Βιοπληροφορική

Διάλεξη 3η :

Στοίχιση αλληλουχιών : εισαγωγή.
Πίνακες βαθμολόγησης (PAM, BLOSUM).
Στοίχιση δύο αλληλουχιών : Dot plots, Σχολαστικοί αλγόριθμοι : Ο αλγόριθμος των Needleman & Wunsch.

Στοίχιση αλληλουχιών

Είναι μια από τις βασικότερες διαδικασίες της μοριακής βιολογίας και η 'εκ των ουκ άνευ' της βιοπληροφορικής. Η μέθοδος είναι αναπόσπαστα δεμένη με την εξέλιξη (κληρονομούμενες αλλαγές πληροφορίας) ασχέτως με το εάν η εξέλιξη αποτελεί τμήμα του προβλήματος

[«πριν από πόσα Myr διαχωρίστηκαν τα είδη αυτά ;»]

ή όχι

[«κωδικοποιεί το γονίδιο αυτό για μια πρωτεάση ;»].

Εφαρμογές στοιχίσεων

- Έρευνα βάσεων δεδομένων για αναζήτηση ομολόγων αλληλουχιών και (δυνητικά) ταυτοποίηση λειτουργίας.
- Συστηματική/ταξινομική αλληλουχιών, γονιδιωμάτων, οργανισμών.
- Συστηματική/ταξινομική οικογενειών πρωτεϊνών και ταυτοποίηση συντηρημένων (και συνεπώς, λειτουργικά/δομικά σημαντικών) καταλοίπων.
- Αναγνώριση και ταυτοποίηση μοτίβων σε πρωτεϊνικές οικογένειες.
- Πρόβλεψη της δομής μίας πρωτεΐνης με βάση την ομολογία της με άλλη πρωτεΐνη γνωστής δομής (homology modelling).

Ομολογία, αναλογία, ομοιότητα

Δυο αλληλουχίες είναι ομόλογες εάν έχουν αποκλίνει (εξελικτικά) από μία κοινή προγονική αλληλουχία.
Η ομολογία είναι μια απόλυτη δήλωση σχέσης : δύο αλληλουχίες ή είναι ή δεν είναι ομόλογες (η ομολογία δεν επιδέχεται ποσοτικούς προσδιορισμούς).

Δυο αλληλουχίες είναι ανάλογες εάν είναι προϊόντα συγκλίνουσας εξέλιξης, δηλαδή εάν έχουν αντίστοιχες δομές ή και λειτουργίες αλλά χωρίς να έχουν προέλθει (εξελικτικά) από κάποια κοινή προγονική αλληλουχία.
Η αναλογία είναι επίσης μια απόλυτη δήλωση απουσίας εξελικτικής σχέσης.

Ομολογία, αναλογία, ομοιότητα

Η ομοιότητα δύο αλληλουχιών είναι ένα ποσοτικό μέτρο των μεταξύ τους κοινών χαρακτηριστικών. Ένα από τα ευρέως χρησιμοποιούμενα μέτρα είναι το επί τοις εκατό ποσοστό ταυτότητας μεταξύ των στοιχισμένων αλληλουχιών (% sequence identity), αν και υπάρχουν και άλλα μέτρα τα οποία λαμβάνουν υπόψη το κατά πόσο οι παρατηρούμενες αλλαγές είναι συντηρητικές ή όχι [το οποίο, για να επιτείνει τη σύγχυση, αναφέρεται από πολλά προγράμματα ως επί τοις εκατό «ομοιότητα» (% similarity), η οποία είναι διακριτή από την επί τοις εκατό ταυτότητα].

Ορθόλογες και παράλογες αλληλουχίες

Ορθόλογες (orthologues) ονομάζονται δύο ομόλογες αλληλουχίες οι οποίες έχουν την ίδια λειτουργία αλλά προέρχονται από διαφορετικά είδη. Οι ορθόλογες αλληλουχίες είναι το κυρίως υλικό για τη δημιουργία φυλογενετικών δένδρων.

Παράλογες (paralogues) ονομάζονται δύο ομόλογες αλληλουχίες ενός και του αυτού είδους οι οποίες έχουν παρόμοιες (αλλά όχι πανομοιότυπες) λειτουργίες. Αυτές δίνουν πληροφορία για την εξελικτική πορεία γονιδίων μετά από φαινόμενα γονιδιακού διπλασιασμού.

Το πρόβλημα

Το ζητούμενο της στοίχισης αλληλουχιών είναι να αποκαλυφθούν οι μεταξύ τους εξελικτικές σχέσεις [δηλ. μεταλλάξεις & εισαγωγές/διαγραφές].
Όντως, μία οποιαδήποτε στοίχιση μεταξύ αλληλουχιών υποδηλώνει ένα εξελικτικό σενάριο π.χ.

```
AGGACTGCG-TAG-TAC
AGG---TCGATAGGCAC
```

ή

```
AGGACT-GCGTA-GTAC
AGG--TCGA-TAGGCAC
```

ή ...

Το πρόβλημα

Δυστυχώς η εξελικτική ιστορία των αλληλουχιών μας είναι άγνωστη (εάν δεν ήταν δεν θα χρειαζόμασταν τη διαδικασία στοίχισης). Αυτό έχει δύο συνέπειες :

1. Δεν γνωρίζουμε εάν όντως οι αλληλουχίες έχουν εξελικτική σχέση (δηλαδή εάν είναι όντως στοιχίσιμες).
2. Εάν έχουν εξελικτική σχέση, το μόνο που συνήθως έχουμε σαν βάση για να συνάγουμε τη στοίχιση είναι η παροντική μορφή των αλληλουχιών (εκτός και εάν υπάρχει διαθέσιμη η αλληλουχία από κάποιο κοινό εξελικτικό πρόγονο).

Η μηδενική υπόθεση

Επειδή δεν γνωρίζουμε εάν όντως οι αλληλουχίες έχουν εξελικτική σχέση, πρέπει να τροποποιήσουμε το στόχο της στοίχισης αλληλουχιών ως εξής :
Στόχος μας είναι να συγκρίνουμε δύο διαφορετικές υποθέσεις. Η μηδενική υπόθεση είναι ότι οι αλληλουχίες δεν έχουν εξελικτική σχέση (ότι δηλ. δεν έχουν προκύψει από κάποιο κοινό εξελικτικό πρόγονο). Αυτή η υπόθεση υπονοεί ότι η οποιαδήποτε ομοιότητα μεταξύ των αλληλουχιών είναι τυχαία και δεν οφείλεται σε εξελικτικές διαδικασίες.
Για παράδειγμα :

Η μηδενική υπόθεση

Με τη βοήθεια μίας γεννήτριας τυχαίων αριθμών παρήχθησαν οι κάτωθι αλληλουχίες A & B :

```
A  :  CAGTATTGCTAGCATTG
B  :  GTATGCGGAACGTATTC
```

Παρότι δεν υπάρχει καμία μεταξύ τους εξελικτική σχέση, οι δυο αυτές αλληλουχίες έχουν ομοιότητες, όπως φαίνεται από την παρακάτω στοίχιση

```
A  :  CAGTATTGC---TAGCATTG
B  :    GTAT-GCGGAACGTATTC
```

Σε αυτό το τεχνητό παράδειγμα γνωρίζουμε εκ των προτέρων ότι αυτές οι ομοιότητες είναι τυχαίες. Αλλά εάν το μόνο που είχαμε σαν δεδομένο ήταν οι καθ'αυτό αλληλουχίες θα συναγάγαμε το ίδιο ;

Η εξελικτική υπόθεση

Η δεύτερη υπόθεση (την οποία επιθυμούμε να συγκρίνουμε με το μοντέλο της τυχαίας ομοιότητας) είναι οι αλληλουχίες να έχουν εξελικτική σχέση, δηλ. να έχουν προκύψει από έναν κοινό εξελικτικό πρόγονο μέσω μεταλλάξεων και indels. Σε αυτή την περίπτωση θα θέλαμε η στοίχιση να αποκαλύπτει αυτά τα εξελικτικά γεγονότα.

Για παράδειγμα :

Εξελικτική υπόθεση : παράδειγμα

Ξεκινώντας από την αλληλουχία ACGTACGT, και χρησιμοποιώντας σταθερούς ρυθμούς μεταλλάξεων και γεγονότων εισαγωγών/διαγραφών, προέκυψε μετά από ~9000 γενιές η αλληλουχία ACACGGTCCTAATAATGGCC. Επανάληψη της ίδιας διαδικασίας (ξεκινώντας από την ACGTACGT και για ίδιο αριθμό γενιών), έδωσε την αλληλουχία CAGGAAGATCTTAGTTC. Επειδή σε αυτή την περίπτωση η ιστορία των αλληλουχιών μας είναι γνωστή, μπορούμε να στοιχίσουμε την αρχική αλληλουχία με κάθε μία από τις τελικές :

Εξελικτική υπόθεση : παράδειγμα

```
--ACG-T-A---CG-T-----
ACACGGTCCTAATAATGGCC
```

και

```
---AC-GTA-C--G-T--
CAG-GAAGATCTTAGTTC
```

και με υπέρθεση να βρούμε την εξελικτικά ορθή στοίχιση μεταξύ των τελικών αλληλουχιών :

```
-ACAC-GGTCCTAAT--AATGGCC
CAG-GAA-G-AT--CTTAGTTC--
   *       *     * *
```

Εξελικτική υπόθεση : παράδειγμα

```
--ACG-T-A---CG-T-----
ACACGGTCCTAATAATGGCC
```

και

```
---AC-GTA-C--G-T--
CAG-GAAGATCTTAGTTC
```

και με υπέρθεση να βρούμε την εξελικτικά ορθή στοίχιση μεταξύ των τελικών αλληλουχιών :

```
-ACAC-GGTCCTAAT--AATGGCC
CAG-GAA-G-AT--CTTAGTTC--
   *       *     * *
```

Μόνο που ένας αλγόριθμος θα έδινε άλλη στοίχιση ...

```
ACACG--GTCCTAATAATGGCC
-CAGGAAGATCT--TAGTT--C
** *  *  **  ** *   *
```

Το πρόβλημα

Άρα το πρόβλημα της στοίχισης αλληλουχιών είναι
1. Να υπολογίσουμε την πιθανότητα η ομοιότητα των αλληλουχιών να οφείλεται σε εξελικτική σχέση, και,
2. Να προσδιορίσουμε τη στοίχιση εκείνη που καλύτερα αναπαριστά την εξελικτική σχέση (ιστορία) των αλληλουχιών.

Δυστυχώς, είναι αδύνατο να υπολογίσουμε το [1ο] χωρίς να έχουμε ήδη κάνει μια στοίχιση. Αυτό συμβαίνει γιατί ο υπολογισμός της στατιστικής σημασίας μιας στοίχισης απαιτεί να υπάρχει η στοίχιση (ώστε να μπορούμε στη συνέχεια να ρωτήσουμε «ποιά είναι η πιθανότητα να έχει προκύψει αυτή η στοίχιση τυχαία;»).

Η λύση

- Υποθέτουμε (a priori) πως οι αλληλουχίες έχουν εξελικτική σχέση.
- Προσδιορίζουμε την στοίχιση η οποία μεγιστοποιεί την μεταξύ τους (εξελικτική) ομοιότητα.
- Ελέγχουμε την στατιστική σημαντικότητα της προκύπτουσας στοίχισης :

  - Εάν η ομοιότητα της στοίχισης δεν είναι στατιστικά σημαντική, καταρρίπτουμε την αρχική μας υπόθεση και συνάγουμε ότι οι αλληλουχίες δεν είναι ομόλογες.

  - Στην αντίθετη περίπτωση πιθανολογούμε πως οι αλληλουχίες είναι ομόλογες και αναζητούμε επιπρόσθετες (μη υπολογιστικές ;) ενδείξεις.

Τα υπολογιστικά προβλήματα

Το πρόβλημα μας τώρα είναι το εξής :
Δεδομένων των αλληλουχιών, πως θα προσδιορίσουμε την στοίχιση εκείνη που μεγιστοποιεί την μεταξύ τους (εξελικτική) ομοιότητα.

Το οποίο, με τη σειρά του, μας φέρνει στο πρόβλημα του πως μετράμε την «εξελικτική ομοιότητα» (ώστε να μπορούμε να την μεγιστοποιήσουμε).

Πίνακες βαθμολόγησης

Πίνακες βαθμολόγησης

Οι πίνακες βαθμολόγησης είναι απόπειρες να κωδικοποιηθούν αριθμητικά οι μέσες (στατιστικά) προτιμήσεις της φυσικής επιλογής σε ό,τι αφορά τις μεταλλάξεις (αμινοξικές αλλαγές).

Οι τιμές που περιέχουν είναι (μεταφορικά) ενδεικτικές για το τι άποψη έχει η φυσική επιλογή για τα πεπραγμένα της εξέλιξης. Δεν κωδικοποιούν τις συχνότητες συγκεκριμένων μεταλλάξεων, αλλά τις συχνότητες με τις οποίες οι διάφορες μεταλλάξεις γίνονται αποδεκτές.

Πίνακες βαθμολόγησης

Επειδή η φυσική επιλογή δρα στο επίπεδο των φαινοτύπων (και όχι των γονοτύπων), θα επικεντρωθούμε από εδώ και πέρα στις στοιχίσεις πρωτεϊνικών αλληλουχιών. Οι διαδικασίες και αλγόριθμοι που θα περιγραφούν μεταφέρονται σχεδόν αυτούσιοι και για αλληλουχίες νουκλεϊκών οξέων.

Πίνακες βαθμολόγησης

Ο πλέον απλοϊκός πίνακας βαθμολόγησης είναι :
Για κάθε θέση που οι αλληλουχίες ταυτίζονται,
αύξησε την βαθμολογία κατά ένα σταθερό θετικό
αριθμό, αλλιώς προχώρησε στην επόμενη θέση.
Αυτό σε μορφή πίνακα θα μπορούσε να γραφτεί :

```
A   1
C   0   1
M   0   0   1
P   0   0   0   1
F   0   0   0   0   1
.....................
    A   C   M   P   F ...
```

'Μοναδιαίοι πίνακες βαθμολόγησης'.

Πίνακες βαθμολόγησης

Ένας τέτοιου τύπου πίνακας έχει ελάχιστη ευαισθησία (διαγνωστική ισχύ) : αγνοεί πλήρως τις σχέσεις συγγένειας μεταξύ των διάφορων ομάδων αμινοξέων (π.χ. πολικά, υδρόφοβα, φορτισμένα, ...) όπως αυτές προκύπτουν από τις φυσικοχημικές τους ιδιότητες (αλλά και από τον γενετικό κώδικα).

Πίνακες βαθμολόγησης

	T		C		A		G		
T	TTT	Phe	TCT	Ser	TAT	Try	TGT	Cys	T
	TTC		TCC		TAC		TGC		C
	TTA	Leu	TCA		TAA	**Stop**	TGA	**Stop**	A
	TTG		TCG		TAG		TGG	Trp	G
C	CTT	Leu	CCT	Pro	CAT	His	CGT	Arg	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	Gln	CGA		A
	CTG		CCG		CAG		CGG		G
A	ATT	Ile	ACT	Thr	AAT	Asn	AGT	Ser	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	Lys	AGA	Arg	A
	ATG	**Met**	ACG		AAG		AGG		G
G	GTT	Val	GCT	Ala	GAT	Asp	GGT	Gly	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	Glu	GGA		A
	GTG		GCG		GAG		GGG		G

Οι πίνακες PAM

PAM 250, Deyhoff, et al. (1978)

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

Οι πίνακες PAM

PAM είναι τα αρχικά του Point Accepted Mutation.
Το PAM είναι ένα μέτρο της εξελικτικής απόκλισης μεταξύ δύο αλληλουχιών τέτοιο ώστε το 1 PAM να είναι η εξελικτική απόσταση που απαιτείται ώστε να μεταλλαχθεί το 1% των αμινοξέων μιας πρωτεΐνης.
Μία απόσταση 100 PAM δεν υπονοεί ότι όλα τα αμινοξέα είναι διαφορετικά. Μερικά θα μεταλλαχθούν περισσότερες από μία φορές, άλλα καμία (ανάλογα με την πίεση της φυσικής επιλογής). Κατά μέσο όρο, μία απόσταση 250 PAM αντιστοιχεί σε ~20% ταυτότητα.

Οι πίνακες PAM

Εάν δεν υπήρχε η πίεση της φυσικής επιλογής (και εάν συνεπώς όλες μεταλλάξεις οι γινόντουσαν αποδεκτές), τότε η συχνότητα εμφάνισης των διάφορων αμινοξικών αλλαγών θα ήταν ανάλογη της συχνότητας εμφάνισης των συγκεκριμένων αμινοξέων στις πρωτεΐνες.

Παράδειγμα : εάν δεν υπήρχε η φυσική επιλογή, και εάν η συχνότητα εμφάνισης της τρυπτοφάνης (W) είναι π.χ. μόνο 1%, τότε οι μεταλλάξεις προς την τρυπτοφάνη θα είναι σχετικά σπάνιες. Οι αναμενόμενες συχνότητες εμφάνισης των μεταλλάξεων με βάση τις συχνότητες εμφάνισης των αμινοξέων (στις πρωτεϊνικές αλληλουχίες) ονομάζονται "συχνότητες υποβάθρου".

Οι πίνακες PAM

Παρουσία της φυσικής επιλογής, οι συχνότητες εμφάνισης των διαφόρων μεταλλάξεων αλλάζουν. Η αλλαγή οφείλεται στο ότι επιλέγονται κατά προτίμηση οι μεταλλάξεις εκείνες οι οποίες είναι συμβατές με την δομή και λειτουργία της πρωτεΐνης (δηλ. εκείνες που δεν μειώνουν την λειτουργικότητα της). Με άλλα λόγια, επιλέγονται σημειακές μεταλλάξεις οι οποίες έγιναν αποδεκτές από τη φυσική επιλογή (εξ ου και 'point accepted mutations'). Οι συχνότητες εμφάνισης μεταλλάξεων όπως τις παρατηρούμε σε ομόλογες πρωτεΐνες (παρουσία της φυσικής επιλογής) ονομάζονται "παρατηρούμενες συχνότητες".

Οι πίνακες PAM

Η κάθε καταχώρηση στον πίνακα της Deyhoff, είναι ο φυσικός λογάριθμος του πηλίκου της παρατηρούμενης συχνότητας δια την συχνότητα υποβάθρου της μετάλλαξης (log-odds ratio). Για παράδειγμα, η μετάλλαξη W <=> K (τρυπτοφάνη-λυσίνη) έχει τιμή -3. Άρα,

ln [Pobs / Pback] = -3

=> Pobs / Pback = exp(-3) = 0.0497

=> Pobs = 0.0497 · Pback

=> Pobs = Pback / 20

το οποίο σημαίνει ότι παρουσία της φυσικής επιλογής, μια μετάλλαξη από τρυπτοφάνη προς λυσίνη συμβαίνει 20 φορές πιο σπάνια απ'ότι θα συνέβαινε κατά τύχη.

Οι πίνακες PAM

Με βάση τις ιδιότητες των λογαρίθμων λοιπόν, όταν κάποια τιμή του πίνακα είναι θετική, τότε η αντίστοιχη μετάλλαξη παρατηρείται πιο συχνά απ'ότι θα περιμέναμε εάν συνέβαινε τυχαία. Άρα, οι θετικές τιμές του πίνακα αντιστοιχούν σε προτιμώμενες (από την φυσική επιλογή) μεταλλάξεις.

Κατ'αναλογία, οι αρνητικές τιμές του πίνακα αντιστοιχούν σε μεταλλάξεις που σπανίως γίνονται αποδεκτές, ενώ μία τιμή ίση με το μηδέν αντιστοιχεί σε ουδέτερες μεταλλάξεις.

Οι πίνακες PAM

PAM 250

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

Οι πίνακες PAM

Ένας φαύλος κύκλος ;

Προκειμένου να υπολογίσουμε τις απαιτούμενες συχνότητες μεταλλάξεων, πρέπει να έχουμε ήδη βρει οικογένειες ομόλογων πρωτεϊνών και να τις έχουμε στοιχίσει. Άρα, για να δημιουργήσουμε τον πίνακα χρειαζόμαστε το προϊόν εφαρμογής του. Η λύση αυτού του φαύλου κύκλου είναι ότι για αλληλουχίες υψηλής ομολογίας (αλληλουχίες που διαχωρίστηκαν πρόσφατα) η στοίχιση είναι προφανής και συνήθως περιορίζεται στην απλή παράθεση των αλληλουχιών. Από αυτές τις στενά συσχετιζόμενες αλληλουχίες μπορούμε εύκολα να υπολογίσουμε πίνακες όπως ο PAM1.

Οι πίνακες PAM

Ένας φαύλος κύκλος ;

Οι πίνακες για πιο μακρινές (εξελικτικά) σχέσεις (π.χ. PAM200 ή PAM250) προκύπτουν με απλό πολλαπλασιασμό πινάκων ξεκινώντας από τους χαμηλότερης 'τάξης' πίνακες. Ένα παράδειγμα για κάποιες υποθετικές αλληλουχίες οι οποίες αποτελούνται από μόνο δύο τύπων μονομερή (α,β) είναι :

PAM (x) **PAM (2x)**

	α	β
α	2	-1
β	-1	3

=>

	α	β
α	5	-5
β	-5	10

Οι πίνακες BLOSUM

Ένας φαύλος κύκλος ;

Το πρόβλημα με αυτήν την προσέγγιση, είναι ότι για την στοίχιση απόμακρων εξελικτικά αλληλουχιών (που είναι και αυτό που περισσότερο μας ενδιαφέρει), δεν χρησιμοποιούμε "παρατηρούμενες" συχνότητες μεταλλάξεων, αλλά συχνότητες που τις συνάγαμε υπολογιστικά από αλληλουχίες με μεγάλη ομοιότητα.

Οι πίνακες BLOSUM

Henikoff & Henikoff (1992)

Οι πίνακες BLOSUM (για BLOcks SUbstitution Matrix) αποφεύγουν το προαναφερθέν πρόβλημα με το να υπολογίζουν τις παρατηρούμενες συχνότητες μεταλλάξεων από εξελικτικά απομακρυσμένες ομόλογες αλληλουχίες.

Το πρόβλημα αρχικής στοίχισης αυτών των αλληλουχιών αποφεύγεται με το μη χρησιμοποιούνται ολόκληρες οι αλληλουχίες, παρά μόνο πρωτεϊνικά μοτίβα (χωρίς κενά) από τη βάση BLOCKS.

Οι πίνακες BLOSUM

```
OPN3_HUMAN/293-309      VSylfAKSNTvyNPviY
OPN3_MOUSE/291-307      VSylfAKSSTvyNPviY
OPN4_HUMAN/334-350      VPaviAKASAihNPiiY
OPN4_MOUSE/331-347      VPaviAKASAihNPiiY
OPS1_CALVI/311-327      WGacfAKSAAcyNPivY
OPS1_DROME/313-329      WGacfAKSAAcyNPivY
OPS1_DROPS/314-330      WGacfAKSAAcyNPivY
OPS1_HEMSA/319-335      LPallAKSCScyNPfvY
OPS1_LIMPO/312-328      WGsvfAKANScyNPivY
OPS1_PATYE/316-332      LPmmlAKSSSmhNPvvY
OPS1_SCHGR/317-333      WGslfAKANAvfNPivY
OPS2_DROME/320-336      WGatfAKTSAvyNPivY
OPS2_DROPS/320-336      WGatfAKTSAvyNPivY
OPS2_HEMSA/319-335      LPallAKSCScyNPfvY
OPS2_LIMPO/312-328      WGsvfAKANScyNPivY
OPS2_PATYE/276-292      LPtlfAKASCayNPfiY
```

BLOSUM62

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4																			
R	-1	5																		
N	-2	0	6																	
D	-2	-2	1	6																
C	0	-3	-3	-3	9															
Q	-1	1	0	0	-3	5														
E	-1	0	0	2	-4	2	5													
G	0	-2	0	-1	-3	-2	-2	6												
H	-2	0	1	-1	-3	0	0	-2	8											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

PAM250

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

Βαθμολόγηση στοιχίσεων

Ο πίνακας βαθμολόγησης είναι ένα σημαντικό βήμα για να μπορούμε να βαθμολογήσουμε μια στοίχιση, άλλα δεν αρκεί. Αυτό που λείπει είναι ένας τρόπος να λάβουμε υπόψη τα γεγονότα εισαγωγών/διαγραφών. Για παράδειγμα, οι αλληλουχίες

```
α GATGGCGTAGCTAGGATTAACA
β AGTGCATTTAGATCAGGATCTTCTATAGC
```

μπορούν να στοιχιθούν έτσι,

```
GACGGCGT--AG--CTAGCAT----TA-A-CA
AGTG-CATTTAG-ATCAGGATCTTCTATCGC-
   * * *   **     ** **     **   *
```

ή και έτσι,

```
GACGGCGTAGCTAGCATTAACA----------
---AGTGCATTTAG-ATCAGGATCTTCTATCGC
   * * *   *** ** *   *
```

Βαθμολόγηση στοιχίσεων

Η πρώτη στοίχιση έχει δύο παραπάνω ταυτότητες βάσεων, αλλά έχει εισαγάγει ένα μάλλον απίθανο αριθμό από εισαγωγές/διαγραφές για να επιτύχει τις δύο επιπλέον ταυτότητες. Η δεύτερη στοίχιση φαίνεται να έχει μεγαλύτερο βιολογικό περιεχόμενο, κυρίως γιατί τα γεγονότα εισαγωγών/διαγραφών που υπονοεί είναι συνεχή και όχι διάσπαρτα και μεμονωμένα.

Βαθμολόγηση στοιχίσεων

Από την άποψη των στοιχίσεων, είναι αδύνατο να διακριθούν τα γεγονότα εισαγωγής από αυτά των διαγραφών. Για το λόγο αυτό, αναφερόμαστε σε αυτά με τον όρο "κενά" (gaps), και στον τρόπο βαθμολόγησης τους ως "κόστος κενών" (gap penalty). Η πλέον συνηθισμένη μέθοδος βαθμολόγησης των κενών αποτελείται από δύο όρους :

- Ο πρώτος όρος αφαιρεί από την βαθμολογία της στοίχισης ένα σταθερό ποσό για κάθε κενό που δημιουργείται ασχέτως του μήκους του κενού.
- Ο δεύτερος όρος αφαιρεί ένα σταθερό ποσό για κάθε επέκταση του μήκους ενός κενού.

Βαθμολόγηση στοιχίσεων

Άρα, για να υπολογίσουμε την βαθμολογία μίας στοίχισης :

- Για κάθε ζεύγος στοιχισμένων καταλοίπων, αυξάνουμε την βαθμολογία κατά όσο αναφέρεται στον πίνακα βαθμολόγησης που χρησιμοποιούμε (PAM, BLOSUM, ...).
- Για κάθε κενό μήκους N, αφαιρούμε N·α όπου α είναι το gap penalty. Εάν χρησιμοποιούμε τον διπλό τρόπο βαθμολόγησης κενών (προηγούμενη διαφάνεια), τότε αφαιρούμε [(N-1)·α + k)] όπου k είναι το κόστος δημιουργίας του κενού. Για απλότητα, από εδώ και πέρα θα υποθέτουμε ότι k=0.

Άρα, το πρόβλημα είναι ...

Με δεδομένες δύο αλληλουχίες A & B, έναν πίνακα βαθμολόγησης K και ένα κόστος εισαγωγής κενού α, βρείτε για ποιά στοίχιση ---από όλες τις δυνατές στοιχίσεις των αλληλουχιών A&B --- η βαθμολογία (της στοίχισης) μεγιστοποιείται.

Πόσες είναι "οι δυνατές στοιχίσεις" ;

Το μέγεθος του προβλήματος

WHAT
WIAT

WHAT	WH-AT	WHA-T	WH-A-T
WIAT	W-IAT	WI-AT	W-I-AT
W-HAT	WHA-T	W--HAT	
WI-AT	W-IAT	WIA--T	...

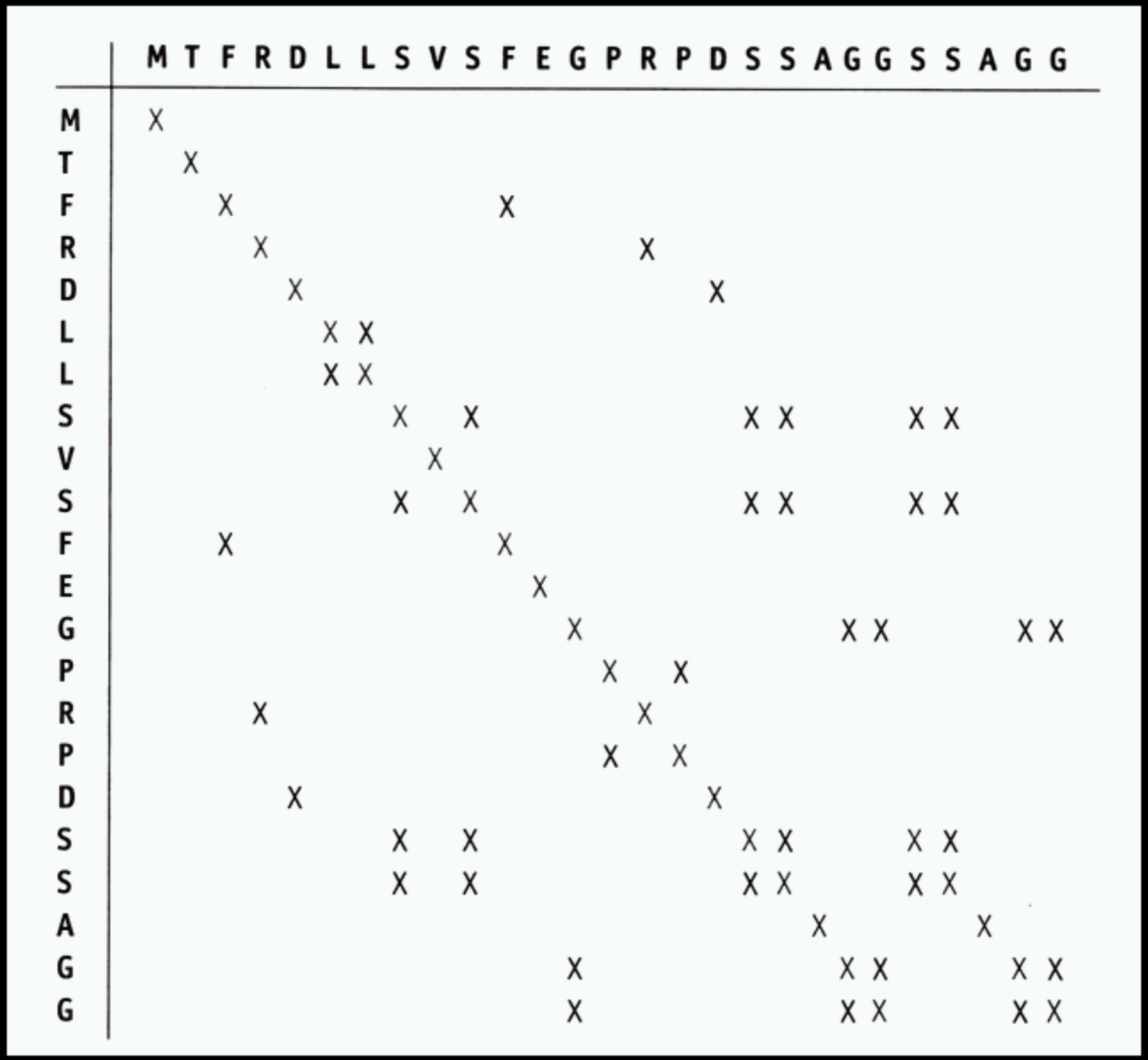
Το μέγεθος του προβλήματος

Για δύο αλληλουχίες A & B, με αντίστοιχα μήκη m & n, η φαινομενική πολυπλοκότητα του προβλήματος είναι ανάλογη του [m^n].
Για μία στοίχιση δύο πρωτεϊνών 200 αμινοξέων εκάστη, και εάν εξετάζαμε 1 δις στοιχίσεις ανά δευτερόλεπτο, μετά από 15 δις χρόνια (την ηλικία του σύμπαντος) θα είχαμε εξετάσει μόνο 10^27 στοιχίσεις (δεν θα είχαμε καν ξεκινήσει).
Στην πραγματικότητα, η πολυπλοκότητα του προβλήματος είναι μόνο [m · n]. Για το παραπάνω πρόβλημα, θα είχαμε τελειώσει με τη στοίχιση σε 40 εκατομμυριοστά του δευτερολέπτου.

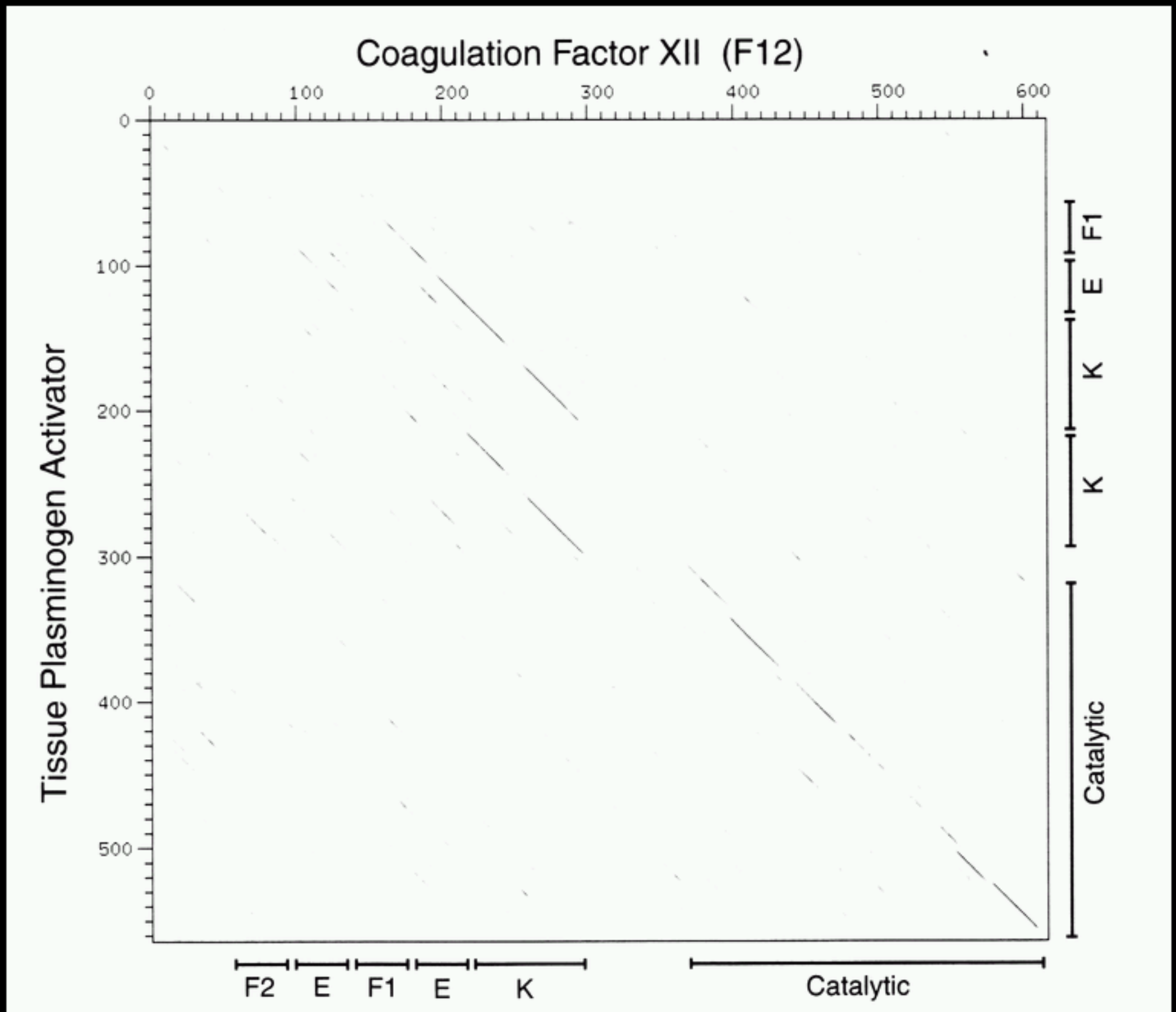
Στοιχίσεις : dot plots

Μια μέθοδος που βοηθάει στην κατανόηση του γιατί η στοίχιση δυο αλληλουχιών έχει πολυπλοκότητα ανάλογη του [m · n] είναι τα λεγόμενα dot plots.
Τα dot plots (σημείο-διαγράμματα ;) δεν είναι ένας αλγόριθμος στοίχισης, αλλά ένα μέσο παρουσίασης των σχέσεων μεταξύ δυο αλληλουχιών. Πρόκειται για μια δισδιάστατη γραφική αναπαράσταση στην οποία κατά μήκος των δυο αξόνων αναπαρίστανται οι δυο αλληλουχίες και στον χώρο που ορίζεται από αυτές τοποθετούνται σημεία των οποίων η φωτεινότητα είναι ανάλογη της σχέσης των αμινοξέων (των αλληλουχιών) που αντιστοιχούν στις συντεταγμένες του σημείου.

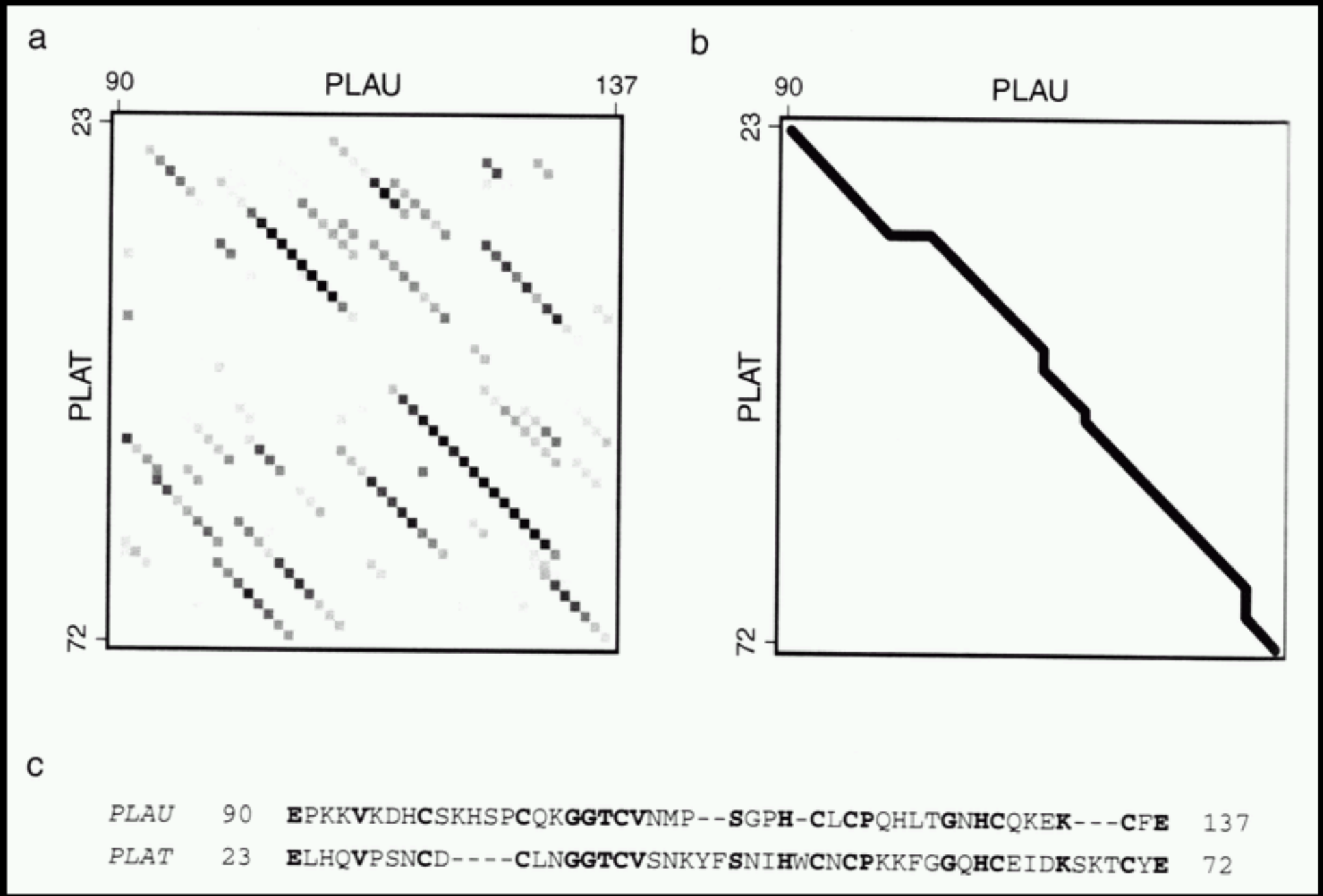
Στοιχίσεις : dot plots



Στοιχίσεις : dot plots



Στοιχίσεις : dot plots



C

```
PLAU   90  EPKKVKDHCSKHSPCQKGGTCVNMP--SGPH-CLCPQHLTGNHCQKEK---CFE  137
PLAT   23  ELHQVPSNCD----CLNGGTCVSNKYFSNIHWCNCPKKFGGQHCEIDKSKTCYE   72
```

Αλγόριθμοι στοίχισης

Άρα, για δύο αλληλουχίες A & B, με αντίστοιχα μήκη m & n, το πρόβλημα είναι να βρεθεί το καλύτερο (υψηλότερης βαθμολογίας) μονοπάτι μέσω ενός πίνακα [mn]. Το μονοπάτι αυτό θα πρέπει να περνάει διαδοχικά από όλα τα αμινοξέα των δυο αλληλουχιών (ολική στοίχιση).

Needleman & Wunsch

Ο αλγόριθμος των Needleman & Wunsch (N & W), ανήκει στην κατηγορία των λεγομένων μεθόδων 'δυναμικού προγραμματισμού' (dynamic programming). Η βασική παρατήρηση στην οποία στηρίζεται ο αλγόριθμος είναι ότι οποιοδήποτε υποσύνολο της βέλτιστης στοίχισης, θα πρέπει επίσης να είναι βέλτιστο (ειδάλλως η στοίχιση θα μπορούσε να βελτιστοποιηθεί και άλλο μέσω βελτίωσης της υποστοίχισης).

N & W : η μέθοδος

Υποθέστε ότι έχουμε ήδη επιτύχει μία βέλτιστη στοίχιση μεταξύ των (κ) πρώτων αμινοξέων από την αλληλουχία Α με τα (μ) πρώτα αμινοξέα από την Β. Εάν βρούμε ένα τρόπο να επεκτείνουμε τη στοίχιση κατά μία θέση προς τα δεξιά έτσι ώστε η προκύπτουσα στοίχιση να είναι επίσης βέλτιστη, έχουμε τελειώσει : εφαρμόζοντας επαναληπτικά την μέθοδο αυτή, θα στοιχίσουμε (βέλτιστα) ολόκληρες τις αλληλουχίες.

N & W : η μέθοδος

Για την επέκταση της στοίχισης κατά μία θέση υπάρχουν μόνο τρεις τρόποι :

- Στοιχίζουμε το επόμενο αμινοξύ της Α, δηλαδή το Α(κ+1), με το επόμενο της Β, το Β(μ+1).
- Προσθέτουμε ένα αμινοξύ από την Α, το Α(κ+1), και εισαγάγουμε ένα κενό στη Β.
- Προσθέτουμε ένα αμινοξύ από την Β, το Β(μ+1), και εισαγάγουμε ένα κενό στη Α.

Αυτό που θέλουμε είναι να διαλέξουμε εκείνο τον τρόπο επέκτασης που μεγιστοποιεί την βαθμολογία της τελικής στοίχισης.

N & W : η μέθοδος

Έστω, λοιπόν, ότι η βαθμολογία της βέλτιστης στοίχισης των (κ) αμινοξέων της Α, με τα (μ) της Β, είναι S(κ,μ).
Τότε :
Εάν στοιχίσουμε το επόμενο αμινοξύ της Α, δηλαδή το Α(κ+1), με το επόμενο της Β, το Β(μ+1), η βαθμολογία της στοίχισης θα αυξηθεί κατά τόσο όσο η βαθμολογία υποκατάστασης των αμινοξέων Α(κ+1) και Β(μ+1) όπως αυτή δίδεται από τον πίνακα βαθμολόγησης Κ. Άρα σε αυτή την περίπτωση :

$$S(\kappa+1,\mu+1) = S(\kappa,\mu) + K[\, A(\kappa+1), B(\mu+1)\,]$$

N & W : η μέθοδος

Εάν προσθέσουμε το επόμενο αμινοξύ της Α, δηλαδή το Α(κ+1), και εισαγάγουμε ένα κενό στη Β, η τελική βαθμολογία θα είναι ίση με τη βαθμολογία της βέλτιστης στοίχισης των (κ+1) αμινοξέων της Α με τα (μ) της Β μείον το gap penalty (α). Δηλαδή :

$$S(\kappa+1,\mu+1) = S(\kappa+1,\mu) - \alpha$$

N & W : η μέθοδος

Εάν προσθέσουμε το επόμενο αμινοξύ της Β, δηλαδή το Β(μ+1), και εισαγάγουμε ένα κενό στη Α, η τελική βαθμολογία θα είναι ίση με τη βαθμολογία της βέλτιστης στοίχισης των (κ) αμινοξέων της Α με τα (μ+1) της Β μείον το gap penalty (α). Δηλαδή :

$$S(\kappa+1,\mu+1) = S(\kappa,\mu+1) - \alpha$$

N & W : η μέθοδος

Η καινούργια βέλτιστη στοίχιση θα είναι αυτή για την οποία η τιμή του S(κ+1,μ+1) μεγιστοποιείται.
Η ολική βαθμολογία της καινούργιας βέλτιστης στοίχισης θα είναι προφανώς :

$$S(\kappa+1,\mu+1) = \max \{ S(\kappa,\mu) + K[A(\kappa+1), B(\mu+1)],$$
$$S(\kappa+1,\mu) - \alpha,$$
$$S(\kappa,\mu+1) - \alpha \quad \}$$

N & W : η μέθοδος

Άρα, εάν έχουμε τιμές για τα S(κ,μ), S(κ+1,μ) και S(κ,μ+1), τότε μπορούμε να υπολογίζουμε το S(κ+1,μ+1) (και την αντίστοιχη στοίχιση) απλά με το να βρίσκουμε το μέγιστο των τριών αριθμητικών εκφράσεων που δίδονται στις αγκύλες της προηγούμενης εξίσωσης.

Γράφοντας αυτά τα αποτελέσματα σε μορφή πίνακα δείχνει ότι εάν έχουμε τιμές βαθμολογίας για την πρώτη γραμμή και την πρώτη στήλη του, μπορούμε επαναληπτικά να τον συμπληρώσουμε ολόκληρο :

N & W : η μέθοδος

S(κ,μ)	S(κ,μ+1)	S(κ,μ+2)	S(κ,μ+3)
S(κ+1,μ)			
S(κ+2,μ)			
S(κ+3,μ)			

N & W : η μέθοδος

S(κ,μ) S(κ,μ+1) S(κ,μ+2) S(κ,μ+3)

S(κ+1,μ) S(κ+1,μ+1)

S(κ+2,μ)

S(κ+3,μ)

$$S(\kappa+1,\mu+1) = \max \{ S(\kappa,\mu) + K[\, A(\kappa+1),\, B(\mu+1) \,],$$
$$S(\kappa+1,\mu) - \alpha,$$
$$S(\kappa,\mu+1) - \alpha \quad \}$$

N & W : η μέθοδος

S(κ,μ) S(κ,μ+1) S(κ,μ+2) S(κ,μ+3)

S(κ+1,μ) S(κ+1,μ+1) S(κ+1,μ+2) S(κ+1,μ+3)

S(κ+2,μ)

S(κ+3,μ)

$$S(\kappa+1,\mu+1) = \max \{ S(\kappa,\mu) + K[\, A(\kappa+1),\, B(\mu+1) \,],$$
$$S(\kappa+1,\mu) - \alpha,$$
$$S(\kappa,\mu+1) - \alpha \quad \}$$

N & W : η μέθοδος

S(κ,μ)	S(κ,μ+1)	S(κ,μ+2)	S(κ,μ+3)
S(κ+1,μ)	S(κ+1,μ+1)	S(κ+1,μ+2)	S(κ+1,μ+3)
S(κ+2,μ)	S(κ+2,μ+1)	S(κ+2,μ+2)	S(κ+2,μ+3)
S(κ+3,μ)	S(κ+3,μ+1)	S(κ+3,μ+2)	S(κ+3,μ+3)

$$S(\kappa+1,\mu+1) = \max \{ S(\kappa,\mu) + K[\, A(\kappa+1),\, B(\mu+1) \,],$$
$$S(\kappa+1,\mu) - \alpha,$$
$$S(\kappa,\mu+1) - \alpha \quad \}$$

N & W : η μέθοδος

Άρα, για να βρούμε την βέλτιστη στοίχιση των δύο αλληλουχιών A & B, με μήκη m & n, χρειαζόμαστε τιμές για τα S(0,0) S(1,0) S(2,0) ... S(m,0) και για τα S(0,1) S(0,2) ... S(0,n).
Για το αλγόριθμο των N & W η αρχικοποίηση των τιμών είναι :

$$S(0,0) = 0$$
$$S(\kappa,0) = -\kappa \cdot \alpha$$
$$S(0,\mu) = -\mu \cdot \alpha$$

N & W : παράδειγμα 1ο

Θέλουμε να στοιχίσουμε τις αλληλουχίες ASPERA και APTERA. Ο (μοναδιαίος) πίνακας βαθμολόγησης είναι :

```
A   3
C   0   3
M   0   0   3
P   0   0   0   3
F   0   0   0   0   3
. . . . . . . . . . . . . . . . . . .
    A   C   M   P   F ...
```

και το gap penalty έχει τιμή 1 (ανά κενό).

N & W : παράδειγμα 1ο

		A	S	P	E	R	A
	0	-1	-2	-3	-4	-5	-6
A	-1						
P	-2						
T	-3						
E	-4						
R	-5						
A	-6						

Αρχικοποίηση του πίνακα.

N & W : παράδειγμα 1ο

		A	S	P	E	R	A
	0	-1	-2	-3	-4	-5	-6
A	-1	3					
P	-2						
T	-3						
E	-4						
R	-5						
A	-6						

Συμπλήρωση.

N & W : παράδειγμα 1ο

		A	S	P	E	R	A
	0	-1	-2	-3	-4	-5	-6
A	-1	3	2	1	0	-1	-2
P	-2	2					
T	-3						
E	-4						
R	-5						
A	-6						

Συμπλήρωση.

N & W : παράδειγμα 1ο

		A	S	P	E	R	A
	0	-1	-2	-3	-4	-5	-6
A	-1	3	2	1	0	-1	-2
P	-2	2	3	5	4	3	2
T	-3	1	2	4	5	4	3
E	-4	0	1	3	7	6	5
R	-5	-1	0	2	6	10	9
A	-6	-2	-1	1	5	9	13

Συμπλήρωση.

N & W : παράδειγμα 1ο

		A	S	P	E	R	A
	0	-1	-2	-3	-4	-5	-6
A	-1	3	2	1	0	-1	-2
P	-2	2	3	5	4	3	2
T	-3	1	2	4	5	4	3
E	-4	0	1	3	7	6	5
R	-5	-1	0	2	6	10*	9
A	-6	-2	-1	1	5	9	13*

Μονοπάτι βέλτιστης στοίχισης.

N & W : παράδειγμα 1ο

		A	S	P	E	R	A
	0	-1	-2	-3	-4	-5	-6
A	-1	3*	2*	1	0	-1	-2
P	-2	2	3	5*	4	3	2
T	-3	1	2	4*	5	4	3
E	-4	0	1	3	7*	6	5
R	-5	-1	0	2	6	10*	9
A	-6	-2	-1	1	5	9	13*

A S P - E R A

A - P T E R A

N & W : παράδειγμα 2ο

Θέλουμε να στοιχίσουμε τις αλληλουχίες ASPERA και APTERA. Ο (μοναδιαίος) πίνακας βαθμολόγησης είναι :

```
A   3
C   0   3
M   0   0   3
P   0   0   0   3
F   0   0   0   0   3
. . . . . . . . . . . . . . . . . . .
    A   C   M   P   F  . . .
```

και το gap penalty έχει τιμή 2 (ανά κενό).

N & W : παράδειγμα 2ο

		A	S	P	E	R	A
	0	-2	-4	-6	-8	-10	-12
A	-2						
P	-4						
T	-6						
E	-8						
R	-10						
A	-12						

N & W : παράδειγμα 2ο

		A	S	P	E	R	A
	0	-2	-4	-6	-8	-10	-12
A	-2	3	1				
P	-4						
T	-6						
E	-8						
R	-10						
A	-12						

N & W : παράδειγμα 2ο

		A	S	P	E	R	A
	0	-2	-4	-6	-8	-10	-12
A	-2	3	1	-1	-3	-5	-7
P	-4	1	3	4	2	0	-2
T	-6	-1	1	3	4	2	0
E	-8	-3	-1	1	6	4	2
R	-10	-5	-3	-1	4	9	7
A	-12	-7	-5	-3	2	7	12

N & W : παράδειγμα 2ο

		A	S	P	E	R	A
	0	-2	-4	-6	-8	-10	-12
A	-2	3*	1	-1	-3	-5	-7
P	-4	1	3*	4	2	0	-2
T	-6	-1	1	3*	4	2	0
E	-8	-3	-1	1	6*	4	2
R	-10	-5	-3	-1	4	9*	7
A	-12	-7	-5	-3	2	7	12*

A S P E R A

A P T E R A