



# Βιοπληροφορική

# ΣΚΟΠΟΣ

- Σύγκριση αλληλουχιών (DNA, RNA, proteins) κατά ζεύγη.
- Βασικές πράξεις μετασχηματισμού αλληλουχιών (Ένθεση Διαγραφή Αντικατάσταση).
- Στοίχιση αλληλουχιών με κενά.
- Ολική και τοπική στοίχιση.
- Συστήματα βαθμονόμησης για σύγκριση DNA και πρωτεϊνών.
- Αντικαταστάσεις αμινοξέων.
- Πίνακες αντικατάστασης PAM.

# Σύγκριση αλληλουχιών (DNA, RNA, proteins)

- **Ερώτηση:** Σχετίζονται δύο αλληλουχίες;
- **Μέθοδος:** Σύγκριση των δύο αλληλουχιών και εξαγωγή συμπεράσματος εάν είναι παρόμοιες.
  
- Παράδειγμα 1: Πάργα & Πράγα.
- Παράδειγμα 2: pear & tear.
  
- **Μεγάλη ομοιότητα μεταξύ των λέξεων, αλλά διαφορετικές έννοιες.**

# Γιατί σύγκριση αλληλουχιών;

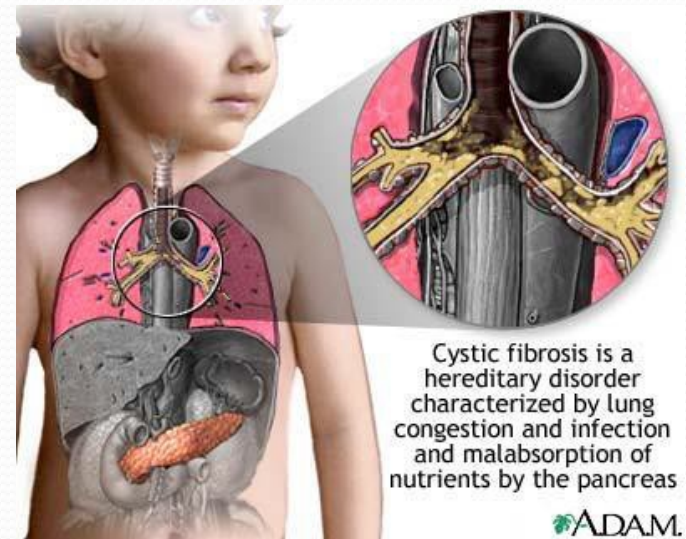
- **Συμπεράσματα για λειτουργία ενός γονιδίου (μίας πρωτεΐνης):**
  - Όταν δύο γονίδια έχουν μεγάλο ποσοστό ομοιότητας, τότε πιθανά κωδικοποιούν πρωτεΐνες με παρόμοια λειτουργία.
- **Συμπεράσματα για σημαντικές περιοχές στην αλληλουχία γονιδίων/πρωτεϊνών.**
  - Όταν πρωτεΐνες με συγκεκριμένο χαρακτηριστικό μοιράζονται κοινή περιοχή, τότε αυτή η κοινή περιοχή είναι πιθανά υπεύθυνη για το συγκεκριμένο χαρακτηριστικό.
- **Συμπεράσματα για την εξελεγκτική απόσταση μεταξύ 3 ειδών.**
  - Όταν η αλληλουχία μίας πρωτεΐνης είναι σχεδόν η ίδια μεταξύ 2 ειδών (ποντίκι και αρουραίος), τότε τα είδη αυτά είναι «κοντά».

# DNA Sequence Comparison: First Success Story

- Η εύρεση ομοιοτήτων με γονίδια γνωστής λειτουργίας είναι ένας τρόπος να εξάγουμε συμπεράσματα για τη λειτουργία ενός νέου αναγνωρισμένου γονιδίου.
- **Παράδειγμα (1):** Το 1984 ο Russell Doolittle (βιοχημικός, Η.Π.Α.) και οι συνεργάτες του βρήκαν ομοιότητες μεταξύ ενός άγνωστου γονιδίου που εμφανίζεται σε πολλούς τύπους καρκίνου και του «φυσιολογικού» γονιδίου του αυξητικού παράγοντα PDGF.

# Άλλο Παράδειγμα: Κυστική ίνωση (Cystic Fibrosis)

- Η **κυστική ίνωση** είναι μία πάθηση της λευκής φυλής που επιφέρει το θάνατο σε νεαρή ηλικία.
- Κύριο χαρακτηριστικό της νόσου είναι η:
  - παραγωγή ιδιαίτερα **πυκνής βλέννας** η οποία φράσσει τα διάφορα όργανα και πόρους του σώματος, κυρίως τους **πνεύμονες** και το **πάγκρεας**,
  - με αποτέλεσμα την βαριά παγκρεατική ανεπάρκεια από πολύ μικρή ηλικία και την εμφάνιση **σοβαρών χρόνιων αναπνευστικών λοιμώξεων** που σταδιακά καταστρέφουν τους πνεύμονες και οδηγούν τον ασθενή σε **αναπνευστική ανεπάρκεια και θάνατο.**



# Μεταλλάξεις - Mutations

Normal DNA sequence

ATC-CCT-AGT-AAA



Mutated DNA sequence

ATC-**CTT**-AGT-**AAG**

Normal protein sequence

Isoleucine – Proline – Serine – Lysine

Mutated protein sequence

Isoleucine – **Leucine** – Serine – **Lysine**

# Κυστική Ίνωση: Κληρονομικότητα

- Στις αρχές του 1980 προτάθηκε ότι η νόσος προκαλείται από μετάλλαξη κάποιου γονιδίου, γεγονός που επαληθεύτηκε.
- Η Κυστική ίνωση δεν είναι μεταδοτική νόσος, αλλά
- κληρονομική. Για να νοσήσει κάποιος πρέπει να έχει δυο γονίδια παθολογικά τα οποία κληρονομεί και από τους δυο γονείς του που είναι φορείς της νόσου, χωρίς να το ξέρουν.
- Στην Ελλάδα σήμερα οι φορείς του παθολογικού γονιδίου που προκαλεί την Κυστική Ίνωση υπολογίζονται σε περισσότερους από 500.000!!!
- Αν και οι δύο γονείς ενός παιδιού είναι φορείς της νόσου, η πιθανότητα να γεννηθεί παιδί με Κυστική Ίνωση είναι 1 στις 4. Κάθε χρόνο στην Ελλάδα γεννιούνται 50 πάσχοντα παιδιά!



# Κυστική ίνωση: Ψάχνοντας το γονίδιο

- Εάν ένα υψηλό ποσοστό ασθενών με κυστική ίνωση έχουν μία συγκεκριμένη μετάλλαξη σε ένα γονίδιο την οποία τα υγιή άτομα δεν την έχουν, τότε αυτό σημαίνει ότι η συγκεκριμένη μετάλλαξη σχετίζεται με την ασθένεια της κυστικής ίνωσης.
- Το 1989 βρέθηκε μία συγκεκριμένη μετάλλαξη στο 70% των ασθενών με κυστική ίνωση, η οποία πλέον χρησιμοποιείται ως διαγνωστικός δείκτης της ασθένειας.

## CFTR Sequence:

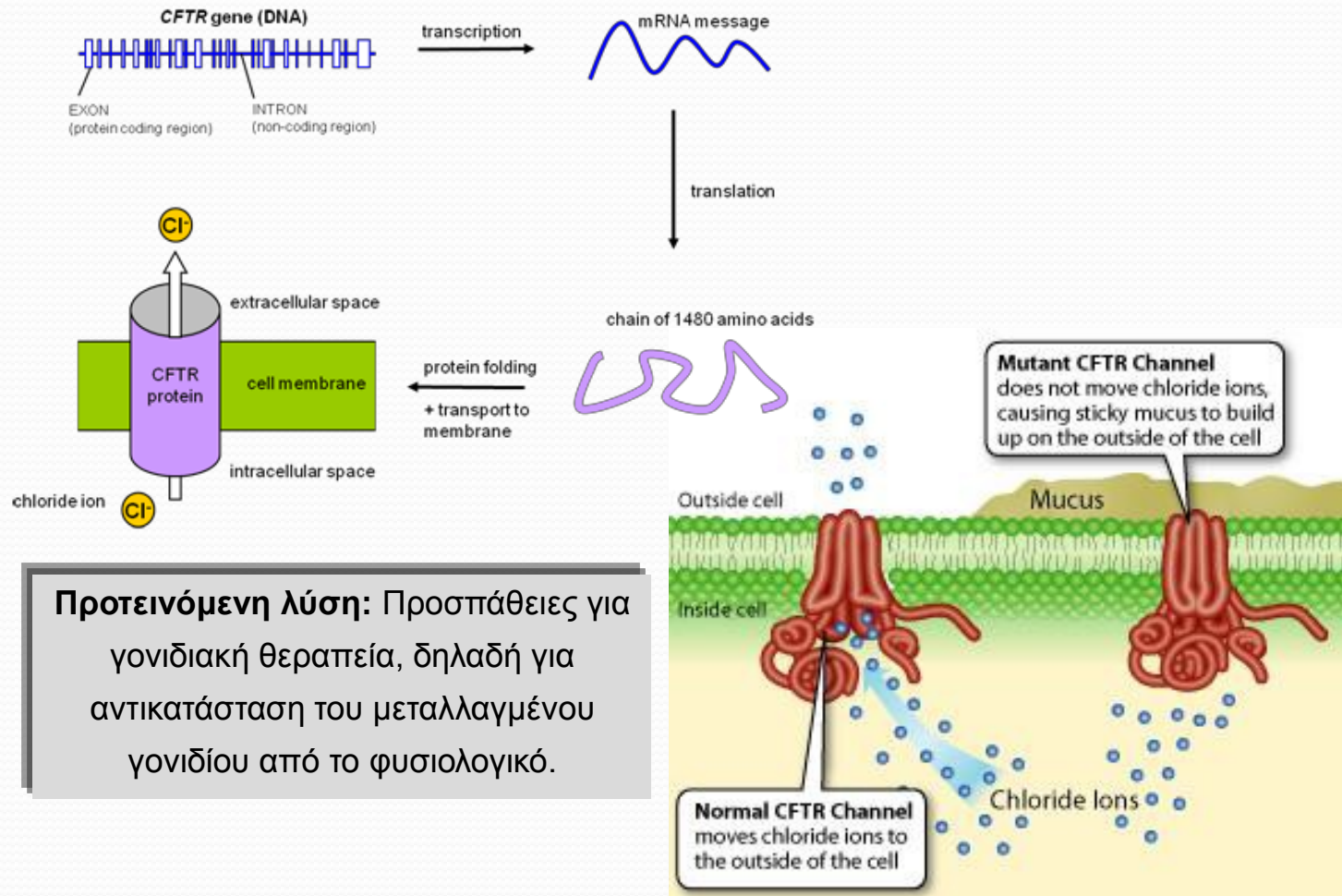
Nucleotide	ATC	AT	C T T	T	GGT	GTT
Amino Acid	Ile	Ile	Phe		Gly	Val
	506		508			510

Deleted in  $\Delta F508$

## $\Delta F508$ CFTR Sequence:

Nucleotide	ATC	ATT	GGT	GTT
Amino Acid	Ile	Ile	Gly	Val
	506			

# Κυστική ίνωση: Μεταλλαγμένη πρωτεΐνη



**Προτεινόμενη λύση:** Προσπάθειες για γονιδιακή θεραπεία, δηλαδή για αντικατάσταση του μεταλλαγμένου γονιδίου από το φυσιολογικό.

# Γιατί σύγκριση αλληλουχιών;;

- **Συμπεράσματα για λειτουργία ενός γονιδίου (μίας πρωτεΐνης).**
  - Όταν δύο γονίδια έχουν μεγάλο ποσοστό ομοιότητας, τότε πιθανά κωδικοποιούν πρωτεΐνες με παρόμοια λειτουργία.
- **Συμπεράσματα για σημαντικές περιοχές στην αλληλουχία γονιδίων/πρωτεϊνών.**
  - Όταν πρωτεΐνες με συγκεκριμένο χαρακτηριστικό μοιράζονται κοινή περιοχή, τότε αυτή η κοινή περιοχή είναι πιθανά υπεύθυνη για το συγκεκριμένο χαρακτηριστικό.
- **Συμπεράσματα για την εξελεγκτική απόσταση μεταξύ 3 ειδών.**
  - Όταν η αλληλουχία μίας πρωτεΐνης είναι σχεδόν η ίδια μεταξύ 2 ειδών (ποντίκι και αρουραίος), τότε τα είδη αυτά είναι «κοντά».

# Γονίδια που κωδικοποιούν πρωτεΐνες με μεμβρανική εντόπιση

## Sequence 1:

TCGCTGCGAAGGACATTTGGGCTGTGTGTGCGACGCGGGTCTGGAGGGGCGAGTCGGGGGAACCGCGAAGAAGCCGAGGAGCCCCGGAGC  
CCCGCGTGACGCTCCTCTCTCAGTCCAAAAGCGGCTTTTGGTTGCGCGCAGAGAGACCCGGGGTCTTCAGGACAGCGATTGTCATTGCT  
GAAGCTTTTCTCGAAAAGCGCCGCCCTGCCCTTGCCCCGAGAACAGACAAAAGAGCACCCGCAGGGCCGATCACGCTGGGGGCGCTGA  
GGCCGGCCATGGTCATGGAAGTGGGCACCCCTGGACGCTGGAGGCCTGCGGGCGCTGCTGGGGGAGCGAGCGGCGCAATGCCTGCTGC  
TGGACTGCCGCTCCTTCTTCGCTTTCAACGCCGGCCACATCGCCGGCTCTGTCAACGTGCGCTTCAGCACCATCGTGCGGCGCCGGGCC  
AAGGGCGCCATGGGCCTGGAGCACATCGTGCCCAACGCCGAGCTCCGCGGTGACGACAGCGATTGTCATTGCTGACCGCCTGCTGGCC  
GGCGCCTACCACGCCGTGGTGTGCTGGACGAGCGCAGCGCCGCCCTGGACGGCGCCAAGCGCGACGGCACCCCTGGCCCTGGCGGCC  
GGCGCGCTCTGCCGCGAGGGCGCGCGCCGCGCAAGTCTTCTCCTCAAAGGAGGATACGAAGCGTTTTTCGGCTTCTGCCCGGAGCTCAG  
GACAGCGATTGTCATTGCTGATGTGCAGCAAACAGTCGACCCCATGGGGCTCAGCCTTCCCCTGAGTACTAGCGTCCCTGACAGCGCGG  
AATCTGGGTGCAGTTCCTGCAGTACCCCACTCTACGATCAGGGTGGCCCGGTGAAATCCTGCCCTTCTGTACCTGGGCAGTGCCTATC  
ACGCTTCCCGCAAGGACATGCTGGATGCCTTGGGCA

## Sequence 2:

TAACTGCCTTGATCAACGTCTCAGCCAATTGTCCCAACCATTTTGGAGGGTCACTACCAGTACAAGAGCATCCCTGTGGAGGACAACCACAA  
GGCAGACATCAGCTCCTGGTTCAACGAGGCCATTGACTTCATAGACTCCATCAAGAATGCTGGAGGAAGGGTGTGTTGCCACTGCCAGGC  
AGGCATTTCCCGGTCAGCCACCATCTGCCTTGCTTACCTTATGAGGACTAATCGAGTCATCAGGACAGCGATTGTCATTGCTGAAGCTGGA  
CGAGGCCTTTGAGTTTGTGAAGCAGAGGCCGAAGCATCATCTCTCCAACTTCAGCTTCATGGGCCAGCTGCTGCAGTTTGTGAGTCCCAGGT  
GCTGGCTCCGCACTGTTGCGCAGAGGCTGGGAGCCCCGCCATGGCTGTGCTCGACCGAGGCACCTCCACCACCACCGTGTCAACTTTC  
AGGACAGCGATTGTCATTGCTGACCCCGTCTCCATCCCTGTCCACTCCACGAACAGTGCCTGAGCTACCTTCAGAGCCCCATTACGACC  
TCTCCCAGCTGCTGAAAGGCCACGGGAGGTGAGGCTCTTACATCCCATTTGGGACTCCATGCTCCTTGAGAGGAGAAATGCAATAACTCT  
GGGAGGGGCTCAGGACAGCGATTGTCATTGCTGATCGAGAGGGTGGTCTTATTTATTTAACTTCACCCGAGTTCCTCTGGGTTTCTAAG  
CAGTTATGGTGATGACTTAGCGTCAAGACATTTGCTGAACCTCAGCACATTCGGGACCAATATATAGTGGGTACATCAAGTCCATCTGACAAA  
ATGGGGCAGAAGAGAAAGGACTCAGTGTGTGATCCGGTTTTCTTTTTGCTCGCCCCTGTTTTTTGTAGAATCTCTTCATGCTTGACATACCTA  
CCAGTATTATTTCCCGACGACACATATACATATGAGAATATACCTTATTTATTTTTGTGTAGGTGTCTGCCTTACAAATGTCATTGTCTACTCC  
TAGAAGAACCAAATACCTCAATTTTTGTTTTGAGTACTGTACTATCCTGTAAATATATCTTAAGCAGTTTTGTTTTCA

# Γονίδια που κωδικοποιούν πρωτεΐνες με μεμβρανική εντόπιση

## Sequence 1:

TCGCTGCGAAGGACATTTGGGCTGTGTGTGCGACGCGGGTCCGGAGGGGCAGTCCGGGGGAACCGCGAAGAAGCCGAGGAGCCCCGGAGCCCCGCGTGAC  
GCTCCTCTCTCAGTCCAAAAGCGGCTTTTGGTTCCGGCGCAGAGAGACCCGGGGTCTTCAGGACAGCGATTGTCATTGCTGAAGCTTTTCTCGAAAAGC  
GCCGCCCTGCCCTTGGCCCCGAGAACAGACAAAGAGCACCCGAGGGCCGATCACGCTGGGGGCGCTGAGGCCGGCCATGGTCATGGAAGTGGGCACC  
CTGGACGCTGGAGGCCTGCGGGCGCTGCTGGGGGAGCGAGCGGCCAATGCCTGCTGCTGGACTGCCGCTCCTTCTTCGCTTTCAACGCCGGCCACAT  
CGCCGGCTCTGTCAACGTGCGCTTCAGCACCATCGTGCGGCGCCGGGCCAAGGGCGCCATGGGCCTGGAGCACATCGTGCCCAACGCCGAGCTCCGCG  
GTCAGGACAGCGATTGTCATTGCTGACCCGCTGCTGGCCGGCGCCTACCACGCCGTGGTGTGCTGGACGAGCGCAGCGCCGCCCTGGACGGCGCCAA  
GCGCGACGGCACCCCTGGCCCTGGCGGCCGGCGCGCTCTGCCGCGAGGGCGCGCGCCGCGCAAGTCTTCTCCTCAAAGGAGGATACGAAGCGTTTTCGG  
CTTCTGCCCGGAGCTCAGGACAGCGATTGTCATTGCTGATGTGCAGCAAACAGTCGACCCCATGGGGCTCAGCCTTCCCCTGAGTACTAGCGTCCCTG  
ACAGCGCGGAATCTGGGTGCAGTTCCTGCAGTACCCACTCTACGATCAGGGTGGCCCGGTGGAAATCCTGCCCTTCTGTACCTGGGCAGTGCATCA  
CGCTTCCCGCAAGGACATGCTGGATGCCTTGGGCA

## Sequence 2:

TAACCTGCCTTGATCAACGTCTCAGCCAATTGTCCCAACCATTTGAGGGTCACTACCAGTACAAGAGCATCCCTGTGGAGGACAACCACAAGGCAGACATC  
AGCTCCTGGTTCAACGAGGCCATTGACTTCATAGACTCCATCAAGAATGCTGGAGGAAGGGTGTGTTGTCCTACTGCCAGGCAGGCATTTCCCGGTCAGCCA  
CCATCTGCCTTGCTTACCTTATGAGGACTAATCGAGTCAATCAGGACAGCGATTGTCATTGCTGAAGCTGGACGAGGCCTTTGAGTTTGTGAAGCAGAGGCG  
AAGCATCATCTCTCCCAACTTCAGCTTCATGGGCCAGCTGCTGCAGTTTGTGAGTCCCAGGTGCTGGCTCCGCACTGTTCCGGCAGAGGCTGGGAGCCCCGC  
CATGGCTGTGCTCGACCGAGGCACCTCCACCACCACCGTGTCAACTTATCAGGACAGCGATTGTCATTGCTGACCCCGTCTCCATCCCTGTCCACTCCAC  
GAACAGTGCCTGAGCTACCTTCAGAGCCCCATTACGACCTCTCCAGCTGCTGAAAGGCCACGGGAGGTGAGGCTCTTCACATCCCATTGGGACTCCAT  
GCTCCTTGAGAGGAGAAATGCAATAACTCTGGGAGGGGCTCAGGACAGCGATTGTCATTGCTGATCGAGAGGGCTGGTCCCTATTTATTTAACTTCAACCG  
AGTTCTCTGGGTTTCTAAGCAGTTATGGTGATGACTTAGCGTCAAGACATTTGCTGAACTCAGCACATTCGGGACCAATATATAGTGGGTACATCAAGTCC  
ATCTGACAAAATGGGGCAGAAGAGAAAGGACTCAGTGTGTGATCCGGTTTCTTTTTGCTCGCCCCTGTTTTTGTAGAATCTTTCATGCTTGACATACCTA  
CCAGTATTATTCGACGACACATATACATATGAGAATATACCTTATTTATTTTTGTGTAGGTGTCTGCCTTACAAAATGTCATTGTCTACTCCTAGAAGAACC  
AAATACCTCAATTTTTGTTTTGAGTACTGTACTATCCTGTAATAATATCTTAAGCAGGTTTGTTTTCA

- **Συμπέρασμα:** Πιθανά η όμοια περιοχή να είναι υπεύθυνη για την κυτταρική εντόπιση των 2 πρωτεϊνών.

# Σύγκριση αλληλουχιών κατά ζεύγη

Scarites	C	T	T	A	G	A	T	C	G	T	A	C	C	A	A	-	-	-	A	A	T	A	T	T	A	C
Carenum	C	T	T	A	G	A	T	C	G	T	A	C	C	A	C	A	-	T	A	C	-	T	T	T	A	C
Pasimachus	A	T	T	A	G	A	T	C	G	T	A	C	C	A	C	T	A	T	A	A	G	T	T	T	A	C
Pheropsophus	C	T	T	A	G	A	T	C	G	T	T	C	C	A	C	-	-	-	A	C	A	T	A	T	A	C
Brachinus armiger	A	T	T	A	G	A	T	C	G	T	A	C	C	A	C	-	-	-	A	T	A	T	A	T	T	C
Brachinus hirsutus	A	T	T	A	G	A	T	C	G	T	A	C	C	A	C	-	-	-	A	T	A	T	A	T	A	C
Aptinus	C	T	T	A	G	A	T	C	G	T	A	C	C	A	C	-	-	-	A	C	A	A	T	T	A	C
Pseudomorpha	C	T	T	A	G	A	T	C	G	T	A	C	C	-	-	-	-	-	A	C	A	A	A	T	A	C

# Σύγκριση αλληλουχιών κατά ζεύγη

- Η σύγκριση μεταξύ δύο ή περισσότερων αλληλουχιών γίνεται με την τοποθέτηση της **μίας αλληλουχίας κάτω από την άλλη** με τέτοιο τρόπο ώστε να δώσουν το καλύτερο αποτέλεσμα.
- **Όμοιοι ή παρόμοιοι** χαρακτήρες τοποθετούνται ο ένας κάτω από τον άλλο.
- **Ανόμοιοι χαρακτήρες** μπορεί να βρίσκονται στην ίδια στήλη ή να στοιχίζονται με κενά διαστήματα.

DUSP1  
SEQ2

```
AGAGCCCCATTACGACCTCTCCCAGCTGCTGAAAGGCCACGGGAGGTGAGGCTCTTCACA  
AGAGCCCCATTACGACCTCTCCCAGCTGC----AGACTCGAGGAGCAAAAAGCTCAT----  
***** ** * **** * **** *
```



# Βασικές Πράξεις

- Βασικές πράξεις μετασχηματισμού:
  - Ένθεση – Insertion (I).
  - Διαγραφή – Deletion (D).
  - Αντικατάσταση συμβόλων – Replacement (R).
- **Απόσταση μετασχηματισμού (Edit distance)** μεταξύ δύο συμβολοσειρών: Το ελάχιστο πλήθος των πράξεων μετασχηματισμού που απαιτούνται για να μετασχηματίσουμε την πρώτη συμβολοσειρά στη δεύτερη.



# Βασικές Πράξεις

1.  $S_1$ : vintner.
  2.  $S_2$ : writers.
- **Σκοπός:** Να μετασχηματίσουμε την  $S_1$  στην  $S_2$ .
  - **Βασικές πράξεις μετασχηματισμού:**
    - a. Να αντικαταστήσουμε το “v” με το “w” (vintner → wintner).
    - b. Να εισάγουμε το “r” (wintner → wrintner).
    - c. Να διαγράψουμε το “n” δύο φορές (wrintner → writer).
    - d. Να εισάγουμε το “s” (writer → writers).
  - **Συνολικά:** 5 βασικές πράξεις μετασχηματισμού.
  - **Edit-distance ( $S_1 \rightarrow S_2$ ) = 5.**

# Βασικές Πράξεις

- **Ακολουθία μετασχηματισμού (Edit transcript):** Η ακολουθία των πράξεων μετασχηματισμού που απαιτούνται για να μετασχηματίσουμε την πρώτη συμβολοσειρά στη δεύτερη.
- Οι βασικές πράξεις μετασχηματισμού αναπαρίστανται ως εξής:
  - Ένθεση (insertion), I.
  - Διαγραφή (deletion), D.
  - Αντικατάσταση (replacement), R.
  - Ταίριασμα (match), M.

R	I	M	D	M	D	M	M	I
V		I	N	T	N	E	R	
W	R	I		T		E	R	S

# Βασικές Πράξεις

- Η ακολουθία μετασχηματισμού αποτελεί μία συμβολοσειρά με τα στοιχεία  $\Sigma = \{D, I, M, R\}$ .
- Για δύο συμβολοσειρές δεν υπάρχει μία μοναδική ακολουθία μετασχηματισμού.
- **Στόχος:** Βέλτιστη ακολουθία (Optimal Edit Transcript), δηλ. αυτή που αντιστοιχεί στον ελάχιστο δυνατό αριθμό πράξεων μετασχηματισμού.

R	I	M	D	M	D	M	M	I
V		I	N	T	N	E	R	
W	R	I		T		E	R	S



Ακολουθία μετασχηματισμού  
 $u = \text{RIMDMDMMI}$

# Βασικές Πράξεις

- Κάθε πράξη μετασχηματισμού έχει συγκεκριμένο κόστος-βάρος.
- **Ζυγισμένη απόσταση μετασχηματισμού (Weighted Edit Distance) μεταξύ δύο συμβολοσειρών:** Το ελάχιστο πλήθος των πράξεων μετασχηματισμού που απαιτούνται για να μετασχηματίσουμε την πρώτη συμβολοσειρά στη δεύτερη.
- Έστω ότι οι βασικές πράξεις μετασχηματισμού έχουν τα ακόλουθα βάρη:
  - Ένθεση ή διαγραφή:  $d$ ,
  - Αντικατάσταση:  $r$
  - Ταίριασμα:  $m$
- **Παράδειγμα:**  $\text{Weighted-edit-distance}(S_1 \rightarrow S_2) = 1r + 4d + 4m$ .

R	I	M	D	M	D	M	M	I
V		I	N	T	N	E	R	
W	R	I		T		E	R	S

# Example - 2

- **A = interestingly.**
- **B = bioinformatics.**

A = - i - - n t e r e s t i n g l y

B = b i o i n f o r m a t i c s - -

---

1 0 1 1 0 1 1 0 1 1 0 0 1 1 1 1

	-	A	C	G	T
-		1	1	1	1
A	1	0	1	1	1
C	1	1	0	1	1
G	1	1	1	0	1
T	1	1	1	1	0

# Βασικές Πράξεις

- **Ζυγισμένη απόσταση μετασχηματισμού βάσει αλφαβήτου (weighted edit distance), βάσει αλφαβήτου:** Το ελάχιστο πλήθος των πράξεων μετασχηματισμού που απαιτούνται για να μετασχηματίσουμε την πρώτη συμβολοσειρά στη δεύτερη. Κάθε πράξη μετασχηματισμού έχει **συγκεκριμένο κόστος-βάρος ανάλογα με το χαρακτήρα που μετασχηματίζουμε.**
- Εφαρμόζεται κυρίως **στα προβλήματα στοίχισης ακολουθιών DNA και πρωτεϊνών**, όπου χρησιμοποιούνται συγκεκριμένοι πίνακες αντικατάστασης, οι οποίοι ορίζουν το κόστος μετασχηματισμού του κάθε χαρακτήρα.

# Στοιχίση αλληλουχιών κατά ζεύγη

- Οι περισσότερες μέθοδοι στοιχίσης επιχειρούν να προσομοιώσουν τους διάφορους εξελεγκτικούς μηχανισμούς.
- Οι στοιχειώδεις αλλαγές σε σχέση με το προγονικό μόριο χαρακτηρίζονται ως:
  - αντικαταστάσεις (replacements),
  - προσθήκες (insertions),
  - εξαλείψεις (deletions),
- αμινοξικών καταλοίπων ή βάσεων.

# Στοιχίση αλληλουχιών κατά ζεύγη

## Αντικατάσταση

AAT	CTC	AAA	CAT	GGC	← DNA 1
N	L	K	H	G	← Protein 1
AGT	CTA	AAA	TAT	GGC	← DNA 2
S	L	K	Y	G	← Protein 2

## Προσθήκη Ένθεση

AAT	CTC	AAA	CAT	GGC	← DNA 1
N	L	K	H	G	← Protein 1
AAT	GCT	CAA	ACA	TGG	← DNA 2
N	A	Q	T	W	← Protein 2

## Εξάλειψη Διαγραφή

AAT	CTC	AAA	CAT	GGC	← DNA 1
N	L	K	H	G	← Protein 1
AAT	CTA	AAC	ATG	GCC	← DNA 2
N	L	N	M	A	← Protein 2

N: Asparagine, L: Leucine, K: Lysine, H: Histidine, G: Glycine, A: Alanine, Q: Glutamine,  
T: Threonine, W: Tryptophan



# Στοιχισή αλληλουχιών κατά ζεύγη

ancestral  
sequence

A C G T C A T C A

mutation

A C G T **G** A T C A

deletion

A **X** G T G **X** T C A

A G T G T C A

insertion

**T** A G T G T C A

derived  
sequence

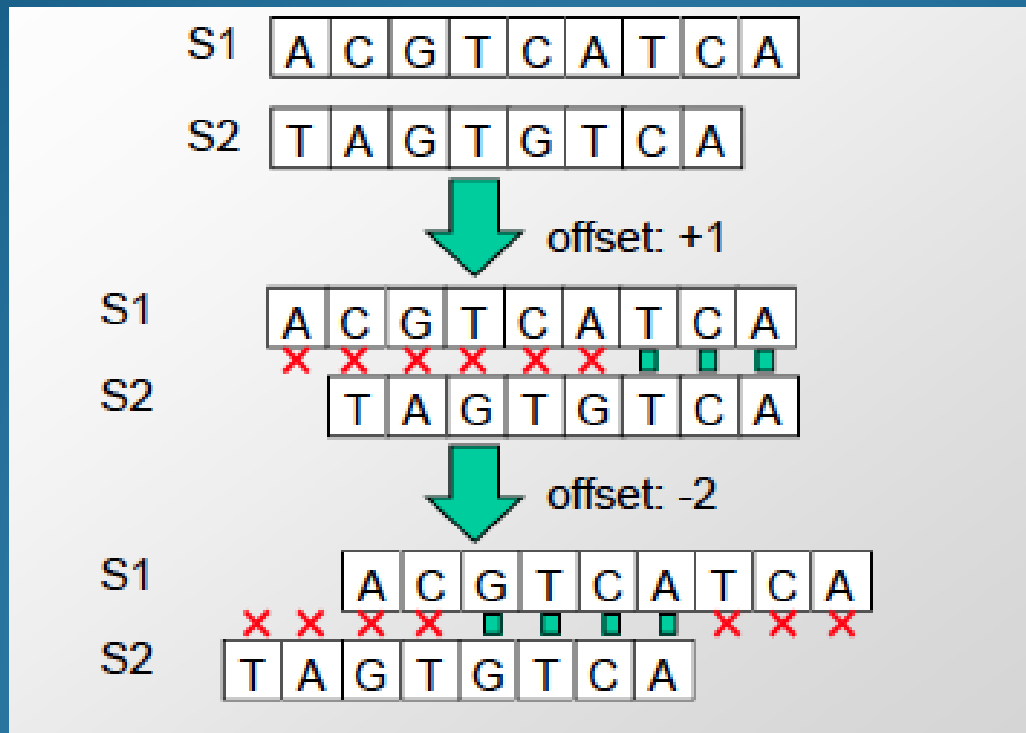
T A G T G T C A

# Στοιχίση αλληλουχιών κατά ζεύγη

- Κατά τη στοιχίση:
  - **ανόμοιοι χαρακτήρες** που έχουν στοιχηθεί, αντιπροσωπεύουν **αντικαταστάσεις** (συνήθως συμβαίνει μεταξύ αμινοξέων με παρόμοιες ιδιότητες),
  - **περιοχές που δεν μπορούν να στοιχηθούν** εμφανίζονται ως κενά διαστήματα σε μία από τις δύο αλληλουχίες (**προσθήκες** στη μία ή **εξαλείψεις** στην άλλη).
- Τα κενά εισέρχονται με τέτοιο τρόπο ώστε να μεγιστοποιείται το ταίριασμα στα προηγούμενα ή επόμενα τμήματα.

# Εισαγωγή κενών

- Longest Common Substring – Μεγαλύτερη κοινή υπο-συμβολοσειρά.



# Στοιχίση αλληλουχιών

Παράδειγμα: DESCRIBING & DESINGING

D E S C R I B I N G

| | |

D E S I N G I N G

Gaps



D E S C R I B I N G

| | | | | | | |

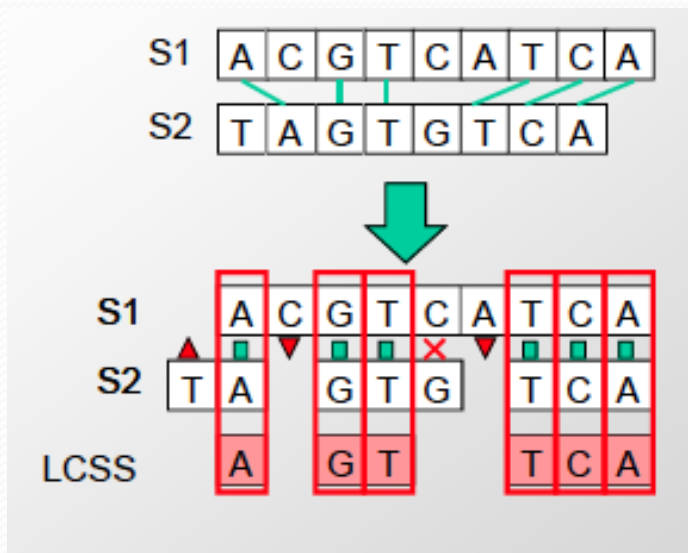
D E S I N G - I N G

# Στοιχίση αλληλουχιών

- Βιολογικές αλληλουχίες:
  - Διαφορετικό μήκος.
  - Περιοχές προσθήκης / διαγραφής.
  - **Παράδειγμα:**
    - Seq 1: CATTTGCTGAACTCAGCACATTCGGGACCAATATATAGTGGGT
    - Seq 2: CATTTTGCTCAGCACATTCGGATATATAG
- **Alignment with gaps**
  - Seq 1: CATTTGCTGAACTCAGCACATTCGGGACCAATATATAGTGGGT
  - Seq 2: CATTT - - TG- - CTCAGCACATTCGG - - - - - ATATATAG

# Εισαγωγή κενών

- **Longest Common Subsequence** (and **not substring**) – Μεγαλύτερη κοινή υποαλληλουχία (και όχι συμβολοσειρά).
- **Definition:** Given a sequence  $X = (x_1, \dots, x_m)$ , we formally define  $Z = (z_1, \dots, z_k)$  to be a subsequence of  $X$  if there exists a strictly increasing sequence  $i_1 < i_2 < \dots < i_k$  of indices of  $X$  such that for all  $j$ ,  $1 < j < k$ , we have  $x_{i_j} = z_j$ .



# Ολική και τοπική στοίχιση

- Sequence 1: C T G T C G C T G C A C G
- Sequence 2: T G C C G T G

1. **Ολική στοίχιση (global alignment):** Έχει στόχο να περιλάβει όσο το δυνατόν περισσότερους χαρακτήρες σε όλο το μήκος των δύο αλληλουχιών.

```
CTGTCGCTGCACG
-TG-C-C-G--TG
```

2. **Τοπική στοίχιση (local alignment):** Δημιουργούνται «νησίδες» στοίχισης από μεμονωμένες περιοχές που εμφανίζουν ομοιότητα, χωρίς να δίνεται ιδιαίτερα βάρος στην επέκταση της στοίχισης σε όλο το μήκος των αλληλουχιών

```
CTGTCGCTGCACG
-TGCCG-TG-----
```

# Τοπική και ολική στοίχιση

- Οι μέθοδοι **τοπικής στοίχισης** είναι ιδιαίτερα χρήσιμες όταν:
  - εξετάζονται δύο αλληλουχίες διαφορετικού μήκους,
  - εξετάζονται αλληλουχίες που **δεν εμφανίζουν ομοιότητα σε όλο το μήκος τους**, αλλά **τοπική ομοιότητα** (π.χ. όταν οι πρωτεΐνες έχουν κοινές αυτοτελείς δομικές ή λειτουργικές περιοχές (domains)).
- Στις περιπτώσεις αυτές οι μέθοδοι ολικής στοίχισης μπορεί να δώσουν αποτελέσματα που **στερούνται βιολογικής σημασίας**.

Seq 1



Seq 2



Ολική στοίχιση



Τοπική στοίχιση

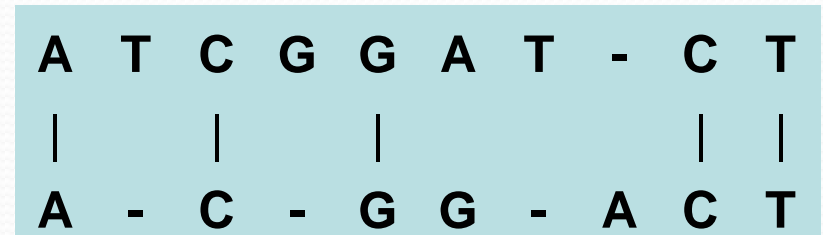
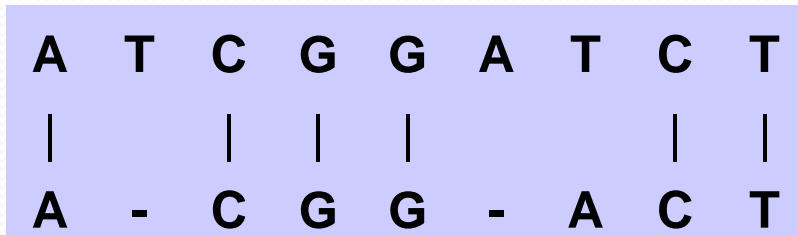


# Στοιχίση αλληλουχιών DNA & RNA

Ποια στοιχίση είναι η καλύτερη;

1. A T C G G A T C T

2. A C G G A C T



# Συστήματα βαθμονόμησης

## - Scoring system

- Κατάλληλο σύστημα βαθμονόμησης (**scoring system**).
- Αποδίδεται συγκεκριμένη βαθμολογία στη:
  - στοίχιση κάθε ζεύγους χαρακτήρων, και
  - στην εισαγωγή κενών.
- Τελικά, απεικονίζεται η στοίχιση με τη μεγαλύτερη βαθμολογία.
- Η επιλογή του συστήματος βαθμονόμησης είναι καθοριστικής σημασίας για το αποτέλεσμα της στοίχισης.

# Πειραματικά γνωρίζουμε ότι:

- Οι ενθέσεις και οι διαγραφές είναι περισσότερες πιθανές από τις τοπικές μεταλλάξεις.
- Ορισμένες μεταλλάξεις είναι περισσότερο πιθανές από μερικές άλλες.
- **Απαραίτητο ένα σύστημα βαθμονόμησης. Ποιο;;**

# Σύστημα βαθμολόγησης στοίχισης - DNA & RNA

Sequence 1: ATCGGATCT

Sequence 2: ACGGACT

ΠΡΟΣΟΧΗ: Συγκρίνουμε νουκλεοτίδια, μόνο 4 πιθανές βάσεις

A	T	C	G	G	A	T	-	C	T	A	T	C	G	G	A	T	C	T
A	-	C	-	G	G	-	A	C	T	A	-	C	G	G	A	-	C	T

- Πιθανό σύστημα βαθμολόγησης:
  - όμοιο κατάλοιπο: +2
  - διαφορετικό κατάλοιπο: -1
  - κενό: -2
- **Alignment 1:**  $5 \times 2 - 1(1) - 4(2) = 10 - 1 - 8 = 1$
- **Alignment 2:**  $7 \times 2 - 0(1) - 2(2) = 14 - 0 - 4 = 10$

# Σύστημα βαθμονόμησης στοίχισης - DNA & RNA

- Πιθανό σύστημα βαθμονόμησης:
  - όμοιο κατάλοιπο: +2
  - διαφορετικό κατάλοιπο: -1
  - κενό: -2

	-	A	C	G	T
-		-2	-2	-2	-2
A	-2	+2	-1	-1	-1
C	-2	-1	+2	-1	-1
G	-2	-1	-1	+2	-1
T	-2	-1	-1	-1	+2

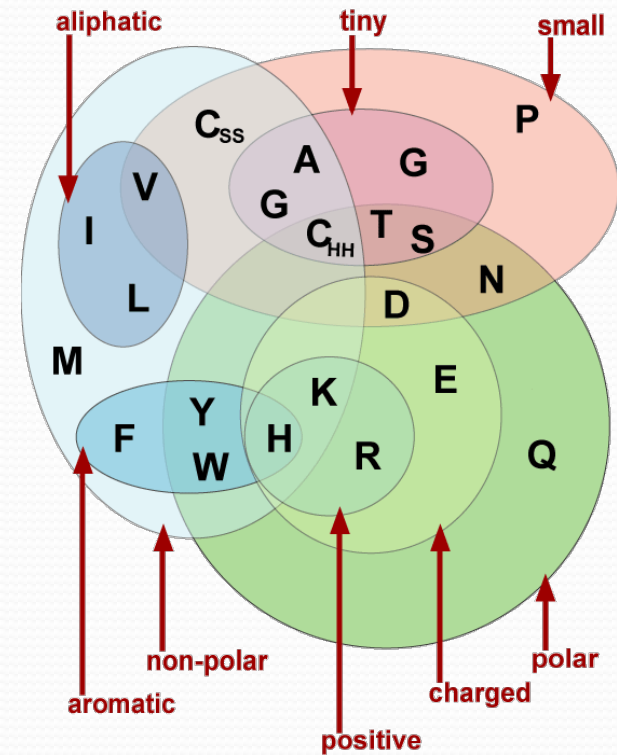
# Στοιχίση πρωτεϊνών

A	S	K	T	M	P	I	A	S	K	T	M	P	I
I	I	?	?	I	I	I	I	I	?	?	I	I	I
A	S	R	H	M	P	I	A	S	Y	H	M	P	I

- Πώς στοιχίζουμε πρωτεΐνες.
- Αμινοξέα με παρόμοιες ιδιότητες.
- Αμινοξέα με διαφορετικές ιδιότητες.

# Σύστημα βαθμονόμησης στοίχισης - Πρωτεΐνες

- Περισσότερο πολύπλοκα συστήματα βαθμονόμησης.
- Η αντικατάσταση αμινοξέων παρόμοιων φυσικοχημικών ιδιοτήτων = **συντηρητική αντικατάσταση**.
- Η συντηρητική αντικατάσταση προτιμάται από τη φύση.
- Επιρροή στη δομή πρωτεΐνης:
  - ✓ Αντικατάσταση υδροφιλικού από υδροφιλικό.
  - ✓ Αντικατάσταση υδροφιλικού από υδροφοβικό.



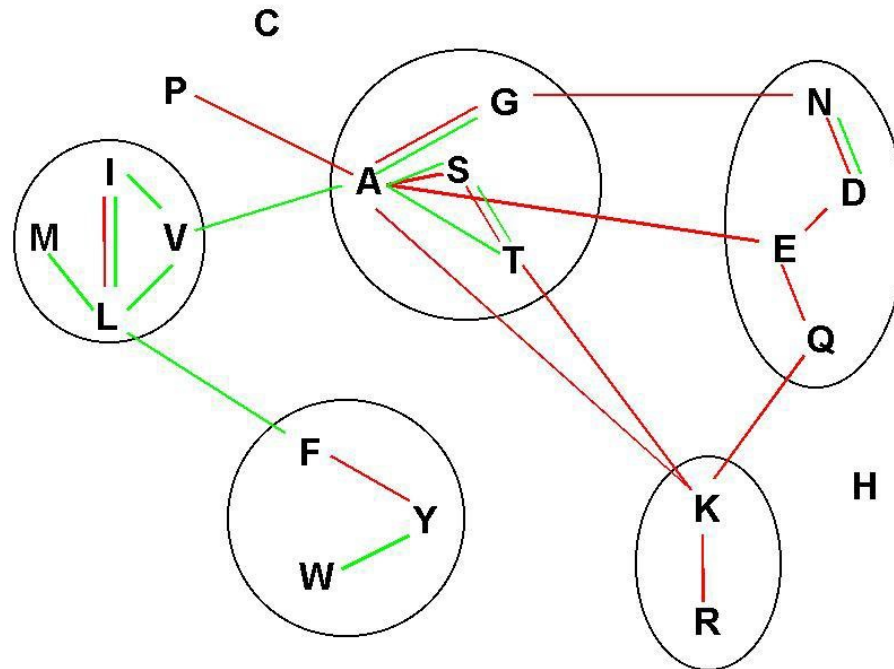
# Σύστημα βαθμονόμησης σ τοίχισης

- Η αλληλουχία που προκύπτει από αντικατάσταση αμινοξέος με παρόμοιες φυσικοχημικές ιδιότητες μοιάζει περισσότερο στην αρχική αλληλουχία.
- Αμινοξέα με **παρόμοιες φυσικοχημικές ιδιότητες** έχουν **περισσότερες** πιθανότητες να **αντικαταστήσουν** το ένα το άλλο κατά τη διάρκεια της εξέλιξης.
- Η **σύγκριση δύο αμινοξέων** (ένα από τη στήλη και ένα από τη γραμμή) **βαθμολογείται** από την **πιθανότητα** που έχει η **αντικατάστασή** τους να παρατηρηθεί στη φύση.



# Αντικαταστάσεις αμινοξέων

**Suggested Amino Acid Substitutions**  
solvent exposed ( $SEA^a > 30 \text{ \AA}^2$ ) / interior ( $SEA^a < 10 \text{ \AA}^2$ )



Amino acids connected by a solid line can be substituted with 95% confidence  
(D. Bordo and P. Argos, J. Mol. Biol. 217(1991)721-729)

<sup>a</sup>SEA=solvent exposed area

# Κατασκευή συστήματος βαθμονόμησης - Κριτήρια

## «Βαθμολογία» ενός ζεύγους αμινοξέων:

1. Πιθανότητα εύρεσης του συγκεκριμένου ζεύγους αμινοξικών καταλοίπων σε στοιχίσεις πρωτεϊνών που σχετίζονται βιολογικά.
2. Πιθανότητα τυχαίου ταιριάσματος του συγκεκριμένου ζεύγους, σύμφωνα με τη συχνότητα εμφάνισης των αμινοξικών καταλοίπων (συχνότητες υποβάθρου).
3. Πιθανότητα να είναι προτιμότερη η εισαγωγή ενός κενού σε μία από τις δύο αλληλουχίες.

Τελικά, το σκορ δείχνει πόσο πιθανό είναι ένα αμινοξύ να στοιχίζεται με ένα άλλο αμινοξύ σε συγγενικές αλληλουχίες.

# Observed amino acid frequency

Amino Acids	Codons	Observed Frequency in Vertebrates
Alanine	GCU, GCA, GCC, GCG	7.4 %
Arginine	CGU, CGA, CGC, CGG, AGA, AGG	4.2 %
Asparagine	AAU, AAC	4.4 %
Aspartic Acid	GAU, GAC	5.9 %
Cysteine	UGU, UGC	3.3 %
Glutamic Acid	GAA, GAG	5.8 %
Glutamine	CAA, CAG	3.7 %
Glycine	GGU, GGA, GGC, GGG	7.4 %
Histidine	CAU, CAC	2.9 %
Isoleucine	AUU, AUA, AUC	3.8 %
Leucine	CUU, CUA, CUC, CUG, UUA, UUG	7.6 %
Lysine	AAA, AAG	7.2 %
Methionine	AUG	1.8 %
Phenylalanine	UUU, UUC	4.0 %
Proline	CCU, CCA, CCC, CCG	5.0 %
Serine	UCU, UCA, UCC, UCG, AGU, AGC	8.1 %
Threonine	ACU, ACA, ACC, ACG	6.2 %
Tryptophan	UGG	1.3 %
Tyrosine	UAU, UAC	3.3 %
Valine	GUU, GUA, GUC, GUG	6.8 %
Stop Codons	UAA, UAG, UGA	---

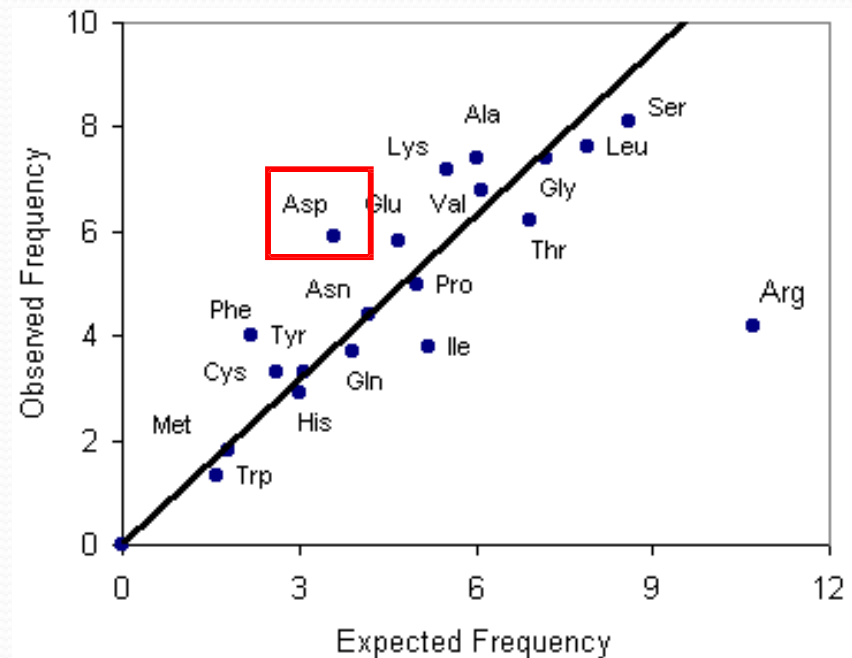
# Expected amino acid frequency

- Συχνότητα εμφάνισης νουκλεοτιδίων:
  - Αδενίνη, A: 30,3% = 0,303
  - Γουανίνη, G: 26.1% = 0.261
  - Κυτοσίνη, C: 21.7 = 0.217
  - Θυμίνη, T ή Ουρακίλη, U: 22.0% = 0,220
- Παράδειγμα: Ασπαρτικό οξύ.
- GAU, GAC.
- **Αναμενόμενη πιθανότητα εμφάνισης:**

$$p = 0.261 \times 0.303 \times 0.220 + 0.261 \times 0.303 \times 0.217$$

$$p = 0.0174 + 0.0172 = 0.03456, p = 3.456\%$$

**Ωστόσο στη φύση το ασπαρτικό οξύ συναντάται με συχνότητα 5,9%.**



# Observed amino acid frequency

Amino Acids	Codons	Observed Frequency in Vertebrates
Alanine	GCU, GCA, GCC, GCG	7.4 %
Arginine	CGU, CGA, CGC, CGG, AGA, AGG	4.2 %
Asparagine	AAU, AAC	4.4 %
Aspartic Acid	GAU, GAC	5.9 %
Cysteine	UGU, UGC	3.3 %
Glutamic Acid	GAA, GAG	5.8 %
Glutamine	CAA, CAG	3.7 %
Glycine	GGU, GGA, GGC, GGG	7.4 %
Histidine	CAU, CAC	2.9 %
Isoleucine	AUU, AUA, AUC	3.8 %
Leucine	CUU, CUA, CUC, CUG, UUA, UUG	7.6 %
Lysine	AAA, AAG	7.2 %
Methionine	AUG	1.8 %
Phenylalanine	UUU, UUC	4.0 %
Proline	CCU, CCA, CCC, CCG	5.0 %
Serine	UCU, UCA, UCC, UCG, AGU, AGC	8.1 %
Threonine	ACU, ACA, ACC, ACG	6.2 %
Tryptophan	UGG	1.3 %
Tyrosine	UAU, UAC	3.3 %
Valine	GUU, GUA, GUC, GUG	6.8 %
Stop Codons	UAA, UAG, UGA	---

%	A	R	N	K
7.4	A	5	-2	-1
4.2	R	-	7	3
4.4	N	-	-	7
7.2	K	-	-	6

## Πίνακας βαθμολογίας: Παράδειγμα

- Η αργινίνη (R) και η λυσίνη (K), παρόλο που είναι διαφορετικά αμινοξέα, έχουν θετική βαθμολογία.
- Γιατί;;
- Είναι αμινοξέα **θετικά φορτισμένα**, οπότε μία αντικατάσταση του ενός από το άλλο **δεν θα αλλάξουν ριζικά τη λειτουργία της πρωτεΐνης.**

A K R A N R

K A A A N K

-1	-1	-2	5	7	3	11
----	----	----	---	---	---	----

# Συντήρηση καταλοίπων

- Τα αμινοξέα αλλάζουν με τέτοιο τρόπο ώστε να διατηρηθούν οι φυσικοχημικές ιδιότητες του αρχικού καταλοίπου.
- Δηλαδή:
  - Πολικό σε πολικό.
    - asparagine → glutamine.
  - Μη πολικό σε μη πολικό.
    - alanine → valine.
  - Όμοιας συμπεριφοράς.
    - arginine → lysine (positive to positive).
    - aspartic acid → glutamic acid (negative to negative).

# Πίνακες βαθμονόμησης (scoring matrices) ή πίνακες αντικατάστασης

- Για την αντικατάσταση αμινοξέων:
  - PAM.
  - BLOSUM.
- Για την αντικατάσταση νουκλεοτιδίων (DNA):
  - Το DNA είναι λιγότερο συντηρημένο από τις πρωτεΐνες.
  - Δεν είναι αποτελεσματικό να συγκρίνουμε κωδικοποιούσες περιοχές, από τις οποίες προκύπτουν οι πρωτεΐνες, σε επίπεδο νουκλεοτιδίων.



# Στοιχισή αλληλουχιών κατά ζεύγη

- **Προσοχή:** Υπάρχουν διαφορετικές τριπλέτες νουκλεοτιδίων που δίνουν το ίδιο αμινοξύ (γενετικός κώδικας). Οπότε, διαφορά στην νουκλεοτιδική αλληλουχία δε σημαίνει αυτόματα και διαφορά στην πρωτεϊνική αλληλουχία.

GCA GAA TTAAA  
| | | \* | | | | \* | | \*  
GCG GAA TTG AAG

75% ομοιότητα

Ala Glu Leu Lys  
| | | |  
Ala Glu Leu Lys

100% ομοιότητα

# Πίνακες αντικατάστασης PAM (Percent Accepted Mutation-PAM or Dayhoff Matrices)

- Μελετήθηκαν από την Margaret Dayhoff.
- Αξιολογούν – βαθμολογούν την αντικατάσταση ενός αμινοξέος από ένα άλλο.
- Αξιολογήθηκαν:
  - 71 πρωτεϊνικά γκρουπ με,
  - 1572 αμινοξικές αντικαταστάσεις,
  - Οι αλληλουχίες είχαν 85% ομοιότητα τουλάχιστον.
- Η κατασκευή τους βασίζεται στις:
  - στοιχίσεις πολλών **όμοιων πρωτεϊνών**, μικρής εξελεγκτικής απόστασης,
  - αποδεκτές σημειακές μεταλλάξεις – accepted mutations (αντικατάσταση αμινοξέος από κάποιο άλλο που δεν επηρεάζει τη λειτουργία της πρωτεΐνης).



# Πίνακες αντικατάστασης PAM (Percent Accepted Mutation-PAM or Dayhoff Matrices)

- Μετρήθηκε ο αριθμός αλλαγής κάθε αμινοξέος σε όλα τα υπόλοιπα αμινοξέα σε κάθε γκρουπ.
  - π.χ. Serine → Threonine, μετράμε πόσες φορές παρατηρήθηκε η αλλαγή από Serine σε Theonine σε στοιχισμένες ακολουθίες ομόλογων αλληλουχιών.
- Διαιρέθηκε με τον παράγοντα «έκθεση σε μετάλλαξη».
- **Παράγοντας «έκθεση σε μετάλλαξη»:** Η συχνότητα εμφάνισης του αμινοξέος στο συγκεκριμένο γκρουπ και η εμφάνιση όλων των αλλαγών των υπολοίπων αμινοξέων σε κάθε 100 αμινοξέα.
- Προστέθηκαν όλοι οι παράγοντες κανονικοποίησης από όλα τα γκρουπ για κάθε αμινοξύ.
- Τετραγωνικοί πίνακες με είκοσι σειρές και είκοσι στήλες (αμινοξέα) δημιουργήθηκαν.

# PAM1 matrix

normalized probabilities multiplied by 10000

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
N	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
D	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
C	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Q	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
E	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
G	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
H	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
I	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
L	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
K	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
M	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
F	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
P	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
S	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
T	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
W	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Y	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
V	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

 : most mutable amino acids  
 : least mutable amino acids

# Πίνακες αντικατάστασης PAM - Παραδοχές

## Παραδοχές κατά την κατασκευή πίνακα PAM:

- Η κάθε αντικατάσταση ενός αμινοξέος από κάποιο άλλο αμινοξύ είναι ανεξάρτητη από οποιαδήποτε άλλη αντικατάσταση που ενδεχομένως μπορεί να έχει συμβεί σε κοντινή περιοχή (διαδικασία Markov – Markov process).
- Όλα τα αμινοξέα έχουν την ίδια πιθανότητα να μεταλλαχθούν
- Βασίζονται σε συγκρίσεις αλληλουχιών **με μικρή εξελεκτική απόσταση** για την εξαγωγή συμπερασμάτων για αλληλουχίες μεγάλης εξελικτικής απόστασης.

# Σειρά πινάκων PAM-n

- Οι τιμές που προκύπτουν στον πίνακα  $M_1$  αντιπροσωπεύουν την πιθανότητα μετάλλαξης ενός αμινοξέος στα 100 αμινοξέα, PAM1 = 1% percent accepted mutation.
- π.χ. PAM60: 60 μεταλλάξεις στα 100 αμινοξέα, PAM250: 250 μεταλλάξεις στα 100 αμινοξέα.

$$\underbrace{M_1 \cdot M_1 \cdot \dots \cdot M_1}_n \Rightarrow M_n$$

$M_n$ : Πίνακας αντικατάστασης πρωτεϊνών που έχουν υποστεί  $n$  μεταλλάξεις

PAM	0	30	80	110	200	250
% identity	100	75	50	60	25	20

Calculate PAM matrix: <http://www.bioinformatics.nl/tools/pam.html>

# Log odds PAMn matrices

- Πιθανότητα μετάλλαξης σύμφωνα με PAM250 Phe  $\rightarrow$  Tyr = 0.15.
- Διαιρώ με συχνότητα εμφάνισης Phe: 0.15 / 0.040 = 3.75.
- Υπολογίζω λογάριθμο με βάση το 10:  
 $\log_{10}3.75 = 0.57$ .
- **Πολλαπλασιάζω με 10:  $0.57 \times 10 = 5.7$ .**
- Πιθανότητα μετάλλαξης σύμφωνα με PAM250 Tyr  $\rightarrow$  Phe = 0.20.
- Διαιρώ με συχνότητα εμφάνισης Tyr: 0.20 / 0.030 = 6.7.
- Υπολογίζω λογάριθμο με βάση το 10:  
 $\log_{10}6.7 = 0.83$ .
- **Πολλαπλασιάζω με 10:  $0.83 \times 10 = 8.3$ .**
- **Average:  $5.7 + 8.3 = 7$ .**

$$\underbrace{M_1 \cdot M_1 \cdot \dots \cdot M_1}_{250} \Rightarrow M_{250} = \text{PAM 250}$$

# The log odds form (the mutation data matrix) of PAM250

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	12																				C
S	0	2																			S
T	-2	1	3																		T
P	-3	1	0	6																	P
A	-2	1	1	1	2																A
G	-3	1	0	-1	1	5															G
N	-4	1	0	-1	0	0	2														N
D	-5	0	0	-1	0	1	2	4													D
E	-5	0	0	-1	0	0	1	3	4												E
Q	-5	-1	-1	0	0	-1	1	2	2	4											Q
H	-3	-1	-1	0	-1	-2	2	1	1	3	6										H
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6									R
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5								K
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6							M
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5						I
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6					L
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4				V
F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9			F
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10		Y
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

Tyrosine: Y

Phenylalanine: F



# Επεξήγηση τιμής πίνακα

- Q (glutamine)→E (glutamic acid), τιμή = 2.
- Επειδή το σκορ του πίνακα πολλαπλασιάστηκε επί 10 για να μην έχουμε δεκαδικά:
  - το σκορ είναι 0.2.

## Συμπεπώς:

- $0.2 = \log_{10}(\text{σχετική αναμενόμενη τιμή μετάλλαξης})$ .
- Σχετική αναμενόμενη τιμή μετάλλαξης =  $10^{0.2} = 1.6$ .
- Πολλαπλασιάζω με τη συχνότητα εμφάνισης του Q,  $3.7\% = 0.037$   
 $1.6 \times 0.037 = 0.0592$ .
- **Ερμηνεία:** Η πιθανότητα μετάλλαξης από Q (glutamine)→E (glutamic acid) είναι 5.92% σύμφωνα με τον πίνακα PAM250.

# Στοίχιση πρωτεϊνών

A	S	K	T	M	P	I	A	S	K	T	M	P	I
		?	?						?	?			
A	S	R	A	M	P	I	M	S	Y	H	M	P	I

- PAM250:

$$S1=2+2+3+1+6+6+5=25$$

$$S2=2+2-4-1+6+6+5=16$$

# Σειρά πινάκων PAM-n

- Σειρά πινάκων PAM-n:
  - όπου n: αποδεκτές σημειακές μεταλλάξεις – εξελεγκτική απόσταση PAM.
- **PAM1**: 1 αποδεκτή μετάλλαξη στα 100 αμινοξέα.
- **PAM250**: 250 αποδεκτές μεταλλάξεις στα 100 αμινοξέα.
- **Μικρό n**: Μικρή εξελεγκτική απόσταση μεταξύ των αλληλουχιών (λίγες αντικαταστάσεις).
- **Μεγάλο n**: Μεγάλη εξελεγκτική απόσταση μεταξύ των αλληλουχιών (πολλές αντικαταστάσεις).

# Πίνακες αντικατάστασης PAM (Percent Accepted Mutation-PAM or Dayhoff Matrices)

- **Πίνακες PAM με μικρό n:** Περιμένουμε οι δύο εξεταζόμενες αλληλουχίες **να έχουν μεγάλο ποσοστό** ομοιότητας (μικρή εξελεγκτική απόσταση).
- **Πίνακες PAM με μεγάλο n:** Περιμένουμε οι δύο εξεταζόμενες αλληλουχίες **να μην έχουν μεγάλο ποσοστό** ομοιότητας (μεγάλη εξελεγκτική απόσταση).

# Το mRNA μεταφράζεται σε πρωτεΐνη

- Το RNA μικραίνει κατά πολύ και εξέρχεται του πυρήνα.
- Στο κυτταρόπλασμα μεταφράζεται σε πρωτεΐνη με βάση το γενετικό κώδικα.
- 4 νουκλεοτίδια (A, U, G, C) οργανώνονται σε τριπλέτες. Πιθανοί συνδυασμοί:  $4^3$ .
- 61 τριπλέτες κωδικοποιούν τα 20 αμινοξέα / 3 μηνύματα τερματισμού.
- Ένα αμινοξύ κωδικοποιείται από διαφορετικές τριπλέτες / κωδικόνια.

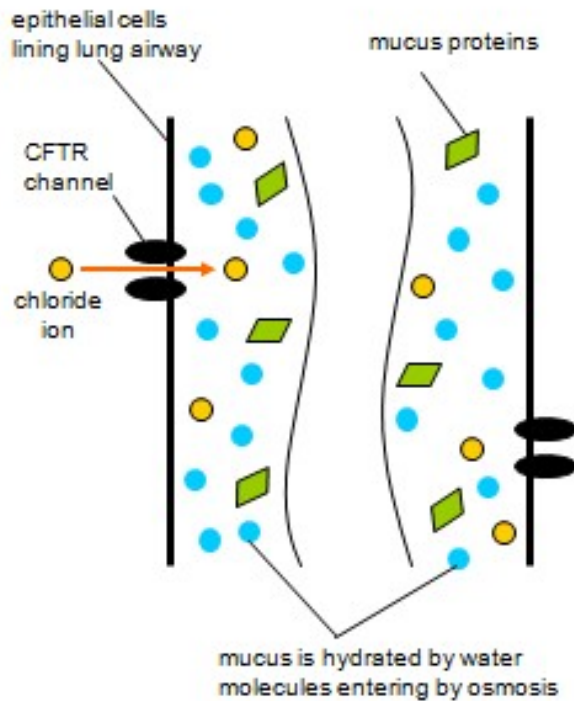
		Second base					
		U	C	A	G		
First base	U	UUU } Phenyl- alanine (phe) UUC } UUA } Leucine (leu) UUG }	UCU } UCC } Serine (ser) UCA } UCG }	UAU } Tyrosine (tyr) UAC } UAA } Stop codon UAG } Stop codon	UGU } Cysteine (cys) UGC } UGA } Stop codon UGG } Tryptophan (trp)	U C A G	
	C	CUU } CUC } Leucine (leu) CUA } CUG }	CCU } CCC } Proline (pro) CCA } CCG }	CAU } Histidine (his) CAC } CAA } Glutamine (glu) CAG }	CGU } CGC } Arginine (arg) CGA } CGG }	U C A G	
	A	AUU } Isoleucine (ile) AUC } AUA } AUG } Methionine (met) Start codon	ACU } ACC } Threonine (thr) ACA } ACG }	AAU } Asparagine (asn) AAC } AAA } Lysine (lys) AAG }	AGU } Serine (ser) AGC } AGA } Arginine (arg) AGG }	U C A G	
	G	GUU } GUC } Valine (val) GUA } GUG }	GCU } GCC } Alanine (ala) GCA } GCG }	GAU } Aspartic acid (asp) GAC } GAA } Glutamic acid (glu) GAG }	GGU } GGC } Glycine (gly) GGA } GGG }	U C A G	

# Amino Acid

Amino Acid			Solvent Exposed area (SEA)		
Name			>30 Å <sup>2</sup> (exposed)	<10 Å <sup>2</sup> (buried)	10-30 Å <sup>2</sup>
Alanine	Ala	A	48%	35%	17%
Arginine	Arg	R	84%	5%	11%
Aspartic Acid	Asp	D	81%	9%	10%
Asparagine	Asn	N	82%	10%	8%
Cysteine	Cys	C	32%	54%	14%
Glutamic Acid	Glu	E	93%	4%	3%
Glutamine	Gln	Q	81%	10%	9%
Glycine	Gly	G	51%	36%	13%
Histidine	His	H	66%	19%	15%
Isoleucine	Ile	I	39%	47%	14%
Leucine	Leu	L	41%	49%	10%
Lysine	Lys	K	93%	2%	5%
Methionine	Met	M	44%	20%	36%
Phenylalanine	Phe	F	42%	42%	16%
Proline	Pro	P	78%	13%	9%
Serine	Ser	S	70%	20%	10%
Threonine	Thr	T	71%	16%	13%
Tryptophan	Trp	W	49%	44%	7%
Tyrosine	Tyr	Y	67%	20%	13%
Valine	Val	V	40%	50%	10%

# Cystic Fibrosis

Lung airway of unaffected person



Lung airway of person with cystic fibrosis

