στο DNA.
- Οριζόντια μεταφορά: Δανεισμός λέξεων μεταξύ γλωσσών & οριζόντια μεταφορά γονιδίων.

Παράδειγμα Συσχέτισης Γλώσσας & DNA

- Μελέτες δείχνουν σχέση μεταξύ γενετικών δεδομένων και γλωσσικών ομάδων.
- Η γλώσσα των Βάσκων δεν έχει γνωστές συγγένειες, ενώ και το DNA τους δείχνει γενετική απομόνωση.
- Και οι δύο κλάδοι χρησιμοποιούν συγκριτικές μεθόδους για την αποκάλυψη της εξέλιξης.
- Η συγκριτική γλωσσολογία αναλύει τη γλωσσική κληρονομιά, ενώ η γονιδιωματική τη βιολογική.
- Ο συνδυασμός τους μας βοηθά να κατανοήσουμε καλύτερα την ιστορία της ανθρωπότητας.

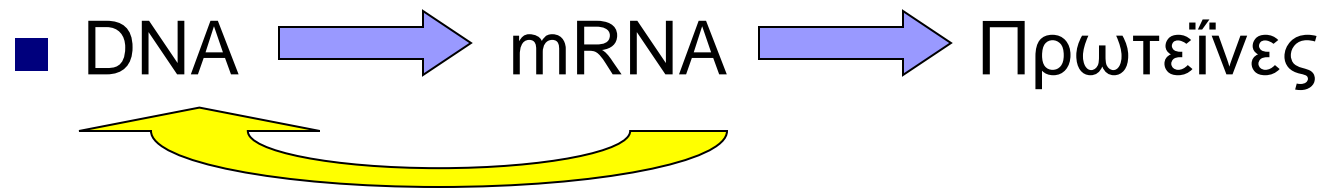
Γραμμική B

- Την ίδια εποχή που αποκρυπτογραφήθηκε η δομή του DNA, ο Michael Ventrís αποκωδικοποίησε τη Γραμμική B χρησιμοποιώντας ειδικές γνώσεις μαθηματικής κρυπτανάλυσης
- Πρόσεξε ότι μερικές λέξεις στη Γραμμική B χρησιμοποιούνταν αποκλειστικά για το νησί, και υπέθεσε ότι αποτελούν ονόματα πόλεων
- Με αυτή τη γνώση, μπόρεσε να αποκωδικοποιήσει τη γραφή, η οποία αποδείχθηκε ότι ήταν η ελληνική γλώσσα με διαφορετικό αλφάβητο

Αποκωδικοποίηση των αμινοξέων

- Νωρίτερα, ένα πείραμα στις αρχές της δεκαετίας του 1900 έδειξε ότι όλες οι πρωτεΐνες αποτελούνται από αλληλουχίες 20 αμινοξέων
- Εξαιτίας αυτού, κάποιοι υπέθεσαν ότι το «σχέδιο» της ζωής βρισκόταν στα πολυπεπτίδια

Κεντρικό δόγμα



- Το DNA στο χρωμόσωμα μεταγράφεται σε mRNA, το οποίο εξάγεται από τον πυρήνα στο κυτόπλασμα. Εκεί, μεταφράζεται σε πρωτεΐνη.
- Αργότερα, ανακαλύφθηκε ότι μπορούμε επίσης να μεταβούμε από το mRNA στο DNA (ρετροϊοί).
- Ακόμη, το mRNA μπορεί να υποβληθεί σε εναλλακτικό μάτισμα, και να προκύψουν έτσι διαφορετικά πρωτεϊνικά προϊόντα.

Από τη δομή στη λειτουργία

- Η οργανική χημεία μας δείχνει ότι η δομή των μορίων προσδιορίζει τις δυνατές αντιδράσεις τους.
- Μία μέθοδος μελέτης των πρωτεϊνών είναι να συμπεράνουμε τη λειτουργία τους με βάση τη δομή τους, ειδικά για τις ενεργές θέσεις.

Δύο σύντομες εφαρμογές της βιοπληροφορικής

- BLAST (Basic Local Alignment Search Tool, Βασικό εργαλείο αναζήτησης τοπικών στοιχίσεων)
- PROSITE (PROtein SITEs and patterns database, Βάση δεδομένων για μοτίβα και θέσεις πρωτεϊνών)



BLAST

- Ένα υπολογιστικό εργαλείο που μας επιτρέπει να συγκρίνουμε δεδομένες αλληλουχίες με τις καταχωρίσεις στις τρέχουσες βιολογικές βάσεις δεδομένων.
- Ένα εξαιρετικό εργαλείο για την πρόβλεψη των λειτουργιών μιας αλληλουχίας που βασίζεται σε ομοιότητες στοίχισης με γνωστά γονίδια.

Ο αρχικός ρόλος της βιοπληροφορικής

- Σύγκριση αλληλουχιών
- Αναζητήσεις σε βάσεις δεδομένων που περιέχουν αλληλουχίες

Σύγκριση βιολογικών αλληλουχιών

■ Needleman- Wunsch, 1970

- Αλγόριθμος δυναμικού προγραμματισμού για τη στοίχιση αλληλουχιών

	A	D	C	N	S	R	Q	C	L	C	R	P	M
A	8	7	6	6	5	4	4	3	3	2	1	0	0
S	7	7	6	6	6	4	4	3	3	2	1	0	0
C	6	6	7	6	5	4	4	4	3	3	1	0	0
S	6	6	6	5	6	4	4	3	3	2	1	0	0
N	5	5	5	6	5	4	4	3	3	2	1	0	0
R	4	4	4	4	4	5	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	4	3	3	1	0	0
K	3	3	3	3	3	3	3	3	3	2	1	0	0
C	2	2	3	2	2	2	2	3	2	3	1	0	0
R	2	1	1	1	1	2	1	1	1	1	2	0	0
D	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

Βιολογικές βάσεις δεδομένων

- Άπειρα βιολογικά δεδομένα και δεδομένα αλληλουχιών είναι διαθέσιμα δωρεάν μέσω των ηλεκτρονικών βάσεων δεδομένων
- Υπολογιστικοί αλγόριθμοι χρησιμοποιούνται για την αποδοτική αποθήκευση τεράστιων ποσοτήτων βιολογικών δεδομένων

Παραδείγματα

- **NCBI GeneBank** <http://ncbi.nih.gov>
Τεράστια συλλογή βάσεων δεδομένων, από τις οποίες ξεχωρίζει η βάση δεδομένων με αλληλουχίες νουκλεοτιδίων
- **Protein Data Bank** <http://www.pdb.org>
Βάση δεδομένων με τριτοταγείς δομές πρωτεϊνών
- **SWISSPROT** <http://www.expasy.org/sprot/>
Βάση δεδομένων με σχολιασμένες αλληλουχίες πρωτεϊνών
- **PROSITE** <http://kr.expasy.org/prosite>
Βάση δεδομένων με μοτίβα ενεργών θέσεων πρωτεϊνών

Βάση δεδομένων PROSITE

- Βάση δεδομένων με ενεργές θέσεις πρωτεϊνών.
- Ένα εξαιρετικό εργαλείο που προβλέπει την ύπαρξη ενεργών θέσεων σε μια άγνωστη πρωτεΐνη με βάση μια πρωτοταγή αλληλουχία.

Ανάλυση αλληλουχιών

- Μερικοί αλγόριθμοι αναλύουν βιολογικές αλληλουχίες για να βρουν μοτίβα
 - Θέσεις ματίσματος του RNA
 - ORFs (ανοιχτά πλαίσια ανάγνωσης)
 - Τάσεις των αμινοξέων σε μια πρωτεΐνη
 - Συντηρημένες περιοχές σε
 - αλληλουχίες AA [πιθανή ενεργή θέση]
 - DNA/RNA [πιθανή θέση πρόσδεσης πρωτεΐνης]
- Άλλοι κάνουν προβλέψεις με βάση την αλληλουχία
 - Αναδίπλωση δευτεροταγούς δομής πρωτεϊνών/RNA

Η αλληλουχία προσδιορίστηκε, τι ακολουθεί μετά;

- Η ανίχνευση φυλογενετικών σχέσεων
 - Βρίσκουμε οικογενειακές σχέσεις μεταξύ των ειδών, παρακολουθώντας τις ομοιότητες μεταξύ τους.
- Σχολιασμός γονιδίων (συνεργατική γονιδιωματική)
 - Σύγκριση παρόμοιων ειδών
- Προσδιορισμός ρυθμιστικών δικτύων
 - Οι μεταβλητές που καθορίζουν τον τρόπο αντίδρασης του σώματος σε ορισμένα ερεθίσματα.
- Πρωτεϊνωματική
 - Από την αλληλουχία DNA σε μια αναδιπλωμένη πρωτεΐνη

Μοντελοποίηση

- Η μοντελοποίηση βιολογικών διεργασιών μας δείχνει αν κατανοούμε μια δεδομένη διεργασία
- Λόγω του μεγάλου πλήθους μεταβλητών που υπάρχουν στα βιολογικά προβλήματα, χρειαζόμαστε πανίσχυρους υπολογιστές για να αναλύσουμε ορισμένα βιολογικά ερωτήματα

Μοντελοποίηση πρωτεϊνών

- Οι αλγόριθμοι απεικόνισης ενεργών θέσεων της κβαντικής χημείας μας επιτρέπουν να δούμε τους πιθανούς μηχανισμούς αντίδρασης και δημιουργίας δεσμών
- Η ομόλογη μοντελοποίηση πρωτεϊνών (homologous protein modeling) είναι μια συγκριτική πρωτεϊνωματική μέθοδος με την οποία προσδιορίζεται η τριτοταγής δομή μιας άγνωστης πρωτεΐνης
- Οι αλγόριθμοι πρόβλεψης τριτοταγούς αναδίπλωσης απέχουν πολύ από το ιδανικό, αλλά μπορούμε να προβλέψουμε τη δευτεροταγή δομή με ακρίβεια ~80%.

Μοντελοποίηση ρυθμιστικών δικτύων

- Τα πειράματα με μικροσυστοιχίες DNA μας επιτρέπουν να συγκρίνουμε τις διαφορές της έκφρασης για δύο διαφορετικές καταστάσεις
- Οι αλγόριθμοι για την ομαδοποίηση των επιπέδων γονιδιακής έκφρασης συμβάλλουν στον εντοπισμό πιθανών ρυθμιστικών δικτύων
- Άλλοι αλγόριθμοι εκτελούν στατιστική ανάλυση για να βελτιώσουν το λόγο σήματος προς θόρυβο

Μοντελοποίηση της βιολογίας των συστημάτων

- Προβλέψεις των αλληλεπιδράσεων ολόκληρων κυττάρων.
 - Διεργασίες οργανιδίων, μοντελοποίηση επιπέδων έκφρασης
- Είναι σήμερα εφικτή για συγκεκριμένες διεργασίες (π.χ., μεταβολισμός στο βακτήριο *E. coli*, απλά κύτταρα)
 - Ανάλυση ισορροπίας ροής (flux balance analysis)

Το μέλλον...

- Έχουμε να μάθουμε ακόμα πολλά σχετικά με το πώς οι πρωτεΐνες μπορούν να χειρίζονται μια αλληλουχία ζευγών βάσεων με τόσο συγκεκριμένο τρόπο, το οποίο έχει σαν αποτέλεσμα έναν πλήρως λειτουργικό οργανισμό.
- Επομένως, πώς μπορούμε να χρησιμοποιήσουμε αυτές τις πληροφορίες προς όφελος της ανθρωπότητας, χωρίς να τις καταχραστούμε;

Πηγές


- Daniel Sam, “Greedy Algorithm”, παρουσίαση.
- Glenn Tesler, “Genome Rearrangements in Mammalian Evolution: Lessons from Human and Mouse Genomes”, παρουσίαση.
- Ernst Mayr, “What evolution is”.
- Neil C. Jones, Pavel A. Pevzner, “Εισαγωγή στους αλγορίθμους βιοπληροφορικής”.
- Alberts, Bruce, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, Peter Walter. Molecular Biology of the Cell. New York: Garland Science. 2002.
- Mount, Ellis, Barbara A. List. Milestones in Science & Technology. Phoenix: The Oryx Press. 1994.
- Voet, Donald, Judith Voet, Charlotte Pratt. Fundamentals of Biochemistry. New Jersey: John Wiley & Sons, Inc. 2002.
- Campbell, Neil. Biology, 3η έκδοση. The Benjamin/Cummings Publishing Company, Inc., 1993.
- Snustad, Peter και Simmons, Michael. Principles of Genetics. John Wiley & Sons, Inc, 2003.

Τεχνικές Ανάλυσης και Σύγκρισης Ακολουθιών Βιολογικών Δεδομένων

- Παραδείγματα Βάσεων Δεδομένων Βιολογικών Ακολουθιών
- Βασικοί Ορισμοί
- Το πρόβλημα του ακριβούς ταιριάσματος προτύπου
 - Απλοϊκή Μέθοδος
 - Αλγόριθμος Knuth-Morris-Pratt
 - Αλγόριθμος Boyer-Moore
 - Αλγόριθμος Shift-Or/Shift And
 - Το Αυτόματο Aho-Corasick
- Εφαρμογές σε Προβλήματα Μοριακής Βιολογίας

Βιολογικές Βάσεις Δεδομένων

- ❑ Γενικευμένες (Generalised) ή Αρχειακές (Archival) βιολογικές βάσεις δεδομένων. Διακρίνονται σε:
 - Πρωτογενείς βάσεις δεδομένων ακολουθιών (Primary Sequence Databases). Περιέχουν νουκλεοτιδικές και αμινοξικές ακολουθίες από γονιδιώματα οργανισμών που είτε έχουν αποκρυπτογραφηθεί πλήρως είτε
 - βάσεις δεδομένων που περιέχουν τρισδιάστατες δομές νουκλεϊνικών οξέων και πρωτεϊνών
- (NCBI, GENBANK, EMBL-Bank, DDJB, Swiss-Prot, PIR-PSD)
- ❑ Βάσεις δεδομένων ακολουθιών νουκλεοτιδικών ακολουθιών
 - ❑ Βάσεις δεδομένων ακολουθιών πρωτεϊνικών ακολουθιών
 - ❑ Βάσεις δεδομένων τρισδιάστατων βιολογικών δομών
 - ❑ Βάσεις δεδομένων γονιδιακής έκφρασης
 - ❑ Βάσεις δεδομένων γενετικής ποικιλομορφίας
 - ❑ Βάσεις δεδομένων βιβλιογραφίας




❑ Δευτερεύουσες (Secondary) βιολογικές βάσεις δεδομένων που προκύπτουν από ανάλυση των δεδομένων που είναι αποθηκευμένα στις αρχειακές βιολογικές βάσεις δεδομένων και διακρίνονται σε:

✓ Δευτερεύουσες ΒΔ ακολουθιών DNA και πρωτεϊνών που προκύπτουν από τις βασικές ΒΔ ακολουθιών και περιλαμβάνουν

(α) ΒΔ ακολουθιών στις οποίες έχουν απομακρυνθεί οι ακολουθίες που έχουν αποθηκευτεί περισσότερες από μία φορές

(β) ΒΔ που καταγράφουν παραλλαγές στις ακολουθίες DNA και πρωτεϊνών

(γ) Γονιδιωματικές ΒΔ που είτε ομαδοποιούν συγγενή ή όχι πλήρως αποκρυπτογραφημένα γονιδιώματα είτε ασχολούνται με γονιδιώματα οργανισμών μοντέλων



✓ ΒΔ που ασχολούνται με τις ιεραρχήσεις ή/και συσχετίσεις μεταξύ βιομορίων όπως οικογένειες πρωτεϊνών, κοινές δομές πρωτεϊνών κοινά μοτίβα ακολουθιών DNA και πρωτεϊνών.

□ Εξειδικευμένες Β.Δ., κατηγορία στην οποία ανήκουν:

✓ Β.Δ. μικροσυστοιχιών που περιλαμβάνουν πληροφορίες για την έκφραση γονιδίων και πρωτεϊνών

✓ Β.Δ. Μεταβολικών μονοπατιών που περιέχουν πληροφορίες για τις χημικές αντιδράσεις που πραγματοποιούνται στο κύτταρο

□ Βιβλιογραφικές βιολογικές βάσεις δεδομένων

□ Βιολογικές βάσεις δεδομένων ιστοσελίδων που περιλαμβάνουν:

✓ Β.Δ. που περιλαμβάνουν ως εγγραφές βιολογικές βάσεις


✓ Συνδέσμους μεταξύ βιολογικών βάσεων δεδομένων.

Τεχνικές Ανάλυσης και Σύγκρισης Ακολουθιών Βιολογικών Δεδομένων

- Παραδείγματα Βάσεων Δεδομένων Βιολογικών Ακολουθιών
- Βασικοί Ορισμοί
- Το πρόβλημα του ακριβούς ταιριάσματος προτύπου
 - Απλοϊκή Μέθοδος
 - Αλγόριθμος Knuth-Morris-Pratt
 - Αλγόριθμος Boyer-Moore
 - Αλγόριθμος Shift-Or/Shift And
 - Το Αυτόματο Aho-Corasick
- Εφαρμογές σε Προβλήματα Μοριακής Βιολογίας

Βιολογικές Βάσεις Δεδομένων

- ❑ Γενικευμένες (Generalised) ή Αρχειακές (Archival) βιολογικές βάσεις δεδομένων. Διακρίνονται σε:
 - Πρωτογενείς βάσεις δεδομένων ακολουθιών (Primary Sequence Databases). Περιέχουν νουκλεοτιδικές και αμινοξικές ακολουθίες από γονιδιώματα οργανισμών που είτε έχουν αποκρυπτογραφηθεί πλήρως είτε
 - βάσεις δεδομένων που περιέχουν τρισδιάστατες δομές νουκλεϊνικών οξέων και πρωτεϊνών
- (NCBI - GENBANK, EMBL-Bank, DDJB, Swiss-Prot, PIR-PSD)
- ❑ Βάσεις δεδομένων ακολουθιών νουκλεοτιδικών ακολουθιών
 - ❑ Βάσεις δεδομένων ακολουθιών πρωτεϊνικών ακολουθιών
 - ❑ Βάσεις δεδομένων τρισδιάστατων βιολογικών δομών
 - ❑ Βάσεις δεδομένων γονιδιακής έκφρασης
 - ❑ Βάσεις δεδομένων γενετικής ποικιλομορφίας
 - ❑ Βάσεις δεδομένων βιβλιογραφίας




❑ Δευτερεύουσες (Secondary) βιολογικές βάσεις δεδομένων που προκύπτουν από ανάλυση των δεδομένων που είναι αποθηκευμένα στις αρχειακές βιολογικές βάσεις δεδομένων και διακρίνονται σε:

✓ Δευτερεύουσες ΒΔ ακολουθιών DNA και πρωτεϊνών που προκύπτουν από τις βασικές ΒΔ ακολουθιών και περιλαμβάνουν

(α) ΒΔ ακολουθιών στις οποίες έχουν απομακρυνθεί οι ακολουθίες που έχουν αποθηκευτεί περισσότερες από μία φορές

(β) ΒΔ που καταγράφουν παραλλαγές στις ακολουθίες DNA και πρωτεϊνών

(γ) Γονιδιωματικές ΒΔ που είτε ομαδοποιούν συγγενή ή όχι πλήρως αποκρυπτογραφημένα γονιδιώματα είτε ασχολούνται με γονιδιώματα οργανισμών μοντέλων

- 
- ✓ ΒΔ που ασχολούνται με τις ιεραρχήσεις ή/και συσχετίσεις μεταξύ βιομορίων όπως οικογένειες πρωτεϊνών, κοινές δομές πρωτεϊνών κοινά μοτίβα ακολουθιών DNA και πρωτεϊνών.
 - Εξειδικευμένες Β.Δ., κατηγορία στην οποία ανήκουν:
 - ✓ Β.Δ. μικροσυστοιχιών που περιλαμβάνουν πληροφορίες για την έκφραση γονιδίων και πρωτεϊνών
 - ✓ Β.Δ. Μεταβολικών μονοπατιών που περιέχουν πληροφορίες για τις χημικές αντιδράσεις που πραγματοποιούνται στο κύτταρο
 - Βιβλιογραφικές βιολογικές βάσεις δεδομένων
 - Βιολογικές βάσεις δεδομένων ιστοσελίδων που περιλαμβάνουν:
 - ✓ Β.Δ. που περιλαμβάνουν ως εγγραφές βιολογικές βάσεις
 - ✓ Συνδέσμους μεταξύ βιολογικών βάσεων δεδομένων.

Βάσεις Βιολογικών Δεδομένων

- GenBank: [NCBI](http://www.ncbi.nlm.nih.gov) (<http://www.ncbi.nlm.nih.gov>)
[GenBank®](http://www.ncbi.nlm.nih.gov) is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences.
(National Center of Biotechnology Information)
- Swiss-Prot + TrEMBL: [Swiss-Prot.htm](http://tw.expasy.org/sprot/) (<http://tw.expasy.org/sprot/>)
Swiss-Prot is a curated protein sequence database which strives to provide a high level of annotation (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases
TrEMBL is a computer-annotated supplement of Swiss-Prot that contains all the translations of EMBL nucleotide sequence entries not yet integrated in Swiss-Prot.
- PROSITE: [Prosites](http://tw.expasy.org/prosite/) (<http://tw.expasy.org/prosite/>)
PROSITE is a database of protein families and domains. It consists of biologically significant sites, patterns and profiles that help to reliably identify to which known protein family (if any) a new sequence belongs.
- PDB-Protein Data Bank: [PDB](https://www.rcsb.org/) (<https://www.rcsb.org/>)
The PDB is the single worldwide repository for the processing and distribution of 3-D structure data of large molecules of proteins and nucleic acids.

Ένα παράδειγμα: PDB data entry example

Title: Structure Of The DNA Binding Domains Of Nfat, Fos and Jun Bound To DNA

Compound: Mol_Id: 1; Molecule: Nfat; Chain: N; Fragment: DNA Binding Domain; Synonym: Nf-At;

Biological_Unit: Monomer

Mol_Id: 2; Molecule: C-Fos; Chain: F; Engineered: Yes; Mutation: C154S

Mol_Id: 3; Molecule: C-Jun; Chain: J; Engineered: Yes; Mutation: C279S

Mol_Id: 4; Molecule: DNA; Chain: A, B; Engineered: Yes

Authors: L. Chen, J. N. M. Glover, P. G. Hogan, A. Rao, S. C. Harrison

Exp. Method: X-ray Diffraction

Classification: Complex (Transcription/Nuclear/Nuclear)

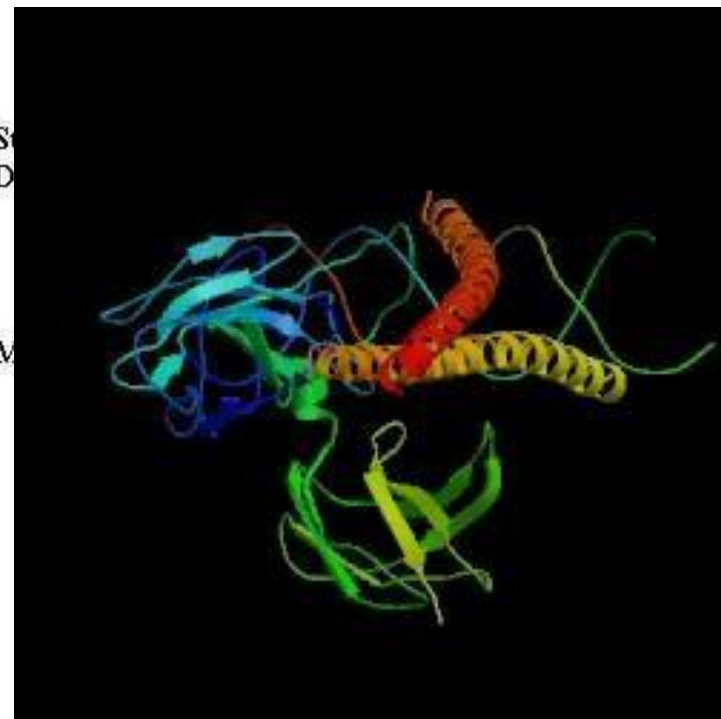
Source: Homo sapiens

Primary Citation: Chen, L., Glover, J. N., Hogan, P. G., Rao, A., Harrison, S. C.: Structure of the DNA binding domains from NFAT, Fos and Jun bound specifically to DNA (1998)



Deposition Date: 08-Dec-1997

Release Date: 27-Mar-1998



■ GenBank ([NCBI \(http://www.ncbi.nlm.nih.gov\)](http://www.ncbi.nlm.nih.gov))

Συλλογή ανοιχτής πρόσβασης, με σχολιασμό αλληλουχιών νουκλεοτιδίων και πρωτεϊνικών μεταφράσεων.

Παράγεται και διατηρείται από το Εθνικό Κέντρο Πληροφοριών Βιοτεχνολογίας (NCBI, μέρος των Εθνικών Ινστιτούτων Υγείας στις Ηνωμένες Πολιτείες) ως μέρος της Διεθνούς Συνεργασίας Βάσεων Δεδομένων Ακολουθιών Νουκλεοτιδίων (INSDC)

Η βάση δεδομένων ξεκίνησε το 1982 από το Εθνικό Εργαστήριο Walter Goad και Los Alamos. Η GenBank έχει γίνει μια σημαντική βάση δεδομένων για την έρευνα σε βιολογικά πεδία και έχει αναπτυχθεί τα τελευταία χρόνια με εκθετικό ρυθμό διπλασιαζόμενος περίπου κάθε 18 μήνες.

Τον Οκτώβριο του 2023, περιείχε πάνω από 268 τρισεκατομμύρια νουκλεοτιδικές βάσεις σε περισσότερες από 3.2 δισεκατομμύρια αλληλουχίες.

- **EMBL** <http://www.ebi.ac.uk/embl/> :

Το Ευρωπαϊκό Εργαστήριο Μοριακής Βιολογίας (EMBL) δημιουργήθηκε το 1974 και χρηματοδοτείται από δημόσια ερευνητικά χρήματα των κρατών μελών του.

Η EMBL Nucleotide Sequence Database (<http://www.ebi.ac.uk/embl/>) αποτελεί τη μεγαλύτερη βάση νουκλεοτιδικών αλληλουχιών στην Ευρώπη, βρίσκεται υπό την αιγίδα του Ευρωπαϊκού Εργαστηρίου Μοριακής Βιολογίας (EMBL) ενώ εδράζεται και συντηρείται από το Ευρωπαϊκό Ινστιτούτο Βιοπληροφορικής (EBI) στο Cambridge, UK.

Η τελευταία έκδοση της Βάσης Δεδομένων Ακολουθιών Νουκλεοτιδίων EMBL περιέχει πάνω από 4,8 δισεκατομμύρια καταχωρήσεις, που αντιπροσωπεύουν ένα εκπληκτικό 1200 τρισεκατομμύρια νουκλεοτίδια



- **DDBJ** (<http://www.ddbj.nig.ac.jp>):

- Η DNA Databank of Japan (DDBJ - <http://www.ddbj.nig.ac.jp/>) είναι η μοναδική διεθνώς αναγνωρισμένη βάση νουκλεοτιδικών αλληλουχιών στην Ιαπωνία. Βασική πηγή δεδομένων της βάσης αποτελούν οι εργασίες των Ιαπώνων ερευνητών. Επιπλέον στην DDBJ είναι διαθέσιμα διάφορα εργαλεία ανάλυσης νουκλεοτιδικών αλληλουχιών.

- As of 2022, DDBJ housed over 6.6 billion nucleotide sequences. This number likely exceeds 7 billion by now.

Στοιχεία από https://en.wikipedia.org/wiki/DNA_Data_Bank_of_Japan

Βάσεις δεδομένων τρισδιάστατων βιολογικών δομών

- Οι βάσεις αυτές περιέχουν δεδομένα που έχουν να κάνουν με την τρισδιάστατη δομή βιολογικών μακρομορίων.
- Οι τρισδιάστατες δομές αποτελούν το τελικό στάδιο μιας επίπονης διαδικασίας η οποία μετά τη χρήση μοριακών τεχνικών (κλωνοποίηση, απομόνωση, κρυστάλλωση κ.ο.κ.), οδηγεί τελικά στην υπολογιστική επίλυση της δομής μέσω της διαδικασίας της κρυσταλλογραφίας ακτίνων X, ή, σε πιο σπάνιες περιπτώσεις με φασματογραφία NMR.
- Το μεγαλύτερο ενδιαφέρον, βέβαια, έχουν οι δομές πρωτεϊνών, καθώς οι πρωτεΐνες είναι τα μακρομόρια των οποίων η μεγάλη ποικιλομορφία της δομής συνδέεται άμεσα με την βιολογική δράση. Η μοναδική βάση αυτού το είδους παγκοσμίως, είναι η PDB, η οποία και αναλύεται παρακάτω.
- Protein Data Bank (PDB - www.rcsb.org)

Protein Data Bank

- Η Protein Data Bank (PDB - www.rcsb.org) είναι η βάση στην οποία περιέχονται τρισδιάστατες δομές βιολογικών μακρομορίων (Kouranov, et al., 2006).
- Η Βάση Δεδομένων Πρωτεϊνών (PDB) [1] είναι μια βάση δεδομένων για τα τρισδιάστατα Τα δεδομένα, που λαμβάνονται συνήθως με κρυσταλλογραφία [ακτίνων X](#), φασματοσκοπία NMR ή, ολοένα και περισσότερο, κρυοηλεκτρονική μικροσκοπία, και υποβάλλονται από βιολόγους και βιοχημικούς από όλο τον κόσμο, είναι ελεύθερα προσβάσιμα στο Διαδίκτυο μέσω των ιστοσελίδων των οργανισμών μελών του (PDBe, [2] PDBj, [3] RCSB, [4] και BMRB[5]). Το PDB εποπτεύεται από έναν οργανισμό που ονομάζεται Worldwide Protein Data Bank, wwPDB.
- Σήμερα (2024) η PDB περιλαμβάνει 187.000 δομές βιομορίων.

Δευτερογενείς βάσεις δεδομένων

- **Βάσεις δεδομένων οικογενειών**
- Οι πρωτεΐνες αποτελούνται από μία ή περισσότερες διακριτές λειτουργικές περιοχές (domains), οι οποίες πολλές φορές είναι και δομικά αυτοτελείς.
- Διαφορετικοί συνδυασμοί τέτοιων περιοχών οδηγούν σε μια μεγάλη ποικιλία των πρωτεϊνών στη φύση.
- Συνεπώς, η ανίχνευση τέτοιων περιοχών είναι σημαντική στην προσπάθεια λειτουργικής ταξινόμησης των πρωτεϊνών.

Βασικοί Ορισμοί (1)

- Συμβολοσειρά-string: $x = x[1]x[2] \dots x[n]$, $x[i] \in \Sigma$ & $|x| = n$
 $x = \text{acgttaaaca}$, $|x| = 10$ & $\Sigma = \{A, C, G, T\}$
(Αδενίνη, Θυμίνη, Κυτοσίνη, Γουανίνη)
- Σ^+ : το σύνολο των συμβολοσειρών που ορίζονται στο αλφάβητο Σ
? Πόσες συμβολοσειρές μήκους « λ » ορίζονται στο $\Sigma = \{a, c, g, t\}$
- Κενή συμβολοσειρά: ε
- Υπο-συμβολοσειρά-substring w : $x = uwn$
- Πρόθεμα –Prefix w : $x = wu$
- Επίθεμα-Suffix w : $x = uw$

Βασικοί Ορισμοί (2)

- Border συμβολοσειράς x : συμβολοσειρά w που είναι πρόθεμα και επίθεμα του x .
- $x^k = \underbrace{x \dots x}_{\text{κ φορές}}$, k -οστή δύναμη του x
- $y = x^k$, $k > 1 \rightarrow$ το y περιοδικό με περίοδο x
- Περίοδος $y =$ η μικρότερη τέτοια συμβολοσειρά
- Primitive (πρωταρχική) συμβολοσειρά
- Κάλυμμα (Cover) συμβολοσειράς
- Φύτρο (Seed) συμβολοσειράς

Προβλήματα Ταυρίσματος Προτύπου (1)

- Ακριβές Ταίριασμα: ενδιαφερόμαστε να εντοπίσουμε όλες τις εμφανίσεις ενός δοσμένου προτύπου (μοτίβου) P (“δομημένου” ή “μη-δομημένου”) σε μια συμβολοσειρά (βιολογική αλληλουχία) T .
- Προσεγγιστικό Ταίριασμα: *Για ένα κείμενο T , ένα μοτίβο P , μια παράμετρο k και μια συνάρτηση ομοιότητας $sim()$, εντόπισε τις θέσεις i, j στο κείμενο, έτσι ώστε $sim(P, T_{i..j}) \geq k$.*

Προβλήματα Ταιριάσματος Προτύπου (2)

- Η διαδικασία σύγκρισης της ομοιότητας δυο ακολουθιών στηρίζεται σε πίνακες που βαθμολογούν τις ομοιότητες (matches) και διαφορές (mismatches) μεταξύ διαδοχικών συμβόλων. Τέτοιου τύπου πίνακες είναι οι: Dayhoff Mutation Data Matrix, BLOSUM κτλ.
- Επίσης η σύγκριση ακολουθιών μπορεί να κατηγοριοποιηθεί σε: α) **τοπική ευθυγράμμιση -local alignment** και β) **ολική ευθυγράμμιση - global alignment**. Στην τοπική ευθυγράμμιση αναζητούμε περιοχές τοπικής ομοιότητας. Γνωστοί τέτοιοι αλγόριθμοι είναι των Smith-Waterman (τοπικοί), Needleman & Wunsch (ολικοί). Και στις δυο περιπτώσεις υπάρχουν παραπάνω από μια δυνατές ευθυγραμμίσεις. Η βέλτιστη λύση πρέπει να ελαχιστοποιεί τις διαφορές ανάμεσα στις δυο ακολουθίες ή διαφορετικά να μεγιστοποιεί τη συνάρτηση ομοιότητας.

Στοιχείση Ακολουθιών

- Συνέκρινε δύο ή περισσότερες ακολουθίες ελέγχοντας για μία ακολουθία ατομικών χαρακτήρων που είναι με την ίδια σειρά στις ακολουθίες.
- Ανακάλυψε λειτουργική, δομική και εξελεκτική πληροφορία.

L G P S S K Q T G K G S - S R I W D N
 |
 L N - I T K S A G K G A I M R L G D A ολική στοίχιση

-- -- -- -- -- -- -- T K G S - -- -- -- -- --
 |
 -- -- - -- -- -- -- A K G A -- -- -- -- -- τοπική στοίχιση

Εύρεση Επαναλήψεων σε Βιολογικές Ακολουθίες

- Οι επαναλήψεις σε βιολογικές ακολουθίες κατηγοριοποιούνται στις εξής 3 βασικές κατηγορίες:
 - επαναλήψεις περιορισμένου μήκους που εμφανίζονται σε τοπικό επίπεδο, και των οποίων η λειτουργία είναι γνωστή,
 - επαναλήψεις περιορισμένου μήκους που εμφανίζονται σε όλο το μήκος της ακολουθίας, και των οποίων η λειτουργία δεν είναι απόλυτα γνωστή,
 - δομημένες επαναλήψεις μεγάλου μήκους των οποίων η λειτουργία δεν έχει προσδιοριστεί.

Παραδείγματα Επαναλήψεων

- 1η κατηγορία:
 - τα **συμπληρωματικά παλίνδρομα** σε ακολουθίες DNA & RNA, που ρυθμίζουν τη μετεγγραφή του DNA,
 - τα **εμφωλευμένα συμπληρωματικά παλίνδρομα** σε ακολουθίες RNA
- 2η κατηγορία:
 - **συνεχόμενες επαναλήψεις- tandem repeats,**
 - **δορυφορικά τμήματα DNA- satellite DNA, (micro & mini satellite DNA)**
- 3η κατηγορία:
 - **SINE-Short Interspersed Nuclear Sequences (π.χ.: *Alu family*)**
 - **LINE-Long Interspersed Nuclear Sequences.**

Πρότυπα

- **Μοτίβα DNA**

TRANSFAC, JASPAR, SCPD, DBTBS, RegulonDB

- **Μοτίβα πρωτεϊνών**

PROSITE, Pfam, ProDom, BLOCKS, TIGRFAM, Interpro

Γονιδιωματικές επαναλήψεις

- Παράδειγμα επαναλήψεων:

□ ATGGTCTAGGTCCTAGTGGTC

- Κίνητρα για την εύρεσή τους:

- Οι γονιδιωματικές αναδιατάξεις σχετίζονται συχνά με επαναλήψεις
- Εξακρίβωση των εξελικτικών μυστικών
- Πολλοί όγκοι χαρακτηρίζονται από μια απότομη και μεγάλη αύξηση των επαναλήψεων

Γονιδιωματικές επαναλήψεις

- Συχνά, το πρόβλημα είναι πιο δύσκολο:

□ ATGGTCTAGGACTAGTGTTC

- Κίνητρα για την εύρεσή τους:
 - Οι γονιδιωματικές αναδιατάξεις σχετίζονται συχνά με επαναλήψεις
 - Εξακρίβωση των εξελικτικών μυστικών
 - Πολλοί όγκοι χαρακτηρίζονται από μια απότομη και μεγάλη αύξηση των επαναλήψεων

Επαναλήψεις l -μερών

- Η εύρεση των επαναλήψεων μεγάλου μήκους είναι δύσκολη
- Η εύρεση των επαναλήψεων μικρού μήκους είναι εύκολη (π.χ., με κατακερματισμό)
- Απλή μέθοδος για την εύρεση επαναλήψεων μεγάλου μήκους:
 - Βρίσκουμε τις ακριβείς επαναλήψεις l -μερών μικρού μήκους (το l είναι συνήθως ίσο με 10 έως 13)
 - Χρησιμοποιούμε τις επαναλήψεις l -μερών για να επεκτείνουμε πιθανώς σε μεγαλύτερου μήκους, *μέγιστες*, επαναλήψεις.

Επαναλήψεις *l*-μερών (συνέχεια)

- Υπάρχουν συνήθως πολλές θέσεις στις οποίες επαναλαμβάνεται ένα *l*-μερές:

GCTTACAGATTTCAGTCTTACAGATGGT

- Το 4-μερές TTAC αρχίζει στις θέσεις 3 και 17

Επέκταση επαναλήψεων *l*-μερών

GC**TTAC**AGATTTCAGTCT**TTAC**AGATGGT

- Επεκτείνουμε τα ταιριάσματα αυτού του 4-μερούς:

GC**TTAC**AGATTTCAGTCT**TTAC**AGATGGT

- Μέγιστη επανάληψη: **TTACAGAT**

Μέγιστες επαναλήψεις

- Για να βρούμε μέγιστες επαναλήψεις με αυτόν τον τρόπο, χρειαζόμαστε ΟΛΕΣ τις θέσεις όλων των *l*-μερών στο γονιδίωμα
- Ο **κατακερματισμός** μας επιτρέπει να βρίσκουμε επαναλήψεις γρήγορα με το συγκεκριμένο τρόπο

Κατακερματισμός: τι είναι;

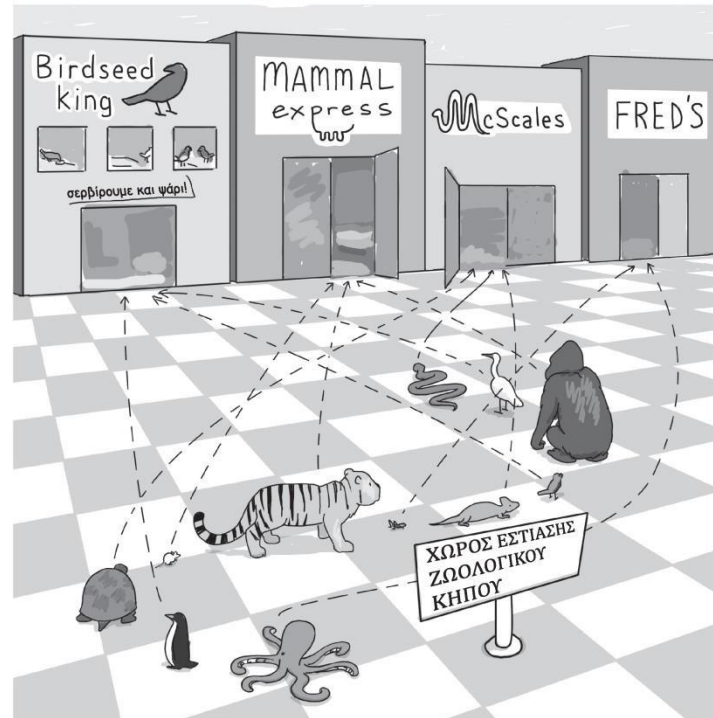
- Τι κάνει ο κατακερματισμός;
 - Για διαφορετικά δεδομένα, παράγει ένα μοναδικό ακέραιο
 - Αποθηκεύει δεδομένα σε έναν πίνακα στη θέση που αντιστοιχεί στο μοναδικό ακέραιο δείκτη που έχει παραχθεί από τα δεδομένα
- Ο κατακερματισμός είναι ένας πολύ αποδοτικός τρόπος για την αποθήκευση και ανάκτηση δεδομένων

Κατακερματισμός: ορισμοί

- Πίνακας κατακερματισμού: η διάταξη που χρησιμοποιείται στον κατακερματισμό
- Εγγραφές: δεδομένα αποθηκευμένα σε έναν *πίνακα κατακερματισμού*
- Κλειδιά: προσδιορίζουν σύνολα εγγραφών
- Συνάρτηση κατακερματισμού: χρησιμοποιεί ένα *κλειδί* για να παραγάγει ένα δείκτη για τη θέση εισαγωγής (δεδομένων) στον *πίνακα κατακερματισμού*
- Σύγκρουση: όταν περισσότερες από μία εγγραφές αντιστοιχίζονται με τον ίδιο δείκτη στον πίνακα κατακερματισμού

Κατακερματισμός: παράδειγμα

- Πού τρώνε τα ζώα;
- Εγγραφές: κάθε ζώο
- Κλειδιά: το πού τρώει κάθε ζώο



Εγγραφές Κλειδιά

x	$h(x)$
Πιγκουΐνος	1
Χταπόδι	4
Χελώνα	3
Ποντίκι	2
Φίδι	3
Ερωδιός	1
Τίγρης	2
Ιγουάνα	3
Πίθηκος	2
Γρύλος	4
Σπουργίτι	1

Κατακερματισμός αλληλουχιών DNA

- Κάθε *l*-μερές μπορεί να μεταφραστεί σε μια δυαδική συμβολοσειρά (οι βάσεις **A**, **T**, **C**, **G** μπορούν να αναπαρασταθούν ως **00**, **01**, **10**, **11**)
- Μετά από την αντιστοίχιση ενός μοναδικού ακεραίου σε κάθε *l*-μερές, είναι εύκολο να βρούμε όλες τις αρχικές θέσεις για κάθε *l*-μερές στο γονιδίωμα

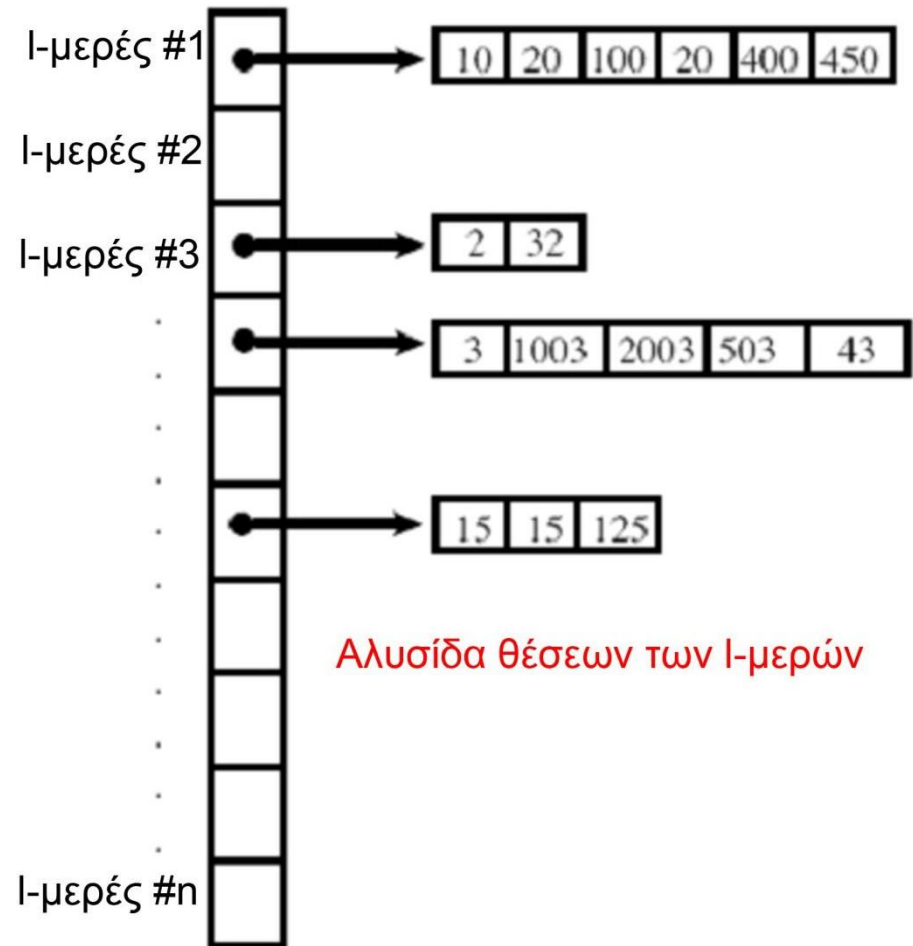
Κατακερματισμός: μέγιστες επαναλήψεις

- Για να βρούμε επαναλήψεις σε ένα γονιδίωμα:
 - Για όλα τα l -μερή στο γονιδίωμα, καταγράφουμε την αρχική θέση και την αλληλουχία
 - Παράγουμε ένα δείκτη του πίνακα κατακερματισμού για την αλληλουχία κάθε μοναδικού l -μερούς
 - Σε κάθε δείκτη του πίνακα κατακερματισμού, αποθηκεύουμε όλες τις θέσεις (στο γονιδίωμα) στις οποίες αρχίζει το l -μερές που παρήγαγε το συγκεκριμένο δείκτη
 - Επεκτείνουμε τις επαναλήψεις του l -μερούς σε μέγιστες επαναλήψεις

Κατακερματισμός: συγκρούσεις

- Αντιμετώπιση συγκρούσεων:

- «Δημιουργία αλυσιδωτής δομής» για όλες τις αρχικές θέσεις των *l*-μερών (συνδεδεμένη λίστα)



Κατακερματισμός: περίληψη

- Κατά την εύρεση γονιδιωματικών επαναλήψεων από *l*-μερή:
 - Παράγουμε ένα δείκτη του πίνακα κατακερματισμού για την αλληλουχία κάθε *l*-μερούς
 - Σε κάθε δείκτη, αποθηκεύουμε όλες τις θέσεις (στο γονιδίωμα) στις οποίες αρχίζει το *l*-μερές που παρήγαγε το συγκεκριμένο δείκτη
 - Επεκτείνουμε τις επαναλήψεις του *l*-μερούς σε μέγιστες επαναλήψεις

Ταίριασμα μοτίβου

- Τι θα συμβεί αν, αντί να βρούμε τις επαναλήψεις σε ένα γονιδίωμα, θέλουμε να βρούμε όλες τις αλληλουχίες σε μια βάση δεδομένων που περιέχουν ένα δεδομένο μοτίβο;
- Αυτό μας οδηγεί σε ένα διαφορετικό πρόβλημα, το ***πρόβλημα του Ταιριάσματος Μοτίβου***

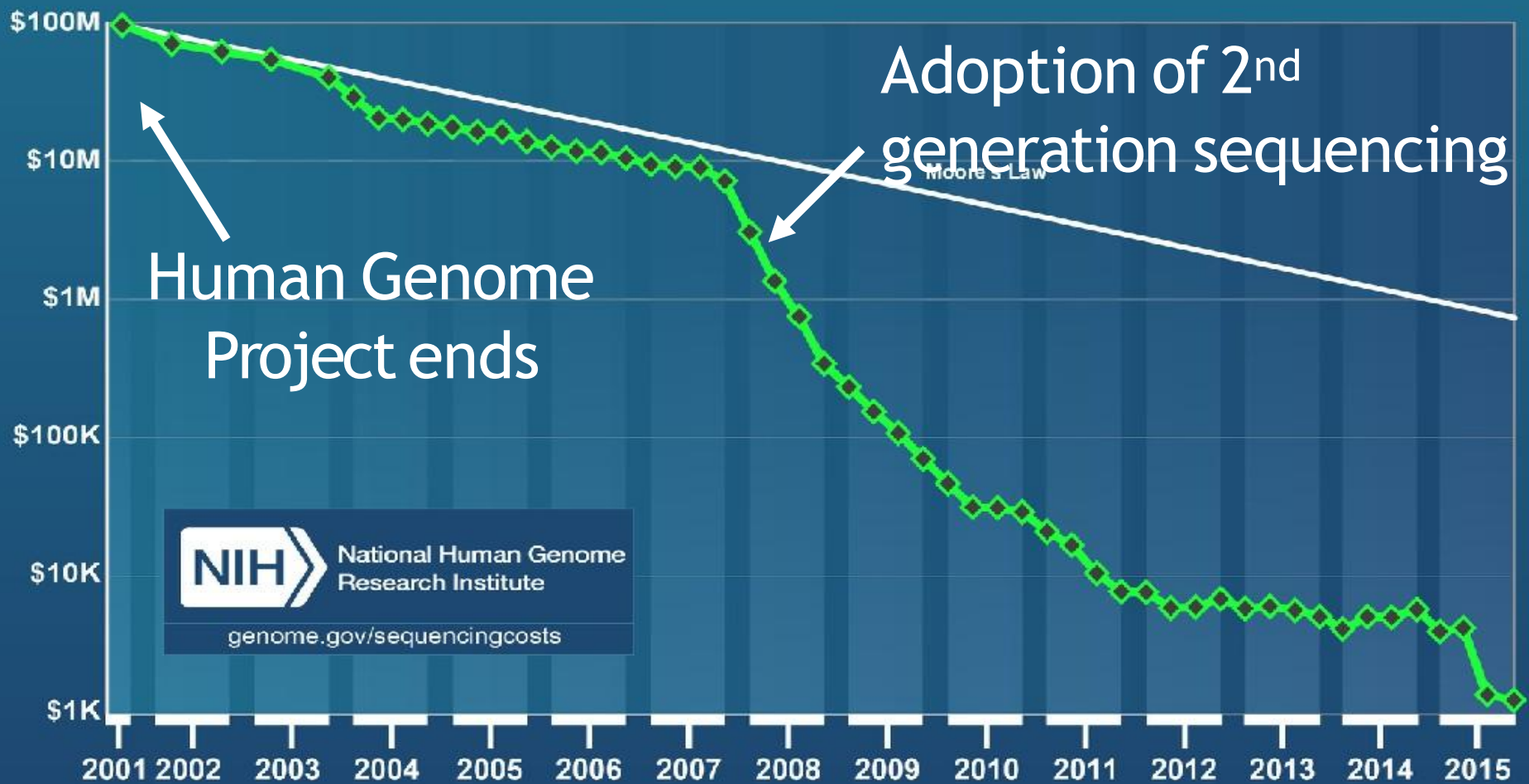
Το πρόβλημα του Ταιριάσματος Μοτίβου

- Στόχος: Βρείτε όλες τις εμφανίσεις ενός *μοτίβου* σε ένα κείμενο
- Είσοδος: Το μοτίβο $\mathbf{p} = p_1 \dots p_n$ και το κείμενο $\mathbf{t} = t_1 \dots t_m$
- Έξοδος: Όλες οι θέσεις $1 \leq i \leq (m - n + 1)$ έτσι ώστε η υποσυμβολοσειρά n γραμμάτων του \mathbf{t} που αρχίζει στη θέση i να ταιριάζει με το \mathbf{p}
- **Κίνητρο**: Αναζήτηση ενός γνωστού μοτίβου σε μια βάση δεδομένων

Ακριβές Ταίριασμα (εφαρμογές)

- Επεξεργαστές κειμένου
- Utilities (grep στο Unix)
- Textual Information Retrieval (Medline, Lexis, Nexis)
- Internet News Readers
- On-line dictionaries και θησαυρούς
- Molecular Biology Databases

Cost per Genome



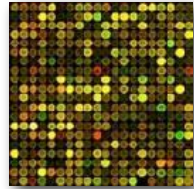
slide from Ben Langmead

Genomics technology



Sanger DNA
sequencing

1977-1990s



DNA Microarrays

Since mid-1990s



2nd-generation DNA
sequencing

Since ~2007



3rd-generation
& single-
molecule DNA
sequencing

Since ~2010



slide from Ben Langmead

Reads

GTATGCACGCGATAG
TAGCATTGCGAGACG
TGTCTTTGATTCCCTG
GACGCTGGAGCCGGA
TATCGCACCTACGTT
CACGGGAGCTCTCCA
GTATGCACGCGATAG
GCGAGACGCTGGAGC
CCTACGTTCAATATT
GACGCTGGAGCCGGA
TATCGCACCTACGTT
CACGGGAGCTCTCCA

TATGTCGCAGTATCT
GGTATGCACGCGATA
CGCGATAGCATTGCG
GCACCCATGTGCGCA
CAATATTCGATCATG
TGCATTTGGTATTTT
ACCTACGTTCAATAT
CTATCACCCATTATTA
GCACCTACGTTCAAT
GCACCCATGTGCGCA
CAATATTCGATCATG
TGCATTTGGTATTTT

CACCCATATGTGCGAG
TGGAGCCGGAGCACC
GCATTGCGAGACGCT
GTATCTGTCTTTGAT
GATCACAGGTCTATC
CGTCTGGGGGGTATG
TATTTATCGCACCTA
CTGTCTTTGATTCCCT
GTCTGGGGGGTATGC
GTATCTGTCTTTGAT
GATCACAGGTCTATC
CGTCTGGGGGGTATG

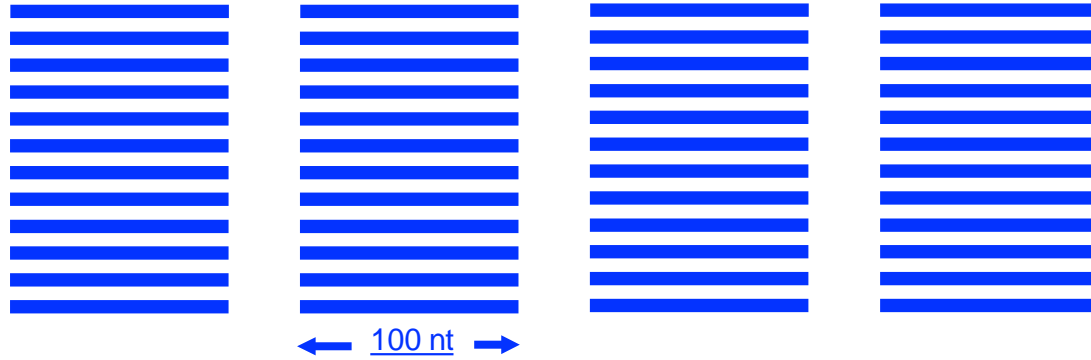
GAGACGCTGGAGCCG
CGCTGGAGCCGGAGC
CCTATGTGCGAGTAT
CCTCATCCTATTATT
ACCCTATTAACCACT
CACGCGATAGCATTG
CCACTCACGGGAGCT
ACTCACGGGAGCTCT
AGCCGGAGCACCCCTA
CCTCATCCTATTATT
ACCCTATTAACCACT
CACGCGATAGCATTG

Your genome

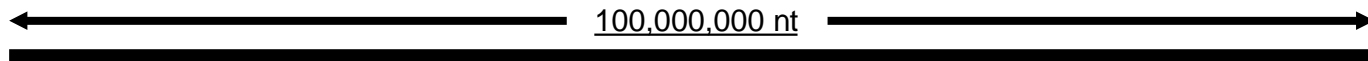
CGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCATATGTGCGAGTATCTGTCTTTGATTCCCTG

slide from Ben Langmead

Read
s



Your genome



slide from Ben Langmead

slide from Ben Langmead

Read

CTCAAACCTCCTGACCTTTGGTGATCCACCCGCCTAGGCCTTC

Reference



GATCACAGGTCTATCACCCCTATTAACCACCTCACGGGAGCTCTCCATGCATTTGGTATTTT
CGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCCTATGTC
GCAGTATCTGTCTTTGATTCTCTGCCTCATCTATTATTATTCGCACCTACGTTCAATATT
ACAGGCGAACATACTTACTAAAGTGTGTTAATTAATTAATGCTTGTAGGACATAATAATA
ACAATTGAATGTCTGCACAGCCACTTTCCACACAGACATCATAACAAAAAATTCCACCA
AACCCCCCTCCCCCGCTTCTG
ACAAAGAACCCTAACACCAGCC
CTCATCAATACAACCCCGCCC
GCAATACACTGACCCGCTCAA
TCACCCTCTAAATCACCACGAT
ACGAAAGTTTAACTAAGCTATA
TCCCCAATAAGCTAAACTCA
TACCCCACTATGCTTAGCCCTA
AGCCTGTTCTGTAATCGATAAA
ACGTTAGTCAAGGTGTAGCCC
AGTAGAGTGCTTAGTTGAACAG AAGTACTTCAAGGACATTT
CGTAACCTCAAACCTCCTGCCTTTG GTGACACCCCGCCTTGGCCTACCTTAAATGAAG
AAGCACCCAACCTTACACTTAGGAG ATTTCACTTAACTTGACCGCTTAACTTAACCTA
GCCCAAACCCCACTCCACCTTACT ACCAGACAACCTTCCATTATTAATTA
AGTATAGGCGATAGAAATTGAAAC CTGGCGCAATAGATATAGTACCGCAAGTATG
AAAAATTATAACCAAGCATAATAT AGCAAGGACTAACCCCTATACCTTCTGCA
TTAACTAGAAAATACTTTGCAAGG AGAGCCAAAGCTAAGACCCCGAAACCAGAC
ACCTAAGAACAGCTAAAGAGCAC ACCCGTCTATGTAGCAAAATAGTGGGAAGATT
GGTAGAGGCGACAAACCTACCGAG CCTGGTGATAGCTGGTTGTCCAAGATAGAATCTTA
TTCAACTTTAAATTTGCCACAGAACCCCTCTAAATCCCTTGTAAATTAACTGTTAGTC
CAAAGAGGAACAGCTCTTTGGACACTAGGAAAAACCTTGTAGAGAGAGTAAAAATTTA

Strings come from somewhere

Processes that give rise to real-world strings are complicated. It helps to understand them.

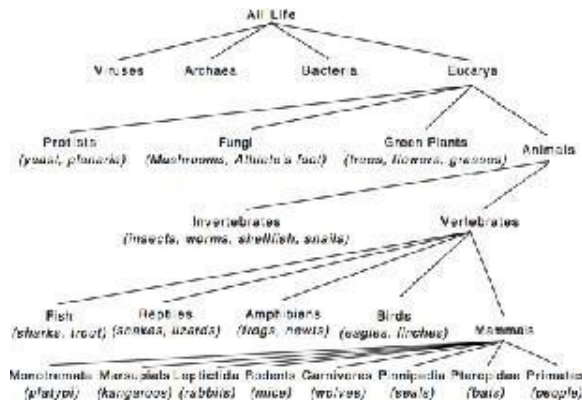


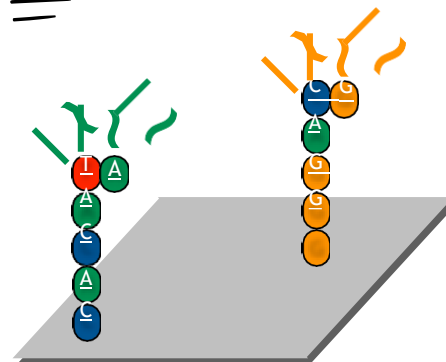
Figure from: Hunter, Lawrence. "Molecular biology for computer scientists." *Artificial intelligence and molecular biology* (1993): 1-46.

Mutation

1. Evolution: Recombination
(Retro)transposition



2. Lab procedures: PCR ✓
Cell line passages



3. Sequencing:

Fragmentation
bias Miscalled
bases

Εφαρμογές εύρεσης προτύπων σε προβλήματα Βιοπληροφορικής

- Αναζήτηση **Sequence-tagged-site (STS) & Expressed Sequence Tags (ESTs)** σε ακολουθίες γονιδιωμάτων.
 - **STS:** τμήματα του DNA μήκους 200-300 νουκλεοτιδίων
 - **ESTs:** τμήματα mRNA & cDNA ακολουθίες που αντιπροσωπεύουν τα τμήματα κωδικοποίησης μιας πρωτεΐνης σε μια ακολουθία γονιδίων.
- Αναζήτηση transcription factors
- Αναζήτηση “**κανονικών εκφράσεων**” (regular expressions)
 - [ED]-[EN]-L-[SAN]-x-x-[DE]-x-E-L \Rightarrow **ENLSSEDEEL****PROSITE**

Ακριβής Εύρεση Προτύπου

The Exact Pattern Matching Problem

Ορισμός: «έστω μια ακολουθία χαρακτήρων T. Αναζητούμε τις θέσεις εμφάνισης του προτύπου/ λέξης P μέσα στην ακολουθία».

P= acgttaaaca

T= tcg[acgttaaaca]ttttaaattt[acgttaaaca]ggggaattcg[acgttaaaca]

↑ 1η εμφάνιση ↑ 2η εμφάνιση ↑ 3η εμφάνιση

Η απλοϊκή μέθοδος επίλυσης – Naive Method

1ο βήμα: Στοιχίζουμε την ακολουθία και το πρότυπο & συγκρίνουμε τους χαρακτήρες

[illegible]

2ο βήμα: Στο πρώτο mismatch – 4η θέση μετατοπίζουμε το πρότυπο κατά 1 θέση

[illegible]

Η απλοϊκή μέθοδος επίλυσης – Naive Method

3ο βήμα: Σε κάθε mismatch μετατοπίζουμε το πρότυπο κατά 1 θέση

[illegible]

4ο βήμα:

[illegible]

Η απλοϊκή μέθοδος επίλυσης – Naive Method

5ο βήμα: 1η εύρεση του προτύπου $L_{\{x\}} = \{5, \dots\}$

[illegible]

6ο βήμα:

[illegible]

Κώδικας Απλοϊκής Μεθόδου

```
void Naïve-Method (char *P, int m, char *T, int n)
{
    int i,j;
    for (j=0; j<=n-m; ++j) {
        for (i=0; i<m && P[i]==T[i+j]; ++i);
        if (i==m) output(j);
    }
}
```

Ανάλυση της απλοϊκής μεθόδου σε χρόνο

- Πολυπλοκότητα μεθόδου: $O(n*m)$, όπου $|T|=n$ & $|P|=m$
 1. Πόσες μετατοπίσεις θα χρειαστεί να γίνουν?
 $|T| - |P| + 1 = n - m + 1$
 2. Πόσες συγκρίσεις πραγματοποιούνται το πολύ κάθε φορά?
 $|P|=m$
 3. Συνολικός χρόνος επεξεργασίας: $(n - m + 1) * m$
 4. Πώς μπορώ να βελτιώσω το χρόνο?

Βασική Προεπεξεργασία (1)

(D. Gusfield)

- $Z_i(S)$ = το μήκος της μεγαλύτερης υποσυμβολοσειράς του S , που αρχίζει στο i και ταιριάζει πρόθεμα του S .
- Z-box at i = το σύνολο χαρακτήρων που αρχίζουν από i και τελειώνουν στη θέση $i+Z_i(S)-1$.
- Για κάθε i , r_i συμβολίζει το δεξιότερο άκρο των Z-boxes που ξεκινά από ή πριν τη θέση i . Διαφορετικά, r_i είναι η μεγαλύτερη τιμή του $j+Z_j-1$ για κάθε $2 \leq j \leq i$. Το αριστερό άκρο (j) το συμβολίζουμε ως l_i .

Βασική Προεπεξεργασία (2)

Δοθέντων Z_i για $i \leq k-1$ και r, l για Z_{k-1} :

1. If $k > r$, find Z_k explicitly. If $Z_k > 0$, $r = k + Z_k - 1$, $l = k$.
2. If $k \leq r$, $S[k..r]$ matches $S[k'..Z_l]$ and the substring at k matches a prefix of S of length $\geq \min(Z_k, r - k + 1)$ ($k' = k - l + 1$)
 - a. If $Z_k < |S[k..r]|$ then $Z_k = Z_k$, r, l remain unchanged
 - b. Compare the characters starting at $r+1$ of S with the characters starting at $|S[k..r]|$ until a mismatch. Say the mismatch occurs at $q \geq r+1$. Then Z_k is $q - k$, r is set to $q - 1$ and l is set to k .

Βασική Προεπεξεργασία (3)

- Εφάρμοσε τον αλγόριθμο στην ακολουθία P\$T
- Κάθε τιμή $Z_i=m$, $i>m$ σηματοδοτεί match στη θέση $i-m-1$.
- Η μέθοδος μπορεί να υλοποιηθεί έτσι ώστε να απαιτεί επιπλέον (του χώρου αποθήκευσης P, T), $O(m)$ χώρο.
- Είναι μέθοδος που οι πολυπλοκότητές της είναι ανεξάρτητες από το μέγεθος του αλφαβήτου (ίδια ιδιότητα έχει ο αλγόριθμος Boyer Moore, Knuth Morris Pratt, όχι όμως ο Shift Or και ο αλγόριθμος Aho-Corasick)