Βιοπληροφορική

Ενότητα 4: Πίνακες αντικατάστασης & οπτική σύγκριση αλληλουχιών

Αν. καθηγητής Αγγελίδης Παντελής
e-mail: paggelidis@uowm.gr
ΕΕΔΙΠ Μπέλλου Σοφία
e-mail: sbellou@uowm.gr
Τμήμα Μηχανικών Πληροφορικής και Τηλεπικοινωνιών

Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.
- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδειας χρήσης, η άδεια χρήσης αναφέρεται ρητώς.

Χρηματοδότηση

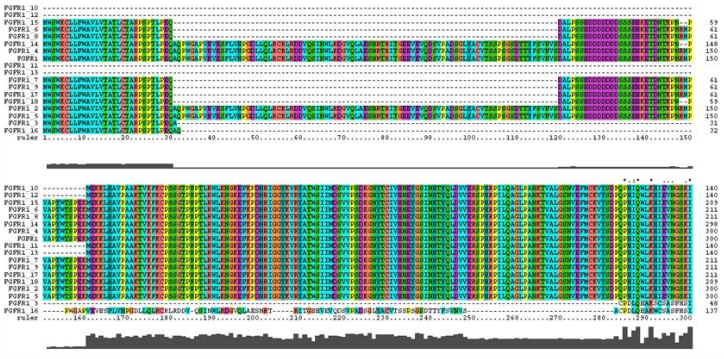
- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «**Ανοικτά Ψηφιακά Μαθήματα στο Πανεπιστήμιο Δυτικής Μακεδονίας**» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

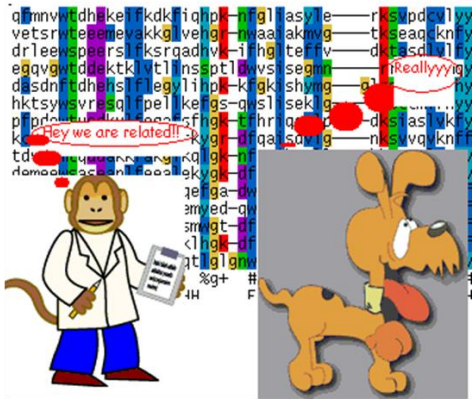
Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Σκοπός του μαθήματος

- 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.

Πίνακες αντικατάστασης & οπτική σύγκριση αλληλουχιών





Σοφία Μπέλλου, sbellou@uowm.gr

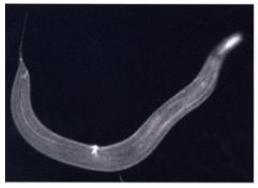
Ολοκληρωμένη αλληλουχία DNA



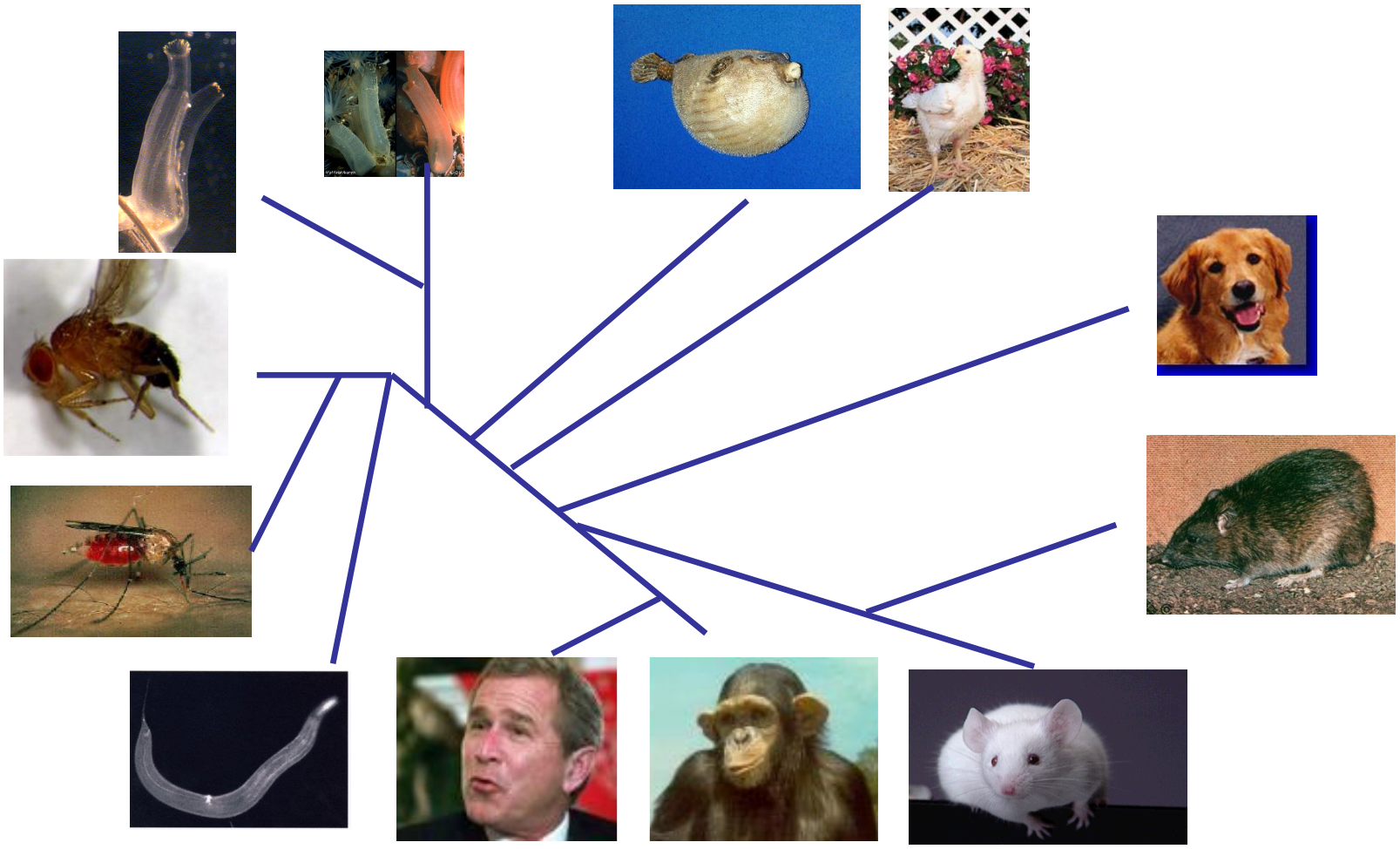



Περισσότερα από 400 γονιδιώματα
έχουν αποκρυπτογραφηθεί πλήρως






Εξέλιξη



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;;;

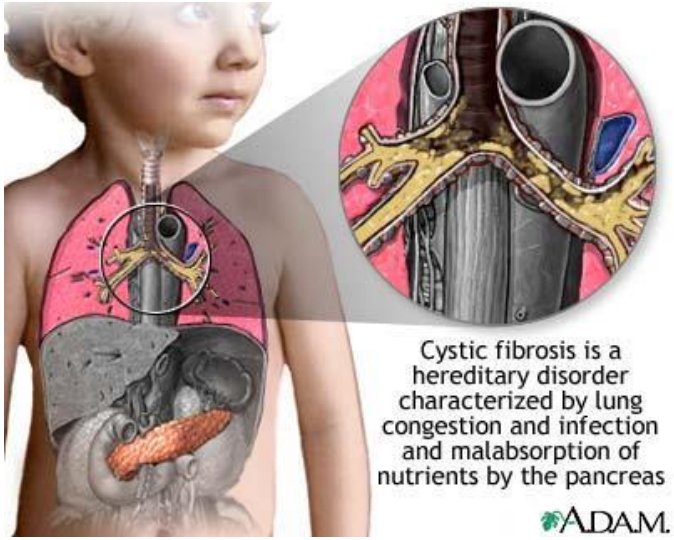
- **Συμπεράσματα για λειτουργία ενός γονιδίου (μίας πρωτεΐνης):**
  - Όταν δύο γονίδια έχουν μεγάλο ποσοστό ομοιότητας, τότε πιθανά κωδικοποιούν πρωτεΐνες με παρόμοια λειτουργία.
- Συμπεράσματα για σημαντικές περιοχές στην αλληλουχία γονιδίων/πρωτεϊνών.
  - Όταν πρωτεΐνες με συγκεκριμένο χαρακτηριστικό 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.

Παράδειγμα 2: 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**

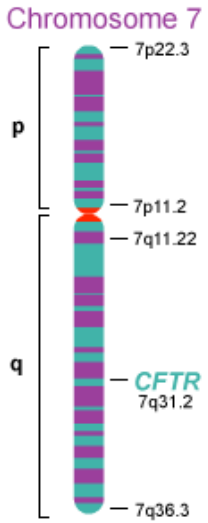
Mutated protein sequence **Isoleucine – Leucine – Serine – Lysine**

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:

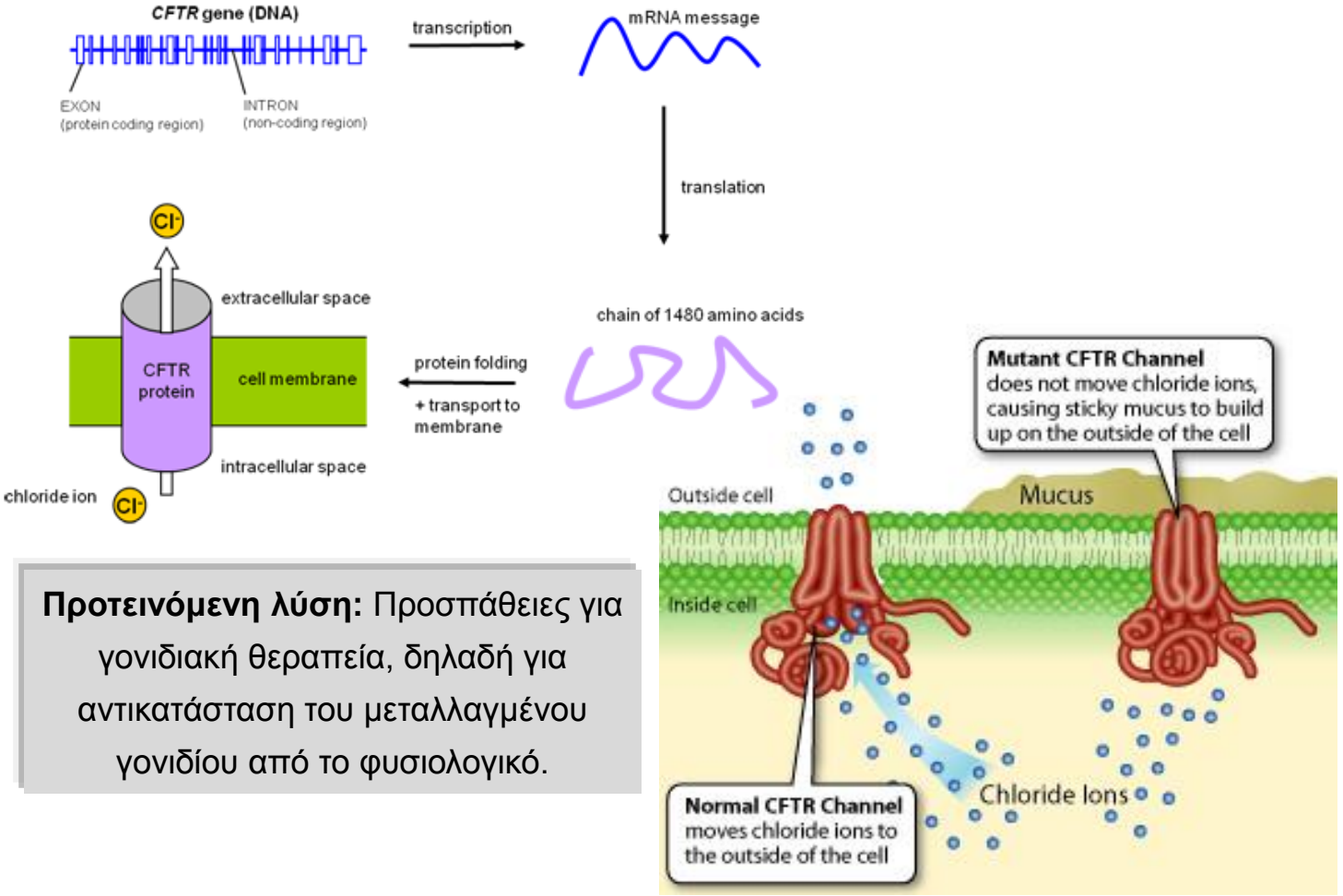
Nucleotide	A	T	C	A	T	C	T	T	T	G	G	T	G	T	T				
Amino Acid	Ile			Ile			Phe			Gly			Val						
			506						508						510				

Deleted in ΔF508

ΔF508 CFTR Sequence:

Nucleotide	A	T	C	A	T	T	G	G	T	G	T	T
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 (1/2)

Sequence1:

TCGCTGCGAAGGACATTTGGGCTGTGTGTGCGACGCGGGTCGGAGGGGCAGTCGGGGGAACCGCGAAGAAGCCGAGGAGCCCGGAGC
CCCGCGTGACGCTCCTCTCTCAGTCCAAAAGCGGCTTTTGGTTCGGCGCAGAGAGACCCGGGGGTCTTCAGGACAGCGATTGTCATTGCT
GAAGCTTTTCCTCGAAAAGCGCCGCCCTGCCCTTGGCCCCGAGAACAGACAAAGAGCACCGCAGGGCCGATCACGCTGGGGGCGCTGA
GGCCGGCCATGGTCATGGAAGTGGGCACCCTGGACGCTGGAGGCCTGCGGGCGCTGCTGGGGGAGCGAGCGGCGCAATGCCTGCTGC
TGGACTGCCGCTCCTTCTTCGCTTTCAACGCCGGCCACATCGCCGGCTCTGTCAACGTGCGCTTCAGCACCATCGTGCGGCGCCGGGCC
AAGGGCGCCATGGGCCTGGAGCACATCGTGCCCAACGCCGAGCTCCGCGGTCAGGACAGCGATTGTCATTGCTGACCGCCTGCTGGCC
GGCGCCTACCACGCCGTGGTGTTGCTGGACGAGCGCAGCGCCGCCCTGGACGGCGCCAAGCGCGACGGCACCCTGGCCCTGGCGGCC
GGCGCGCTCTGCCGCGAGGCGCGCGCCGCGCAAGTCTTCTTCCTCAAAGGAGGATACGAAGCGTTTTCGGCTTCCTGCCCGGAGCTCAG
GACAGCGATTGTCATTGCTGATGTGCAGCAAACAGTCGACCCCCATGGGGCTCAGCCTTCCCCTGAGTACTAGCGTCCCTGACAGCGCGG
AATCTGGGTGCAGTTCCTGCAGTACCCCACTCTACGATCAGGGTGGCCCGGTGGAAATCCTGCCCTTTCTGTACCTGGGCAGTGCGTATC
ACGCTTCCCGCAAGGACATGCTGGATGCCTTGGGCA

Sequence 2:

TAACTGCCTTGATCAACGTCTCAGCCAATTGTCCCAACCATTTTGAGGGTCACTACCAGTACAAGAGCATCCCTGTGGAGGACAACCACAA
GGCAGACATCAGCTCCTGGTTCAACGAGGCCATTGACTTCATAGACTCCATCAAGAATGCTGGAGGAAGGGTGTTTGTCCACTGCCAGGC
AGGCATTTCCCGGTCAGCCACCATCTGCCTTGCTTACCTTATGAGGACTAATCGAGTCATCAGGACAGCGATTGTCATTGCTGAAGCTGGA
CGAGGCCTTTGAGTTTGTGAAGCAGAGGCGAAGCATCATCTCTCCCAACTTCAGCTTCATGGGCCAGCTGCTGCAGTTTGAGTCCCAGGT
GCTGGCTCCGCACTGTTCGGCAGAGGCTGGGAGCCCCGCCATGGCTGTGCTCGACCGAGGCACCTCCACCACCACCGTGTTCAACTTTC
AGGACAGCGATTGTCATTGCTGACCCCGTCTCCATCCCTGTCCACTCCACGAACAGTGCGCTGAGCTACCTTCAGAGCCCCATTACGACC
TCTCCCAGCTGCTGAAAGGCCACGGGAGGTGAGGCTCTTCACATCCCATTGGGACTCCATGCTCCTTGAGAGGAGAAATGCAATAACTCT
GGGAGGGGCTCAGGACAGCGATTGTCATTGCTGATCGAGAGGGCTGGTCCTTATTTATTTAACTTCACCCGAGTTCCTCTGGGTTTCTAAG
CAGTTATGGTGATGACTTAGCGTCAAGACATTTGCTGAACTCAGCACATTCGGGACCAATATATAGTGGGTACATCAAGTCCATCTGACAAA
ATGGGGCAGAAGAGAAAGGACTCAGTGTGTGATCCGGTTTCTTTTTGCTCGCCCCTGTTTTTTGTAGAATCTCTTCATGCTTGACATACCTA
CCAGTATTATTCCCGACGACACATATACATATGAGAATATACCTTATTTATTTTTGTGTAGGTGTCTGCCTTCACAAATGTCATTGTCTACTCC
TAGAAGAACCAAATACCTCAATTTTTGTTTTTGAGTACTGTACTATCCTGTAAATATATCTTAAGCAGGTTTGTTTTCA

Γονίδια που κωδικοποιούν πρωτεΐνες με μεμβρανική εντόπιση (2/2)

Sequence1:

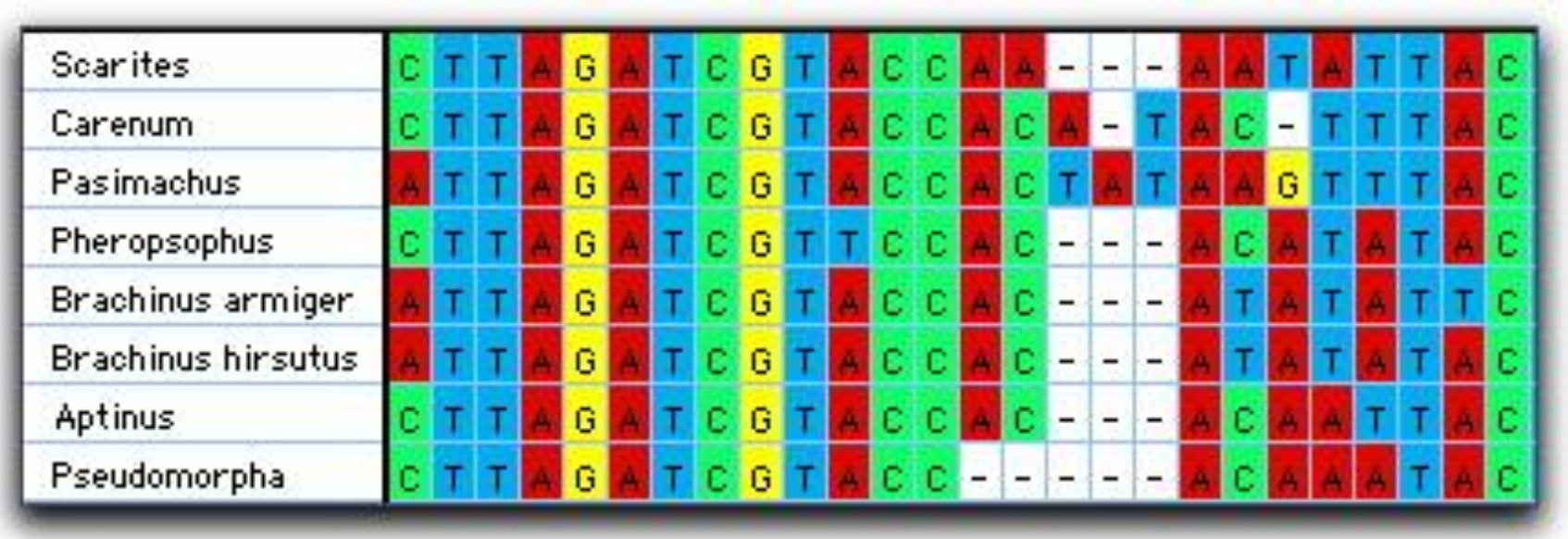
TCGCTGCGAAGGACATTTGGGCTGTGTGTGCGACGCGGGTCGGAGGGGCAGTCGGGGGAACCGCGAAGAAGCCGAGGAGCCCGGAGCCCCGCGTGAC
GCTCCTCTCTCAGTCCAAAAGCGGCTTTTGGTTCGGCGCAGAGAGACCCGGGGGTCT**TCAGGACAGCGATTGTCATTGCTGA**AGCTTTTCCTCGAAAAGC
GCCGCCCTGCCCTTGGCCCCGAGAACAGACAAAGAGCACCGCAGGGCCGATCACGCTGGGGGCGCTGAGGCCGGCCATGGTCATGGAAGTGGGCACC
CTGGACGCTGGAGGCCTGCGGGCGCTGCTGGGGGAGCGAGCGGCGCAATGCCTGCTGCTGGACTGCCGCTCCTTCTTCGCTTTCAACGCCGGCCACAT
CGCCGGCTCTGTCAACGTGCGCTTCAGCACCATCGTGCGGCGCCGGGCCAAGGGCGCCATGGGCCTGGAGCACATCGTGCCCAACGCCGAGCTCCGCG
G**TCAGGACAGCGATTGTCATTGCTGA**CCGCCTGCTGGCCGGCGCCTACCACGCCGTGGTGTTGCTGGACGAGCGCAGCGCCGCCCTGGACGGCGCCAA
GCGCGACGGCACCCTGGCCCTGGCGGCCGGCGCGCTCTGCCGCGAGGCGCGCGCCGCGCAAGTCTTCTTCCTCAAAGGAGGATACGAAGCGTTTTCGG
CTTCCTGCCCGGAGC**TCAGGACAGCGATTGTCATTGCTGA**TGTGCAGCAAACAGTCGACCCCCATGGGGCTCAGCCTTCCCCTGAGTACTAGCGTCCCTG
ACAGCGCGGAATCTGGGTGCAGTTCCTGCAGTACCCCACTCTACGATCAGGGTGGCCCGGTGGAAATCCTGCCCTTTCTGTACCTGGGCAGTGCGTATCA
CGCTTCCCGCAAGGACATGCTGGATGCCTTGGGCA

Sequence 2:

TAACTGCCTTGATCAACGTCTCAGCCAATTGTCCCAACCATTTTGAGGGTCACTACCAGTACAAGAGCATCCCTGTGGAGGACAACCACAAGGCAGACATC
AGCTCCTGGTTCAACGAGGCCATTGACTTCATAGACTCCATCAAGAATGCTGGAGGAAGGGTGTTTGTCCACTGCCAGGCAGGCATTTCCCGGTCAGCCA
CCATCTGCCTTGCTTACCTTATGAGGACTAATCGAGTCA**TCAGGACAGCGATTGTCATTGCTGA**AGCTGGACGAGGCCTTTGAGTTTGTGAAGCAGAGGCG
AAGCATCATCTCTCCCAACTTCAGCTTCATGGGCCAGCTGCTGCAGTTTGAGTCCCAGGTGCTGGCTCCGCACTGTTCGGCAGAGGCTGGGAGCCCCGC
CATGGCTGTGCTCGACCGAGGCACCTCCACCACCACCGTGTTCAACTTA**TCAGGACAGCGATTGTCATTGCTGA**CCCCGTCTCCATCCCTGTCCACTCCAC
GAACAGTGCGCTGAGCTACCTTCAGAGCCCCATTACGACCTCTCCCAGCTGCTGAAAGGCCACGGGAGGTGAGGCTCTTCACATCCCATTGGGACTCCAT
GCTCCTTGAGAGGAGAAATGCAATAACTCTGGGAGGGGC**TCAGGACAGCGATTGTCATTGCTGA**TCGAGAGGGCTGGTCCTTATTTATTTAACTTCACCCG
AGTTCCTCTGGGTTTCTAAGCAGTTATGGTGATGACTTAGCGTCAAGACATTTGCTGAACTCAGCACATTCGGGACCAATATATAGTGGGTACATCAAGTCC
ATCTGACAAAATGGGGCAGAAGAGAAAGGACTCAGTGTGTGATCCGGTTTCTTTTTGCTCGCCCCTGTTTTTTGTAGAATCTCTTCATGCTTGACATACCTA
CCAGTATTATTCCCGACGACACATATACATATGAGAATATACCTTATTTATTTTTGTGTAGGTGTCTGCCTTCACAAATGTCATTGTCTACTCCTAGAAGAACC
AAATACCTCAATTTTTGTTTTTGAGTACTGTACTATCCTGTAAATATATCTTAAGCAGGTTTGTTTTCA

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

Σύγκριση αλληλουχιών κατά ζεύγη (1/2)



Σύγκριση αλληλουχιών κατά ζεύγη (2/2)

- 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     AGAGCCCCATTACGACCTCTCCCAGCTGCTGAAAGGCCACGGGAGGTGAGGCTCTTCACA
SEQ2      AGAGCCCCATTACGACCTCTCCCAGCTGC----AGACTCGAGGAGCAAAAGCTCAT----
          ****************************    ** *    ****   * **** *
```

Βασικοί ορισμοί - 1

- 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.

Βασικοί ορισμοί - 1, Παράδειγμα

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” (**v**intner → **w**intner).
  b. Να εισάγουμε το “r” (wintner → w**r**intner).
  c. Να διαγράψουμε το “n” δύο φορές (wri**n**t**n**er → writer).
  d. Να εισάγουμε το “s” (writer → writer**s**).
- **Συνολικά:** 5 βασικές πράξεις μετασχηματισμού.
- **Edit-distance (S_1 → S_2) = 5.**

Βασικοί ορισμοί - 2

- **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

Βασικοί ορισμοί - 2 (συν.)

- Η ακολουθία μετασχηματισμού αποτελεί μία συμβολοσειρά με τα στοιχεία Σ={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=RIMDMDMMI

Βασικοί ορισμοί - 3

- 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

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

- **Παράδειγμα:** Weighted-edit-distance($S_1 \rightarrow S_2$)**=1r+4d+4m.**

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

Βασικοί ορισμοί - 4

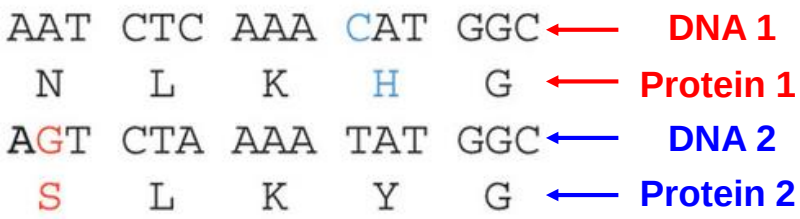
- **Ζυγισμένη απόσταση μετασχηματισμού βάσει αλφαβήτου (weighted edit distance), βάσει αλφαβήτου:** Το ελάχιστο πλήθος των πράξεων μετασχηματισμού που απαιτούνται για να μετασχηματίσουμε την πρώτη συμβολοσειρά στη δεύτερη. Κάθε πράξη μετασχηματισμού έχει **συγκεκριμένο κόστος-βάρος ανάλογα με το χαρακτήρα 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.

Στοίχιση αλληλουχιών κατά ζεύγη – Βασικοί κανόνες (1/4)

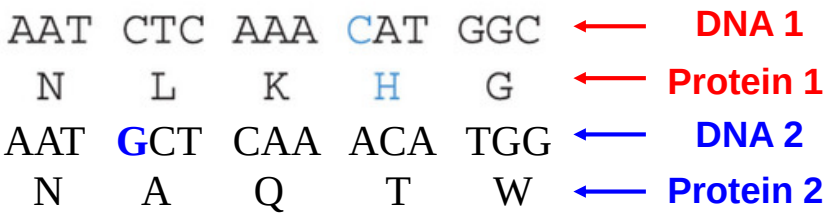
- Οι περισσότερες μέθοδοι στοίχισης επιχειρούν να προσομοιώσουν τους διάφορους εξελεγκτικούς μηχανισμούς.
- Οι στοιχειώδεις αλλαγές σε σχέση με το προγονικό μόριο χαρακτηρίζονται ως:
  - αντικαταστάσεις (replacements),
  - προσθήκες (insertions),
  - εξαλείψεις (deletions),
- αμινοξικών καταλοίπων ή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

Στοίχιση αλληλουχιών κατά ζεύγη – Βασικοί κανόνες (2/4)

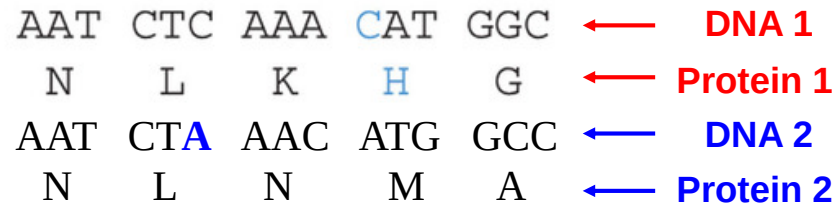
Αντικατάσταση



Προσθήκη Ένθεση

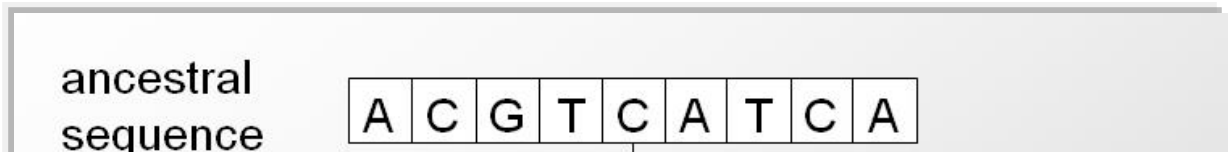


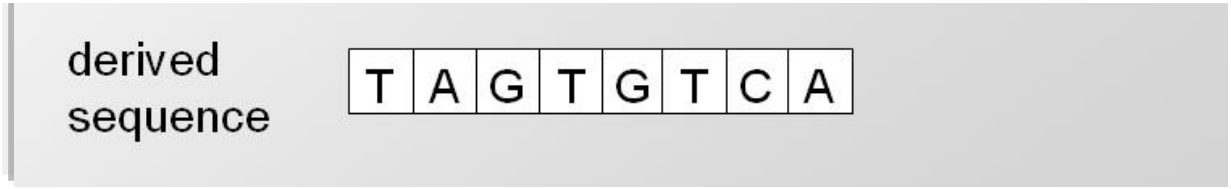
Εξάλειψη Διαγραφή



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

Στοίχιση αλληλουχιών κατά ζεύγη – Βασικοί κανόνες (3/4)





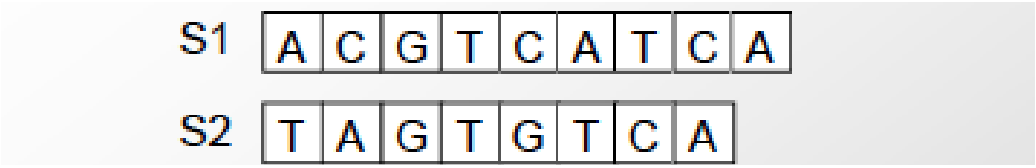
Computational Biology: Genomes, Networks, Evolution, MIT course 6.047/6.878

Στοίχιση αλληλουχιών κατά ζεύγη – Βασικοί κανόνες (4/4)

- Κατά τη στοίχιση:
  - **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.

Είναι η εισαγωγή κενών απαραίτητη;

- Longest Common Substring – Μεγαλύτερη κοινή υπο-συμβολοσειρά.



Computational Biology: Genomes, Networks, Evolution, MIT course 6.047/6.878

Στοίχιση αλληλουχιών με κενά (1/2)

Παράδειγμα: DESCRIBING & DESINGING

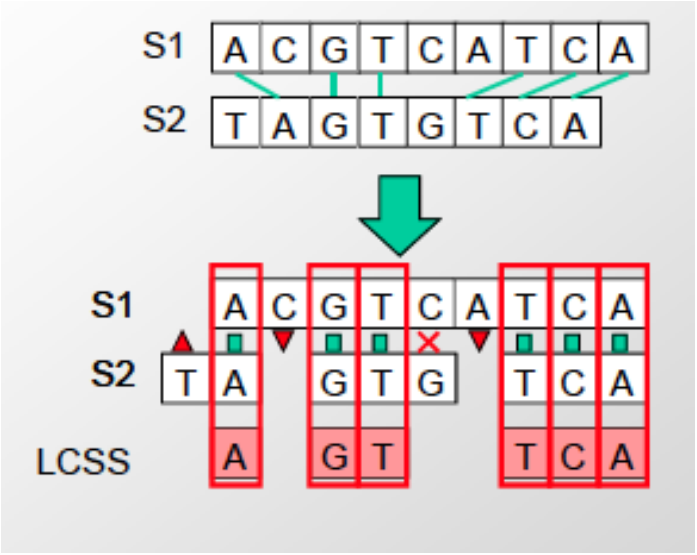
```
D E S C R I B I N G          D E S C R I B I N G
| | |          Gaps           | | | |     | | |
D E S I N G I N G            D E S I N G - I N G
```

Στοίχιση αλληλουχιών με κενά (2/2)

- 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

Η εισαγωγή κενών είναι τελικά απαραίτητη

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



Computational Biology: Genomes, Networks, Evolution, MIT course 6.047/6.878

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.

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

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

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

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

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

Τοπική vs. 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.**

Στοίχιση αλληλουχιών κατά ζεύγη - DNA & RNA

Ποια στοίχιση είναι η καλύτερη;

1. **A T C G G A T C T**
2. **A C G G A C T**

```
A T C G G A T C T
|   | | |     | |
A - C G G - A C T
```

```
A T C G G A T - C T
|   |   |       | |
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 βαθμονόμησης στοίχισης - DNA & RNA (1/2)

Sequence 1: ATCGGATCT

Sequence 2: ACGGACT **ΠΡΟΣΟΧΗ: Συγκρίνουμε νουκλεοτίδια, μόνο 4 πιθανές βάσεις**

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

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

- Πιθανό σύστημα βαθμονόμησης:
  - όμοιο κατάλοιπο: +2
  - διαφορετικό κατάλοιπο: -1
  - κενό: –2
- **Alignment 1:** 5 x 2 – 1(1) – 4(2) = 10 – 1 – 8 = 1
- **Alignment 2:** 7 x 2 – 0(1) – 2 (2) = 14 – 0 – 4 = 10

Σύστημα βαθμονόμησης στοίχισης - DNA & RNA (2/2)

- 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

Στοίχιση πρωτεϊνών;;

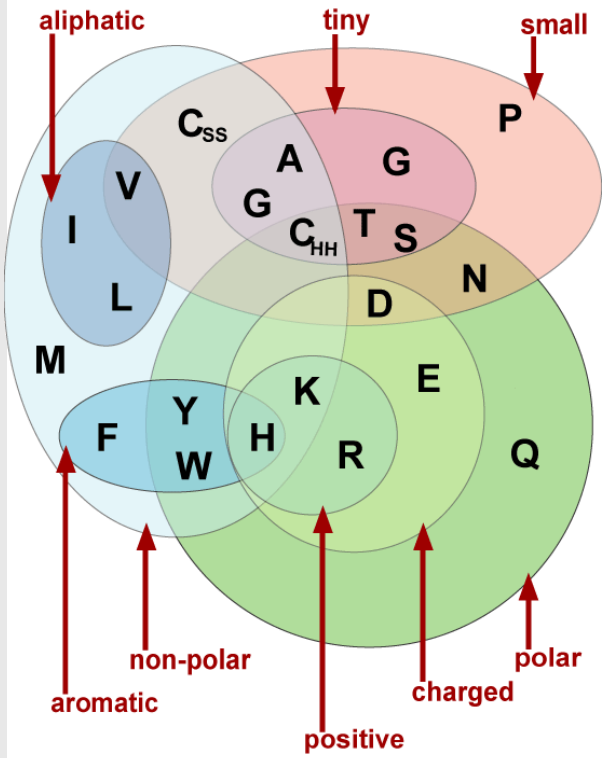
A	**S**	**K**	**T**	**M**	**P**	**I**
I	I	?	?	I	I	I
A	**S**	**R**	**H**	**M**	**P**	**I**

A	**S**	**K**	**T**	**M**	**P**	**I**
I	I	?	?	I	I	|
A	**S**	**Y**	**H**	**M**	**P**	**I**

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

Σύστημα βαθμονόμησης στοίχισης - Πρωτεΐνες

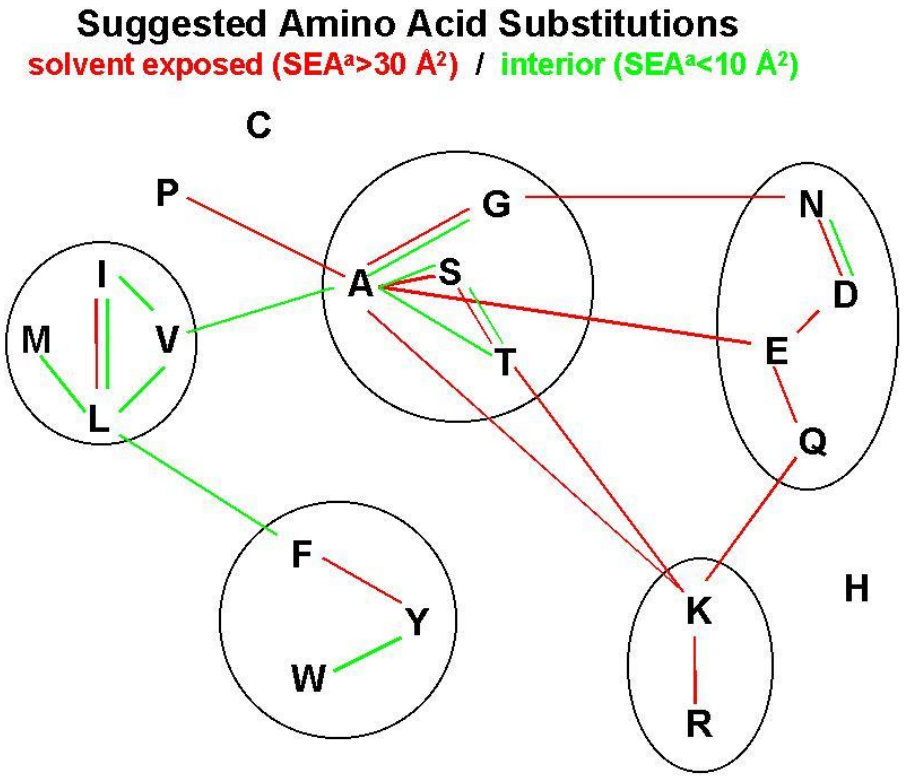
- Περισσότερο πολύπλοκα συστήματα βαθμονόμησης.
- Η αντικατάσταση αμινοξέων παρόμοιων φυσικοχημικών ιδιοτήτων = **συντηρητική αντικατάσταση.**
- Η συντηρητική αντικατάσταση προτιμάται από τη φύση.
- Επιρροή στη δομή πρωτεΐνης:
  - ✓ Αντικατάσταση υδροφιλικού από υδροφιλικό.
  - ✓ Αντικατάσταση υδροφιλικού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



Σύστημα βαθμονόμησης στοίχισης

- 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



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)

[a]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 (1/2)

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	---

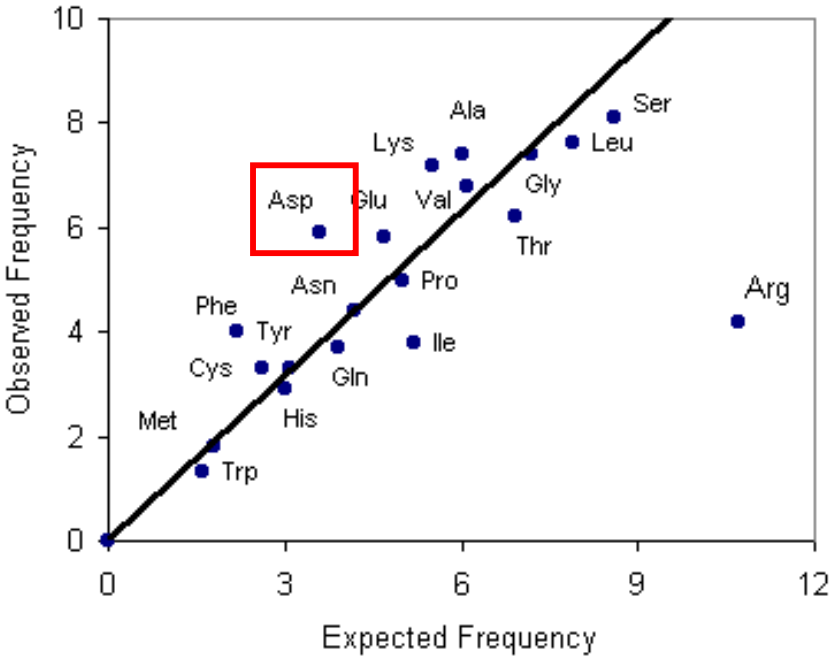
Expected amino acid frequency

- 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 = 0.261x0.303x0.220 + 0.261x0.303x0.217
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 (2/2)

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	---

Πίνακας βαθμονόμησης: Παράδειγμα

%		A	R	N	K
7.4	A	5	-2	-1	-1
4.2	R	-	7	-1	3
4.4	N	-	-	7	0
7.2	K	-	-	-	6

A K R A N R

K A A A N K

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

- 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.**

Συντήρηση καταλοίπων

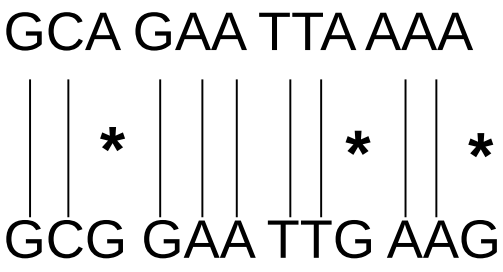
- Τα αμινοξέα αλλάζουν με τέτοιο τρόπο ώστε να διατηρηθούν οι φυσικοχημικές ιδιότητες του αρχικού καταλοίπου.
- Δηλαδή:
  - Πολικό σε πολικό.
    - asparagine → glutamine.
  - Μη πολικό σε μη πολικό.
    - alanine → valine.
  - Όμοιας συμπεριφοράς.
    - arginine → lysine (positive to positive).
   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

Πίνακες βαθμονόμησης (scoring matrices) ή πίνακες αντικατάστασης

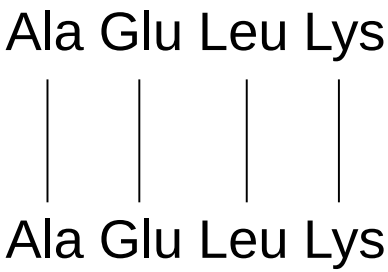
- Για την αντικατάσταση αμινοξέων:
  - PAM.
  - BLOSUM.
- Για την αντικατάσταση νουκλεοτιδίων (DNA):
  - Το DNA είναι λιγότερο συντηρημένο από τις πρωτεΐνες.
  - Δεν είναι αποτελεσματικό να συγκρίνουμε κωδικοποιούσες περιοχές, από τις οποίες προκύπτουν οι πρωτεΐνες, σε επίπεδο νουκελοτιδίων.

Παράδειγμα στοίχισης 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.



75% ομοιότητα



100% ομοιότητα

Πίνακες αντικατάστασης PAM (Percent Accepted Mutation-PAM or Dayhoff Matrices) (1/2)

- 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) (2/2)

- 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 Marcov – Marcov 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 (1/2)

- 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.

$$\underbrace{M_1 \cdot M_1 \cdot \ldots \cdot M_1}_{n} \Rightarrow M_n$$

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 → 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: 0.57x10 = 5.7.**
- Πιθανότητα μετάλλαξης σύμφωνα με PAM250 Tyr → 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: 0.83x10 = 8.3.**
- **Average: 5.7 + 8.3 = 7.**

$$\underbrace{M_1 \cdot M_1 \cdot \ldots \cdot M_1}_{250} \Rightarrow M_{250} = PAM250$$

Amino acid change	PAM1	PAM250
Phe to Ala	0.0002	0.04
Phe to Arg	0.0001	0.01
Phe to Asn	0.0001	0.02
Phe to Asp	0.0000	0.01
Phe to Cys	0.0000	0.01
Phe to Gln	0.0000	0.01
Phe to Glu	0.0000	0.01
Phe to Gly	0.0001	0.03
Phe to His	0.0002	0.02
Phe to Ile	0.0007	0.05
Phe to Leu	0.0013	0.13
Phe to Lys	0.0000	0.02
Phe to Met	0.0001	0.02
Phe to Phe	0.9946	0.32
Phe to Pro	0.0001	0.02
Phe to Ser	0.0003	0.03
Phe to Thr	0.0001	0.03
Phe to Trp	0.0001	0.01
Phe to Tyr	0.0021	0.15
Phe to Val	0.0001	0.05
SUM[a]	1.0000	1.00

[a]Approximate since scores are rounded off.

The multiplication of two PAM1 matrices to give a PAM2 matrix. Only three rows and columns are shown for illustrative purposes.

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.

Συνεπώς:

- $0.2 = \log_{10}$(σχετική αναμενόμενη τιμή μετάλλαξης).
- Σχετική αναμενόμενη τιμή μετάλλαξης = $10^{0.2} = 1.6$.
- Πολλαπλασιάζω με τη συχνότητα εμφάνισης του Q, 3.7% = 0.037 1.6x0.037 = 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
I	I	?	?	I	I	I
A	S	R	A	M	P	I

A	S	K	T	M	P	I
I	I	?	?	I	I	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 (2/2)

- 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).

Μετάφραση: Το mRNA μεταφράζεται σε πρωτεΐνη

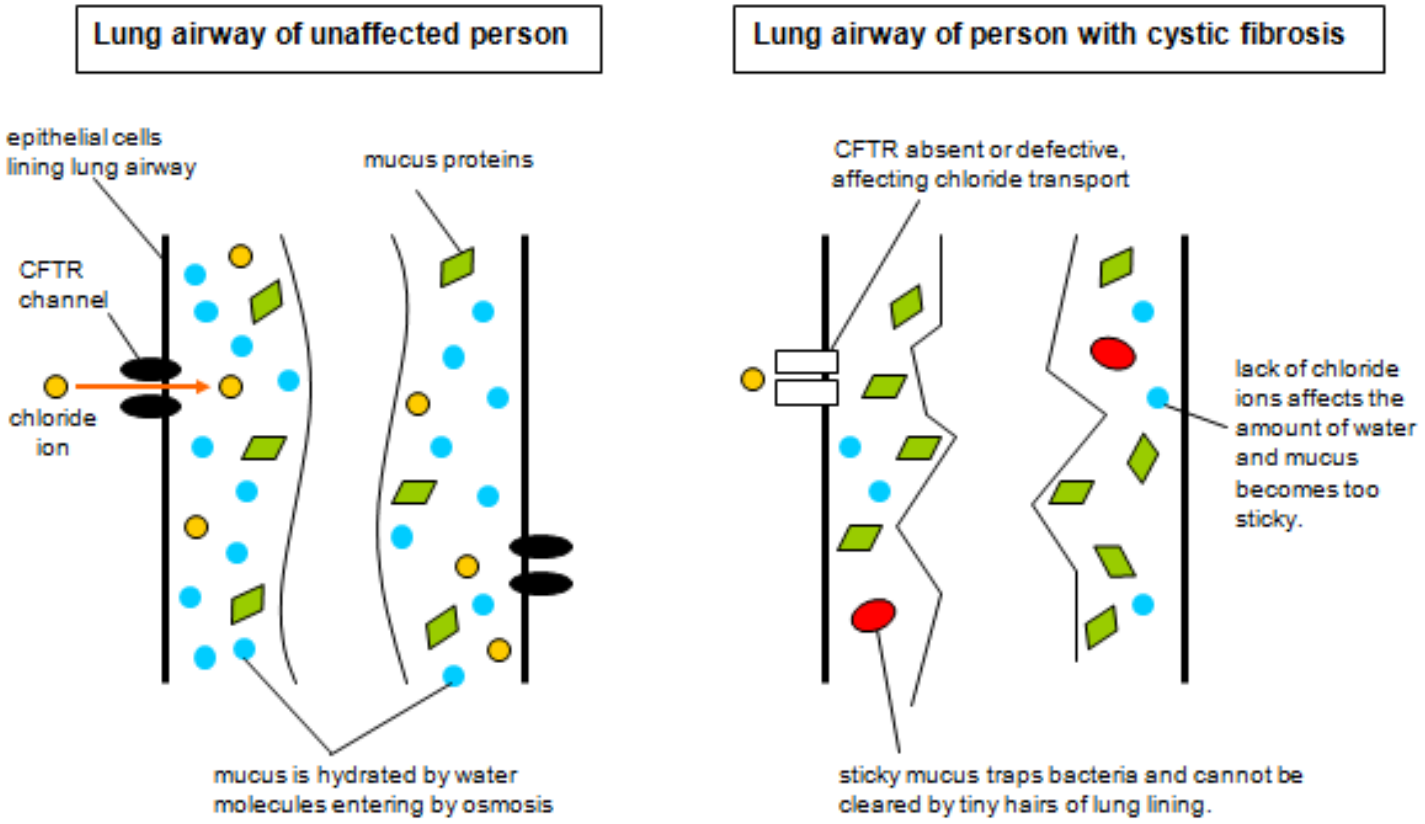
- Το RNA μικραίνει κατά πολύ και εξέρχεται του πυρήνα.
- Στο κυτταρόπλασμα μεταφράζεται σε πρωτεΐνη με βάση το γενετικό κώδικα.
- 4 νουκλεοτίδια (A, U, G, C) οργανώνονται σε τριπλέτες. Πιθανοί συνδυασμοί: 4^3.
- 61 τριπλέτες κωδικοποιούν τα 20 αμινοξέα / 3 μηνύματα τερματισμού.
- Ένα αμινοξύ κωδικοποιείται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

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

Amino Acid

Amino Acid			Solvent Exposed area (SEA)		
Name			>30 Å^2 (exposed)	<10 Å^2 (buried)	10-30 Å^2
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%

Παράδειγμα Cystic Fibrosis



Τέλος Ενότητας

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Σημείωμα Αναφοράς

Σημείωμα Αδειοδότησης

Διατήρηση Σημειωμάτων

Οποιαδήποτε αναπαραγωγή ή διασκευή του υλικού θα πρέπει να συμπεριλαμβάνει:

- το Σημείωμα Αναφοράς
- το Σημείωμα Αδειοδότησης
- τη δήλωση Διατήρησης Σημειωμάτων
- το Σημείωμα Χρήσης Έργων Τρίτων (εφόσον υπάρχει)

μαζί με τους συνοδευόμενους υπερσυνδέσμους.

Σημείωμα Χρήσης Έργων Τρίτων

Το Έργο αυτό κάνει χρήση των ακόλουθων έργων:

Εικόνες:

- http: //blog.com.mk/send/121903
- http: //foter.com/Cmyk/
- http: //www.ebi.ac.uk/training/online/course/introduction-protein-classification-ebi/what-are-protein-signatures/signature-types/what-ar-0
- http: //contentinacottage.blogspot.ca/2012_01_29_archive.html
- https: //www.cartoonstock.com/