Σημειώσεις Βιοπληροφορικής

Πολλαπλή Στοίχιση Ακολουθιών

Βασικές Έννοιες
Γενίκευση των Αλγορίθμων Στοίχισης Κατά Ζεύγη
Προοδευτική Πολλαπλή Στοίχιση – CLUSTALW
Πρακτικά Ζητήματα

ΒΑΣΙΛΗΣ ΠΡΟΜΠΟΝΑΣ

ΑΘΗΝΑ 2004-2005, ΛΕΥΚΩΣΙΑ 2006

1. Εισαγωγή στην Πολλαπλή Στοίχιση Ακολουθιών

Βασικές Έννοιες

Μια από τις πιο σημαντικές συνεισφορές της μοριακής βιολογίας στη μελέτη της εξέλιξης των ειδών βασίζεται στην παρατήρηση ότι οι ακολουθίες του DNA διαφορετικών οργανισμών συχνά σχετίζονται. Οι ομοιότητες αυτές είναι δυνατόν να εντοπίζονται τόσο σε κωδικές περιοχές (οπότε προφανώς συνεπάγονται και ομοιότητα στο επίπεδο των προϊόντων των αντίστοιχων γονιδίων) όσο και σε ρυθμιστικές ή μη κωδικές περιοχές του DNA.

Είναι συχνή η περίπτωση όπου γονίδια με συντηρημένες νουκλεοτιδικές αλληλουχίες εμφανίζονται σε οργανισμούς οι οποίοι μορφολογικά είναι τελείως διαφορετικοί και αναμένουμε ότι έχουν απομακρυνθεί σημαντικά μεταξύ τους κατά τη διάρκεια της εξελικτικής διαδικασίας. Τα προϊόντα αυτών των συντηρημένων κατά την εξέλιξη γονιδίων εκτελούν παρόμοιες (ή ορισμένες φορές ταυτόσημες) κυτταρικές λειτουργίες, ή σε άλλες περιπτώσεις μεταλλάσσονται ή αναδιατάσσονται στο επίπεδο της αλληλουχίας τους, ώστε να πραγματοποιούν διαφορετικές λειτουργίες οι οποίες παγιώνονται μέσα στα πλαίσια της φυσικής επιλογής[1].

Οι μέθοδοι σύγκρισης ακολουθιών κατά ζεύγη (τόσο οι ακριβείς όσο και οι ευριστικές) στις οποίες έχουμε ήδη αναφερθεί είναι προφανές ότι μπορούν να χρησιμοποιηθούν για να αναδείξουν τέτοιες ομοιότητες. Παρόλα αυτά, είναι λογικό να σκεφτεί κανείς ότι η δυνατότητα της ταυτόχρονης στοίχισης περισσότερων των δύο ακολουθιών θα μπορούσε να δώσει περισσότερες πληροφορίες, τόσο για την υποκείμενη εξελικτική διαδικασία (την οποία δυστυχώς δεν γνωρίζουμε) όσο και για πιθανά δομικά-λειτουργικά χαρακτηριστικά

[1] Προφανώς, οι περιπτώσεις γονιδιακών προϊόντων με διαφοροποιημένες λειτουργίες που δεν είναι ευνοϊκές για την επιβίωση ενός οργανισμού δεν σταθεροποιούνται με την πάροδο του χρόνου και οι οργανισμοί που τις φέρουν εκλείπουν.

των εξεταζόμενων μορίων. Συνεπώς, η δυνατότητα για την Πολλαπλή Στοίχιση Ακολουθιών (Multiple Sequence Alignment, ή απλά MSA) αποτέλεσε αφενός μια αναγκαιότητα για τη μελέτη των μοριακών δεδομένων, αφετέρου δε υποβοηθήθηκε από την ύπαρξη μεθόδων για τη στοίχιση ακολουθιών κατά ζεύγη. Επακόλουθο ήταν να εστιαστεί μεγάλο μέρος της Βιοπληροφορικής έρευνας στην ανάπτυξη μεθόδων MSA η οποία σημειωτέον συνεχίζεται και στις ημέρες μας.

Στις ακόλουθες παραγράφους δεν επιχειρείται η εξαντλητική επισκόπηση της μεγάλης σχετιζόμενης με το θέμα βιβλιογραφίας αλλά η ανάπτυξη των βασικών εννοιών και η παρουσίαση θεμελιωδών αλγορίθμων που δίνουν λύσεις στο δύσκολο (όχι μόνο από υπολογιστικής σκοπιάς) πρόβλημα της Πολλαπλής Στοίχισης Ακολουθιών βιολογικών μακρομορίων.

Η χρησιμότητα των Πολλαπλών Στοιχίσεων Ακολουθιών

Η σημαντικότητα των Πολλαπλών Στοιχίσεων Ακολουθιών θα μπορούσε να αναλυθεί σε σελίδες επί σελίδων κειμένου, με κατάλληλες αναφορές σε σημαντικές πρακτικές εφαρμογές. Θα μπορούσε να συνοψίσει κανείς τη μεγάλη τους σημασία με δύο φράσεις, οι οποίες αντικατοπτρίζουν αυτό το πρόβλημα:

"*... two strings good, four strings better ...*" (Gusfield, 1997, σελίδα 332)

και

"*One or two homologous sequences whisper ... a full multiple alignment shouts out loud.*" (Hubbard *et al.*, 1996)

Παρά τον έμμεσο συσχετισμό των πολλαπλών στοιχίσεων με τις στοιχίσεις κατά ζεύγη (π.χ. για την αποκάλυψη κοινών συντηρημένων μοτίβων), η χρήση τους είναι δυνατόν (υπό ορισμένες προϋποθέσεις) να παρέχει απαντήσεις σε τελείως διαφορετικά πρακτικά προβλήματα. Για παράδειγμα, μια στοίχιση κατά ζεύγη είναι δυνατόν να μας βοηθήσει να «ψαρέψουμε» ακολουθίες οι οποίες μοιάζουν μεταξύ τους και με βάση το επίπεδο ομοιότητάς τους (και την εκτίμηση της στατιστικής σημαντικότητας) να αποφανθούμε για μια πιθανή λειτουργική-δομική-εξελικτική μεταξύ τους σχέση. Αντίστροφα, μια πολλαπλή στοίχιση είναι δυνατόν να αποκαλύψει άγνωστες συντηρημένες περιοχές ακολουθιών για τις οποίες είναι δυνατόν να έχουμε εκ των προτέρων υποψίες (συχνά μετά από συσχέτιση με πειραματικά δεδομένα που αφορούν λειτουργία-δομή-φυλογένεση) για τη βιολογική τους σχέση.

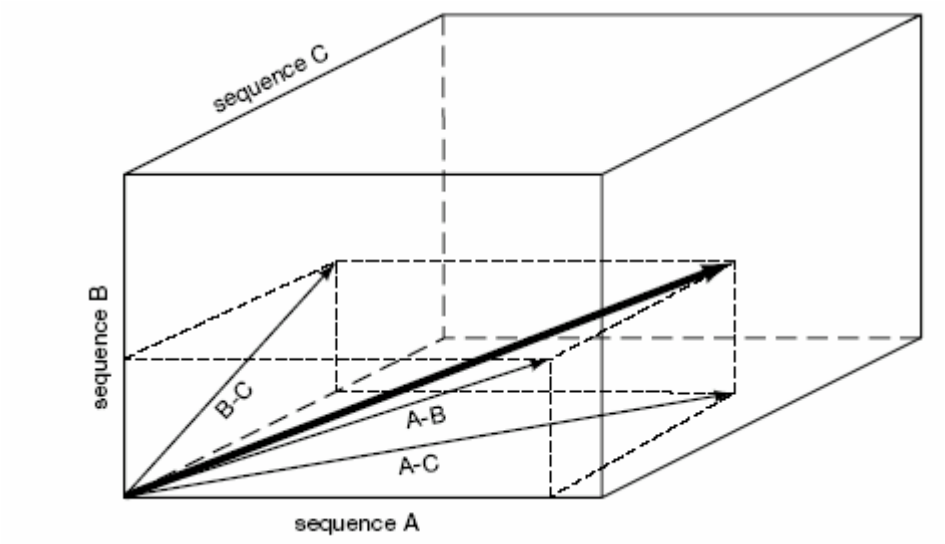
Πολλαπλή Στοίχιση Ακολουθιών και Δυναμικός Προγραμματισμός

Μια προφανής προσέγγιση στο πρόβλημα της Πολλαπλής Στοίχισης Ακολουθιών προκύπτει από τη σκέψη ότι, ιδανικά, οι αλγόριθμοι δυναμικού προγραμματισμού είναι δυνατόν να επεκταθούν και για τη στοίχιση περισσότερων από δύο ακολουθιών.

Εάν βασιστούμε στην ιδέα ότι όλες οι δυνατές στοιχίσεις δύο ακολουθιών μπορούν να αντιστοιχηθούν με διαδρομές οι οποίες διέρχονται από τους κόμβους του πλέγματος δύο διαστάσεων που προκύπτει από την ορθογώνια διάταξη των ακολουθιών, η επέκταση στη στοίχιση τριών ακολουθιών είναι προφανής. Αρκεί να διατάξουμε με αντίστοιχο τρόπο τις τρεις ακολουθίες που επιθυμούμε να στοιχίσουμε κατά τους άξονες ενός τρισορθογώνιου συστήματος (Εικόνα 1).

Σύμφωνα με όσα έχουμε ήδη συζητήσει, γίνεται προφανές ότι για τη στοίχιση τριών ακολουθιών με δυναμικό προγραμματισμό απαιτούνται $O(L_1*L_2*L_3)$ υπολογιστικά βήματα, όπου L_1, L_2, L_3 τα μήκη των τριών ακολουθιών. Μπορεί να αποδειχθεί στη γενική των περιπτώσεων ότι η εφαρμογή της επέκτασης των αλγορίθμων δυναμικού προγραμματισμού για την πολλαπλή στοίχιση N το πλήθος ακολουθιών με μήκη L_1, L_2, …, L_N έχει υπολογιστικές απαιτήσεις $O(L_1*L_2* \ldots *L_N)$ σε μνήμη και $O(2^N* L_1*L_2* \ldots *L_N)$ σε χρόνο. Αυτές οι υπολογιστικές απαιτήσεις (π.χ. χρόνος εκθετικός του πλήθους των ακολουθιών) καθιστούν την εφαρμογή των αλγορίθμων δυναμικού προγραμματισμού για την στοίχιση περισσότερων των δύο ακολουθιών πρακτικά ασύμφορη. Αναλογιστείτε το παράδειγμα κατά το οποίο επιθυμούμε να στοιχίσουμε τρεις αμινοξικές ακολουθίες μήκους $L_1=L_2=L_3=300$. Για απλούστευση των υπολογισμών ας υποθέσουμε ότι αγνοούμε τους υπολογισμούς για την εισαγωγή κενών. Τότε το πλήθος των υπολογισμών που απαιτούνται είναι κατά προσέγγιση $300^3 = 2.7 * 10^7$, ο οποίος πρακτικά επιτρέπει την εφαρμογή του δυναμικού προγραμματισμού. Η στοίχιση όμως ακολουθιών μεγαλύτερου μήκους ή πολύ περισσότερων από τρεις το πλήθος (όπως τις περισσότερες φορές είναι η περίπτωση) δεν είναι πρακτικά εφικτή, ούτε και με τους ταχύτερους υπολογιστές που έχουμε σήμερα διαθέσιμους[2].

[2] Πιθανότατα ούτε και με τους πιο γρήγορους υπολογιστές που θα κατασκευαστούν και σε μερικές δεκαετίες ...

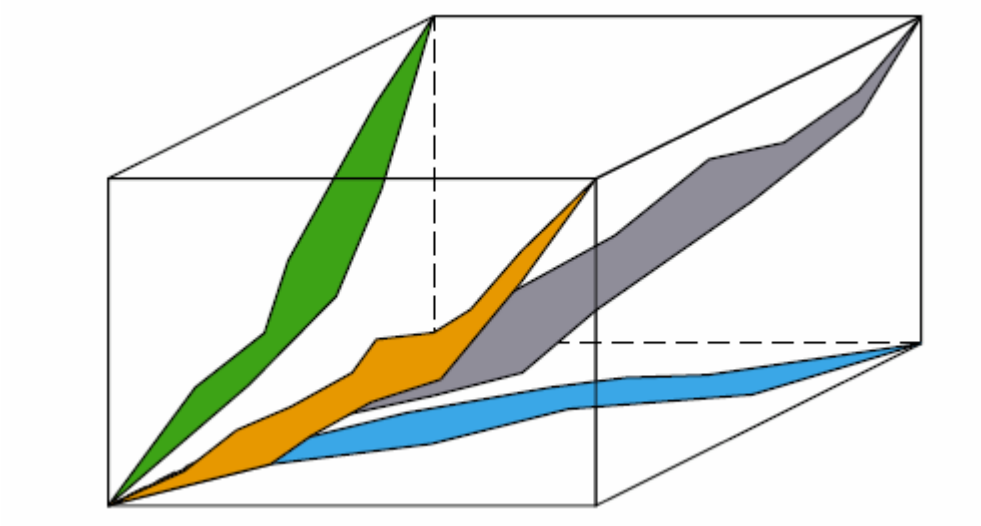


Εικόνα 1: Επέκταση του αλγόριθμου δυναμικού προγραμματισμού για τη στοίχιση τριών ακολουθιών.

Τα βέλη στις επιφάνειες του ορθογωνίου παραλληλεπιπέδου (A-B, A-C, B-C) υποδεικνύουν τη φορά κατά την οποία θα γινόταν ο υπολογισμός των τιμών για τη στοίχιση κατά ζεύγη των αντίστοιχων ακολουθιών. Η στοίχιση των τριών ακολουθιών απαιτεί τον υπολογισμό τιμών για τους στοιχειώδεις κύβους οι οποίοι ορίζονται από τα κατάλοιπα των τριών ακολουθιών, με τις βέλτιστες τιμές scores, ακολουθώντας την ίδια διαδικασία με την στοίχιση κατά ζεύγη. Συγκεκριμένα, το score σε κάθε κύβο του πλέγματος υπολογίζεται (με βάση κάποιο δεδομένο πίνακα αντικατάστασης και ποινή εισαγωγής κενών) λαμβάνοντας υπόψη τα scores που προκύπτουν από όλες τις πιθανές μετακινήσεις προς εκείνο το κελί. Επιλέγεται η μετακίνηση η οποία οδηγεί στο μεγαλύτερο (βέλτιστο) score, και η διαδρομή σημειώνεται αντίστοιχα σε ένα τρισδιάστατο πίνακα-ιχνηθέτη. [Εικόνα από Mount, 2001].

Εξαιτίας της μεγάλης χρησιμότητας για την κατασκευή πολλαπλών στοιχίσεων από πολύ νωρίς άρχισαν προσπάθειες για την ανάπτυξη ευριστικών μεθόδων (Waterman and Perlwitz, 1984). Ακολουθώντας αντίστοιχη πορεία με την έρευνα για τη σύγκριση/στοίχιση ακολουθιών κατά ζεύγη, οι Carrillo και Lipman (Carrillo and Lipman, 1988) πρότειναν μια ευριστική μέθοδο (υλοποίηση της οποίας πραγματοποιήθηκε στο λογισμικό MSA, Lipman *et al.*, 1989), που στοχεύει στη μείωση του αριθμού των υπολογισμών με την ελάχιστη

δυνατή απόκλιση από τη βέλτιστη (μαθηματικά) σύγκριση. Η μέθοδος MSA βασίστηκε στην εισαγωγή ενός αντικειμενικού κριτηρίου για την αξιολόγηση των διαφορετικών πολλαπλών στοιχίσεων, το οποίο ονομάστηκε Sum-of-Pairs Score (SP-score, δείτε την επόμενη παράγραφο).



Εικόνα 2: Περιορισμός του χώρου αναζήτησης στον οποίο θα προσδιοριστεί μια πολλαπλή στοίχιση που βελτιστοποιεί το SP-score, από το πρόγραμμα MSA.

Ο χώρος στον οποίο περιορίζεται η εκτέλεση του αλγόριθμου δυναμικού προγραμματισμού αντιστοιχεί στη γκρι περιοχή του ορθογώνιου παραλληλεπιπέδου και αποτελεί σημαντικά μικρό μέρος του συνολικού χώρου αναζήτησης (ολόκληρο το ορθογώνιο παραλληλεπίπεδο). Αυτή η υποπεριοχή εντοπίζεται από τους περιορισμούς που προκύπτουν από όλες τις βέλτιστες στοιχίσεις κατά ζεύγη μεταξύ των τριών ακολουθιών καθώς και από την ευριστική προσεγγιστική πολλαπλή στοίχιση των ακολουθιών. Οι περιοχές με πράσινο, μπλε και πορτοκαλί χρώμα στις πλευρές του ορθογωνίου παραλληλεπιπέδου αποτελούν τις ορθές προβολές της περιοχής στην οποία εκτελείται ο δυναμικός προγραμματισμός.

Παρότι η μέθοδος MSA συνέχισε να βελτιώνεται (Gupta *et al.*, 1995) σύμφωνα με τους απαιτούμενους υπολογιστικούς πόρους (μνήμη – χρόνος), η χρήση του δεν είναι πρακτική παρά μόνο για λίγες

ακολουθίες μικρού μάλιστα μήκους (τυπικά ~5-7 ακολουθίες με 100-200 κατάλοιπα η κάθε μία).

Παρότι πρακτικά ασύμφορη, η μελέτη των αλγορίθμων δυναμικού προγραμματισμού για πολλαπλή στοίχιση ακολουθιών έχουν (τουλάχιστον) θεωρητικό ενδιαφέρον. Η επέκταση για περισσότερες από τρεις ακολουθίες (N) πραγματοποιείται διαισθητικά με την διάταξη των προς στοίχιση ακολουθιών στις ακμές ενός N-διάστατου υπερ-κύβου και στον υπολογισμό της διαδρομής εκείνης που αντιστοιχεί στο βέλτιστο score.

Βαθμονόμηση πολλαπλών στοιχίσεων – Sum-of-Pairs score

Όπως ακριβώς και στη στοίχιση ακολουθιών κατά ζεύγη, έτσι και κατά την πολλαπλή στοίχιση έπρεπε να αναζητηθεί η ποσότητα εκείνη την οποία οφείλουμε να μεγιστοποιήσουμε προκειμένου να επιτύχουμε μια «καλή» στοίχιση.

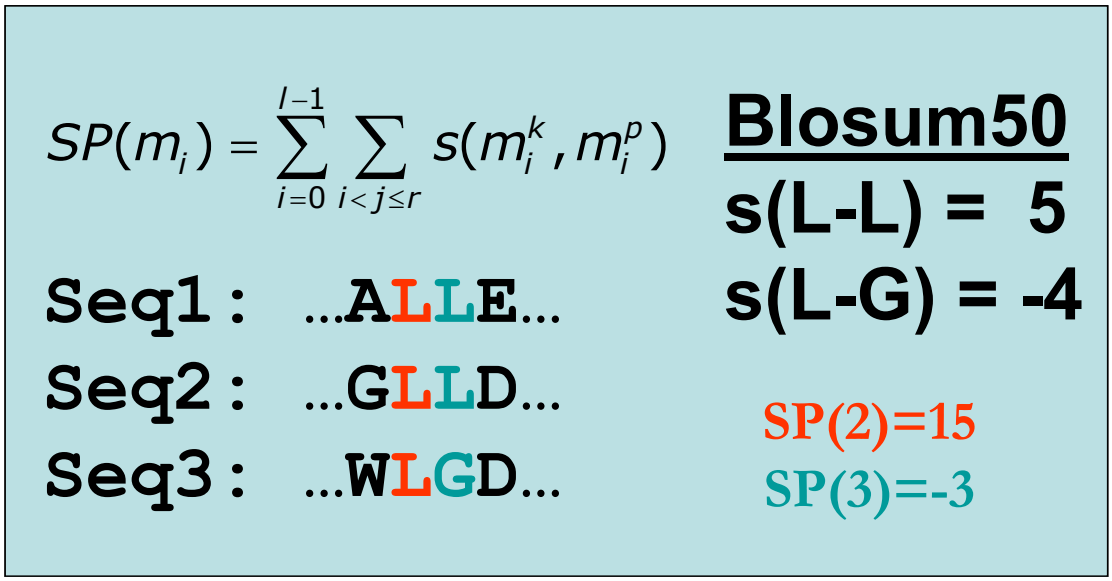
Μια διαισθητική προσέγγιση, η οποία μπορεί να σχετισθεί με ένα πιθανό μοντέλο εξέλιξης είναι το λεγόμενο Sum-of-Pairs score (SP-score). Για τον υπολογισμό του SP-score μιας πολλαπλής στοίχισης αυτό που απαιτείται είναι να αθροιστούν τα scores όλων των δυνατών ζευγών καταλοίπων που καταλαμβάνουν μια στήλη της στοίχισης, με βάση έναν πίνακα αντικατάστασης και δεδομένες ποινές για την εισαγωγή κενών. Έτσι, κατά την πολλαπλή στοίχιση N ακολουθιών απαιτείται για κάθε στήλη της στοίχισης ο υπολογισμός των *N(N-1)/2* scores και η άθροισή τους ώστε να υπολογιστεί το score μιας στήλης (Εικόνα 3). Η άθροιση των scores όλων των στηλών οδηγεί στο τελικό score της πολλαπλής στοίχισης[3].

Από υπολογιστικής πλευράς η βέλτιστη πολλαπλή στοίχιση ακολουθιών με βάση το SP-score αποδεικνύεται (Murata *et al.*, 1985) ότι είναι δυνατόν να υπολογιστεί με τη χρήση της επέκτασης

[3] Από εξελικτικής σκοπιάς αυτή η προσέγγιση θεωρεί ότι εν δυνάμει κάθε ακολουθία θα μπορούσε να θεωρηθεί πιθανός πρόγονος όλων των υπόλοιπων.

των αλγορίθμων δυναμικού προγραμματισμού που συζητήσαμε στα προηγούμενα. Παρόλα αυτά η εκθετική πολυπλοκότητα δεν επιτρέπει να έχουμε τέτοιες εφαρμογές οι οποίες να είναι πρακτικές για προβλήματα που συναντούμε στην πράξη.

Μια παραλλαγή αυτής της προσέγγισης (δείτε τα επόμενα) είναι δυνατόν να λαμβάνει υπόψη σταθμισμένα scores για τα διάφορα κατάλοιπα ώστε να μετριάζεται καταρχήν η επανειλημμένη συνεισφορά πολύ όμοιων ακολουθιών στο αποτέλεσμα της τελικής στοίχισης.



Εικόνα 3: Υπολογισμός SP-score για στήλες μιας πολλαπλής στοίχισης.

Ο υπολογισμός για τη στήλη i πραγματοποιείται με την άθροιση των επιμέρους scores για όλες τις κατά ζεύγη στοιχίσεις στη στήλη αυτή για τις l πλήθος ακολουθίες. Η άθροιση των SP-scores που προκύπτει για όλες τις στήλες της πολλαπλής στοίχισης δίνει το τελικό SP-score για την προκύπτουσα πολλαπλή στοίχιση και αποτελεί το μέτρο το οποίο επιθυμούμε να βελτιστοποιήσουμε. Να σημειωθεί ότι κάθε ζεύγος καταλοίπων συνεισφέρει μόνο μία φορά στον υπολογισμό.

Η προσέγγιση αυτή παρουσιάζει σημαντικά μειονεκτήματα (για λεπτομέρειες αλλά και εναλλακτικούς τρόπους βαθμονόμησης πολλαπλών στοιχίσεων συμβουλευτείτε το σύγγραμμα Mount, 2001,

σελίδες 151-152). Στην πράξη, όμως, είναι η ευρύτερα χρησιμοποιούμενη στους αλγορίθμους πολλαπλής στοίχισης. Αυτό συμβαίνει γιατί, αφενός μεν ο υπολογισμός του SP-score είναι πολύ απλός-ταχύς, αφετέρου δε κάποιες βελτιώσεις (όπως π.χ. ο υπολογισμός παραγόντων στάθμισης της συνεισφοράς των ακολουθιών που στοιχίζουμε – δείτε επόμενες παραγράφους) αμβλύνουν σημαντικά τα προβλήματα.

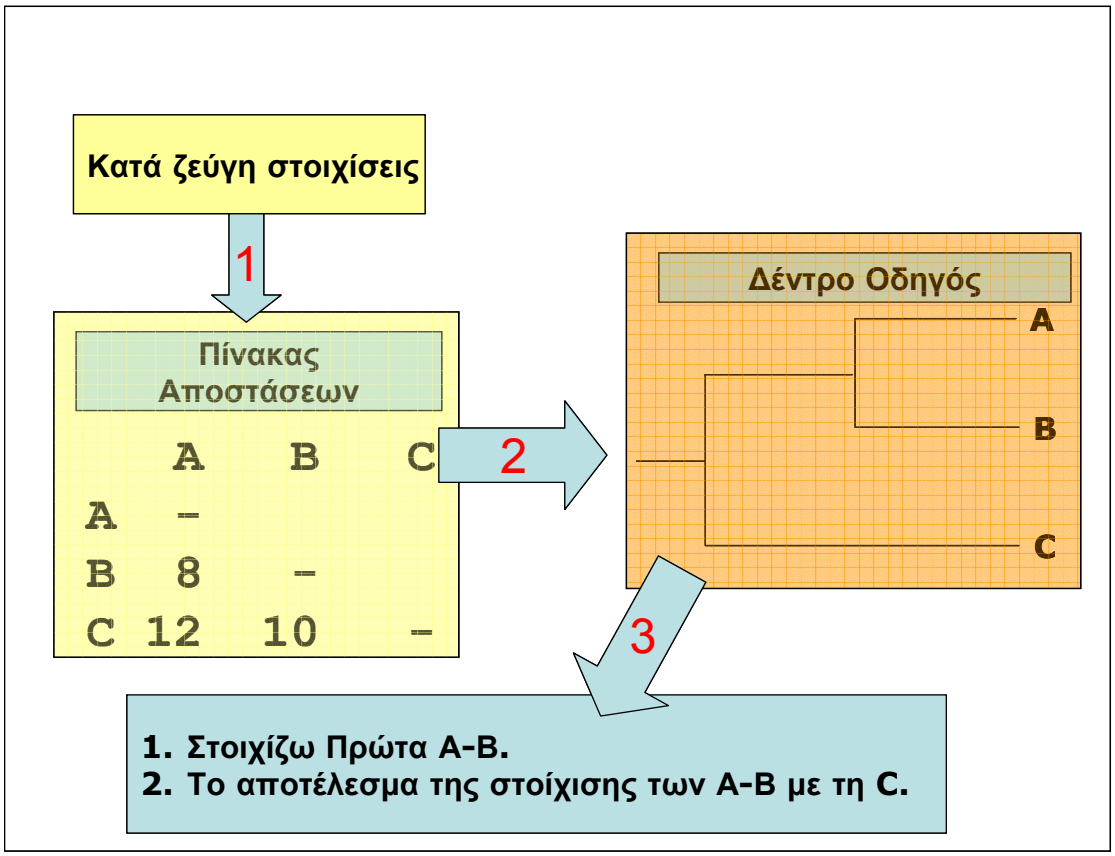
2. Προοδευτική Πολλαπλή Στοίχιση Ακολουθιών

Γενικά

Οι μέθοδοι Προοδευτικής Πολλαπλής Στοίχισης Ακολουθιών (Progressive Multiple Sequence Alignment) έδωσαν σημαντική ώθηση στη χρήση των πολλαπλών στοιχίσεων στη μελέτη οικογενειών βιολογικών μακρομορίων. Αυτό συνέβη καθώς αυτές οι προσεγγίσεις έκαναν εφικτή τη στοίχιση μεγάλου πλήθους ακολουθιών (συχνά με μεγάλα μήκη) σε χρονικά διαστήματα τα οποία κρίνονται ανεκτά.

Η βασική ιδέα για την ανάπτυξη μεθοδολογιών αυτού του τύπου βασίζεται στη χρήση της πληροφορίας, η οποία προκύπτει από όλες τις δυνατές κατά ζεύγη στοιχίσεις των ακολουθιών των οποίων επιθυμούμε την πολλαπλή στοίχιση. Συγκεκριμένα, έχοντας προσδιορίσει την ομοιότητα μεταξύ όλων των ζευγών ακολουθιών, οι αλγόριθμοι αυτοί προσπαθούν να κατασκευάσουν προοδευτικά την πολλαπλή στοίχιση στοιχίζοντας αρχικά τις πιο όμοιες ακολουθίες, και προσθέτοντας *προοδευτικά* στις στοιχίσεις αυτές (από εδώ προκύπτει και ο όρος «προοδευτική») τις πιο απομακρυσμένες από τις ακολουθίες του συνόλου που μας ενδιαφέρει.

Αυτή η προσέγγιση βασίζεται ουσιαστικά στην πεποίθηση που έχουμε ότι οι περισσότερο όμοιες ακολουθίες θα στοιχίζονται με όμοιο τρόπο στην τελική πολλαπλή στοίχιση. Έτσι, δίνεται ιδιαίτερη βαρύτητα σε αυτές και, στην ουσία, οι κατά ζεύγη στοιχίσεις των πιο όμοιων ακολουθιών αποτελούν το ικρίωμα πάνω στο οποίο θα χτιστεί η τελική πολλαπλή στοίχιση. Μια σχηματική απεικόνιση των βασικών βημάτων που απαιτούνται σε μια διαδικασία προοδευτικής πολλαπλής στοίχισης παρέχεται στην Εικόνα 4.



Εικόνα 4: Σχηματική απεικόνιση των βημάτων κατά την προοδευτική πολλαπλή στοίχιση ακολουθιών.

Η διαδικασία ξεκινά με όλες τις δυνατές κατά ζεύγη στοιχίσεις, με βάση τις οποίες υπολογίζεται ένας τριγωνικός κάτω πίνακας αποστάσεων D_{ij}(1). Από τις τιμές του πίνακα αποστάσεων δημιουργείται ένα δέντρο-οδηγός4, συνήθως με μια μέθοδο clustering (2). Οι ακολουθίες στοιχίζονται με τη σειρά με την οποία εισήλθαν στο δέντρο δίνοντας προοδευτικά την τελική πολλαπλή στοίχιση (3). Προφανώς, τα διαφορετικά βήματα είναι δυνατόν να υλοποιηθούν με διαφορετικούς τρόπους.

[4] Προσοχή!! Το δέντρο-οδηγός ΔΕΝ αποτελεί σε καμία περίπτωση φυλογενετικό δέντρο.

Η μέθοδος των Feng και Doolittle

Η πρώτη μεθοδολογία αυτού του τύπου προτάθηκε το 1987 από τους Feng και Doolittle (Feng and Doolittle, 1987). Η πρώτη μορφή της μεθόδου στηρίχτηκε στη χρήση της μεθόδου Needleman-Wunsch για ολική στοίχιση κατά ζεύγη. Με επαναληπτικό τρόπο στοιχίζονται όλες οι (N το πλήθος) ακολουθίες μεταξύ τους με αρχικό σκοπό την κατασκευή ενός (πρόχειρου) φυλογενετικού δέντρου, το οποίο ονομάζεται *δέντρο-οδηγός* (*guide-tree*). Τα scores ομοιότητας μεταξύ δύο τυχαίων ακολουθιών s_i, s_j ακολουθιών μετατρέπονται σε ένα μέτρο απόστασης D_{ij} (i,j=1,2,…,N, με $i<j$) μεταξύ τους σύμφωνα με τη σχέση:

$$D_{ij} = -C \ln S_{eff}$$

όπου:

S_{eff}: Το *ενεργό score* (*effective score*) της στοίχισης μεταξύ των δύο ακολουθιών

C: Μια σταθερά (συνήθως, C=100)

Το ενεργό score S_{eff} της στοίχισης κατά ζεύγη υπολογίζεται με μια διαδικασία σχετικά χρονοβόρα:

$$S_{eff} = \frac{S_{ij,obs} - S_{rand}}{S_{\max} - S_{rand}}$$

$S_{ij,obs}$: Το score της στοίχισης μεταξύ των δύο ακολουθιών

S_{rand}: Το μέσο score που προκύπτει από στοιχίσεις τυχαίων ακολουθιών με τα ίδια μήκη και την ίδια αμινοξική σύσταση με τις s_i, s_j[5]

S_{max}: Το μέγιστο score για στοιχίσεις των δύο ακολουθιών, όπως προφανώς μπορεί να υπολογιστεί από τη στοίχιση κάθε ακολουθίας με τον εαυτό της και εν συνεχεία υπολογίζοντας το μέσο όρο

Καθώς η τιμή S_{rand} μεγαλώνει (και αυτό προφανώς συμβαίνει για ακολουθίες που η μεταξύ τους εξελικτική απόσταση μεγαλώνει) η S_{eff} μικραίνει. Λογαριθμίζοντας το ενεργό score η σχέση μεταξύ D_{ij} και της εξελικτικής απόστασης γίνεται προσεγγιστικά γραμμική.[6]

Αυτός ο τρόπος προσδιορισμού αποστάσεων απαιτεί δύο σημαντικές παραδοχές (οι οποίες γνωρίζουμε πολύ καλά ότι δεν ισχύουν):

- Όλα τα κατάλοιπα μιας ακολουθίας έχουν την ίδια πιθανότητα να μεταλλαχθούν

- Κάθε τύπος αμινοξικού καταλοίπου έχει την ίδια πιθανότητα να αντικατασταθεί από οποιοδήποτε άλλο τύπο καταλοίπου

Ο υπολογισμός των $N(N-1)/2$ τιμών D_{ij} μπορεί εύκολα να αντιστοιχηθεί με την κατασκευή ενός τριγωνικού κάτω πίνακα (Βήμα 1, Εικόνα 4). Η διαδικασία συνεχίζεται με την κατασκευή του δέντρού-οδηγού. Οι Feng και Doolittle εφάρμοσαν τη διαδικασία που είχαν παλιότερα προτείνει οι Fitch και Margoliash (Fitch and

[5] Μπορεί να υπολογιστεί ως η μέση τιμή της κατανομής των scores που προκύπτουν μετά από επαναληπτικό τυχαίο «ανακάτεμα» (random shuffling) των καταλοίπων των ακολουθιών, διατηρώντας τα μήκη και την αμινοξική τους σύσταση αμετάβλητα. Το πλήθος των τυχαίων στοιχίσεων σε αυτό το στάδιο είναι μερικές εκατοντάδες (διαπιστώστε το κόστος σε χρόνο!!).

[6] Για εναλλακτικές μεθοδολογίες μετατροπής των scores των στοιχίσεων σε αποστάσεις μεταξύ των ακολουθιών συμβουλευθείτε το βιβλίο του Mount, σελίδες 264-269.

Margoliash, 1967). Η μέθοδος αυτή είναι μια κλασική μέθοδος ιεραρχικού clustering (hierarchical ή agglomerative clustering)[7].
Το δέντρο-οδηγός που προκύπτει υποδεικνύει προσεγγιστικά μόνο τις εξελικτικές σχέσεις μεταξύ των ακολουθιών[8], αλλά είναι αυτό που καθοδηγεί τη σειρά με την οποία θα προστεθούν οι ακολουθίες στην πολλαπλή στοίχιση.
Η πολλαπλή στοίχιση ξεκινά με την κατά ζεύγη στοίχιση των 2 πιο όμοιων ακολουθιών. Στα επόμενα βήματα θα απαιτηθεί η στοίχιση είτε ακολουθίας με υπάρχουσα στοίχιση είτε στοίχισης με στοίχιση. Στην απλούστερη των περιπτώσεων, αυτή της στοίχισης ακολουθίας με προϋπάρχουσα στοίχιση, οι Feng και Doolittle πρότειναν την εξής διαδικασία:

1. Η ακολουθία στοιχίζεται με ΟΛΕΣ τις ακολουθίες της στοίχισης (με το γνωστό αλγόριθμο δυναμικού προγραμματισμού)

2. Η ακολουθία προστίθεται στην πολλαπλή στοίχιση βε βάση τη στοίχιση κατά ζεύγη του βήματος 1 που είχε το μεγαλύτερο score.

Είναι προφανής η δυνατότητα επέκτασης αυτής της ιδέας για στοίχιση μεταξύ δύο στοιχίσεων:

[7] Οι μέθοδοι clustering έχουν ως σκοπό την ομαδοποίηση οντοτήτων (στη συγκεκριμένη περίπτωση οι ακολουθίες που στοιχίζουμε) με βάση τις μεταξύ τους αποστάσεις. Οι ιεραρχικές μέθοδοι ξεκινούν φτιάχνοντας μια ομάδα για κάθε οντότητα και συνενώνουν ιεραρχικά ομάδες μεταξύ τους λαμβάνοντας υπόψη τις αποστάσεις μεταξύ των μελών των ομάδων. Μια εισαγωγική παρουσίαση τέτοιων μεθόδων μπορείτε να βρείτε online στο URL:
http://biodiver.bio.ub.es/vegana/resources/help/ginkgo/Agglomerative.html
Τεχνικές λεπτομέρειες σχετικά με τη μέθοδο Fitch-Margoliash μπορείτε να βρείτε επίσης στο βιβλίο του Mount, σελίδες 256-260.

[8] Περιττό να σημειώσουμε εδώ ότι οι μεθοδολογίες πολλαπλής στοίχισης θεωρούν εκ των προτέρων δεδομένη τη συσχέτιση των ακολουθιών που επιχειρούμε να στοιχίσουμε. Αυτό προϋποθέτει (από την πλευρά του χρήστη) προσεκτική επιλογή των ακολουθιών που θα στοιχίσει. Να θυμάστε πάντα ότι κατά την πολλαπλή στοίχιση ακολουθιών ισχύει ο θεμελιώδης μνημονικός κανόνας Garbage-IN=>Garbage-Out. Η μέθοδος επιλογής των ακολουθιών που θα στοιχίσουμε εξαρτάται από τη φύση της μελέτης μας (συζήτηση γύρω από το ζήτημα αυτό πραγματοποιήθηκε κατά τη διάλεξη ...)

1. Οι ακολουθίες των δύο στοιχίσεων στοιχίζονται κατά ζεύγη με ΟΛΟΥΣ τους δυνατούς τρόπους

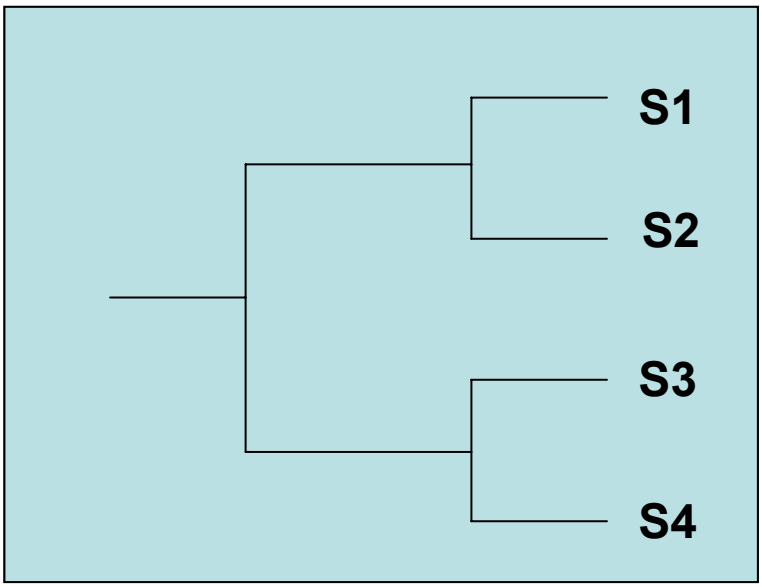
2. Οι στοιχίσεις ενοποιούνται με βάση τη στοίχιση κατά ζεύγη του βήματος 1 που είχε το μεγαλύτερο score.

Παράδειγμα Προοδευτικής Πολλαπλής Στοίχισης με βάση τη μέθοδο Feng-Doolittle

Έστω, οι ακολουθίες:
`S1: AAATCGG, S2: AAACCGG, S3: ATACCCTG, S4: ATACCCGG`
και ας υποθέσουμε ότι με κάποιο σύστημα βαθμονόμησης για τις κατά ζεύγη στοιχίσεις προκύπτει ο παρακάτω πίνακας αποστάσεων:

	S1	**S2**	**S3**	**S4**
S1	-			
S2	1	-		
S3	3	2	-	
S4	2	2	1	-

Ένα πιθανό δέντρο-οδηγός θα ήταν λοιπόν το:



Η πληροφορία που μας χρειάζεται για τη δημιουργία της πολλαπλής στοίχισης (με βάση το δέντρο-οδηγό) συνοψίζεται στις παρακάτω οδηγίες:

	ΟΔΗΓΙΑ	**ΑΠΟΤΕΛΕΣΜΑ**
1	Στοίχισε S1-S2 => $S_{1,2}$	AAATCGG ||| ||| AAACCGG
2	Στοίχισε S3-S4 => $S_{3,4}$	ATACCCTG ||||| | ATACCCGG
3	Στοίχισε $S_{1,2}$ – $S_{3,4}$	Δείτε παρακάτω ...

Για τη στοίχιση $S_{1,2}$ – $S_{3,4}$, πρέπει να στοιχίσω S1-S3, S1-S4, S2-S3, S3-S4.[9]

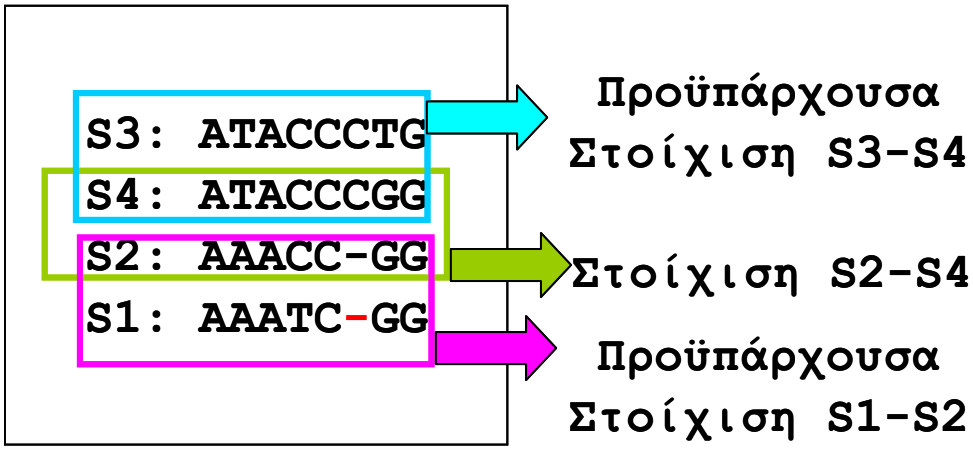
	Στοίχιση	Score
S1-S3	AAATC-GG | | | | ATACCCTG	3

[9] Καλή πρακτική για την εξοικονόμηση χρόνου είναι να έχω αποθηκεύσει τις στοιχίσεις αυτές (και τα αντίστοιχα scores) από το πρώτο βήμα της διαδικασίας. Στην περίπτωση βέβαια που έχουμε πολλές ακολουθίες και η αποθήκευση όλων των κατά ζεύγη στοιχίσεών τους δεν είναι πρακτική, μπορώ να φυλάξω μόνο τα scores και να χρειαστεί να υπολογίσω ξανά μόνο εκείνη τη στοίχιση με το μέγιστο score.

S1-S4	AAATCGG | | | | ATACCCG	4
S2-S3	AAACCG-G | ||| | ATACCCTG	4
S2-S4	AAACC-GG | ||| || ATACCCGG	5

Σύμφωνα με τα προηγούμενα, η ζητούμενη στοίχιση μεταξύ $S_{1,2}$ – $S_{3,4}$ θα πραγματοποιηθεί με βάση τη στοίχιση S2-S4.

Επομένως:



Εικόνα 5: Η πολλαπλή στοίχιση που υπολογίστηκε για τις ακολουθίες του παραδείγματος.

Παρατηρήστε με κόκκινο χρώμα το κενό που εισάγουμε στην ακολουθία S1, λαμβάνοντας υπόψη ότι υπήρχε κενό στην αντίστοιχη θέση της στοίχισης S2-S4 με βάση την οποία «ενώθηκαν» οι επιμέρους στοιχίσεις.

Η μέθοδος CLUSTAL

Η διαδικασία που ακολούθησαν οι Feng και Doolittle, είναι προφανές ότι μπορεί να έχει διάφορες παραλλαγές, και αυτό γιατί κάθε ένα από τα ξεχωριστά της στάδια είναι δυνατόν να υλοποιηθεί με διαφορετικές μεθοδολογίες. Παρόλα αυτά αποτελεί εξαιρετικό παράδειγμα για την εισαγωγή των εννοιών πίσω από τη μεθοδολογία της προοδευτικής πολλαπλής στοίχισης εξαιτίας της απλότητάς του. Πέρα από τις παραδοχές που έχει εγγενώς η προοδευτική πολλαπλή στοίχιση, η μεθοδολογία αυτή εμφάνιζε ένα σημαντικό μειονέκτημα. Η ανάγκη για εφαρμογή δυναμικού προγραμματισμού για μεγάλο πλήθος στοιχίσεων κατά ζεύγη καθιστά σημαντικές τις υπολογιστικές απαιτήσεις σε χρόνο CPU.

Λίγο καιρό αργότερα, οι Higgins και Sharp (Higgins and Sharp, 1988) πρότειναν μια εναλλακτική μέθοδο (CLUSTAL) η οποία (αν και είχε την ίδια φιλοσοφία) διέφερε σε δύο σημαντικά σημεία της βασικής διαδικασίας σε σχέση με τη μέθοδο Feng-Doolittle:

1. Οι κατά ζεύγη στοιχίσεις ακολουθίας-ακολουθίας ήταν δυνατόν να πραγματοποιηθούν πολύ ταχύτερα με βάση τους νέους ευριστικούς αλγορίθμους οι οποίοι είχαν ήδη προταθεί (δείτε FASTA, Pearson and Lipman, 1988). Το πλήθος των διαφορών των ακολουθιών σε αυτές τις στοιχίσεις (mismatches, indels) χρησιμοποιούνται για τη δημιουργία του πίνακα αποστάσεων $D_{ij.}$ Εναλλακτικά, μπορεί να εφαρμοστεί πλήρης δυναμικός προγραμματισμός ή η βελτιωμένη έκδοση που προτάθηκε από τους Myers και Miller (Myers and Miller, 1988) η οποία εκτελεί ΔΠ με γραμμική απαίτηση σε χρόνο O(N)
2. Η δημιουργία του δέντρου-οδηγού γίνεται με μια διαφορετική μέθοδο clustering (Neighbor-joining, Saitou and Nei, 1987)[10]

Σχολιάζοντας τις βασικές αυτές τροποποιήσεις της μεθοδολογίας από τους Higgins και Sharp μπορούμε εύκολα να σκεφτούμε ότι, στην

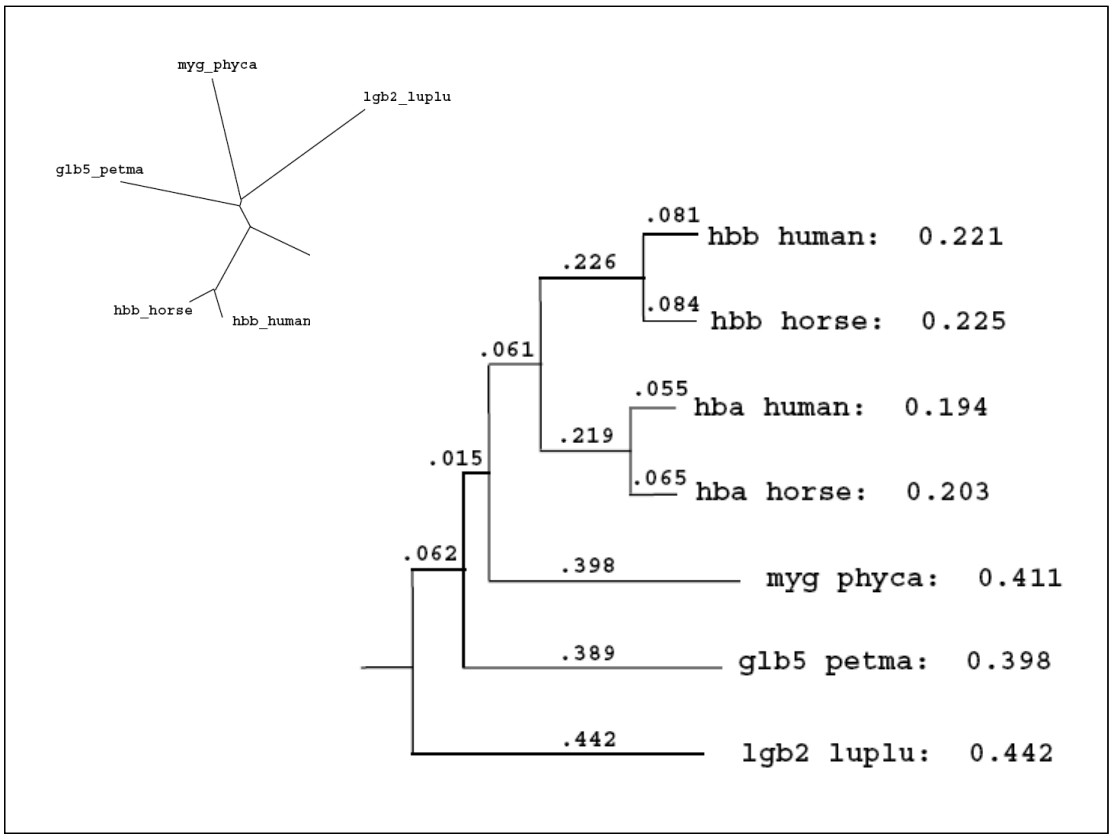
[10] Τεχνικές λεπτομέρειες στο βιβλίο του Mount, σελίδες 260-261

πραγματικότητα, η χρησιμότητα των στοιχίσεων κατά ζεύγη έγκειται στον υπολογισμό των scores για την κατασκευή του δέντρου-οδηγού. Με δεδομένο ότι η απόλυτη ακρίβεια στον υπολογισμό του δέντρου αυτού δεν είναι το ζητούμενο, η χρήση προσεγγιστικών στοιχίσεων δεν αναμένουμε να επηρεάσει σημαντικά το τελικό αποτέλεσμα.

Η μέθοδος CLUSTAL συνέχισε να αναπτύσσεται συστηματικά και σύντομα υπήρξαν βελτιώσεις της (CLUSTALV, Higgins *et al.*, 1992 και CLUSTALW, Thompson *et al.*, 1994). Η μέθοδος αυτή έχει κατά κόρον χρησιμοποιηθεί σε πρακτικές εφαρμογές και αποτελεί μέτρο σύγκρισης για οποιαδήποτε νέα μέθοδο πολλαπλής στοίχισης ακολουθιών, τόσο για την ταχύτητά της όσο και για την ποιότητα των στοιχίσεων που παράγει. Αρκετές καινοτομίες βελτίωσαν σημαντικά την ποιότητα των αποτελεσμάτων του CLUSTAL τόσο έναντι παλαιότερων μεθόδων όσο και έναντι προγενέστερων εκδόσεων του, και τις σημαντικότερες από αυτές θα συζητήσουμε στα επόμενα. Συγκεκριμένα, θα μας απασχολήσουν η μέθοδος στάθμισης των ακολουθιών βάσει του δέντρου-οδηγού, η διαδικασία του profile-alignment και η εισαγωγή κενών με ποινές εξαρτώμενες από τη θέση.

Στάθμιση Ακολουθιών

Οι συνεισφορές ακολουθιών με μεγάλη μεταξύ τους ομοιότητα σταθμίζονται έτσι ώστε να αποφευχθεί το πιθανό γεγονός κατά το οποίο η τελική πολλαπλή στοίχιση θα εξαρτάται σε μεγάλο βαθμό από αυτές τις ακολουθίες. Για κάθε ακολουθία, υπολογίζεται ένας παράγοντας στάθμισης (weight) με βάση τις αποστάσεις στο των ακολουθιών στο δέντρο-οδηγό.



Εικόνα 6: Στάθμιση Ακολουθιών με βάση το δέντρο-οδηγό

Από το μήκη των κλαδιών του δέντρου-οδηγού υπολογίζεται ο παράγοντας στάθμισης για κάθε ακολουθία εφαρμόζοντας την παρακάτω διαδικασία: Από το «φύλλο» του δέντρου που αντιστοιχεί στην κάθε ακολουθία ακολουθούμε τη διαδρομή μέχρι τη «ρίζα» του δέντρου αθροίζοντας το μήκος κάθε «κλάδου» διαιρεμένο με το πλήθος των ακολουθιών που βρίσκονται κάτω από αυτόν. Για παράδειγμα, για την ακολουθία hba_human του σχήματος: w = 0.055/1 + 0.219/2 + 0.061/4 + 0.015/5 + 0.062/6 = 0.055 + 0.110 + 0.016 + 0.003 + 0.010 = 0.194

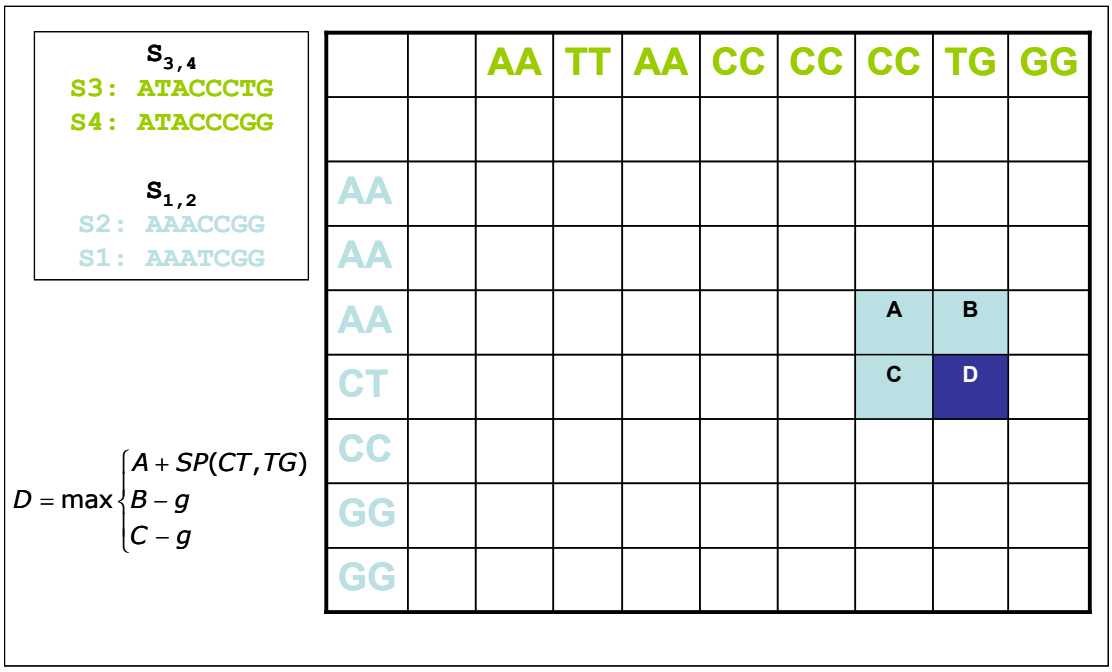
Προφανώς, ο παράγοντας στάθμισης για μια ακολουθία η οποία βρίσκεται σε ένα κλάδο που είναι απευθείας προσαρτημένος στη ρίζα του δέντρου ισούται με το μήκος του κλάδου αυτού.

Σημείωση: Στην περίπτωση που με την παραπάνω (ή κάποια άλλη διαδικασία) έχουμε υπολογίσει παράγοντες στάθμισης για τις ακολουθίες που στοιχίζουμε μπορούμε αντί του SP-score να υπολογίζουμε για κάθε στήλη ένα σταθμισμένο SP-score (weighted SP-score) πολλαπλασιάζοντας το score που προκύπτει από τον

πίνακα αντικατάστασης για κάθε στήλη καταλοίπων με τα «βάρη» των αντίστοιχων ακολουθιών.

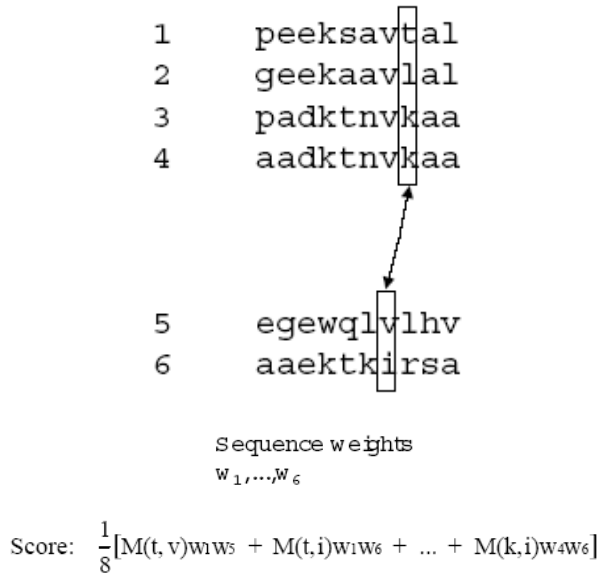
Profile Alignment

Η διαδικασία της προοδευτικής στοίχισης προϋποθέτει ότι σε κάποιο στάδιο της διαδικασίας θα πραγματοποιηθεί στοίχιση μεταξύ κάποιας προϋπάρχουσας στοίχισης και μιας ακολουθίας ή άλλης στοίχισης. Η ομάδα του David Eisenberg (Gribskov *et al.*, 1987) ήταν από τους πρωτοπόρους στη χρήση αποτελεσμάτων πολλαπλών στοιχίσεων ακολουθιών για την κατασκευή προφίλ (profiles, ή πίνακες ομοιότητας εξαρτώμενους από τη θέση, Position-Specific Scoring Matrix, PSSM) πρωτεϊνικών οικογενειών.
Παράλληλα, πρότειναν κατάλληλη μεθοδολογία, η οποία είναι επέκταση των αλγορίθμων δυναμικού προγραμματισμού (πάλι!!) για τη σύγκριση/στοίχιση μιας ακολουθίας με ένα προφίλ που αντιστοιχεί σε μια πρωτεϊνική οικογένεια. Με μια απλή τροποποίηση η μέθοδός τους είναι δυνατόν να χρησιμοποιηθεί για τη στοίχιση δύο διαφορετικών profiles. Ένα profile αποτελεί μια απλή πιθανοκρατική περιγραφή μιας πολλαπλής στοίχισης, όπου κάθε στήλη της στοίχισης περιγράφεται από τη συχνότητα εμφάνισης κάθε διαφορετικού τύπου καταλοίπου.
Η στοίχιση μεταξύ δύο στοιχίσεων μπορεί να υπολογιστεί με βάση τους κλασικούς αλγορίθμους δυναμικού προγραμματισμού εάν σκεφτούμε ότι στις γραμμές και τις στήλες του πίνακα δυναμικού προγραμματισμού δεν τοποθετούμε δύο ακολουθίες αλλά τις στήλες κάθε μιας στοίχισης. Για τη στοίχιση $S_{1,2}$ – $S_{3,4}$ που είδαμε στο παράδειγμα της μεθόδου Feng-Doolittle, η στοίχιση με τη μέθοδο profile alignment θα γινόταν ως εξής (Εικόνα 7):



Εικόνα 7: Επέκταση του αλγόριθμου δυναμικού προγραμματισμού για profile-alignment

Βλέπουμε την περίπτωση που στοιχίζονται δύο κατά ζεύγη στοιχίσεις. Η τιμή D του score στο μπλε κελί του πίνακα υπολογίζεται από τα scores των γραμμοσκιασμένων γειτονικών κελιών του με τη γνωστή διαδικασία μεγιστοποίησης. A, B, C είναι τα ήδη υπολογισμένα scores από προηγούμενα βήματα του δυναμικού προγραμματισμού, ενώ g είναι η ποινή εισαγωγής κενού. Το SP(CT, TG) είναι το SP-score για την στοίχιση των δύο στηλών των στοιχίσεων με τα κατάλοιπα CT και TG αντίστοιχα. Στην περίπτωση που επικρατεί κάποιο από τα B-g, C-g το κενό εισάγεται στη συγκεκριμένη θέση για κάθε ακολουθία της αντίστοιχης προϋπάρχουσας στοίχισης. Να σημειωθεί ότι αντί για το SP-score μπορεί να χρησιμοποιηθεί το σταθμισμένο SP-score (Εικόνα 8).



Εικόνα 8: Υπολογισμός σταθμισμένου SP-score.

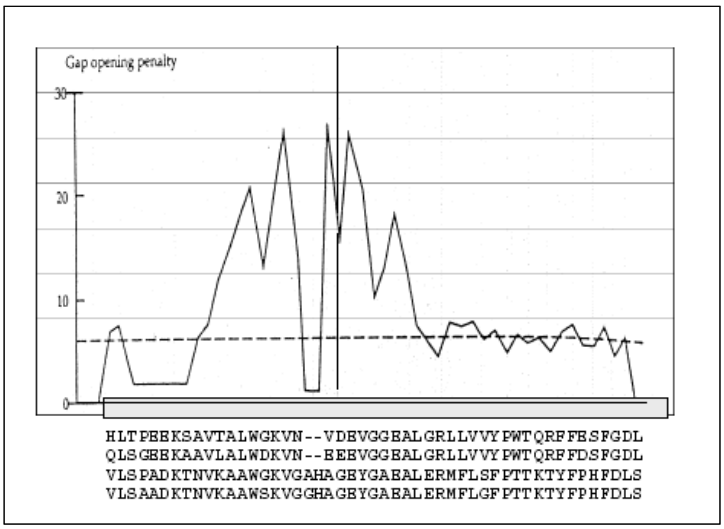
Κατά τη στοίχιση των σημειωμένων στηλών από τις προυπάρχουσες στοιχίσεις τα scores από τον πίνακα ομοιότητας M(a,b) [score για ταίριασμα καταλοίπων a, b] πολλαπλασιάζεται με τα βάρη των αντίστοιχων ακολουθιών και μετά πραγματοποιείται η άθροιση όπως είδαμε στην Εικόνα 3. Ο παράγοντας 1/8 κανονικοποιεί το σταθμισμένο SP-score ανά ζεύγος καταλοίπων.

Ποινές κενών εξαρτώμενες από τη θέση

Το ζήτημα της εισαγωγής κενών φάνηκε από την περίπτωση της στοίχισης ακολουθιών κατά ζεύγη ότι απαιτεί προσεκτικό χειρισμό, ενώ παράλληλα δεν υπάρχει μια κοινά αποδεκτή προσέγγιση η οποία να έχει κάποιο ισχυρό βιολογικό-θεωρητικό υπόβαθρο. Ειδικότερα, στην περίπτωση της προοδευτικής πολλαπλής στοίχισης, είναι προφανές ότι εάν σε κάποια φάση της διαδικασίας των επιμέρους στοιχίσεων εισαχθεί κενό σε κάποια θέση αυτό το κενό παραμένει «παγωμένο». Επομένως, η όσο το δυνατόν καλύτερη τοποθέτηση των κενών από τα αρχικά κιόλας στάδια της πολλαπλής στοίχισης

αποτελεί κρίσιμο στοιχείο για την ποιότητα της τελικής πολλαπλής στοίχισης.

Έχοντας αυτά υπόψη, η ομάδα ανάπτυξης της μεθόδου CLUSTAL επινόησε ένα σύστημα για την εισαγωγή κενών το οποίο να αντιμετωπίζει με διαφορετικό τρόπο τις διαφορετικές θέσεις των ακολουθιών, λαμβάνοντας υπόψη την πληροφορία η οποία έχει ήδη ενσωματωθεί σε κάθε επιμέρους στοίχιση. Ειδικότερα, προκειμένου για τη στοίχιση με τη μέθοδο profile-alignment η εισαγωγή κενών σε μια στοίχιση δεν «τιμωρείται» με σταθερή ποινή, αλλά με ποινή η οποία είναι εξαρτώμενη από τη θέση, και το περιεχόμενο της στοίχισης στη θέση αυτή.



Εικόνα 9: Ποινές κενών εξαρτώμενες από τη θέση

Διαισθητικά (για περισσότερες λεπτομέρειες δείτε Higgins *et al.*, 1996) μπορείτε να φανταστείτε ότι οι ποινές για την εισαγωγή κενών είναι μικρές σε στήλες της στοίχισης που περιέχουν ήδη κενά, ενώ είναι μεγαλύτερες στις γειτονικές περιοχές ή σε περιοχές που εμφανίζουν συντήρηση. Ομάδες στηλών με πολλά υδρόφιλα κατάλοιπα (οι οποίες πιθανότατα αντιστοιχούν σε loops στην επιφάνεια της τρισδιάστατης δομής της πρωτεΐνης) έχουν επίσης μειωμένες ποινές. Ένα παράδειγμα δίνεται στην Εικόνα 9.

3. Ερωτήσεις

1. Δίνεται ένα τμήμα μίας πολλαπλής στοίχισης αμινοξικών ακολουθιών:

```
Seq1: DDRTFRYGP
Seq2: DEKSFRFGP
Seq3: NDKLFKYGG
Seq4: NQHTFRWGG
```

Να υπολογίσετε:
A. Το SP score κάθε στήλης
B. Το σταθμισμένο SP score κάθε στήλης
Σημείωση: Να χρησιμοποιήσετε τον πίνακα αντικατάστασης BLOSUM62
Δίνονται οι παράγοντες στάθμισης w1=0.031, w2 = 0.025, w3 = 0.101, και w4 = 0.133

2. Ένας προπτυχιακός φοιτητής βιολογίας ετοιμάζοντας το σεμινάριο που του ανέθεσαν με θέμα τις αιμοσφαιρίνες, σκέφτηκε ότι θα ήταν πολύ καλό να παρουσιάσει μια πολλαπλή στοίχιση των ακολουθιών που θα μπορούσε να βρει στις βάσεις δεδομένων. Ο στόχος του διττός: αφενός μεν για να εντυπωσιάσει το ακροατήριο, αφετέρου για να δείξει (αφού πρώτα δει και ο ίδιος) τα διατηρημένα χαρακτηριστικά στο επίπεδο της αμινοξικής ακολουθίας.
Η διαδικασία που ακολούθησε ήταν η παρακάτω:

- Στο δικτυακό τόπο http://www.expasy.org/sprot πραγματοποίησε αναζήτηση στις κύριες βάσεις (Swiss-Prot/TrEMBL) με τον όρο *Hemoglobin*

- Ο φοιτητής, τρισευτυχισμένος από το μεγάλο πλήθος ακολουθιών που του επέστρεψε η αναζήτηση στις βάσεις δεδομένων, άρχισε καρτερικά να αποθηκεύει τις ακολουθίες στον υπολογιστή του.

- Χρησιμοποιώντας το CLUSTALW μέσω του διαδικτύου (http://www.ebi.ac.uk/clustalw/index.html), κατασκεύασε μια πολλαπλή στοίχιση με όλες τις ακολουθίες.

Να απαντήσετε στα παρακάτω ερωτήματα:

Α. Συμφωνείτε με τις ενέργειες του συναδέλφου σας; Να δικαιολογήσετε αναλυτικά τις απόψεις σας.

Β. Από το σύνολο των εγγραφών που προκύπτουν από την αναζήτηση το συναδέλφου σας να επικεντρώσετε την προσοχή σας σε εκείνες με Uniprot AC: P18974, Q7M3B8, Q9XSN2, P20244, Q9XSK1, Q9PVM4, P83124, Q865F8. Χρησιμοποιήστε τους σχολιασμούς από τη βάση δεδομένων και όποια σχετική βιβλιογραφία μπορείτε να βρείτε για να κατασκευάσετε μια (ή περισσότερες) πολλαπλές στοιχίσεις οι οποίες να έχουν (βιολογικά) κάποιο νόημα. Χρησιμοποιήστε κάποιο από τα εργαλεία που προτείνονται στην ενότητα 4 για την οπτικοποίηση – παρουσίαση των αποτελεσμάτων.

Γ. Χρησιμοποιείστε το εργαλείο BLAST του NCBI ώστε να εντοπίσετε εάν για κάποια (ες) ακολουθία (ες) που περιλαμβάνονται στην πολλαπλή στοίχιση που κατασκευάσατε υπάρχει κάποια λυμένη δομή στην Protein Data Bank (PDB, http://www.rcsb.org/pdb) με σημαντική ομοιότητα στο επίπεδο της ακολουθίας. Σημειώστε στην πολλαπλή στοίχιση τη θέση των στοιχείων δευτεροταγούς δομής με βάση την πειραματικά προσδιορισμένη δομή. Παραθέστε τα σχόλιά σας.

4. Συμπληρωματικό Υλικό

Χρήσιμες Πηγές στο Διαδίκτυο

1. Εργαλεία πολλαπλής στοίχισης ακολουθιών:

CLUSTALW: http://www.ebi.ac.uk/clustalw/index.html
T-COFFEE: http://igs-server.cnrs-mrs.fr/Tcoffee/tcoffee_cgi/index.cgi
DIALIGN: http://bioweb.pasteur.fr/seqanal/interfaces/dialign2-simple.html
SAGA: http://igs-server.cnrs-mrs.fr/~cnotred/Projects_home_page/saga_home_page.html
PRALINE: http://www.ibivu.cs.vu.nl/programs/pralinewww/

2. Εργαλεία οπτικοποίησης πολλαπλών στοιχίσεων:

BOXSHADE: http://www.ch.embnet.org/software/BOX_form.html
CINEMA: http://www.biochem.ucl.ac.uk/bsm/dbbrowser/CINEMA2.02/index2.html
ESPript: http://espript.ibcp.fr/ESPript/cgi-bin/ESPript.cgi

Βιβλιογραφία

Carrillo, H. and D. Lipman (1988). The multiple sequence alignment problem in biology. *SIAM J. Appl. Math.*, **48**: 197-209.

Feng, D. F. and R. F. Doolittle (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J Mol Evol*, **25**(4): 351-60.

Fitch, W. M. and E. Margoliash (1967). Construction of phylogenetic trees. *Science*, **155**(760): 279-84.

Gribskov, M., A. D. McLachlan and D. Eisenberg (1987). Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci U S A*, **84**(13): 4355-8.

Gupta, S. K., J. D. Kececioglu and A. A. Schèaffer (1995). Improving the practical space and time efficiency of the shortest-paths approach to sum-of-pairs multiple sequence alignment. *J Comput Biol*, **2**(3): 459-72.

Gusfield, D. (1997). Algorithms on strings, trees, and sequences : computer science and computational biology. Cambridge [England] ; New York, Cambridge University Press.

Higgins, D. G., A. J. Bleasby and R. Fuchs (1992). CLUSTAL V: improved software for multiple sequence alignment. *Comput Appl Biosci*, **8**(2): 189-91.

Higgins, D. G. and P. M. Sharp (1988). CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene*, **73**(1): 237-44.

Higgins, D. G., J. D. Thompson and T. J. Gibson (1996). Using CLUSTAL for multiple sequence alignments. *Methods Enzymol*, **266**: 383-402.

Hubbard, T. J., A. M. Lesk and A. Tramontano (1996). Gathering them in to the fold. *Nat Struct Biol*, **3**(4): 313.

Lipman, D. J., S. F. Altschul and J. D. Kececioglu (1989). A tool for multiple sequence alignment. *Proc Natl Acad Sci U S A*, **86**(12): 4412-5.

Mount, D. W. (2001). Bioinformatics : sequence and genome analysis. Cold Spring Harbor, N.Y., Cold Spring Harbor Laboratory Press.

Murata, M., J. Richardson and J. Sussman (1985). Simultaneous comparison of three protein sequences. *Proc Natl Acad Sci U S A*, **82**: 3073-77.

Myers, E. W. and W. Miller (1988). Optimal alignments in linear space. *Comput Appl Biosci*, **4**(1): 11-7.

Pearson, W. R. and D. J. Lipman (1988). Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, **85**(8): 2444-8.

Saitou, N. and M. Nei (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, **4**(4): 406-25.

Thompson, J. D., D. G. Higgins and T. J. Gibson (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, **22**(22): 4673-80.

Waterman, M. and M. Perlwitz (1984). Line Geometries for Sequence Comparisons. *Bulletin of Mathematical Biology*, **46**(4): 567-577.