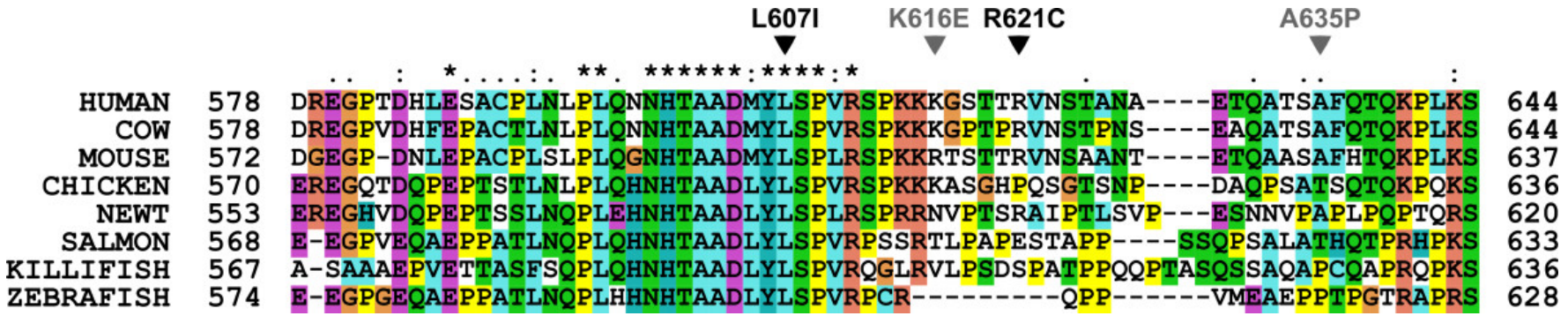
Πολλαπλή στοίχιση
multiple sequence alignment
(MSA)

MSA: Τι είναι

- Στοίχιση για 3 ή περισσότερες ακολουθίες.
- Αποκαλύπτονται οι συντηρημένες περιοχές μεταξύ των ακολουθιών μιας οικογένειας.
- Χρειάζεται για:
  - Δημιουργία profiles/motifs που χαρακτηρίζουν μια επικράτεια (domain).
  - Ανίχνευση συντηρημένων DNA-binding sites σε προμότορες γονιδίων
  - Φυλογένεση.
  - Πρόβλεψη δευτεροταγούς και τριτοταγούς δομής πρωτεϊνών.
  - Σχεδιασμό εκφυλισμένων εκκινητών PCR

MSA


L607I
K616E
R621C
A635P
HUMAN 578 DREGPTDHLESACPLNLPLQNNHTAADMYLSPVRSPKKKGSTTRVNSTANA----ETQATSAFQTQKPLKS 644
COW 578 DREGPVDHFEPACTLNLPLQNNHTAADMYLSPVRSPKKKGPTPRVNSTPNS----EAQATSAFQTQKPLKS 644
MOUSE 572 DGEGP-DNLEPACPLSLPLQGNHTAADMYLSPLRSPKKRTSTTRVNSAANT----ETQAASAFHTQKPLKS 637
CHICKEN 570 EREGQTDQPEPTSTLNLPLQHNHTAADLYLSPVRSPKKKASGHPQSGTSNP----DAQPSATSQTQKPQKS 636
NEWT 553 EREGHVDQPEPTSSLNQPLEHNHTAADLYLSPLRSPRRNVPTSRAIPTLSVP---ESNNVPAPLPQPTQRS 620
SALMON 568 E-EGPVEQAEPPATLNQPLQHNHTAADLYLSPVRPSSRTLPAPESTAPP----SSQPSALATHQTPRHPKS 633
KILLIFISH 567 A-SAAAEPVETTASFSQPLQHNHTAADLYLSPVRQGLRVLPSDSPATPPQQPTASQSSAQAPCQAPRQPKS 636
ZEBRAFISH 574 E-EGPGEQAEPPATLNQPLHHNHTAADLYLSPVRPCR---------QPP------VMEAEPPTPGTRAPRS 628

MSA

- Sum of pairs
- Σκοπός: η μεγιστοποίηση αυτού του score

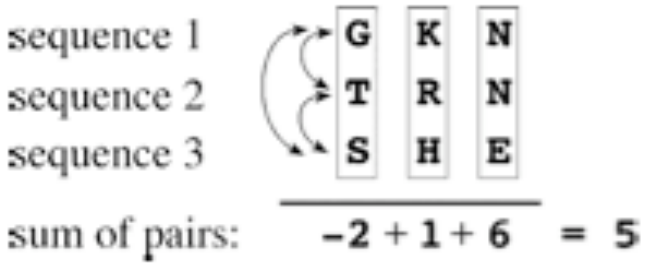


Figure 5.1: Given a multiple alignment of three sequences, the sum of scores is calculated as the sum of the similarity scores of every pair of sequences at each position. The scoring is based on the BLOSUM62 matrix (see Chapter 3). The total score for the alignment is 5, which means that the alignment is $2^5 = 32$ times more likely to occur among homologous sequences than by random chance.

MSA

- Πολλαπλή στοίχιση με:
  - Δυναμικό προγραμματισμό (dynamic programming).
  - Με ευρετικές μεθόδους (heuristics).
    - Προοδευτική στοίχιση (progressive alignment)
    - Στοίχιση με διαδοχικές βελτιώσεις (iterative alignment)
    - Στοίχιση βασισμένη σε blocks

MSA - δυναμικός προγραμματισμός (DP)

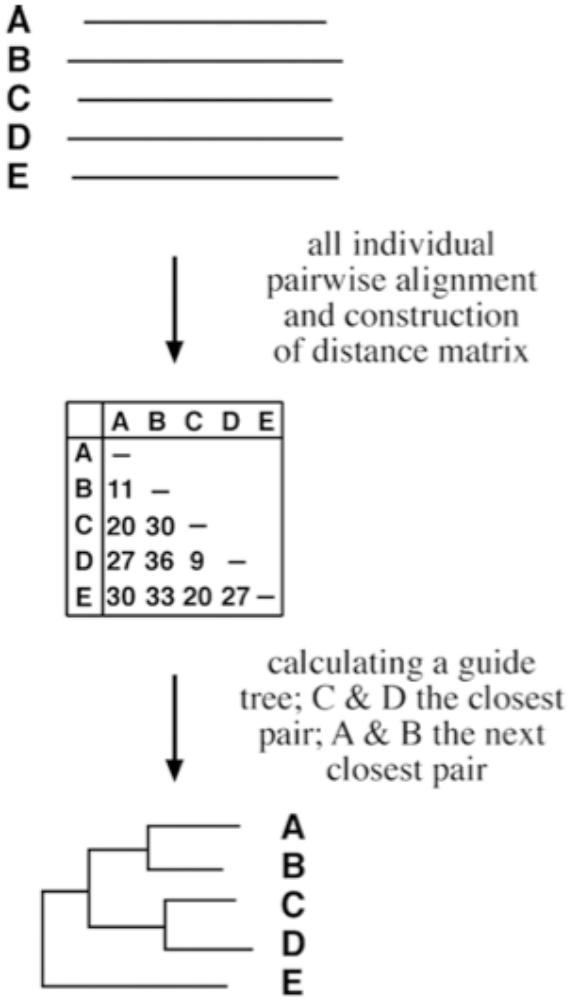
- Για στοίχιση 2 ακολουθιών δημιουργείται ένας πίνακας 2 διαστάσεων.
- Για στοίχιση 3 ακολουθιών δημιουργείται πίνακας 3 διαστάσεων.
- Για στοίχιση N ακολουθιών δημιουργείται πίνακας N διαστάσεων.
- Το υπολογιστικό κόστος αυξάνεται εκθετικά, για κάθε ακολουθία που πρέπει να ενταχθεί στην πολλαπλή στοίχιση.
- Πρακτικά, DP μπορεί να γίνει για λίγες μόνο ακολουθίες, μικρού μήκους.

MSA-ευρετικές μέθοδοι

- Προοδευτική στοίχιση (progressive)
  - ClustalW
- Επαναλαμβανόμενη στοίχιση (Iterative)
- Block-based

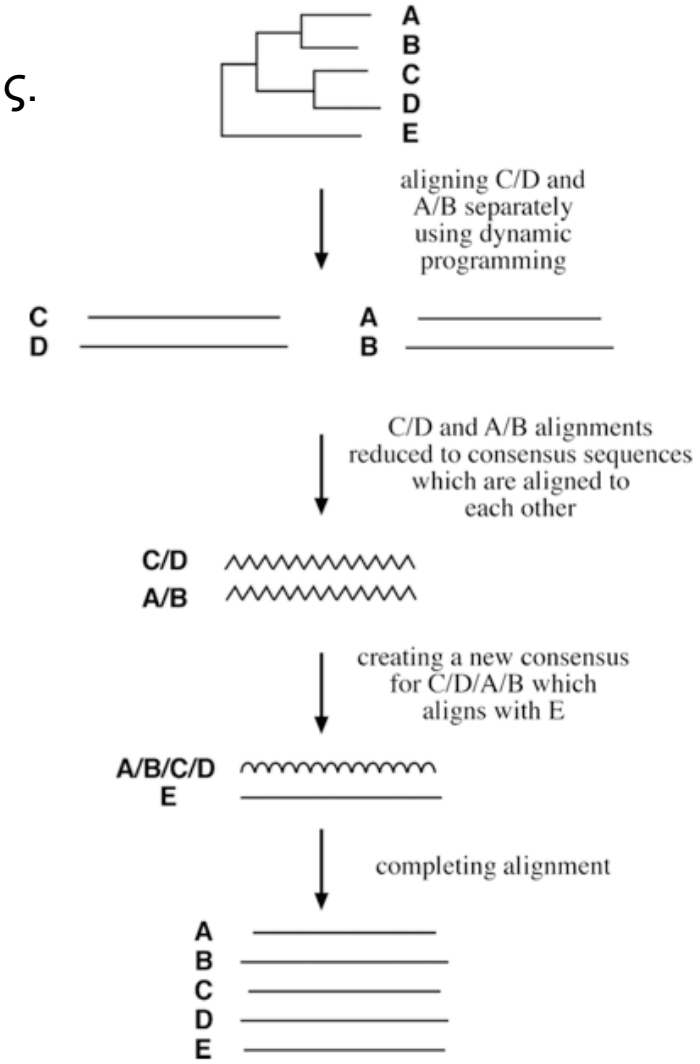
ClustalW (i)

- Ολική στοίχιση (Needlman-Wunsch) κάθε πιθανού ζεύγους
- Πίνακας αποστάσεων (identities ή πίνακες Blossum/PAM).
- Μετατροπή των αποστάσεων σε εξελικτικές αποστάσεις.
- Δημιουργία φυλογενετικού δένδρου - οδηγού (guide tree) (neighbor joining).
  - Χαμηλότερης εμπιστοσύνης από ένα κανονικό φυλογενετικό δένδρο, ωστόσο καταδεικνύει ικανοποιητικά τις βασικές σχέσεις



ClustalW (ii)

- Οι 2 κοντινότερες ακολουθίες στοιχίζονται και δημιουργείται μια ακολουθία συναίνεσης.
- Με βάση το δένδρο-οδηγό, η ακολουθία συναίνεσης στοιχίζεται (δυναμικός προγραμματισμός) με την επόμενη πιο κοντινή ακολουθία ή την επόμενη πιο κοντινή ακολουθία συναίνεσης.
- Η διαδικασία επαναλαμβάνεται έως ότου στοιχιθούν όλες οι ακολουθίες.



ClustalW (iii)

- Ανάλογα με την απόσταση 2 ακολουθιών στο δένδρο-οδηγό, χρησιμοποιείται και ο κατάλληλος πίνακας αντικατάστασης (Blossum62, Blossum 45) για την ολική στοίχιση κατά ζεύγη .

- Οι ποινές των κενών προσαρμόζονται ανάλογα με την παρατηρούμενη συντήρηση μιας περιοχής και ανάλογα με την δευτεροταγή δομή.

- Συντελεστής βαρύτητας ανάλογα με την εξελικτική απόσταση 2 ακολουθιών

Προβλήματα της προοδευτικής στοίχισης

- Δεν ενδείκνυται για ακολουθίες με πολύ διαφορετικά μήκη (λόγω ολικής στοίχισης).
- Η τελική πολλαπλή στοίχιση εξαρτάται από τη σειρά με την οποία θα γίνουν οι επιμέρους στοιχίσεις κατά ζεύγη.
- Ένα αρχικό λάθος θα επηρεάσει τα υπόλοιπα στάδια της πολλαπλής στοίχισης.

T-coffee

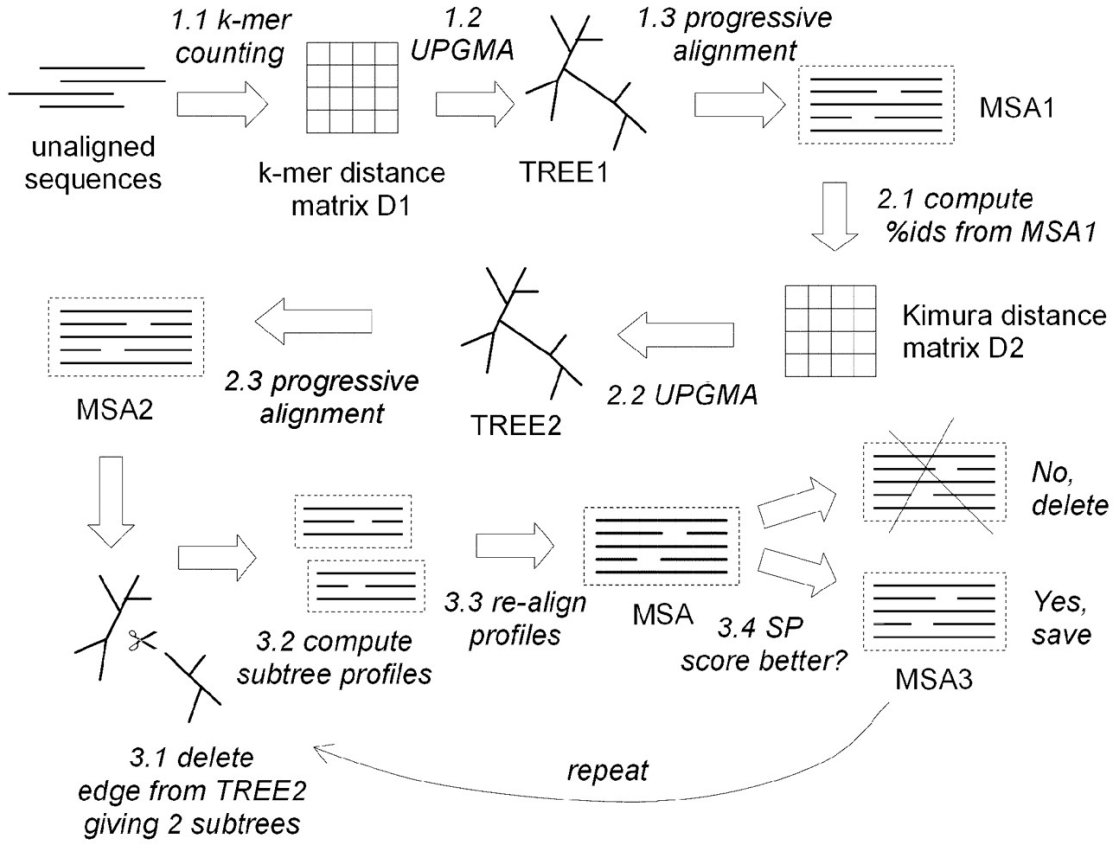
- Προοδευτική στοίχιση.
- Όταν στοιχίζει ένα ζεύγος ακολουθιών, δεν κάνει μόνο ολική στοίχιση, αλλά και τοπικές στοιχίσεις (δημιουργείται μια βιβλιοθήκη στοιχίσεων).
- Υπολογίζεται ένα σκορ συμφωνίας (consistency score) από τις επιμέρους στοιχίσεις (ολική και τοπικές).

- Σε σχέση με το Clustal:
  - Πολύ καλύτερης ποιότητας πολλαπλές στοιχίσεις.
  - Πολύ πιο αργός υπολογισμός.

Muscle

- Προοδευτική στοίχιση.
- Δύο υπολογισμοί δένδρου-οδηγού (UPGMA)
  - Kmer
  - Kimura distance

- Κυκλική λογική
  - Δένδρο-> πολλαπλή στοίχιση-> βελτιωμένο δένδρο -> βελτιωμένη στοίχιση



Edgar R C Nucl. Acids Res. 2004;32:1792-1797

Nucleic Acids Research

Επαναλαμβανόμενη πολλαπλή στοίχιση (iterative)

- Αρχικά δημιουργείται μια πολλαπλή στοίχιση χαμηλής ποιότητας.
- Η πολλαπλή στοίχιση βελτιώνεται σε επαναλαμβανόμενα στάδια.
- Ευρετική μέθοδος.
- Δεν επηρεάζεται από αρχικά λάθη.
- Προγράμματα:
  - PRRN

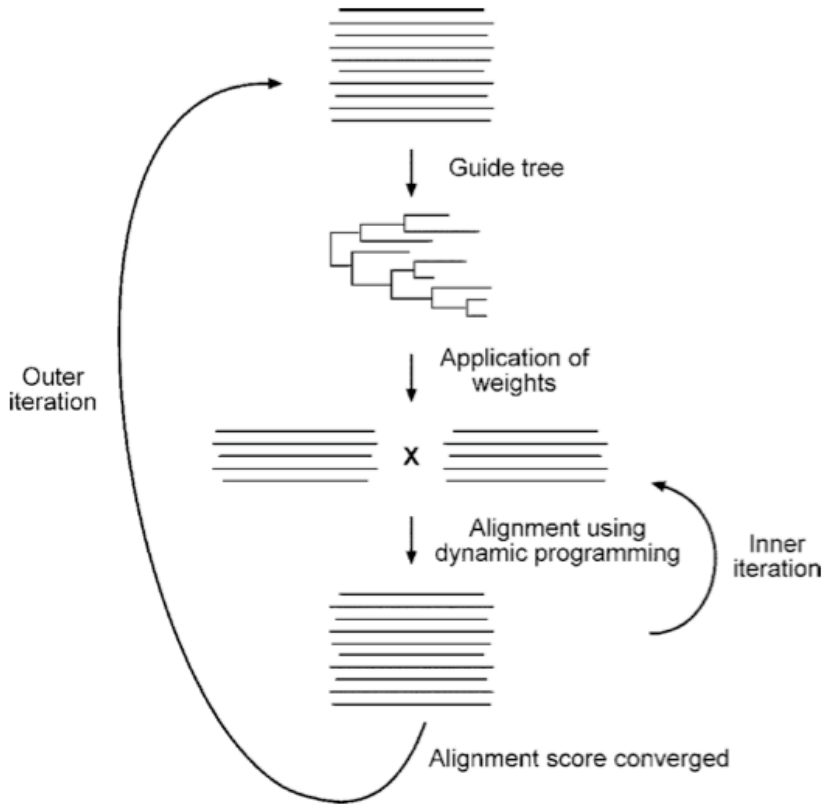


Figure 5.4: Schematic of iterative alignment procedure for PRRN, which involves two sets of iterations.

Block-based

- Ενδείκνυται για πολλαπλή στοίχιση ακολουθιών που έχουν αποκλείνει αρκετά και έχει απομείνει συντηρημένη μια μικρή περιοχή τους.
- Dialign

Πολλαπλή στοίχιση για DNA & πρωτεΐνες

- Revtrans
  - Παίρνει πολλαπλή στοίχιση των ακολουθιών σε επιπέδο πρωτεϊνών και βάση αυτής, στοιχίζει τις ακολουθίες σε επίπεδο DNA

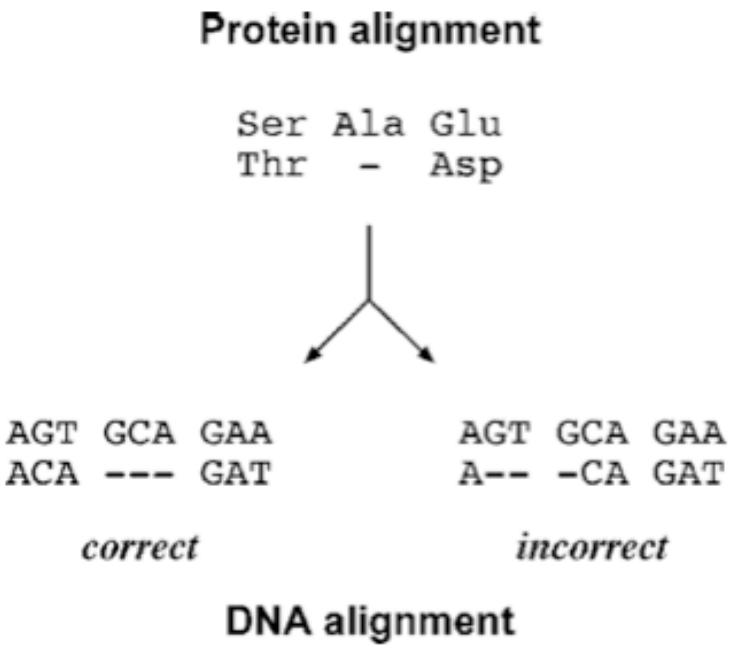
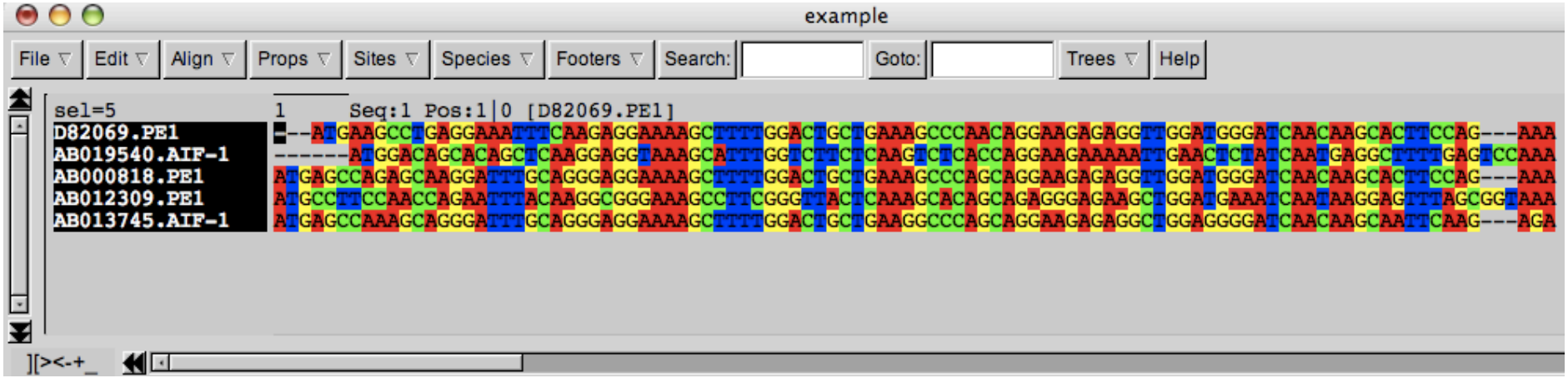


Figure 5.5: Comparison of alignment at the protein level and DNA level. The DNA alignment on the left is the correct one and consistent with amino acid sequence alignment, whereas the DNA alignment on the right, albeit more optimal in matching similar residues, is incorrect because it disregards the codon boundaries.

Χειροκίνητη τροποποίηση/βελτίωση πολλαπλής στοίχισης

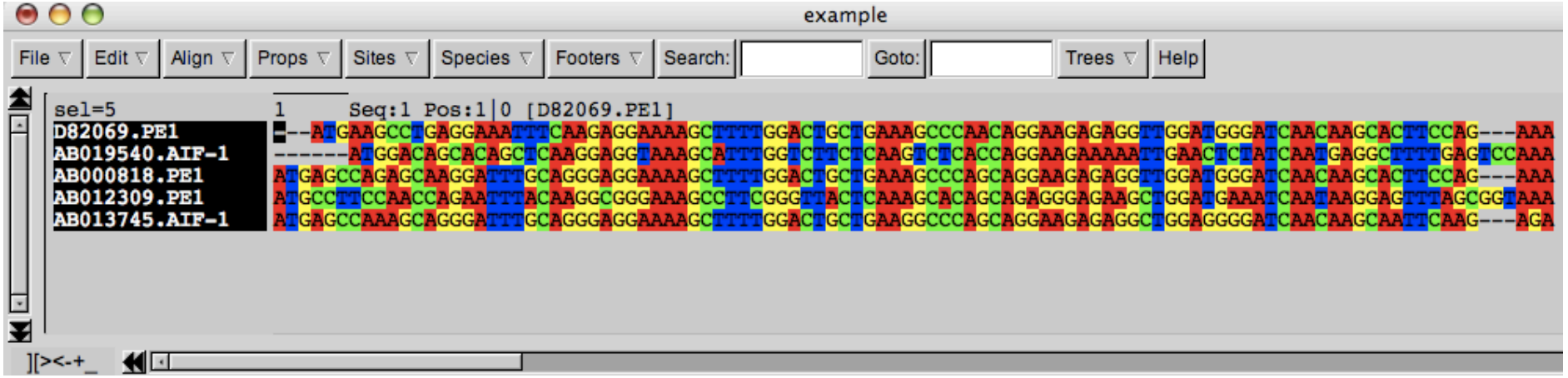
- Τα προγράμματα δεν παράγουν την βέλτιστη στοίχιση.
- Βελτίωση της στοίχισης χειροκίνητα
- alignment editors
  - Seaview
  - Bioedit



Alignment formats

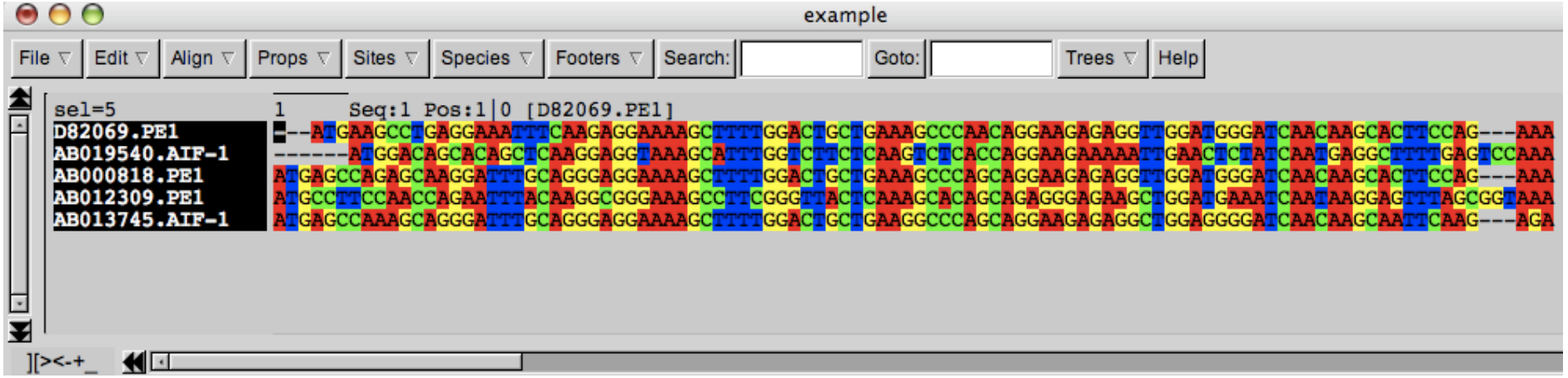
- FASTA (.fa ή .fasta ή .fst)
- Clustal (.aln)
- Phylip (.phy ή .phylip)
- MSF (.msf)
- Mase (.mase)
- Nexus (.nxs)
- Συνήθως, τα alignment editors μπορούν να μετατρέψουν το ένα format σε άλλο.
- Readseq
  - http://www.ebi.ac.uk/cgi-bin/readseq.cgi

Fasta format




```
>D82069.PE1 D82069.PE1 CDS /codon_start=1 /product="iba1, ionized calcium binding adapter mo
---atgaagcctgaggaaatttcaagaggaaaagcttttggactgctgaaagcccaacag
gaagagaggttggatgggatcaacaagcacttccag---aaa
>AB019540.AIF-1 AB019540.AIF-1 CDS /codon_start=1 /transl_table=1 /gene="AIF-1" /product="allogr
------atggacagcacagctcaaggaggtaaagcatttggtcttctcaagtctcaccag
gaagaaaaattgaactctatcaatgaggcttttgagtccaaa
>AB000818.PE1 AB000818.PE1 CDS /codon_start=1 /transl_table=1 /product="MRF-1" /db_xref="GOA:P
atgagccagagcaaggatttgcagggaggaaaagcttttggactgctgaaagcccagcag
gaagagaggttggatgggatcaacaagcacttccag---aaa
>AB012309.PE1 AB012309.PE1 CDS /codon_start=1 /transl_table=1 /product="allograft inflammatory
atgccttccaaccagaatttacaaggcgggaaagccttcgggttactcaaagcacagcag
agggagaagctggatgaaatcaataaggagtttagcggtaaa
>AB013745.AIF-1 AB013745.AIF-1 CDS /codon_start=1 /transl_table=1 /gene="AIF-1" /product="allogr
atgagccaaagcagggatttgcagggaggaaaagcttttggactgctgaaggcccagcag
gaagagaggctggaggggatcaacaagcaattcaag---aga
```

Clustal format



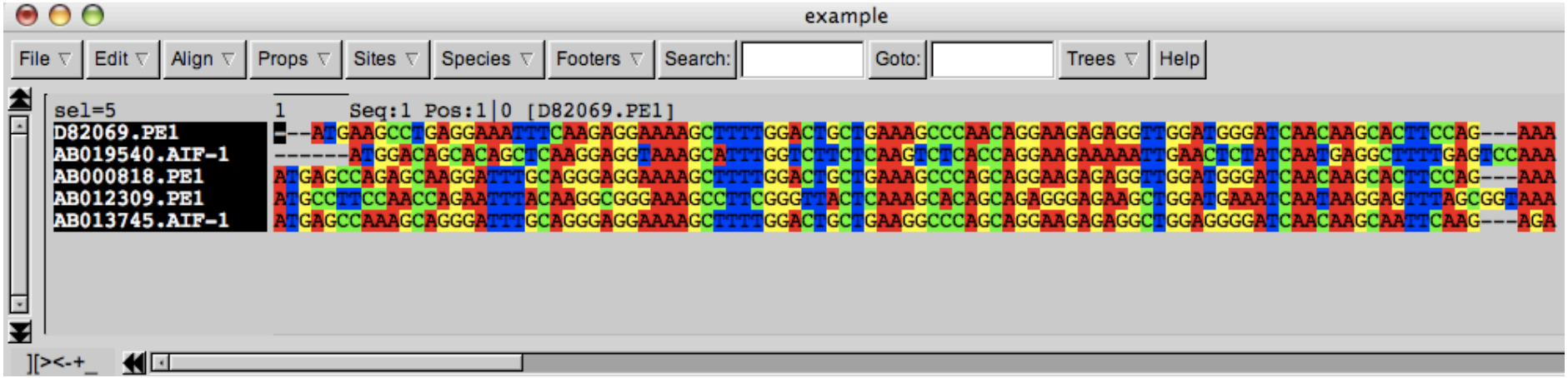

```
CLUSTAL W (1.7) multiple sequence alignment


D82069.PE1        ---atgaagcctgaggaaatttcaagaggaaaagcttttggactgctgaaagcccaacag
AB019540.AIF-1    ------atggacagcacagctcaaggaggtaaagcatttggtcttctcaagtctcaccag
AB000818.PE1      atgagccagagcaaggatttgcagggaggaaaagcttttggactgctgaaagcccagcag
AB012309.PE1      atgccttccaaccagaatttacaaggcgggaaagccttcgggttactcaaagcacagcag
AB013745.AIF-1    atgagccaaagcagggatttgcagggaggaaaagcttttggactgctgaaggcccagcag

D82069.PE1        gaagagaggttggatgggatcaacaagcacttccag---aaa
AB019540.AIF-1    gaagaaaaattgaactctatcaatgaggcttttgagtccaaa
AB000818.PE1      gaagagaggttggatgggatcaacaagcacttccag---aaa
AB012309.PE1      agggagaagctggatgaaatcaataaggagtttagcggtaaa
AB013745.AIF-1    gaagagaggctggaggggatcaacaagcaattcaag---aga
```

Phylip format

- Χρησιμοποιείται στο πρόγραμμα phylip για φυλογένεση




```
5   102
D82069.PE1      ---atgaagc ctgaggaaat ttcaagagga aaagcttttg gactgctgaa agcccaacag
AB019540.AIF-1  ------atgg acagcacagc tcaaggaggt aaagcatttg gtcttctcaa gtctcaccag
AB000818.PE1    atgagccaga gcaaggattt gcagggagga aaagcttttg gactgctgaa agcccagcag
AB012309.PE1    atgccttcca accagaattt acaaggcggg aaagccttcg ggttactcaa agcacagcag
AB013745.AIF-1  atgagccaaa gcagggattt gcagggagga aaagcttttg gactgctgaa ggcccagcag

                gaagagaggt tggatgggat caacaagcac ttccag---a aa
                gaagaaaaat tgaactctat caatgaggct tttgagtcca aa
                gaagagaggt tggatgggat caacaagcac ttccag---a aa
                agggagaagc tggatgaaat caataaggag tttagcggta aa
                gaagagaggc tggaggggat caacaagcaa ttcaag---a ga
```

Πολλαπλή στοίχιση ακολουθιών & profiles

- Ακολουθίες Χ ακολουθίες
- Ακολουθίες Χ profile
- Profile Χ profile

Χρήσεις πολλαπλής στοίχισης

- Δημιουργία:
  - Πινάκων θέσης (Position specific scoring matrices - PSSMs).
  - Profiles.
  - Μαρκοβιανών μοντέλων (Hidden markov models - HMMs).

- Είναι στατιστικά μοντέλα που δείχνουν τη συχνότητα εμφάνισης αμινοξέων/νουκλεοτιδίων για κάθε θέση μιας πολλαπλής στοίχισης.
- Επιπλέον, προβλέπουν τη συχνότητα χαρακτήρων που δεν παρατηρήθηκαν στην πολλαπλή στοίχιση.
- Χρησιμοποιούνται για την ανίχνευση μακρινών ομόλογων ακολουθιών μιας οικογένειας.

PSSMs

- Πολλαπλή στοίχιση χωρίς κενά
- Πίνακας συχνοτήτων για την κάθε θέση

Position	1	2	3	4	5	6
Sequence 1	A	T	G	T	C	G
Sequence 2	A	A	G	A	C	T
Sequence 3	T	A	C	T	C	A
Sequence 4	C	G	G	A	G	G
Sequence 5	A	A	C	C	T	G

Convert multiple alignment to a raw frequency table

Pos.	1	2	3	4	5	6	Overall freq.
A	0.6	0.6	—	0.4	—	0.2	0.30
T	0.2	0.2	—	0.4	0.2	0.2	0.20
G	—	0.2	0.6	—	0.2	0.6	0.27
C	0.2	—	0.4	0.2	0.6	—	0.23

PSSMs

- Κανονικοποίηση του πίνακα συχνοτήτων.
- Μετατροπή των τιμών σε log2

Pos.	1	2	3	4	5	6	Overall freq.
A	0.6	0.6	—	0.4	—	0.2	0.30
T	0.2	0.2	—	0.4	0.2	0.2	0.20
G	—	0.2	0.6	—	0.2	0.6	0.27
C	0.2	—	0.4	0.2	0.6	—	0.23

Normalize the values by dividing them by overall freq.

Pos.	1	2	3	4	5	6	Overall freq.
A	2.0	2.0	—	1.33	—	0.67	0.30
T	1.0	1.0	—	2.0	1.0	1.0	0.20
G	—	0.74	2.22	—	0.74	2.22	0.27
C	0.87	—	1.74	0.87	2.61	—	0.23

Convert the values to log to base of 2

Pos.	1	2	3	4	5	6
A	1.0	1.0	—	0.41	—	-0.58
T	0.0	0.0	—	1.0	0.0	0.0
G	—	-0.43	1.15	—	-0.43	1.15
C	-0.2	—	0.8	-0.2	1.38	—

PSSM

- Τιμή log-odd 1 για ένα χαρακτήρα A στην θέση 1:
  - $2^1=2$: Στην οικογένεια που μελετάμε, η συχνότητα του χαρακτήρα A στην θέση 1 είναι 2 φορές μεγαλύτερη από την συχνότητα υποβάθρου.

- Τιμή log-odd -1 για ένα χαρακτήρα C στην θέση 1:
  - $2^{-1}=1/2$: Στην οικογένεια που μελετάμε, η συχνότητα του χαρακτήρα C στην θέση 1 είναι 2 φορές μικρότερη από την συχνότητα υποβάθρου.

- Τιμή log-odd 0 για ένα χαρακτήρα G στην θέση 1:
  - $2^0=1$: Στην οικογένεια που μελετάμε, η συχνότητα του χαρακτήρα G στην θέση 1 είναι ίδια με την συχνότητα υποβάθρου.

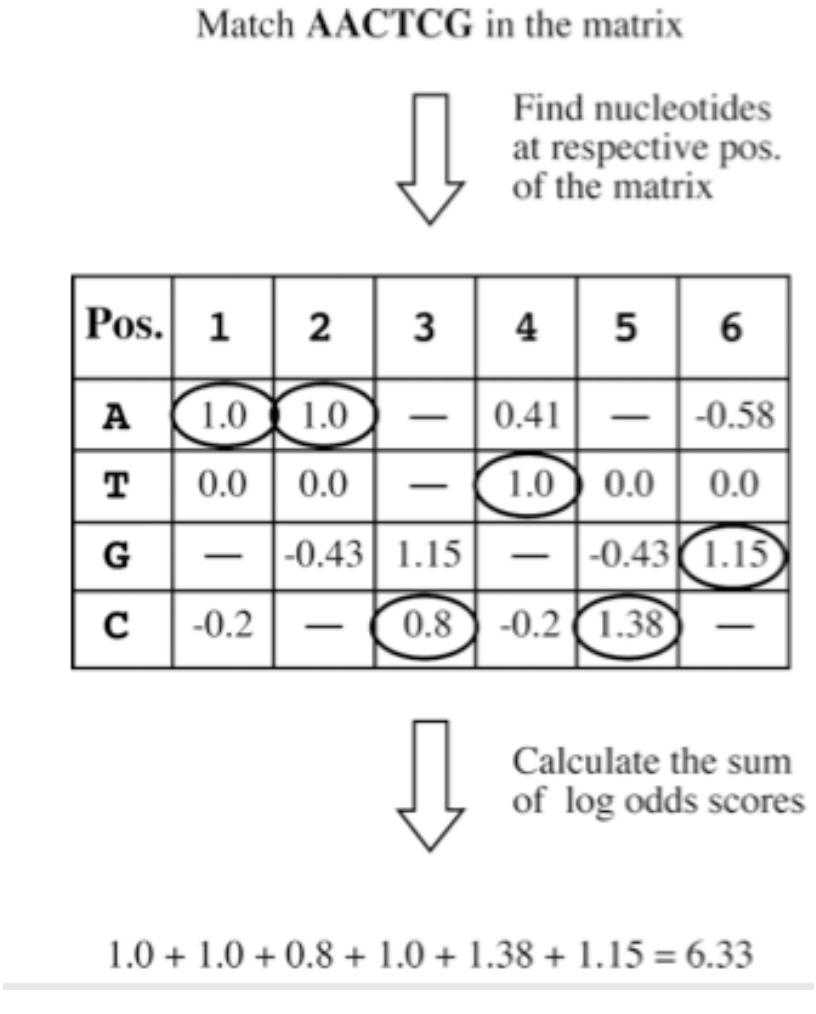
–

PSSM

- Χρησιμοποιείται για
  - Αναζήτηση μακρινών ομόλογων σε βάση δεδομένων.
  - Να υπολογίσουμε πόσο καλά ταιριάζει μια ακολουθία στην οικογένεια.
  - Στοίχιση με ακολουθίες

PSSM

- Πόσο καλά ταιριάζει η ακολουθία AACTCG στον πίνακα θέσης;
- $2^{6.33} = 80$
- Πιθανότητα να ταιριάζει αυτή η ακολουθία στον πίνακα θέσης (ομόλογη) είναι 80 φορές μεγαλύτερη από ότι θα περιμέναμε από μια τυχαία ακολουθία



Pos.	1	2	3	4	5	6
A	1.0	1.0	—	0.41	—	-0.58
T	0.0	0.0	—	1.0	0.0	0.0
G	—	-0.43	1.15	—	-0.43	1.15
C	-0.2	—	0.8	-0.2	1.38	—

PSSM

- Στην πράξη, όταν υπολογίζουμε τις συχνότητες των χαρακτήρων χρησιμοποιούμε συντελεστή βαρύτητας που εξαρτάται από το πόσο όμοιες είναι οι ακολουθίες.
  - Χαμηλός συντελεστής για πολύ όμοιες ακολουθίες.
  - Υψηλός συντελεστής για απομακρυσμένες ακολουθίες.

Profile

Είναι PSSM που περιλαμβάνει και κενά.

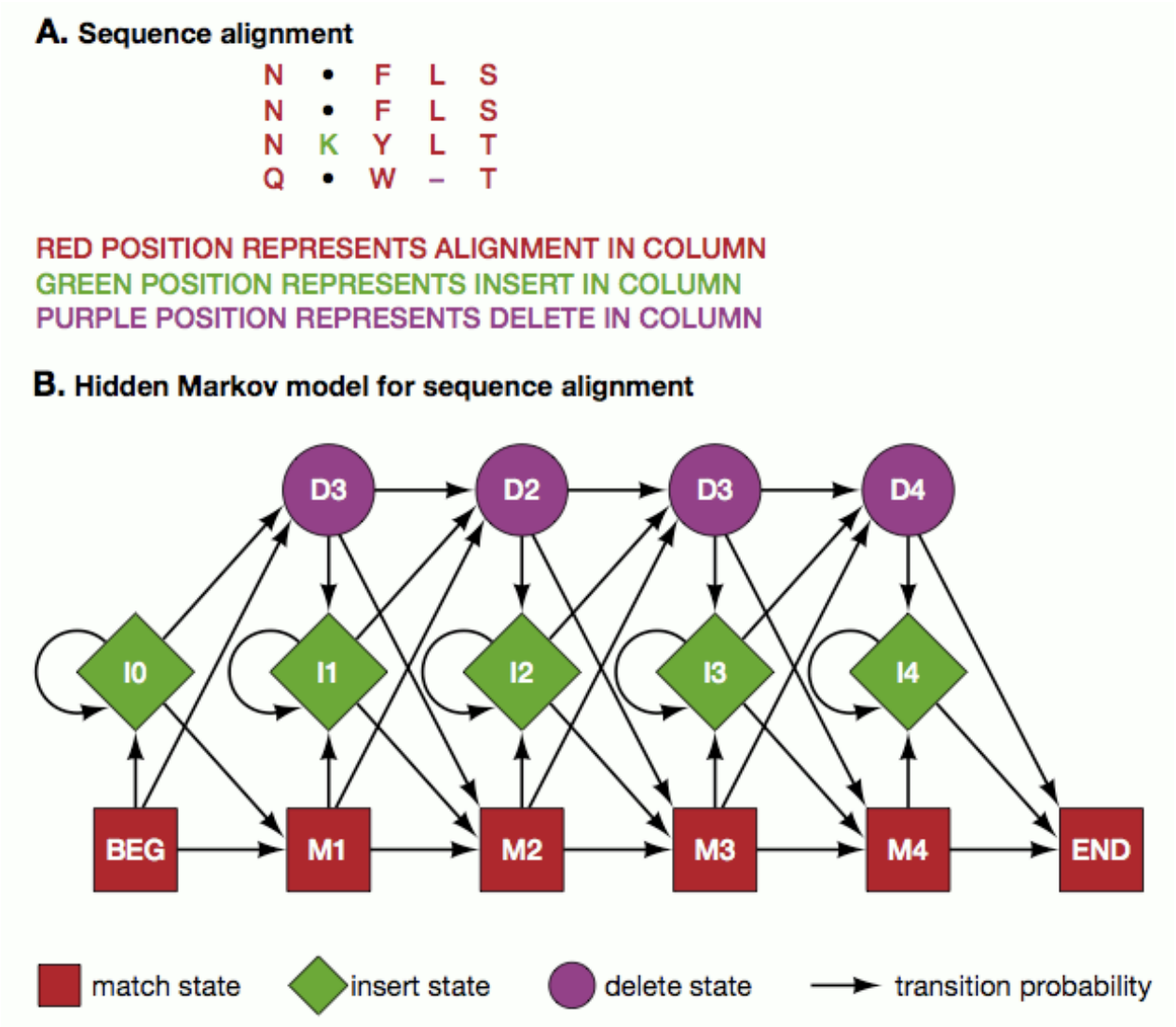
Profile Hidden Markov Models (HMMs)

- Markov models αρχικά χρησιμοποιήθηκαν στην αναγνώριση φωνής.
- Παρόμοια με τα PSSM/profiles.
- Πιο κατάλληλο σύστημα βαθμολόγισης για τα κενά (εισδοχές/ απαλείψεις).
  - Όχι ad hoc, αλλά βασισμένο στις πιθανότητες.
- Για μακρινές ομολογίες, είναι πιο ευαίσθητα από τα profiles.

HMMs

- Χρησιμοποιούνται για:
  - Αναζήτηση ομόλογων ακολουθιών σε Β.Δ.
  - Πολλαπλή στοίχιση ακολουθιών.
  - Κατηγοριοποίηση σε οικογένειες γονιδίων/πρωτεϊνών.
  - Πρόβλεψη γονιδίων (όρια εξονίων/ιντρονίων)
  - Πρόβλεψη διαμεμβρανικών περιοχών πρωτεϊνών.

Profile HMMs



- Στοίχιση του μοντέλου με την ακολουθία μέσω του αλγόριθμου Viterbi (σαν το δυναμικό προγραμματισμό)

HMMs

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

- Regular expression [AT] [CG] [AC] [ACGT]* A [TG] [GC]

The problem with the above regular expression is that it does not in any way distinguish between the highly implausible sequence

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

which has the exceptional character in each position, and the consensus sequence

```
A C A C - - A T C
```

HMMs

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

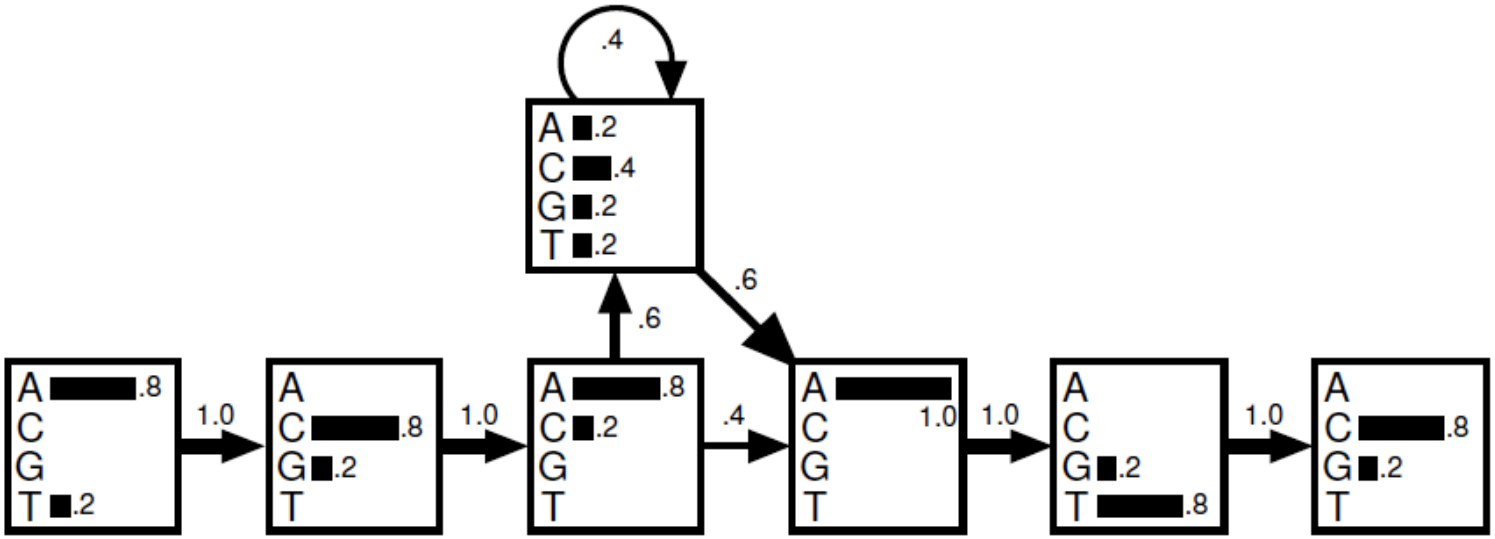
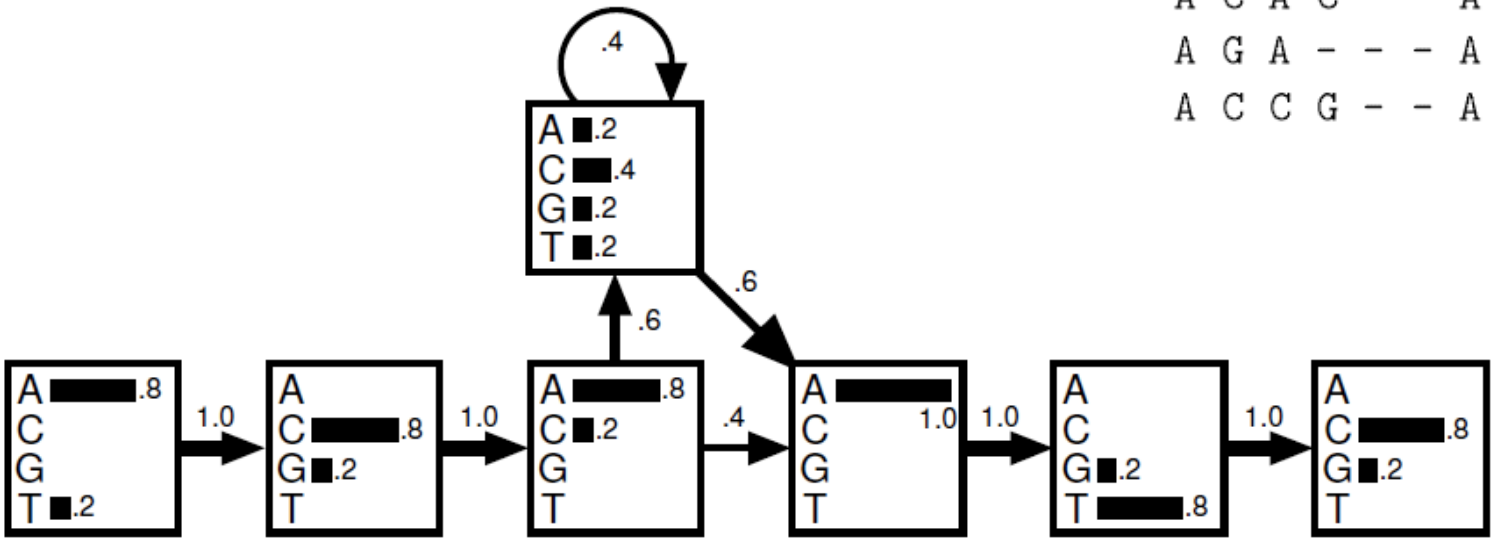


Figure 4.1: A hidden Markov model derived from the alignment discussed in the text. The transitions are shown with arrows whose thickness indicate their probability. In each state the histogram shows the probabilities of the four nucleotides.

HMMs

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



$$
\begin{aligned}
P(\texttt{ACACATC}) &= 0.8 \times 1 \times 0.8 \times 1 \times 0.8 \times 0.6 \times \\
&\quad 0.4 \times 0.6 \times 1 \times 1 \times 0.8 \times 1 \times 0.8 \\
&\simeq 4.7 \times 10^{-2}.
\end{aligned}
$$

HMMs

- Null model: Θεωρεί ότι μια ακολουθία είναι τυχαία.
- Αν θεωρήσουμε ότι και τα 4 νουκλεοτίδια εμφανίζονται με την ίδια συχνότητα (0.25), τότε η πιθανότητα μιας τυχαίας ακολουθίας μήκους L είναι 0.25^L.

- Υπολογίζουμε το log-odds της ακολουθίας:

$$\text{log-odds for sequence } S = \log \frac{P(S)}{0.25^L} = \log P(S) - L \log 0.25$$

HMMs

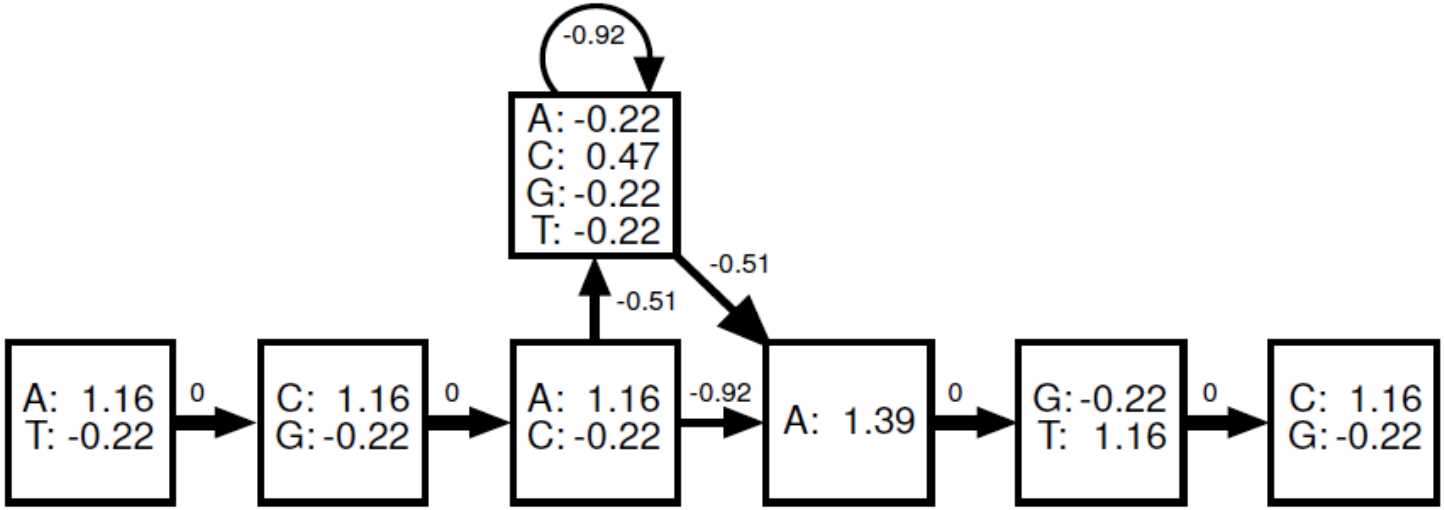


Figure 4.2: The probabilities of the model in Fig. 4.1 have been turned into log-odds by taking the logarithm of each nucleotide probability and subtracting $\log(0.25)$. The transition probabilities have been converted to simple logs.

$$\begin{aligned} \text{log-odds}(\texttt{ACACATC}) &= 1.16+0+1.16+0+1.16-0.51+ \\ &\quad 0.47-0.51+1.39+01.16+0+1.16 \\ &= 6.64. \end{aligned}$$

HMMs

- Overfitting: όταν οι συχνότητες χαρακτήρων υπολογίζονται από ένα μικρό αριθμό ακολουθιών, οι παρατηρούμενες συχνότητες είναι στρεβλωμένες.

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

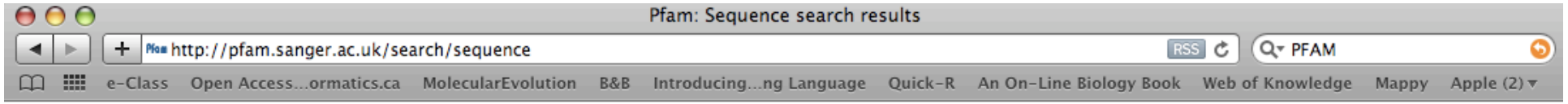
- Pseudocounts: Εξομαλύνουν την παρατηρούμενη συχνότητα χαρακτήρων, με βάση κάποια στατιστικά μοντέλα.
  - Π.χ. Dirichlet mixture (από τις κατανομές αμινοξέων σε domains)

PFAM

- Β.Δ. HMMs για domains (11912).
  - PFAM-A: πολλαπλές στοιχίσεις γνωστών domains που ελέγχθηκαν από ειδικούς
  - PFAM-B: βασίζεται σε συντηρημένες περιοχές πρωτεϊνών που εντοπίστηκαν με αυτόματες μεθόδους και δεν γνωρίζουμε τη λειτουργία τους

- Clans: ομαδοποίηση HMMs (PFAM-A) για ομόλογα domains.
  - Μπορούμε να δημιουργήσουμε ένα HMM που θα χαρακτηρίζει όλη την οικογένεια, ή να δημιουργήσουμε μια σειρά από HMMs, ένα για κάθε υπο-οικογένεια. Όλα μαζί αποτελούν ένα Clan.

PFAM



wellcome trust sanger institute

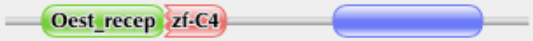
HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam keyword search Go

Sequence search results

Show the detailed description of this results page.

We found **3** Pfam-A matches to your search sequence (**all** significant) but we did not find any Pfam-B matches.



Show the search options and sequence that you submitted.

Return to the search form to look for Pfam domains on a new sequence.

Significant Pfam-A Matches

Show or hide all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To				
Oest_recep	Oestrogen receptor	Family	n/a	42	181	42	181	1	139	172.8	3.1e-51	n/a	Show
zf-C4	Zinc finger, C4 type (two domains)	Domain	n/a	183	252	184	252	2	70	107.8	1.8e-31	n/a	Show
Hormone_recep	Ligand-binding domain of nuclear hormone receptor	Family	n/a	374	543	374	543	1	164	117.7	2.3e-34	n/a	Show

Comments or questions on the site? Send a mail to **pfam-help@sanger.ac.uk**

The Wellcome Trust

PFAM

- Domain architectures
- trees

Motif - Domain

- Motifs:
  - μικρές και συντηρημένες περιοχές που επιτελούν μια συγκεκριμμένη λειτουργία.
- Domains:
  - Συντηρημένες περιοχές, μεγαλύτερες από motifs, συνήθως ώς αυτόνομες λειτουργικές και δομικές μονάδες.
  - 40αα> domain >700aa
  - μέσο μήκος ~100αα

- Συνήθως, οι πρωτεΐνες επιτελούν περισσότερες από μια λειτουργίες. Για κάθε μια είναι υπεύθυνο ένα motif ή domain. Άρα, πρέπει να εξετάζουμε τις επιμέρους βασικές λειτουργικές μονάδες (motifs/ domains), για να κατανοήσουμε όλες τις λειτουργίες μιας πρωτεΐνης.

Regular expressions

Regular expression

- Σχετικά άκαμπτη μέθοδος.
- Λιγότερο ευαίσθητη από ένα στατιστικό μοντέλο.
- Exact matching:
  - Πολλά ψευδώς αρνητικά αποτελέσματα.
- Fuzzy matching:
  - Επιτρέπει αμινοξέα με παρόμοιες φυσικοχημικές ιδιότητες, ακόμα και αν δεν παρατηρήθηκαν στην πολλαπλή στοίχιση.
  - Αυξάνεται ο θόρυβος (ψευδώς θετικά).

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

[AT] [CG] [AC] [ACGT]* A [TG] [GC]

The problem with the above regular expression is that it does not in any way distinguish between the highly implausible sequence

T G C T - - A G G

which has the exceptional character in each position, and the consensus sequence

A C A C - - A T C

Regular expression DBs.

- PROSITE:
  - Η πρώτη Β.Δ. του είδους της.
  - Τα regular expressions δημιουργούνται από πολλαπλές στοιχίσεις συντηρημένων περιοχών.
  - Exact matches.
  - Επίσης δημιουργεί και profiles.

- Emotif:
  - Πολλαπλές στοιχίσεις από τις ΒΔ BLOCKS & PRINTS.
  - Μεγαλύτερη συλλογή πολλαπλών στοιχίσεων από την PROSITE.
  - Fuzzy matching.

Στατιστικά μοντέλα

- PSSM (position specific scoring matrices).
- Profiles.
- HMMs (hidden markov models).

- Επιτρέπουν μερικό ταίριασμα.
- Pseudocounts.

ΒΔ πολλαπλών στοιχίσεων motifs/domains

- PRINTS:
  - Fingerprints: περιοχές της πολλαπλής στοίχισης, βαθειά συντηρημένες και χωρίς κενά.
  - PSSMs (δίχως συντελεστή βαρύτητας) για τα fingerprints.
  - Ένα motif αποτελείται από >1 fingerprints (δεν υπάρχει αλληλεπικάλυψη).
  - Το motif θεωρείται υπάρχων σε μια πρωτεΐνη όταν η πλειοψηφία των fingerprints που το απαρτίζουν έχει ανιχνευθεί.
  - Ορισμός των fingerprints & motifs γίνεται από βιοεπιστήμονες/βιοπληροφορικούς.
  - Σχετικά μικρός αριθμός motifs στη ΒΔ.

ΒΔ πολλαπλών στοιχίσεων motifs/domains

- BLOCKS:
  - Αυτοματοποιημένη πολλαπλή στοίχιση πρωτεϊνικών οικογενειών, όπου χρησιμοποιούνται οι πιο συντηρημένες περιοχές, δίχως κενά (blocks).
  - Για κάθε block δημιουργείται PSSM (με συντελεστή βαρύτητας) και εφαρμόζονται pseudocounts.
  - Οι πίνακες αντικατάστασης BLOSSUM υπολογίζονται από τη ΒΔ BLOCKS.

ΒΔ πολλαπλών στοιχίσεων motifs/domains

- ProDom:
  - Δημιουργεί domains εφαρμόζοντας PSI-Blast σε ακολουθίες από την SWISSPROT & TrEMBL.
  - Η λειτουργία των domains μπορεί να είναι άγνωστη.

ΒΔ πολλαπλών στοιχίσεων motifs/domains

- SMART:
  - Profile HMMs που δημιουργήθηκαν από πολλαπλές στοιχίσεις, ελεγμένες από ειδικούς.
  - Οι στοιχίσεις είτε βασίζονται σε τρισδιάστατες δομές είτε σε profiles που δημιουργεί το PSI-Blast.
  - Και οι στοιχίσεις και ο σχολιασμός των profile HMMs γίνεται από ειδικούς.
  - Συμπληρωματικότητα με την PFAM.

ΒΔ πολλαπλών στοιχίσεων motifs/domains

- INTERPRO:
  - Λόγω ανομοιογένειας στις μεθοδολογίες και στις ακολουθίες που χρησιμοποιούνται, υπάρχει μερική αλληλοεπικάλυψη αλλά και συμπληρωματικότητα μεταξύ των επιμέρους ΒΔ motifs/domains.
  - Η INTERPRO ενσωματώνει αλληλοεπικαλυπτόμενα motifs/ domains που βρίσκονται ταυτόχρονα και στις 5 παρακάτω ΒΔ:
    - PROSITE
    - PFAM
    - PRINTS
    - ProDOM
    - SMART

ΒΔ πολλαπλών στοιχίσεων motifs/domains

- Reverse-Blast (RPS-Blast):
  - Συλλογή profiles που δημιουργήθηκαν από PSI-Blast.

- CDART:
  - Τμήμα του BLAST.
  - ενσωματώνει τις
    - RPS-Blast
    - PFAM
    - SMART

Γραφική απεικόνιση motifs/ profiles: LOGOs

- Weblogo

GDLGAGKTT
GDLGAGKTT
GPLGAGKTS
GDLGAGKTS
GDLGAGKTT
GDLGAGKTT
GEVGSGKTT
GELGAGKTT
GDLGAGKTT
GNLGAGKTT
GELGAGKTT
GTLGAGKTT
GDLGAGKTT
GDLGAGKTT
GDLGAGKTT
GDLGAGKTT
GDLGAGKTT