Σημειώσεις Βιοπληροφορικής

Αναζήτηση Ομοιοτήτων σε Βάσεις Δεδομένων Ακολουθιών

Βάσεις Δεδομένων Βιολογικών Ακολουθιών - Πρακτικά Ζητήματα
Προσεγγιστικοί Ευριστικοί Αλγόριθμοι
Στατιστική Σημαντικότητα
Εφαρμογές σε πρακτικά προβλήματα

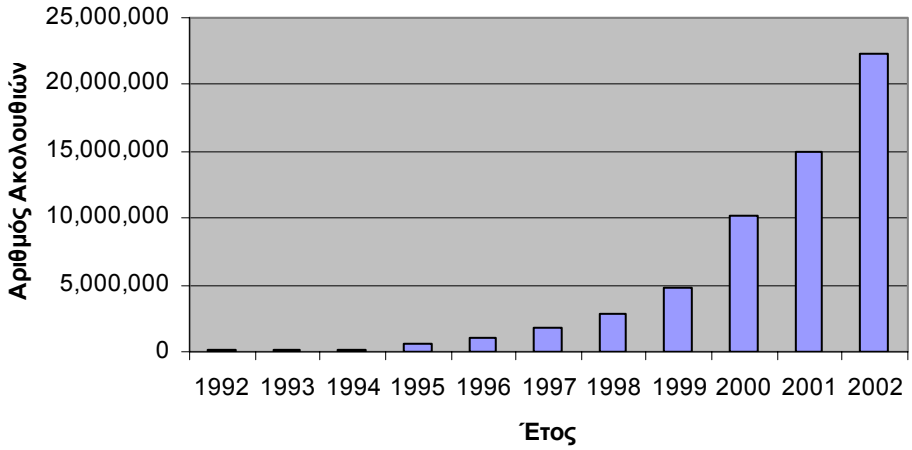
ΒΑΣΙΛΗΣ ΠΡΟΜΠΟΝΑΣ

ΑΘΗΝΑ 2004-2005, ΛΕΥΚΩΣΙΑ 2006

1 ΒΑΣΕΙΣ ΒΙΟΛΟΓΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

1.1 Γενικά

Τα σημαντικά επιτεύγματα στην τεχνολογία της Γονιδιωματικής αλλά και η αλματώδης αύξηση της γνώσης της Βιολογίας των ευκαρυωτικών και των προκαρυωτικών οργανισμών, οι οποίες συντελέσθηκαν τις τελευταίες δεκαετίες, ώθησαν στην υιοθέτηση μιας νέας φιλοσοφίας διεξαγωγής της Βιολογικής έρευνας. Από τη δυνατότητα μελέτης μεμονωμένων γονιδίων περάσαμε στην εποχή που είναι εφικτή η μελέτη ολοκληρωμένων γονιδιωμάτων (ακόμη και στην περίπτωση πολύπλοκων οργανισμών, όπως ο άνθρωπος) (Altman and Dugan, 2003). Οι σύγχρονες τεχνικές προσδιορισμού νουκλεοτιδικών αλληλουχιών, σε συνδυασμό με την κατακόρυφη αύξηση της διαθέσιμης υπολογιστικής ισχύος, έχουν ως αποτέλεσμα τη διαρκή κατάθεση ακολουθιών στις δημόσιες βάσεις δεδομένων (GenBank, http://www.ncbi.nlm.nih.gov/Genbank/index.html, Benson *et al.*, 2003, EMBL-Bank, http://www.ebi.ac.uk/embl/, Stoesser *et al.*, 2003 και DDBJ, http://www.ddbj.nig.ac.jp/, Miyazaki *et al.*, 2003) με ρυθμούς οι οποίοι θα θεωρούνταν ασύλληπτοι, ακόμη και στις αρχές της προηγούμενης δεκαετίας. Ενδεικτικά, στην Εικόνα 1 απεικονίζεται το πλήθος των κατατεθειμένων στη GenBank ακολουθιών ως συνάρτηση του χρόνου. Είναι εμφανές ότι ο αριθμός αυτός ακολουθεί εκθετικό ρυθμό αύξησης, ενώ αντίστοιχος είναι και ο ρυθμός αύξησης των νουκλεοτιδικών βάσεων.



Εικόνα 1: Αριθμός Νουκλεοτιδικών Ακολουθιών κατατεθειμένων στη Βάση Δεδομένων GenBank.

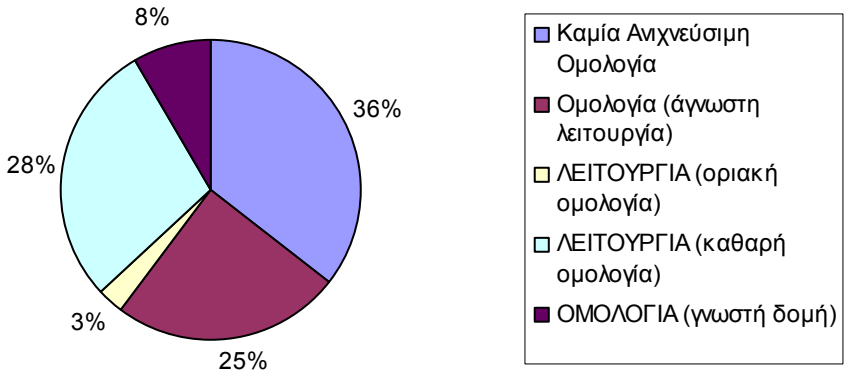
Τα δεδομένα προέρχονται από το δικτυακό τόπο της GenBank, URL: http://www.ncbi.nih.gov/Genbank/GenbankOverview.html, Σεπτέμβριος 2003. Δεδομένα προηγουμένων ετών (1982-1991) δεν συμπεριλήφθηκαν λόγω του πολύ μικρού αριθμού ακολουθιών.

Τις τελευταίες δύο δεκαετίες περισσότερες από 130.000 πρωτεϊνικές ακολουθίες είναι κατατεθειμένες σε βάσεις δεδομένων, όπως είναι η SWISS-PROT (http://us.expasy.org/sprot/, Boeckmann *et al.*, 2003) και η PIR (http://pir.georgetown.edu/, Wu *et al.*, 2003), με σχόλια για τη λειτουργία και τα χαρακτηριστικά κάθε πολυπεπτιδικής αλυσίδας. Παράλληλα, διαρκώς προκύπτει μεγάλος αριθμός νέων ακολουθιών, κυρίως από μετάφραση των κωδικών περιοχών που εντοπίζονται από τα προγράμματα προσδιορισμού της πρωτοταγούς δομής γονιδιωμάτων. Αξιοσημείωτο είναι το γεγονός, ότι για την ταυτοποίηση του τεράστιου όγκου των γονιδίων και των προϊόντων τους, δηλαδή την εύρεση της λειτουργίας τους, επιστρατεύεται πληθώρα υπολογιστικών μεθόδων. Αυτές οι μέθοδοι βασίζονται κυρίως στην αναζήτηση ομοιοτήτων σε επίπεδο νουκλεοτιδικών ή αμινοξικών ακολουθιών. Παρά τη μεγάλη πρόοδο που έχει σημειωθεί στην ταχύτητα και την απόδοση των χρησιμοποιούμενων αλγορίθμων, για ποσοστό ακολουθιών πλήρως προσδιορισμένων γονιδιωμάτων που κυμαίνεται από 20% έως και πάνω από 40%, δεν είναι δυνατόν να προσδιοριστεί κάποια πιθανή λειτουργία. Για παράδειγμα, κατά την ανάλυση του γονιδιώματος του υπερθερμοφιλικού

αρχαιοβακτηρίου *Aeropyrum pernix (K1)*, το αυτοματοποιημένο σύστημα *GeneQuiz* (Andrade *et al.*, 1999; Iliopoulos *et al.*, 2001) απέδωσε λειτουργία (με βάση την ομοιότητα σε επίπεδο αμινοξικών ακολουθιών) στο 39% των πρωτεϊνών (Εικόνα 2).

Σε πλήρη αντίθεση με την πληθώρα των δεδομένων που αφορούν στις πρωτεϊνικές ακολουθίες, τα δεδομένα σχετικά με τις τρισδιάστατες δομές τους είναι σημαντικά λιγότερα. Η τρισδιάστατη δομή περίπου 20,000 πρωτεϊνών σε ατομική ή περίπου ατομική διακριτικότητα, έχει αποκαλυφθεί με μεθόδους κρυσταλλογραφίας ακτίνων-Χ μονοκρυστάλλων ή φασματοσκοπία Πυρηνικού Μαγνητικού Συντονισμού, NMR, (Berman *et al.*, 2000). Στην πραγματικότητα, αν λάβει κανείς υπόψη του ότι πολλές από τις δομές αυτές αναφέρονται σε ίδια πρωτεϊνικά μόρια (π.χ. με διαφορετικούς υποκαταστάτες ή με σημειακές μεταλλαγές), ο αριθμός των αντιπροσωπευτικών δομών είναι αρκετά μικρότερος. Η δυσκολία στον προσδιορισμό της δομής των πρωτεϊνών οφείλεται στο ότι οι πειραματικές μέθοδοι που χρησιμοποιούνται είναι δαπανηρές, επίπονες και χρονοβόρες. Επίσης απαιτούν τη χρήση μονοκρυστάλλων, που για αρκετές πρωτεΐνες δεν δημιουργούνται εύκολα.

Παρά τη μεγάλη πρόοδο στις αντίστοιχες πειραματικές μεθοδολογίες, η λειτουργική και δομική μελέτη των πρωτεϊνών σε κλίμακα αντίστοιχη με αυτή του προσδιορισμού των ακολουθιών τους δεν είναι ακόμη εφικτή. Είναι προφανές, λοιπόν, ότι, προκειμένου να συμπληρωθεί το υπάρχον κενό, η ανάγκη απόκτησης πληροφοριών για το δίπλωμα και τη λειτουργία των πρωτεϊνών με εναλλακτικές μεθόδους είναι επιτακτική.



Εικόνα 2: Απόδοση Λειτουργίας για το πλήρες γονιδίωμα του A. pernix (K1) από το αυτοματοποιημένο σύστημα GeneQuiz.

Τα δεδομένα προέρχονται από το δικτυακό τόπο του GeneQuiz στο Ευρωπαϊκό Ινστιτούτο Βιοπληροφορικής (European Bioinformatics Institute, EBI:

http://jura.ebi.ac.uk:8765/ext-genequiz/genomes/ap0004/index.html).

1.2 Βάσεις δεδομένων πρωτεϊνικών ακολουθιών

Ο τεράστιος όγκος δεδομένων ακολουθιών οργανώνεται συστηματικά σε Υπολογιστικά Συστήματα διαχείρισης δεδομένων, με σκοπό την εύκολη και συστηματοποιημένη πρόσβαση στα δεδομένα. Το μεγαλύτερο ποσοστό από τις πρωτεϊνικές ακολουθίες που είναι σήμερα διαθέσιμες έχει προκύψει από μετάφραση νουκλεοτιδικών ακολουθιών, οι οποίες ως επί το πλείστον κατατίθενται στις δημόσιες βάσεις δεδομένων νουκλεοτιδικών ακολουθιών (GenBank-EMBL-DDBJ). Κατά αντιστοιχία, οι πρωτεϊνικές ακολουθίες καταχωρούνται σε εξειδικευμένες βάσεις δεδομένων. Σημαντικότερες από αυτές θεωρούνται η SwissProt και η PIR.

Για κάθε εγγραφή τους, η οποία αντιστοιχεί σε μια πρωτεϊνική ακολουθία, πέρα από την κύρια πληροφορία της αμινοξικής

ακολουθίας, οι βάσεις δεδομένων πρωτεϊνικών ακολουθιών αποκτούν προστιθέμενη αξία με την ύπαρξη σχολίων σε διάφορα εξειδικευμένα πεδία κάθε εγγραφής. Τα σχόλια αυτά, τα οποία αναφέρονται σε χαρακτηριστικά που αντιστοιχούν στο λειτουργικό μόριο το οποίο αντιπροσωπεύει η κάθε ακολουθία, μπορεί να αφορούν: στη δομή της πρωτεΐνης, όταν αυτή είναι διαθέσιμη (π.χ. στοιχεία δευτεροταγούς ή υπερδευτεροταγούς δομής), στη φυσιολογική της λειτουργία (π.χ. ενζυμική δράση, ενεργό κέντρο, ενδοκυτταρικός εντοπισμός), σε γνωστά συντηρημένα μοτίβα ακολουθίας που χαρακτηρίζουν πρωτεϊνικές οικογένειες, αλλά και γενικότερες πληροφορίες, οι οποίες είναι δυνατόν να σχετίζονται με δυσλειτουργίες οφειλόμενες σε μεταλλάξεις, προϊόντα εναλλακτικής συρραφής (σε ευκαρυωτικούς οργανισμούς), μεταμεταφραστικές τροποποιήσεις και άλλες πολλές.

Η ραγδαία ανάπτυξη της Επιστήμης της Πληροφορικής και η ώθηση που έχει δοθεί τα τελευταία χρόνια στις Διαδικτυακές τεχνολογίες και εφαρμογές, βοηθούν στην ανάπτυξη εύχρηστων συστημάτων ανάκτησης πληροφορίας μέσω του Διαδικτύου. Τέτοια εξειδικευμένα συστήματα έχουν αναπτυχθεί και για την πρόσβαση στις βάσεις δεδομένων ακολουθιών. Χαρακτηριστικό παράδειγμα αποτελεί το σύστημα ανάκτησης ακολουθιών SRS (Sequence Retrieval System, http://srs.ebi.ac.uk, Etzold *et al.*, 1996). Το σύστημα SRS αρχικά σχεδιάστηκε με σκοπό να παρέχει πρόσβαση στις εγγραφές της SwissProt και διασυνδέσεις σε άλλες βάσεις πρωτεϊνικών δεδομένων με βάση τα στοιχεία που υπήρχαν στις εγγραφές αυτές. Σήμερα έχει εξελιχθεί σε ένα δυναμικό εργαλείο ανάκτησης πληροφοριών με τη δυνατότητα να πραγματοποιεί ταυτόχρονα επερωτήσεις (queries) σε δεκάδες διαφορετικές βάσεις δεδομένων με εντελώς διαφορετικό περιεχόμενο.

Επιπλέον, η ευρεία χρήση της υπηρεσίας του Παγκόσμιου Ιστού, καθιστά εξαιρετικά σημαντική, αλλά και ευέλικτη, πρακτική τη διασύνδεση μεταξύ ετερογενών βάσεων δεδομένων μέσω απλών

συνδέσμων (links). Συγκεκριμένα, η SwissProt παρέχει συνδέσμους μεταξύ των άλλων στην κύρια βάση δομικών πρωτεϊνικών δεδομένων (PDB), σε βάσεις με χαρακτηριστικά πρωτεϊνικά μοτίβα ακολουθιών (πιθανοθεωρητικά ή μη), στις βάσεις νουκλεοτιδικών δεδομένων από τις οποίες συνήθως προέρχονται οι εγγραφές της, καθώς και σε βιβλιογραφικές αναφορές.
Κατά αυτόν τον τρόπο, όλη η πληροφορία οργανώνεται με ένα δυναμικό τρόπο και οι δυνατότητες που παρέχονται στην ανάκτηση δεδομένων περιορίζονται, θεωρητικά, μόνο από την ταχύτητα πρόσβασης στο Διαδίκτυο.

1.3 Ποιότητα του σχολιασμού των εγγραφών στις βάσεις δεδομένων (πρωτεϊνικών) ακολουθιών

Οι διαρκώς αυξανόμενες πληροφορίες σχετικά με τις προσδιορισμένες πρωτεϊνικές ακολουθίες καταγράφονται καθημερινά στις δημόσιες βάσεις δεδομένων. Η προσθήκη των πεδίων, πέραν αυτού της ακολουθίας, σε μια εγγραφή των βάσεων δεδομένων πρωτεϊνικών ακολουθιών είναι μια διαδικασία η οποία πραγματοποιείται κάτω τον έλεγχο της ομάδας των φροντιστών της βάσης. Τα πεδία αυτά είναι δυνατόν να περιέχουν στοιχεία σχετικά με τη δομή και τη λειτουργία κάθε πρωτεΐνης, με διαφορετική λεπτομέρεια, ανάλογα με την περίπτωση.

Ορισμένες από τις διαδικασίες πραγματοποιούνται με πλήρως αυτοματοποιημένο τρόπο (π.χ. εύρεση χαρακτηριστικών μοτίβων στην ακολουθία), ενώ άλλες απαιτούν συνδυασμό αυτοματοποιημένων μεθόδων με την ειδική βιολογική γνώση των φροντιστών. Ουσιαστικά, η ποιότητα και η αξία των εγγραφών μιας βάσης δεδομένων βιολογικών ακολουθιών χαρακτηρίζεται από την ποιότητα του σχολιασμού τους.

Η μεγάλη αύξηση του πλήθους των γνωστών ακολουθιών δεν ήταν δυνατόν να ακολουθηθεί από αντίστοιχη αύξηση των αξιόπιστων πειραματικών δεδομένων σχετικά με την πρωτεϊνική δομή και λειτουργία. Επομένως, για μεγάλο αριθμό ακολουθιών, οι επιπλέον πληροφορίες που καταγράφονται στις εγγραφές των βάσεων δεδομένων συνάγονται μετά από τον εντοπισμό ομόλογων τους ακολουθιών, μετά από αναζητήσεις ομοιότητας. Στην περίπτωση που οι ομοιότητες αυτές είναι οριακές, υπάρχει ο κίνδυνος να παρεισφρήσουν στα σχόλια των εγγραφών των βάσεων δεδομένων στοιχεία τα οποία δεν είναι ακριβή (Bork and Koonin, 1998). Τα σφάλματα είναι δυνατόν να εμφανιστούν με διάφορες μορφές, οι πιο συνηθισμένες εκ των οποίων είναι (Galperin and Koonin, 1998; Tsoka *et al.*, 1999; Iliopoulos *et al.*, 2003):

1. Με εντελώς ***λανθασμένες ταυτοποιήσεις*** ακολουθιών, όπως για παράδειγμα σε περιπτώσεις που οι ομοιότητες εμφανίζονται σε περιοχές άσχετες με τη λειτουργία των πρωτεϊνών

2. Με "***αισιόδοξες***" προγνώσεις, δηλαδή σε περιπτώσεις που η ομοιότητα δικαιολογεί ένα γενικό δομικό ή λειτουργικό χαρακτηριστικό (π.χ. μια ακολουθία στην οποία ανιχνεύεται το χαρακτηριστικό μοτίβο δακτύλων ψευδαργύρου, το οποίο προσδένεται στο DNA δεν είναι βέβαιο ότι λειτουργεί ως μεταγραφικός παράγοντας)

3. Με "***συντηρητικές***" προγνώσεις (αντίθετο του 2)

4. Με την εισαγωγή "πληροφοριών" οι οποίες προέρχονται από προγνωστικές μεθόδους ***χαμηλής*** ή αδιευκρίνιστης ***αξιοπιστίας***

5. Από τη (λανθασμένη) ταυτοποίηση με βάση κάποια ομοιότητα με μια λανθασμένα ταυτοποιημένη ακολουθία

6. Ακόμη και από απλά ***τυπογραφικά λάθη***!!

Το πρόβλημα αυτό, οδηγεί σε ταχεία διάδοση σφαλμάτων στις βάσεις δεδομένων, η οποία περιγράφηκε με τον όρο "***database-explosion***" (Bhatia *et al.*, 1997). Σε πρόσφατη μελέτη (Gilks *et al.*, 2002), δημιουργήθηκε ένα δυναμικό στοχαστικό μοντέλο που, προσομοιάζοντας το φαινόμενο με τη διάχυση ενός υγρού μέσα από μια πορώδη επιφάνεια, προσπαθεί να περιγράψει την αλυσιδωτή διάδοση των σφαλμάτων. Σε αυτή την εργασία αποδεικνύεται ότι το η 'διάχυση' των σφαλμάτων σε σύντομο χρονικό διάστημα οδηγεί σε πλήρη υποβάθμιση της ποιότητας των διαθέσιμων δεδομένων.
Απαιτείται, λοιπόν, η προσεκτική χρήση των πληροφοριών που ανακτώνται από τις βάσεις δεδομένων. Όταν η χρήση υπολογιστικών μεθόδων είναι επιβεβλημένη, οφείλουμε να γνωρίζουμε εκ των προτέρων την αξιοπιστία τους και να ελέγχουμε τα αποτελέσματά τους με κριτική άποψη, όπως άλλοτε συμβαίνει και με κάθε πειραματική διαδικασία.

2 Αναζήτηση ομοιοτήτων σε βάσεις δεδομένων ακολουθιών

2.1 Το Πρόβλημα ...

Ένα σημαντικό πρόβλημα που καλούμαστε συχνά να λύσουμε είναι όταν έχουμε στη διάθεσή μας μια ακολουθία ενός βιολογικού μακρομορίου για το οποίο δε διαθέτουμε άλλες πληροφορίες. Στην περίπτωση αυτή, επιθυμούμε να εντοπίσουμε μέσα στις βάσεις δεδομένων άλλες ακολουθίες, οι οποίες λόγω της ομοιότητάς τους με την εξεταζόμενη ακολουθία να μπορούμε με σχετική ασφάλεια να θεωρήσουμε ότι είναι ομόλογές της. Στην ευτυχή αυτή περίπτωση μπορούμε επαγωγικά να αποφανθούμε για την πιθανή δομή-λειτουργία του "άγνωστου" μορίου.

Οι αλγόριθμοι Δυναμικού Προγραμματισμού που έχουν ήδη περιγραφεί, υπολογίζουν με ακριβή τρόπο τις βέλτιστες (μαθηματικά) στοιχίσεις μεταξύ δύο ακολουθιών. Παρότι η χρονική πολυπλοκότητά τους (O(MxN), όπου M, N τα μήκη των συγκρινόμενων ακολουθιών) είναι ικανοποιητική σε περιπτώσεις που έχουμε ακολουθίες συνηθισμένου μεγέθους, είναι μη πρακτική στην περίπτωση που επιθυμούμε να τις χρησιμοποιήσουμε σε σειριακή αναζήτηση έναντι μιας (μεγάλης) βάσης δεδομένων ακολουθιών. Για την αντιμετώπιση αυτού του προβλήματος άλλα και ισοδύναμων παρεμφερών προβλημάτων[1], υπήρξε η ανάγκη ανάπτυξης εναλλακτικής μεθοδολογίας. Πραγματοποιώντας ορισμένες συμβάσεις και "θυσιάζοντας" ένα μέρος της ακρίβειας για την ελαχιστοποίηση των απαιτούμενων υπολογιστικών πόρων, δημιουργήθηκαν ευριστικοί αλγόριθμοι, οι οποίοι μέσα σε σημαντικά μικρότερο χρονικό διάστημα καταφέρνουν να συγκρίνουν μια ακολουθία με μια ολόκληρη βάση δεδομένων όπως η GenBank μέσα σε λογικό χρόνο.

Στα επόμενα θα παρουσιαστούν τα βασικά χαρακτηριστικά αυτών των μεθόδων αλλά και συμπληρωματική μεθοδολογία που αναπτύχθηκε για τη στατιστική εκτίμηση των αποτελεσμάτων της σύγκρισης, καθώς και για την

[1] όπως η σύγκριση μιας μικρής με μια πολύ μεγάλη ακολουθία ή δύο πολύ μεγάλων ακολουθιών (π.χ. των ακολουθιών δύο ολόκληρων χρωμοσωμάτων)

εξάλειψη κάποιων προβλημάτων που αντιμετωπίζονται σε αναζητήσεις αυτού του είδους.

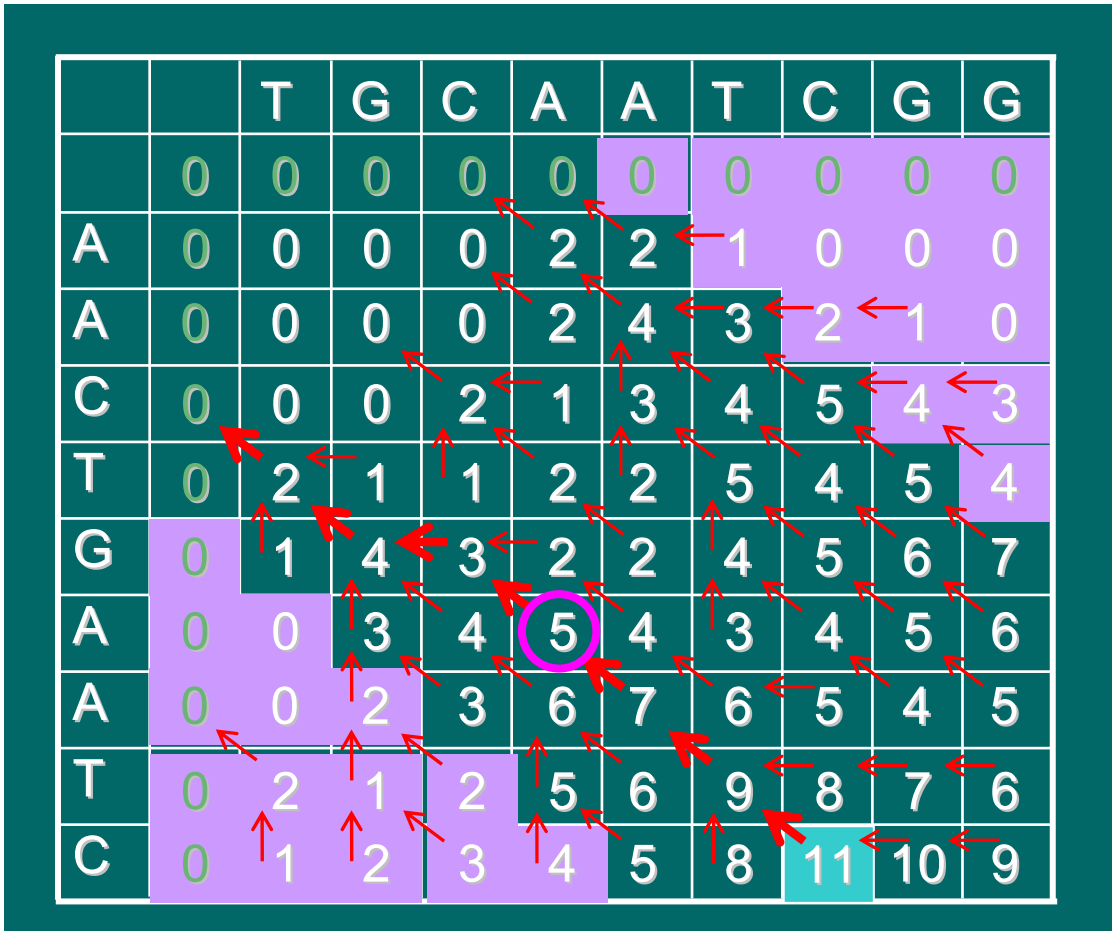
2.2 Ευριστικές Μέθοδοι

Η λύση στα προβλήματα που αναφέρθηκαν θα πρέπει να αναζητηθεί έτσι ώστε η επίτευξη μιας αποδεκτής ταχύτητας στην αναζήτηση ομοιοτήτων (αλλά και η ταυτόχρονη ανάγκη περιορισμού της απαιτούμενης μνήμης) να μην επηρεάζει σε μεγάλο βαθμό την «ποιότητα» των αποτελεσμάτων. Με άλλα λόγια, οι λύσεις στις οποίες θα καταλήγαμε θα πρέπει να έχουν τα παρακάτω χαρακτηριστικά:

1. Να μη διαφέρουν σημαντικά από τις «ακριβείς» (μαθηματικά βέλτιστες) λύσεις των μεθόδων δυναμικού προγραμματισμού.
2. Να μην αποκλείουν βιολογικά πιθανές λύσεις.

Μέθοδος «Κοπής Γωνιών»[2]: Με βάση αυτά τα χαρακτηριστικά, προτάθηκε μια πρώτη ευριστική μέθοδος, η οποία βασίζεται στο Δυναμικό Προγραμματισμό, αυτή που αναφέρεται συχνά ως μέθοδος «Κοπής Γωνιών» (Sankoff and Kruskal, 1983). Αυτή είναι ίσως η απλούστερη «βελτίωση» που θα μπορούσε να σκεφτεί κανείς. Η ιδέα είναι πραγματικά πολύ έξυπνη και απλή, περιορίζοντας στην ουσία τους υπολογισμούς των πινάκων Δυναμικού Προγραμματισμού σε μια «ζώνη» γύρω από τη διαγώνιο του πίνακα (*Εικόνα 3*). Όπως γίνεται εμφανές, η επιλογή του πλάτους της ζώνης στην οποία θα εκτελεστούν οι υπολογισμοί επηρεάζει άμεσα την εξοικονόμηση πόρων κατά τη στοίχιση ακολουθιών.

[2] Η μέθοδος αυτή στη γενικευμένη μορφή της ονομάζεται «kBand alignment», όπου οι απαραίτητοι υπολογισμοί πραγματοποιούνται γύρω από μια «ζώνη» πλάτους k γύρω από δοθείσα διαγώνιο. Η απλή τροποποίηση του αλγορίθμου μπορεί να βρεθεί στο Setubal, J. C. and J. Meidanis (1997). Introduction to computational molecular biology. Boston, PWS Pub..



Εικόνα 3: Πίνακας Δυναμικού Προγραμματισμού για τη μέθοδο «Κοπής Γονιών». Οι τιμές όλων των κελιών έχουν τοποθετηθεί στα κελιά (δείτε σημειώσεις προηγούμενης διάλεξης). Με τη μέθοδο αυτή, υποθέτουμε ότι ένα «καλό μονοπάτι» (δηλ. μια καλή στοίχιση) δεν αναμένουμε να διέρχεται από τις σκιασμένες περιοχές του πίνακα (πάνω δεξιά και κάτω αριστερή γωνία). Με τα παχιά βέλη υποδηλώνεται η βέλτιστη διαδρομή, όπως υπολογίζεται με τον κλασικό Δυναμικό Προγραμματισμό. Παρατηρήστε ότι από τα 100 (=10*10) κελιά του πίνακα απαιτείται το γέμισμα μόνο των 70, κερδίζοντας έτσι 30% σε μνήμη (και χρόνο).

Βασικό μειονέκτημα της μεθόδου είναι ότι λειτουργεί ικανοποιητικά σε περιπτώσεις που γνωρίζουμε ότι οι ακολουθίες που μελετάμε εμφανίζουν πραγματικά υψηλή ομοιότητα, χωρίς να είναι απαραίτητη η εισαγωγή μεγάλου πλήθους κενών[3]. Επίσης, η μέθοδος μπορεί να εφαρμοστεί (και στην πράξη

[3] Θυμηθείτε ότι η εισαγωγή κενών σε μια ακολουθία οδηγεί τη στοίχιση σε διαφορετική διαγώνιο. Επομένως η εισαγωγή μεγάλου πλήθους κενών σε μια από τις δύο (ή και στις 2 ακολουθίες) μπορεί να οδηγήσει το «μονοπάτι» της βέλτιστης στοίχισης έξω από τη ζώνη στην οποία περιορίζουμε την αναζήτησή μας, δίνοντας έτσι στοιχίσεις πολύ διαφορετικές από τη βέλτιστη.

αυτό συμβαίνει) σε συνδυασμό με ταχύτερες μεθόδους (δείτε τα επόμενα), οι οποίες σε μικρό χρόνο εντοπίζουν τις πιθανές «ομόλογες» περιοχές των ακολουθιών (άρα και διαγωνίους στον πίνακα), και στη συνέχεια εκτελείται κατάλληλη παραλλαγή του αλγορίθμου Δυναμικού Προγραμματισμού (δείτε υποσημείωση 2).

Μέθοδοι με τη δημιουργία «ευρετηρίων» (Hashing):

Ο Αλγόριθμος FASTA

Ο πρώτος ευριστικός αλγόριθμος για την ταχύτατη σύγκριση ακολουθιών βιολογικών μακρομορίων ήταν ο αλγόριθμος FASTA (Pearson and Lipman, 1988; Pearson, 1990). Η διαδικασία σύγκρισης-στοίχισης που ακολουθείται πραγματοποιείται σε ξεχωριστά βήματα, ξεκινώντας με βάση το χαρακτηριστικό των βέλτιστων στοιχίσεων να περιέχουν μικρές περιοχές χωρίς κενά (*k-tuples* ή *words*, κ-πλέτες/πολα-πλέτες στα Ελληνικά;;) στις οποίες οι δύο ακολουθίες εμφανίζουν πλήρη ταύτιση.
Τα βήματα αυτά είναι:

1. Η βασική ιδέα έγκειται στη δημιουργία ενός ευρετηρίου με τις θέσεις όλων των k-tuples (τυπικό μήκος για αμινοξικές ακολουθίες 1 ή 2) που υπάρχουν και στις δύο ακολουθίες (*Εικόνα 4*, αριστερά).
2. Από τη διαφορά των θέσεών τους στις δύο ακολουθίες εντοπίζεται η διαγώνιος στην οποία βρίσκονται (*Εικόνα 4*, δεξιά), οπότε στο επόμενο βήμα εντοπίζονται οι διαγώνιες με τα περισσότερα k-tuples.
3. Ακολούθως, αυτές οι περιοχές ταύτισης συνενώνονται επιτρέποντας την εισαγωγή κενών με τον υπολογισμό της αντίστοιχης ποινής (*Εικόνα 5*), και
4. Τελικά πραγματοποιείται η διαδικασία πλήρους δυναμικού προγραμματισμού (με τον επιλεγμένο πίνακα αντικατάστασης),

περιορισμένου σε μια ταινία γύρω από τις συγκεκριμένες διαγωνίους (*Εικόνα 5*).

K-tuple (K=1)	position in SEQ1	position in SEQ2	offset SEQ1-SEQ2
A	1	8	-7
C	2	4	-2
D	-	2	X
G	3	5	-2
I	7	-	X
K	8	7	1
L	5	-	X
V	3	6	-3
W	-	1	X
Y	4	6	-2

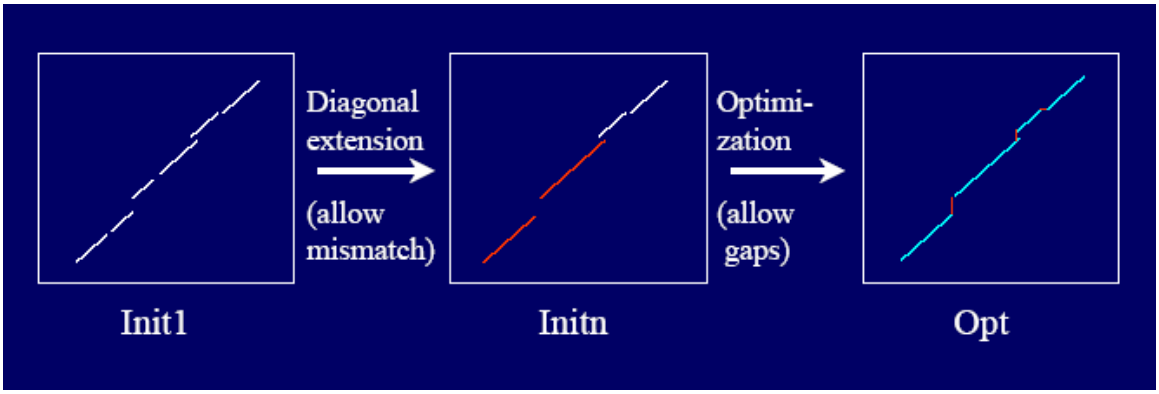
0	1	2	3	4	5	6	7	8	
-1		A	C	G	Y	L	V	I	K
-2	W								
-3	D								
-4	V								
-5	C								
-6	G								
-7	Y								
-8	K								
	A								

Εικόνα 4: Αριστερά - Κατασκευή ευρετηρίου για k-tuples, με k=1 για τις ακολουθίες SEQ1:ACGYLVIK και SEQ2:WDVCGYKA. Για απλότητα οι ακολουθίες του παραδείγματος περιέχουν από μία φορά κάθε k-tuple. Η διαφορά της θέσης μιας k-tuple στη μία ακολουθία με μιά ταυτόσημή της στην άλλη εκφράζει ένα μέτρο «μετατόπισης» της μιας ακολουθίας ως προς την άλλη για να στοιχιθούν μεταξύ τους τα συγκεκριμένα k-tuples.

Δεξιά - Όλες οι διαγώνιες του πίνακα προσδιορίζονται συμβολικά σε σχέση με την κυρία διαγώνιο με τους κόκκινους αριθμούς οι οποίοι φαίνονται στην 1η γραμμή και πρώτη στήλη αντίστοιχα. Οι αριθμοί αυτοί αντιστοιχούν στην κοινή διαφορά (offset) που έχουν οι δείκτες i, j οι οποίοι ορίζουν τα στοιχεία της συγκεκριμένης διαγωνίου. Προφανώς, η κύρια διαγώνιος (για κάθε κελί της οποίας ισχύει i=j) αντιστοιχεί σε offset 0.

Στην πράξη οι 10 καλύτερες διαγώνιες εντοπίζονται με αυτόν τον τρόπο.

Σημείωση: Ο πίνακας του σχήματος αντιστοιχεί με Dot Matrix Plot, εύκολα μπορεί η ιδέα να αποτυπωθεί σε ένα πίνακα δυναμικού προγραμματισμού.



Εικόνα 5: Σχηματική αναπαράσταση των σταδίων για τη σύκριση ακολουθιών με τη μέθοδο FASTA. Μόνο οι «καλύτερες» διαγώνιοι (όπως προέκυψαν μετά τη δημιουργία ευρετηριών) καθορίζουν την περιοχή στην οποία θα υπολογιστεί τελικά η στοίχιση.

Στο πρώτο στάδιο ενοποιούνται περιοχές της ίδιας διαγωνίου επιτρέποντας την στοίχιση και ανόμοιων καταλοίπων (mismatch) αλλά ΟΧΙ την εισαγωγή κενών. Οι περιοχές που φαίνονται στο αριστερό διάγραμμα ονομάζονται Βέλτιστες Αρχικές Περιοχές (Best Initial Regions) και επιλέγονται με κριτήριο να έχουν βαθμολογία για τη στοίχισή τους (με τη χρήση μόνο ενός πίνακα αντικατάστασης) μεγαλύτερη από μια αρχική τιμή κατωφλίου Init1. Σε παραλλαγές της μεθόδου μπορεί να επιλεχθεί να διαλέγουμε συγκεκριμένο πλήθος από τις τοπικές στοιχίσεις με τα μεγαλύτερα scores ανεξάρτητα από το εάν αυτά ξεπερνούν την τιμή Init1.

Στο επόμενο στάδιο ενοποιούνται περιοχές οι οποίες δεν ανήκουν υποχρεωτικά στην ίδια διαγώνιο. Αυτό προφανώς επιβάλει την εισαγωγή κενών σε κάποια από τις ακολουθίες (πιθανότατα και στις 2). Για την εισαγωγή κενών αφαιρείται μια ποινή για την εισαγωγή κάθε κενού. Τώρα πλέον οι στοιχίσεις επεκτείνονται όσο η τιμή του score παραμένει μεγαλύτερη από μια δεύτερη τιμή κατωφλίου InitN.

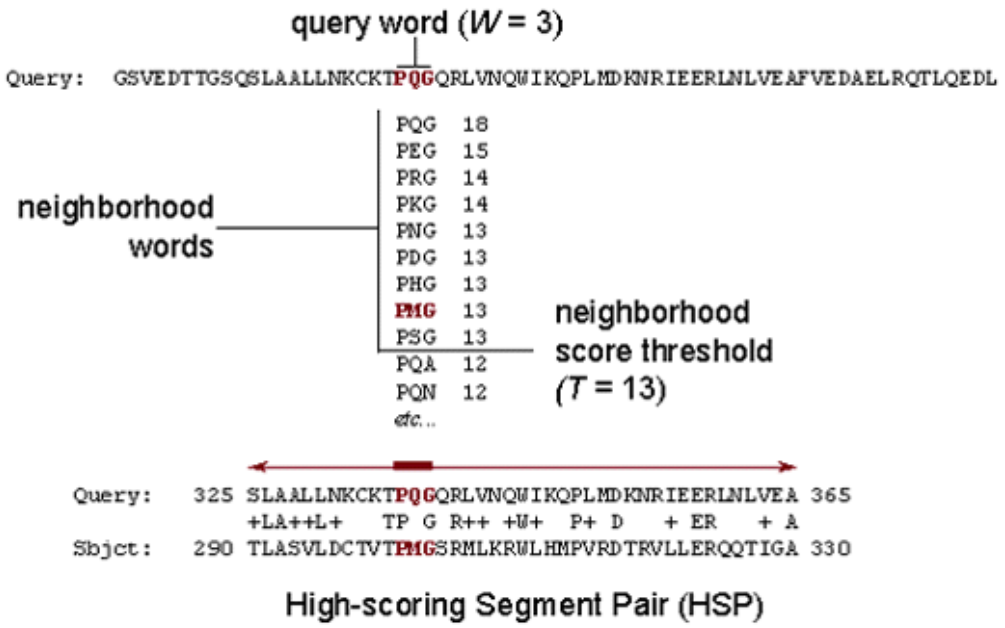
Στο τελευταίο στάδιο πραγματοποιείται ΠΛΗΡΗΣ τοπική στοίχιση με δυναμικό προγραμματισμό, περιορίζοντας όμως τους υπολογισμούς σε μια περιοχή γύρω από την προσεγγιστική στοίχιση που έχουμε από το προηγούμενο βήμα, η οποία δίνει το λεγόμενο Opt score της ευριστικής στοίχισης των δύο ακολουθιών.

Ο Αλγόριθμος BLAST

Ο αλγόριθμος BLAST (Altschul *et al.*, 1990; Altschul *et al.*, 1997) βασίζεται σε ιδέα παρόμοια με αυτή του FASTA. Η βασική θεώρηση είναι ότι οι βέλτιστες στοιχίσεις περιέχουν περιοχές (*λέξεις*, *words*), όπου οι βαθμολογίες (scores) μεταξύ των δύο ακολουθιών (με κάποιον πίνακα αντικατάστασης) είναι υψηλότερες από μια τιμή κατωφλίου *T (γειτονικές λέξεις*, *neighborhood words)*. Το προκαθορισμένο μήκος των *λέξεων* είναι *W*=3 στην περίπτωση των αμινοξικών ακολουθιών.

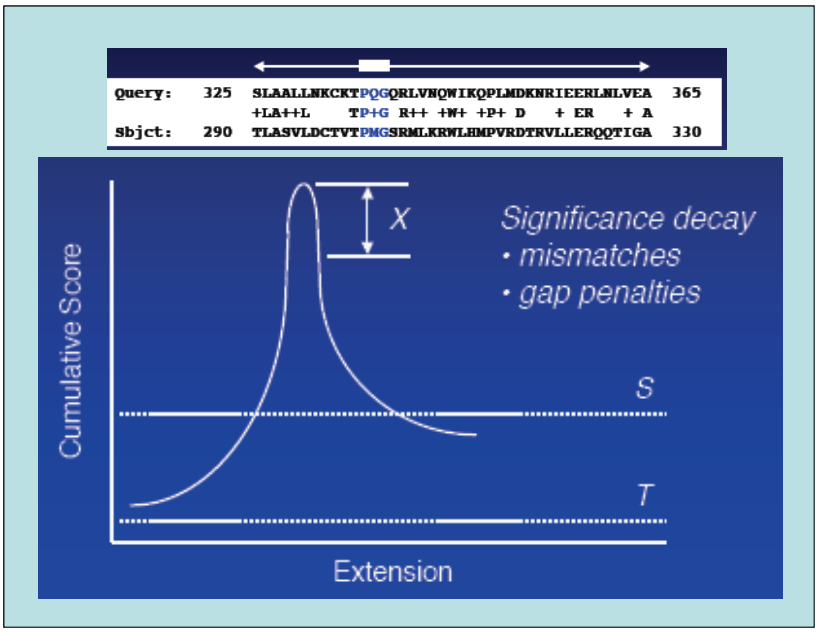
Η διαδικασία της σύγκρισης ξεκινά με την κατασκευή ενός καταλόγου όλων των *λέξεων* που θα ταίριαζαν με κάποια *λέξη* της άγνωστης ακολουθίας ξεπερνώντας την τιμή κατωφλίου (προκαθορισμένη τιμή για πρωτεϊνικές ακολουθίες T=13, *Εικόνα 6*).

Στη συνέχεια, ο αλγόριθμος αναζητά αυτές τις λέξεις στις ακολουθίες της βάσης δεδομένων και κάθε φορά που εντοπίζει κάποια ξεκινάει μια διαδικασία επέκτασης του 'ευρήματος' προς τις δύο κατευθύνσεις, όσο η βαθμολογία συνεχίζει και αυξάνει. Οι περιοχές μέγιστης βαθμολογίας που εντοπίζονται σε αυτό το στάδιο είναι οι υποψήφιες περιοχές ομοιότητας (*HSPs*, *high scoring pairs*). Από όλα τα HSPs αναφέρονται στα αποτελέσματα εκείνες οι περιοχές στις οποίες η βαθμολογία υπερβαίνει μια δεύτερη τιμή κατωφλίου *S (Εικόνα 7)*. Τελικά, επιλέγονται να αναφερθούν εκείνες μόνο οι τοπικές ομοιότητες οι οποίες εμφανίζουν υψηλή στατιστική σημαντικότητα, ο προσδιορισμός της οποίας περιγράφεται στην επόμενη ενότητα.



Εικόνα 6: Κατασκευή καταλόγου «λέξεων» που οδηγούν σε «καλές» στοιχίσεις.

Για κάθε λέξη (αντίστοιχο του k-tuple) της ακολουθίας με την οποία πραγματοποιείται η αναζήτηση (query) προϋπολογίζονται εκείνες οι λέξεις οι οποίες στοιχιζόμενες έναντι αυτής θα έδιναν score (με βάση το επιλεγμένο σύστημα βαθμονόμησης) μεγαλύτερο ή ίσο από την προκαθορισμένη τιμή κατωφλίου T. Αυτές οι k-tuples όταν εντοπισθούν σε ακολουθίες της βάσης δεδομένων δίνουν πιθανά σημεία έναρξης μιας καλής τοπικής στοίχισης. Στην ουσία, η διαδικασία αυτή είναι γενίκευση της διαδικασίας εντοπισμού «καλών» διαγωνίων του αλγορίθμου FASTA. Το πλεονέκτημα αυτής της μεθόδου είναι ότι επιτρέπονται στο αρχικό αυτό στάδιο αντιστοιχίσεις και μεταξύ τμημάτων των ακολουθιών που δεν εμφανίζουν απόλυτη ταύτιση.



Εικόνα 7: Επέκταση ενός HSP.

Στο συγκεκριμένο παράδειγμα επιδεικνύεται η κατασκευή καταλόγου «ευνοϊκών λέξεων» για την τριπλέτα PQG της query ακολουθίας του προηγούμενου παραδείγματος. Ένα ζεύγος περιοχών που στοιχιζόμενα δίνουν υψηλό score (HSP, High-scoring Segment Pair) κατασκευάζεται με την επέκταση της στοίχισης στην περιοχή εκατέρωθεν των «ευνοϊκών λέξεων» και όσο το αθροιστικό score που υπολογίζεται υπερβαίνει την προκαθορισμένη τιμή S.

Στην αρχική υλοποίηση του λογισμικού (Altschul *et al.*, 1990) παράγονταν στοιχίσεις χωρίς κενά, ενώ νεότερες εκδόσεις (Altschul and Gish, 1996; Altschul *et al.*, 1997) επιτρέπουν την εισαγωγή κενών. Για την εισαγωγή κενών εφαρμόζονται οι αρχές που ήδη αναφέρθηκαν και ο αλγόριθμος δέχεται ως παραμέτρους τιμές ποινής για την εισαγωγή και την επέκταση των κενών, οι οποίες επηρεάζουν σημαντικά τις παραγόμενες στοιχίσεις.

Προσδιορισμός της Στατιστικής σημαντικότητα των βαθμολογιών (scores)

Οι αλγόριθμοι στοίχισης ακολουθιών πάντα θα παράγουν μια (βέλτιστη) στοίχιση, ανεξάρτητα από το εάν οι ακολουθίες έχουν πραγματική βιολογική σχέση ή αν μια στοίχιση με την ίδια (ή μεγαλύτερη) βαθμολογία θα μπορούσε να προκύψει κατά τύχη. Μια μέθοδος διαχωρισμού των στοιχίσεων εκείνων που, πιθανόν, να έχουν βιολογικό (δομικό ή / και λειτουργικό) αντίκρισμα είναι με την εκτίμηση της στατιστικής σημαντικότητας της βαθμολογίας που αντιστοιχείται στη δεδομένη στοίχιση.
Ένας απλός τρόπος για την αντιμετώπιση αυτού του ζητήματος με κλασική στατιστική ανάλυση, είναι ο έλεγχος δύο αντικρουόμενων υποθέσεων, δεδομένης της βαθμολογίας *s* μιας στοίχισης:

- **Υπόθεση 1:** Οι δύο ακολουθίες σχετίζονται μεταξύ τους

- **Υπόθεση 2 (μηδενική υπόθεση):** Οι δύο ακολουθίες δεν σχετίζονται

Στις επόμενες παραγράφους παρουσιάζονται οι δύο κυριότερες μεθοδολογίες για τον προσδιορισμό της στατιστικής σημαντικότητας μιας στοίχισης με βάση το score που έχει υπολογιστεί. Ελπίζουμε ότι ένα στατιστικά σημαντικό αποτέλεσμα θα μα δώσει μεγαλύτερη ασφάλεια ώστε να καταλήξουμε σε βιολογικά συμπεράσματα τα οποία να έχουν κάποιο νόημα.

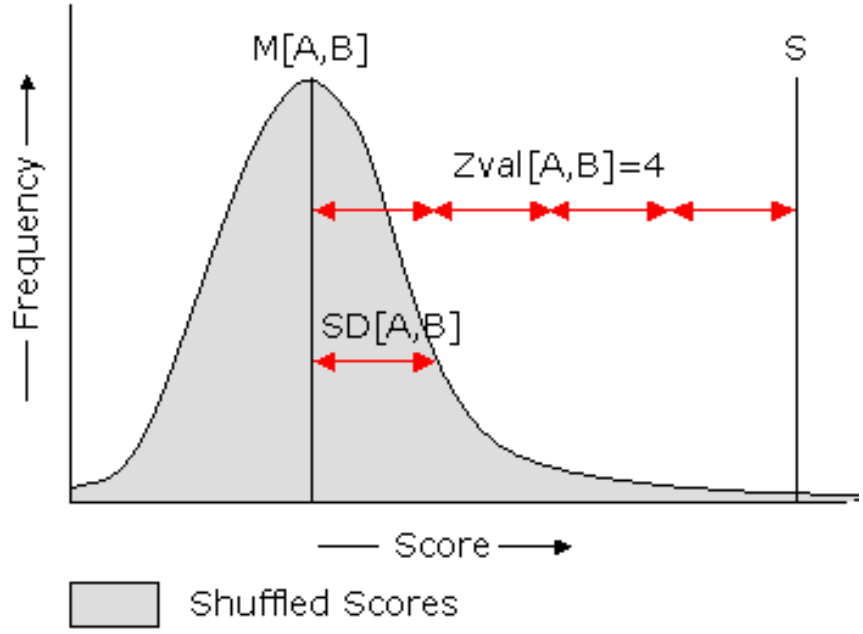
Σύγκριση με τυχαίες ακολουθίες

Η απλούστερη προσέγγιση που θα μπορούσαμε να ακολουθήσουμε μπορεί να βασιστεί στη σύγκριση του score το οποίο θέλουμε να αξιολογήσουμε με τα scores εκείνα που θα προέκυπταν από τη

στοίχιση τυχαίων ακολουθιών η (αμινοξική) σύσταση των οποίων αντικατοπτρίζει τη σύσταση των ακολουθιών της βάσης δεδομένων αλλά και της ακολουθίας με την οποία πραγματοποιείται η αναζήτηση.

Αυτή η διαδικασία μπορεί να πραγματοποιηθεί ως εξής:

1. «Ανακατεύοντας» τυχαία τα κατάλοιπα της ακολουθίας query, μπορούμε επαναληπτικά να συγκρίνουμε τις τυχαίες ακολουθίες με τη βάση δεδομένων. Με τον τρόπο αυτό, αποκτούμε ένα μεγάλο δείγμα από scores (*Εικόνα 8*) για το οποίο μπορούμε να υπολογίζουμε περιγραφικούς στατιστικούς δείκτες, όπως τη *μέση τιμή* (*M*) και την *τυπική απόκλιση (SD)*.



Εικόνα 8: Κατανομή των scores μετά από «ανακάτεμα» (shuffling) της ακολουθίας query.

Επιδεικνύεται η θέση ενός score που απέχει 4 τυπικές αποκλίσεις από το μέσο της κατανομής.

2. Με δεδομένο ότι οι «ανακατεμένες» ακολουθίες δεν περιμένουμε να έχουν καμία σχέση με την ακολουθία query (πέρα από την ίδια σύσταση σε κατάλοιπα), μπορούμε με αρκετή ασφάλεια να θεωρήσουμε ότι αυτά τα scores έχουν

προέλθει από ταίριασμα ασυσχέτιστων (στη γενική περίπτωση) ακολουθιών.

3. Το score της στοίχισης που μας ενδιαφέρει να αξιολογήσουμε συγκρίνεται σχετικά με τη θέση του ως προς την κατανομή των τυχαίων scores. Συγκεκριμένα, υπολογίζεται η σχετική απόκλιση του score από τη μέση τιμή με μέτρο την υπολογισμένη τυπική απόκλιση (Z=(score-M)/SD). Εάν βρεθεί σημαντικά μεγαλύτερο από τη μέση τιμή που υπολογίστηκε (Z>10) τότε το score της στοίχισης θεωρείται στατιστικά σημαντικό, γεγονός που συνεπάγεται μεγαλύτερη ασφάλεια στο να συμπεράνουμε ότι οι δύο ακολουθίες σχετίζονται βιολογικά. Σε μικρότερες τιμές του Z, οι υποθέσεις μας για τη συσχέτιση των ακολουθιών έχουν μεγαλύτερη αμφιβολία και απαιτούν μεγάλη προσοχή και προσεκτική μελέτη των στοιχίσεων. Κάποιοι γενικοί κανόνες παρουσιάζονται στον επόμενο πίνακα.

Τιμή Z	***Συσχέτιση Ακολουθιών***
Z < 3 SD	Απίθανη
3 SD < Z < 5 SD	Οριακά Πιθανή
5 SD < Z < 10 SD	Πιθανή
Z > 10 SD	Βέβαια (;;;)

Πίνακας 1: Κανόνες για την εκτίμηση της αξιοπιστίας μιας στοίχισης με βάση την κατανομή των scores τυχαίων στοιχίσεων.

Σημείωση: Με δεδομένο ότι η διαδικασία παραγωγής στοιχίσεων με (πολλές) τυχαίες ακολουθίες είναι σχετικά χρονοβόρα, μια παραλλαγή πραγματοποιείται χρησιμοποιώντας την κατανομή των scores που προκύπτουν από την ίδια την query ακολουθία έναντι της βάσης δεδομένων.

Η Κατανομή Ακραίων Τιμών

Η προηγούμενη προσέγγιση βασίζεται σιωπηρά στην παραδοχή ότι τα scores των τυχαίων στοιχίσεων ακολουθούν κανονική κατανομή. Για την περίπτωση των τοπικών στοιχίσεων χωρίς κενά έχει αποδειχθεί (Altschul and Gish, 1996) ότι, αν θεωρήσουμε την τυχαία μεταβλητή *S* για τα scores των στοιχίσεων, αυτή ακολουθεί ασυμπτωτικά την *κατανομή Gumbel* ή *κατανομή ακραίων τιμών* (*extreme value distribution-EVD,*Gumbel, 1958), με αθροιστική συνάρτηση κατανομής:

$$P(S \leq s) \simeq e^{-Kmne^{\lambda s}}$$

με παραμέτρους της κατανομής *K* και *λ* και *m*, *n* το μήκος της εξεταζόμενης ακολουθίας και της βάσης δεδομένων αντίστοιχα. Η πιθανότητα να προκύψει στοίχιση με βαθμολογία μεγαλύτερη ή ίση του *s* θα δίνεται:

$$P\text{-value} \equiv P(S \geq s) = 1 - e^{-Kmne^{\lambda s}}$$

Η διαδικασία μεγιστοποίησης της βαθμολογίας (score) που ακολουθούν οι μέθοδοι στοίχισης συνεπάγεται ότι το πλήθος των μη-σχετιζόμενων στοιχίσεων με βαθμολογίες μεγαλύτερες του s ακολουθεί προσεγγιστικά κατανομή Poisson με αναμενόμενη τιμή:

$$E(S \geq s) = Kmne^{-\lambda s}$$

Οι βαθμολογίες κανονικοποιούνται, ώστε να μην εξαρτώνται από το συγκεκριμένο σύστημα βαθμονόμησης (πίνακας αντικατάστασης, ποινές εισαγωγής κενών), οπότε προκύπτει ένα νέο *bit-score*

$s_{bit} = \frac{\lambda s - \ln K}{\ln 2} \Leftrightarrow s = \frac{s_{bit} \ln 2 + \ln K}{\lambda}$, οπότε το αναμενόμενο πλήθος στοιχίσεων με κανονικοποιημένη βαθμολογία μεγαλύτερη του S_{bit} προκύπτει με αντικατάσταση:

$$
\begin{aligned}
E\text{-value} &\equiv E(S_{bit} > s_{bit}) \\
&= Kmne^{-\lambda s} = Kmne^{-(s_{bit} \ln 2 + \ln K)} \\
&= Kmne^{-s_{bit} \ln 2} e^{-\ln K} = Kmne^{\ln 2^{-s_{bit}}} \frac{1}{K} \cdot \\
&= mn\ 2^{-s_{bit}}
\end{aligned}
$$

Προφανώς, όσο πλησιάζει η τιμή *E*-value στο μηδέν τόσο μεγαλύτερη η πιθανότητα πραγματικού συσχετισμού των ακολουθιών.

Προκειμένου για αμινοξικές ακολουθίες, σε πρακτικές εφαρμογές ομοιότητες με E-value < 10^{-10} θεωρούνται στατιστικά σημαντικές ώστε η ομοιότητα σε επίπεδο ακολουθίας να αντιστοιχίζεται με εξελικτική σχέση. Τιμές 10^{-6}>E-value>10^{-10} αντιστοιχούνται σε οριακές ομοιότητες, οι οποίες δεν παρέχουν ασφάλεια στην αντιστοίχιση κάποιας εξελικτικής σχέσης μεταξύ δύο ακολουθιών, ενώ στην περίπτωση που E-value>10^{-6} οι ομοιότητες βρίσκονται στο επίπεδο του "θορύβου" που προέρχεται από στοιχίσεις μη σχετιζόμενων ακολουθιών.

Σημειώνεται ότι τα κριτήρια που μόλις αναφέρθηκαν δεν πρέπει να ακολουθούνται τυφλά για την ταυτοποίηση της λειτουργίας πρωτεϊνικών ακολουθιών, διότι πρέπει να λαμβάνονται υπόψη και τα ειδικά χαρακτηριστικά κάθε πρωτεΐνης ή πρωτεϊνικής οικογένειας. Συχνά απαιτείται ο έλεγχος αυστηρότερων κριτηρίων ώστε να διασφαλιστεί η ορθή ταυτοποίηση ενός πρωτεϊνικού μορίου με βάση τον προσδιορισμό ομοιότητας με μια χαρακτηρισμένη πρωτεϊνική ακολουθία. Τα κριτήρια αυτά είναι:

- Η εμφάνιση της ομοιότητας σε επαρκές μήκος των ακολουθιών (π.χ. 80 κατάλοιπα, που αντιστοιχούν σε συνήθη μήκη δομικών και λειτουργικών περιοχών),

- Η ύπαρξη σημαντικού ποσοστού ταυτόσημων καταλοίπων (π.χ. ταυτότητα 30% συνεπάγεται, συνήθως, παρόμοιο δίπλωμα) στην περιοχή της στοίχισης, και

- Η ταυτόχρονη εμφάνιση χαρακτηριστικών δομικών-λειτουργικών μοτίβων

- Η ταύτιση καταλοίπων τα οποία είναι γνωστό πειραματικά ότι είναι σημαντικά για τη δομή-λειτουργία

Τελικά, η ποιότητα και η βιολογική σημασία των παραγόμενων στοιχίσεων, ακόμη και εκείνων με υψηλή στατιστική σημαντικότητα, μπορούν να αξιολογηθούν μόνο μετά από προσεκτική επόπτευση από έμπειρους ερευνητές. Μια περίπτωση στην οποία οι στοιχίσεις που προκύπτουν μπορεί να είναι παραπλανητικές εμφανίζεται όταν η ακολουθία με την οποία πραγματοποιούμε αναζήτηση περιέχει κάποια περιοχή χαμηλής πολυπλοκότητας (Δείτε επόμενη παράγραφο). Η σύσταση αυτών των περιοχών δεν είναι προϊόν του μοντέλου των τυχαίων ακολουθιών που χρησιμοποιείται για τον υπολογισμό της στατιστικής σημαντικότητας. Στην περίπτωση αυτή, αν αυτές οι περιοχές δεν φιλτραριστούν κατάλληλα, οι αλγόριθμοι αναζήτησης κατασκευάζουν παραπλανητικές στοιχίσεις με υψηλή στατιστική σημαντικότητα.

Περιοχές Ακολουθιών με Χαμηλή Πολυπλοκότητα

Οι υπολογιστικές μέθοδοι που βασίζονται στην αναζήτηση ομοιοτήτων, συγκρίνουν μια εξεταζόμενη πρωτεϊνική ακολουθία

έναντι μιας βάσης δεδομένων με σκοπό τον εντοπισμό ακολουθιών που βάσει της ομοιότητάς τους είναι δυνατόν να υποτεθεί ότι σχετίζονται εξελικτικά με αυτή. Όπως αναφέρθηκε παραπάνω, έχουν αναπτυχθεί ταχύτατες και ευαίσθητες μέθοδοι για το σκοπό αυτό (Altschul *et al.*, 1990; Pearson, 1990; Altschul *et al.*, 1997) και τα αντίστοιχα εργαλεία που διατίθενται αποτελούν, αναμφισβήτητα, τα εργαλεία Βιοπληροφορικής που χρησιμοποιούνται ευρύτερα. Κατά την αναζήτηση ομοιότητας σε βάσεις δεδομένων ακολουθιών, όταν ανιχνευτεί υψηλή ομοιότητα μεταξύ μίας 'άγνωστης' ακολουθίας και μιας καλά χαρακτηρισμένης εγγραφής της βάσης δεδομένων, είναι δυνατόν να επιτευχθεί αξιόπιστη πρόγνωση για χαρακτηριστικά της λειτουργίας και της δομής της άγνωστης πρωτεΐνης.

Παρόλα αυτά, ορισμένες ακολουθίες είναι δυνατόν να περιέχουν περιοχές με 'ασυνήθιστη' αμινοξική σύσταση, δείχνοντας τοπικά κατά μήκος της ακολουθίας προτίμηση στην εμφάνιση ενός ή περισσότερων καταλοίπων. Τέτοιες περιοχές, οι οποίες ονομάζονται περιοχές χαμηλής πολυπλοκότητας (Wootton, 1994), είναι δυνατόν να δημιουργήσουν προβλήματα στις αναζητήσεις ομοιότητας, όχι κατά την εκτέλεση του αντίστοιχου λογισμικού αλλά κατά την εκτίμηση των αποτελεσμάτων.

Για παράδειγμα, εάν επιχειρήσουμε μια αναζήτηση με τον αλγόριθμο BLAST έναντι της πρωτεϊνικής βάσης δεδομένων με άγνωστη ακολουθία η οποία αποτελείται μόνο από ένα τύπο αμινοξικού καταλοίπου (π.χ. ένα ομοπολυμερές αργινίνης) στα αποτελέσματα θα συμπεριλαμβάνονται όλες οι ακολουθίες της βάσης που περιέχουν τουλάχιστον μία περιοχή πλούσια σε αυτό το κατάλοιπο (Εικόνα 9). Προφανώς, στην ακραία αυτή περίπτωση, δεν υπάρχει κανένας λόγος να θεωρήσουμε ότι υπάρχει η παραμικρή εξελικτική ή λειτουργική σχέση ανάμεσα στην ακολουθία με την οποία τροφοδοτήσαμε το λογισμικό και στις ακολουθίες που εντόπισε ο αλγόριθμος.

```
Length = 810
Score = 91.3 bits (223), Expect = 2e-18
Identities=53/137(38%),Positives=70/137(50%),Gaps = 17/137 (12%)
Q: 1   RRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRR 60
       +RRR   R  RRRR R R R R R RR    R  +R +  R+RR + RRRR + +R R R
S: 132 QRRREHEREERRRRERERERERGRGRRDENERDPKREQEERQRREQERRRREQEQRERER 191

Q: 61  RRRR-------------RRRRRRRRRRRRRRRRRRRRRRRRRRRRRRR----RR 103
       R  R             RR + RR + RRRR + + RR R+RR  R      +R    RR
S: 192 RGERDEEDDENQRDPDWRREQERREQERRRREQEQERRERQRRGGRDDEDENQRDPDWRR 251

Q: 104 RRRRRRRRRRRRRRRR 120   'Q: QUERY SEQUENCE'  [poly-R polymer]
        ++RR + + RR R RR       'IDENTICAL/SIMILAR RESIDUES'
S: 252 EQKRREQEQERRERERR 268   'S: DATABASE TARGET' [gi|3808061]
```

Εικόνα 9: Αναζήτηση ομοιότητας με τον Αλγόριθμο BLAST για ένα ομοπολυμερές πολυαργινίνης (πολυ-R).

Η αναζήτηση έγινε με είσοδο μια ακολουθία πολυ-R 120 καταλοίπων με το πρόγραμμα blastp 2.0.8 και πίνακα αντικατάστασης τον BLOSUM62 έναντι της βάσης nrdb του NCBI (20 Απριλίου 1999). Περισσότερες από 80 ακολουθίες της βάσης αναφέρθηκε να περιέχουν μια περιοχή την οποία ο αλγόριθμος ταίριαζε με την εκφυλισμένη ακολουθία, με στατιστικό δείκτη e-value<10^{-11}. Ξεκάθαρα, οποιαδήποτε πρωτεϊνική ακολουθία με κάποια περιοχή πλούσια σε R θα δίνει υψηλή βαθμολογία κατά τη σύγκριση με αυτήν την τεχνητή εκφυλισμένη ακολουθία. Προφανώς, κατά την εκτέλεση που επιδεικνύεται το φίλτρο του BLAST είχε απενεργοποιηθεί.

Στην περίπτωση μιας πραγματικής αναζήτησης με ακολουθία η οποία περιέχει μία περιοχή χαμηλής πολυπλοκότητας, στα αποτελέσματα θα αναφέρονται 'ψευδώς θετικά' ευρήματα και, μάλιστα, αρκετά συχνά μέσα στα αποτελέσματα με καλούς στατιστικούς δείκτες σημαντικότητας. Αυτές οι περιπτώσεις δεν είναι πάντα εύκολο να ανιχνευθούν μέσα στον τεράστιο όγκο των αποτελεσμάτων. Ιδιαίτερα, δε, στις περιπτώσεις που αναζητούμε δυσδιάκριτες απομακρυσμένες ομολογίες οι περιοχές χαμηλής πολυπλοκότητας είναι δυνατόν να κυριαρχούν στον υπολογισμό της βαθμολογίας (score) της ομοιότητας και έτσι οι πραγματικά ομόλογες πρωτεΐνες να 'χάνονται' μέσα σε πλήθος μη σχετιζόμενων ακολουθιών.

Αρκετά συχνά, το πρόβλημα του εντοπισμού περιοχών χαμηλής πολυπλοκότητας συνδέεται με το επίσης σημαντικό πρόβλημα του εντοπισμού (όχι απόλυτα περιοδικών) επαναληπτικών μοτίβων σε

ακολουθίες. Μια τέτοια προσέγγιση απαιτεί χρήση διαφορετικών, μαθηματικών και υπολογιστικών μεθόδων, όπως για παράδειγμα μεθόδους συμπίεσης ψηφιακών δεδομένων (αντίστοιχες με αυτές που χρησιμοποιούνται για συμπίεση αρχείων δεδομένων σε ηλεκτρονικό υπολογιστή, Allison et al., 2000). Ορισμένες, μάλιστα φορές, ξεκινώντας από εντελώς διαφορετικό μαθηματικό φορμαλισμό, αναπτύχθηκαν μαθηματικά ισοδύναμες μέθοδοι (Μέθοδος Αλγοριθμικής Σημαντικότητας, Milosavljevic, 1999, Μέθοδος Ελαχίστου Μήκους Κωδικοποίησης, Wise, 2001).

Αλγόριθμοι φιλτραρίσματος (masking)

Το πρόβλημα αυτό εντοπίστηκε αρκετά νωρίς και γρήγορα αναπτύχθηκαν αλγόριθμοι (Claverie and States, 1993; Wootton and Federhen, 1993) για τον εντοπισμό περιοχών χαμηλής πολυπλοκότητας σε ακολουθίες πριν την εκτέλεση αναζητήσεων ομοιότητας. Στη συνέχεια, αυτές οι περιοχές φιλτράρονται με κατάλληλο τρόπο, με τη διαδικασία που ονομάζεται 'masking', ώστε να μην επηρεάζουν τα αποτελέσματα των προγραμμάτων αναζήτησης ομοιότητας.

Ο αλγόριθμος XNU εντοπίζει επαναλήψεις στην 'άγνωστη' ακολουθία πραγματοποιώντας σύγκριση της ακολουθίας με τον εαυτό της με ένα αλγόριθμο δυναμικού προγραμματισμού. Αντίθετα, ο αλγόριθμος SEG χρησιμοποιεί ένα μαθηματικό μέτρο για την πολυπλοκότητα σε ένα 'παράθυρο' της ακολουθίας με συγκεκριμένο μήκος. Το μέτρο αυτό βασίζεται στη συνδυαστική και τη θεωρία πληροφορίας και χρησιμοποιεί ένα εμπειρικό κατώφλι, κάτω από το οποίο θεωρείται ότι ένα τμήμα μιας ακολουθίας έχει περιεχόμενο μειωμένης πληροφορίας και θεωρείται περιοχή χαμηλής πολυπλοκότητας. Η πολυπλοκότητα *K* που οφείλεται στη σύσταση ενός τμήματος ακολουθίας μήκους *L* υπολογίζεται από τη σχέση:

$$K = \frac{1}{L}\log_a \frac{L!}{\prod_{i=1}^{N} r_i}$$

όπου N=4 ή 20 για νουκλεοτιδικές ή αμινοξικές ακολουθίες αντίστοιχα και r_i το πλήθος καταλοίπων τύπου i στο εξεταζόμενο τμήμα της ακολουθίας. Για δεδομένο μήκος παραθύρου L, προφανώς, η K εξαρτάται μόνο από το γινόμενο $\prod_{i=1}^{N} r_i$. Η βάση του λογαρίθμου στις αρχικές υλοποιήσεις ήταν ίση με N, ενώ σε νεότερες υλοποιήσεις, όπου $a = 2$, η πολυπλοκότητα μετράται σε μονάδες bits.

```
Query: 1   MPSTVAPIKGQDHFLNLVFPERVAAAYMSPLAQKYPKAALSIASLAGFLLGILKLITFPV 60
           MPSTVAPIKGQDHFLNLVFPERVAAAYMSPLAQKYPKAALSIASLAGFLLGILKLITFPV
Sbjct: 1   MPSTVAPIKGQDHFLNLVFPERVAAAYMSPLAQKYPKAALSIASLAGFLLGILKLITFPV 60

Query: 61  LCAAGLFVFPIRGLISCLFHKSFQGCSGYVXXXXXXXXXXXXXIVGIVSCITWAPGFIFP 120
           LCAAGLFVFPIRGLISCLFHKSFQGCSGYV             IVGIVSCITWAPGFIFP
Sbjct: 61  LCAAGLFVFPIRGLISCLFHKSFQGCSGYVLATFLSLFSLALTIVGIVSCITWAPGFIFP 120

Query: 121 MISVSIAFATVETCFQIYTHLFPALEHKPSSSLKIEIAAAKLPRSSSAPDLNYPSLPTQS 180
           MISVSIAFATVETCFQIYTHLFPALEHKPSSSLKIEIAAAKLPRSSSAPDLNYPSLPTQS
Sbjct: 121 MISVSIAFATVETCFQIYTHLFPALEHKPSSSLKIEIAAAKLPRSSSAPDLNYPSLPTQS 180

Query: 181 ASPSQRFSA 189
           ASPSQRFSA
Sbjct: 181 ASPSQRFSA 189
```

Εικόνα 10: Φιλτράρισμα (masking) με τον Αλγόριθμο SEG.

Εφαρμογή του SEG σε μια υποθετική πρωτεΐνη (gi\|3328394) από το γονιδίωμα της Chlamydia trachomatis. Η αναπαράσταση των αποτελεσμάτων έχει γίνει με τη μορφή στοίχισης του λογισμικού BLAST. Ο αλγόριθμος εντόπισε μια περιοχή χαμηλής πολυπλοκότητας (η υπογραμμισμένη περιοχή της ακολουθίας μεταξύ των καταλοίπων 91-103) και τα αντίστοιχα κατάλοιπα έχουν αντικατασταθεί με 'X'.

Παρά την εντελώς διαφορετική προσέγγιση των δύο μεθοδολογιών, τα αποτελέσματα τους είναι συγκρίσιμα. Ο εγγενής χειρισμός των εντοπιζόμενων περιοχών χαμηλής πολυπλοκότητας από το XNU και το SEG είναι ταυτόσημος (Εικόνα 10). Αυτές οι περιοχές

αντικαθίστανται με κατάλοιπα τύπου ‘X’ για αμινοξικές ακολουθίες (στις οποίες και θα εστιάσουμε) ή ‘N’ για νουκλεοτιδικές ακολουθίες. Ο χαρακτήρας ‘X’ χρησιμοποιείται συχνά για αμινοξικά κατάλοιπα άγνωστου τύπου τα οποία συμπεριφέρονται ουδέτερα κατά την αναζήτηση ομοιοτήτων χωρίς να συνεισφέρουν καθόλου στην τελική βαθμολογία της ομοιότητας.

Παρόλο, λοιπόν, που και οι δύο αλγόριθμοι χρησιμοποιούνται κατά κόρον και με πολύ μεγάλη επιτυχία, παρουσιάζουν και κάποια αξιοπρόσεκτα μειονεκτήματα. Οι μέθοδοι αυτές, από το σχεδιασμό τους, εντοπίζουν τις υποψήφιες για φιλτράρισμα περιοχές εστιάζοντας σε επικαλυπτόμενα παράθυρα κατά μήκος της ακολουθίας. Στη συνέχεια, γειτονικές περιοχές συνενώνονται μέχρι τον εντοπισμό της μέγιστης δυνατής περιοχής χαμηλής πολυπλοκότητας. Συνεπώς, ως αποτέλεσμα της διαδικασίας αυτής φιλτράρονται συνεχείς περιοχές των υπό εξέταση ακολουθιών, στις οποίες είναι πιθανόν να υπάρχουν κατάλοιπα που δε συνεισφέρουν στο φαινόμενο και είναι δυνατόν να συμμετέχουν σε κάποιο λειτουργικό ή δομικό μοτίβο που βρίσκεται εκατέρωθεν ή επικαλύπτεται με την περιοχή χαμηλής πολυπλοκότητας.

Αξίζει να σημειωθεί ότι ο αλγόριθμος SEG έχει ενσωματωθεί στο λογισμικό BLAST και κατά τη συνήθη χρήση του αλγορίθμου το φίλτρο είναι ενεργοποιημένο. Για την επίδειξη των αποτελεσμάτων στην Εικόνα 9 χρησιμοποιήσαμε το λογισμικό BLAST απενεργοποιώντας κατάλληλα το φίλτρο.

Ο αλγόριθμος CAST

Μια νέα μέθοδος αναπτύχθηκε (CAST, Promponas et al., 2000) για την ανίχνευση και το φιλτράρισμα των περιοχών χαμηλής πολυπλοκότητας. Η νέα μέθοδος ήταν επιθυμητό να έχει αντίστοιχη ευαισθησία με τις ήδη υπάρχουσες αλλά να επιτυγχάνει μεγαλύτερη

εξειδίκευση, ώστε να πραγματοποιείται επιλεκτικό φιλτράρισμα και, συνεπώς, μικρότερη απώλεια πληροφορίας πριν την πραγματοποίηση των αναζητήσεων ομοιότητας.

Θεμελίωση του αλγορίθμου

Οι βάσεις για την ανάπτυξη της μεθόδου CAST τέθηκαν με ένα απλό και περισσότερο διαισθητικό ορισμό των περιοχών χαμηλής πολυπλοκότητας, γεγονός που την κάνει να διαφέρει από τις ήδη υπάρχουσες μεθόδους. Θεωρώντας ότι οι περιοχές χαμηλής πολυπλοκότητας έχουν ασυνήθιστα υψηλή εμφάνιση ενός ή περισσότερων καταλοίπων, όπως δείχθηκε και προηγούμενα (Εικόνα 9), επιτυγχάνουν μεγάλες βαθμολογίες (scores) σε αναζητήσεις ομοιότητας έναντι οποιουδήποτε ομοπολυμερούς περιέχει το συγκεκριμένο τύπο καταλοίπου. Η παρατήρηση αυτή προέκυψε από τη γενικότερη εμπειρία στις αναζητήσεις ομοιότητας, κατά τις οποίες ακόμη και οι πιο εκφυλισμένες ακολουθίες (π.χ. ομοπολυμερή) ταιριάζουν με υψηλή βαθμολογία ομοιότητας με περιοχές με προτίμηση σε συγκεκριμένους τύπους καταλοίπων. Το γεγονός αυτό είχε αναφερθεί παλαιότερα (Robison et al., 1994), όταν κατά την προσπάθεια ανάπτυξης μεθοδολογίας για το μαζικό και αποτελεσματικό εντοπισμό βακτηριακών ανοικτών πλαισίων ανάγνωσης χρησιμοποιήθηκε μια συλλογή γνωστών πρωτεϊνικών ακολουθιών με ακραία σύσταση για τον εντοπισμό αντίστοιχων βακτηριακών ακολουθιών. Παρόλα αυτά η ιδέα δεν αναπτύχθηκε περαιτέρω.

Με βάση αυτήν την ακραία διατύπωση, κατά τη δική μας προσέγγιση είναι δυνατόν να χρησιμοποιήσουμε πεπτίδια-ομοπολυμερή (με ένα μόνο αμινοξικό τύπο το καθένα) για τον εντοπισμό περιοχών με προτιμήσεις στην αμινοξική τους σύσταση. Εξ ορισμού, τέτοιες εκφυλισμένες ακολουθίες μπορούν να χαρακτηριστούν με δύο μόνο

παραμέτρους: τον τύπο του μονομερούς και το μήκος τους. Προφανώς, μια τέτοια εκφυλισμένη ακολουθία θα εμφανίζει υψηλή ομοιότητα με περιοχές πρωτεϊνών παρόμοιας αμινοξικής σύστασης, ανεξάρτητα από την ακολουθία τους.

Με το σκεπτικό αυτό είναι δυνατόν να εντοπιστούν τέτοιες περιοχές με χρήση των δοκιμασμένων αλγορίθμων σύγκρισης ακολουθιών. Η σημαντική διαφορά αυτής της προσέγγισης έγκειται στο γεγονός ότι όχι μόνο εντοπίζονται οι περιοχές της ακολουθίας με χαμηλό πληροφοριακό περιεχόμενο αλλά, παράλληλα, χαρακτηρίζονται ανάλογα με τον τύπο του καταλοίπου που είναι υπεύθυνο για το γεγονός αυτό. Κατά συνέπεια, δίνεται η δυνατότητα για επιλεκτικό φιλτράρισμα μόνο των καταλοίπων εκείνων που θα δημιουργούσαν πρόβλημα σε αναζητήσεις ομοιότητας, αφήνοντας το υπόλοιπο της περιοχής αναλλοίωτο, επιτρέποντας κατά αυτόν τον τρόπο περισσότερη ευαισθησία στην αναζήτηση ομοιότητας. Η τελική υλοποίηση της μεθόδου της επιτρέπει να εντοπίζει (σε αντίθεση με τις προϋπάρχουσες μεθόδους) περιοχές χαμηλής πολυπλοκότητας οι οποίες μπορεί να εκτείνονται και σε όλο το μήκος της ακολουθίας.

Μετά το νέο διαισθητικό ορισμό για τις περιοχές χαμηλής πολυπλοκότητας, οι όροι: ‘περιοχές χαμηλής πολυπλοκότητας’, ‘περιοχές με ακραία αμινοξική σύσταση’ και ‘περιοχές με προτίμηση στο κατάλοιπο X’ θα χρησιμοποιούνται χωρίς διάκριση στο υπόλοιπο του κειμένου.

Θεωρητικό Υπόβαθρο

Έστω ότι Q και T είναι δύο ακολουθίες, η πρώτη αυτή με την οποία πραγματοποιούμε την αναζήτηση και η δεύτερη από τη βάση δεδομένων στην οποία πραγματοποιούμε την αναζήτηση. Ας υποθέσουμε ότι τα ποσοστά (συχνότητες) εμφάνισης δύο διαφορετικών καταλοίπων α και β στις ακολουθίες Q και T είναι

στατιστικά ανεξάρτητες μεταβλητές τότε συνεπάγεται ότι η πιθανότητα ταιριάσματος δυο καταλοίπων αυτών των τύπων σε μια στοίχιση μπορεί να υπολογιστεί από τις ανεξάρτητες συχνότητες σύμφωνα με την εξίσωση:

$$P_{\alpha\beta} = f_a P_\beta \quad (1)$$

όπου f_a, P_β είναι οι συχνότητες εμφάνισης των αμινοξικών καταλοίπων τύπου *α* και *β* στις ακολουθίες Q και T αντίστοιχα. Ο υπολογισμός της βαθμολογίας για ταίριασμα οποιωνδήποτε τύπων καταλοίπων, όπως τα *α* και *β*, μπορεί να γίνει με χρήση ενός συμμετρικού πίνακα αντικατάστασης **M**, όπως ακριβώς συμβαίνει στις μεθόδους στοίχισης ακολουθιών. Το στοιχείο $\mathbf{M}_{\alpha,\beta}$ αυτού του πίνακα δίνει τη βαθμολογία που αντιστοιχούμε κατά τη στοίχιση ενός καταλοίπου τύπου *α* της μιας ακολουθίας με ένα κατάλοιπο τύπου *β* της άλλης. Τότε, η μέση αναμενόμενη βαθμολογία σε μια περιοχή μήκους *l* των δύο ακολουθιών θα δίνεται από την παράσταση:

$$l\sum_{\alpha,\beta} P_{\alpha\beta}\mathbf{M}_{\alpha\beta} \quad (2)$$

η οποία με απλή αντικατάσταση από την (1) γίνεται:

$$l\sum_{\alpha,\beta} (f_\alpha P_\beta \mathbf{M}_{\alpha,\beta}) \quad (3)$$

Έχοντας υποθέσει τυχαίο (αλλά κοινό για τις δύο ακολουθίες) μήκος περιοχής *l* είναι δυνατόν να αγνοήσουμε τον παράγοντα *l*. Κατά συνέπεια, οι συχνότητες f_a και P_β αντιστοιχούν στις τοπικές συχνότητες των καταλοίπων στην περιοχή των δύο ακολουθιών που συγκρίνεται. Αν θεωρήσουμε μια συγκεκριμένη περιοχή της εξεταζόμενης ακολουθίας Q, η αμινοξική σύσταση σε αυτή είναι

σταθερή οπότε είναι δυνατόν να υπολογίσουμε το άθροισμα της σχέσης (3) για όλους τους τύπους καταλοίπων. Επομένως, όπως φαίνεται στην (4), η μοναδική μεταβλητή που παραμένει είναι η αμινοξική σύσταση fa της ακολουθίας Q.

$$\sum_{\alpha,\beta}(P_{\alpha\beta}\mathbf{M}_{\alpha,\beta}) = \sum_{\alpha,\beta}(f_\alpha P_\beta \mathbf{M}_{\alpha,\beta}) = \sum_\alpha f_\alpha \sum_\beta (P_\beta \mathbf{M}_{\alpha,\beta}) = \sum_\alpha (f_\alpha C_\alpha) \quad (4)$$

όπου η παράμετρος C_a που εισάγεται είναι ξεκάθαρο ότι σχετίζεται μόνο με τον τύπο του αμινοξικού καταλοίπου α της ακολουθίας Q.
Οι τιμές των συχνοτήτων f_a, προφανώς ικανοποιούν τις βασικές σχέσεις:

$$0 \le f_\alpha \le 1 \quad (5)$$

$$\sum_\alpha f_\alpha = 1 \quad (6)$$

Επομένως, το δεξιά μέλος της σχέσης (4) αποτελεί παρεμβολή μεταξύ των 20 πιθανών τιμών C_a (μία για κάθε δυνατό τύπο καταλοίπου) και οι τιμές που είναι δυνατόν να πάρει περιορίζονται μεταξύ της μικρότερης και μεγαλύτερης τιμής των C_a. Η μέγιστη δυνατή τιμή του αθροίσματος μπορεί να είναι C_1=max(C_a). Η γενική περίπτωση κατά την οποία πάντα επιτυγχάνεται η μεγαλύτερη βαθμολογία είναι εκείνη κατά την οποία η συχνότητα εμφάνισης του αντίστοιχου καταλοίπου f_1 ισούται με 1. Προφανώς, αυτή είναι η περίπτωση του ομοπολυμερούς. Με το σκεπτικό αυτό, ένα από τα 20 δυνατά ομοπολυμερή θα δίνει πάντα την υψηλότερη βαθμολογία που είναι δυνατόν να επιτευχθεί από οποιαδήποτε άσχετη ακολουθία.
Το παραπάνω επιχείρημα σαφώς δεν ισχύει εάν οι δύο ακολουθίες πέρα από παρόμοια σύσταση έχουν και πραγματική ομοιότητα. Αυτή η ομοιότητα αντικατοπτρίζεται από την υψηλή βαθμολογία που υπολογίζεται. Για την περίπτωση αυτή, έχει θεμελιωθεί και

αναπτυχθεί στατιστική θεωρία η οποία επιτρέπει τον υπολογισμό της πιθανοφάνειας (likelihood) ώστε τέτοιες ομοιότητες να προκύπτουν τυχαία (Karlin and Altschul, 1990).

CAST: εντοπισμός περιοχών χαμηλής πολυπλοκότητας

Με βάση τις ιδέες που παρουσιάστηκαν στην προηγούμενη ενότητα, το πρόβλημα της εύρεσης περιοχών με ακραία σύσταση ή χαμηλή πολυπλοκότητα, μπορεί να επιλυθεί αλγοριθμικά ακολουθώντας τα παρακάτω βήματα:

α) ***Κατασκευάζουμε μια βάση δεδομένων η οποία αποτελείται από 20 εκφυλισμένες πρωτεϊνικές ακολουθίες:***

Η κάθε ακολουθία είναι ένα ομοπολυμερές το οποίο βασίζεται σε ένα τύπο αμινοξικού καταλοίπου. Όπως αναφέρθηκε και νωρίτερα, το μήκος των ακολουθιών μπορεί να επιλεχθεί αυθαίρετα.

β) ***Η προς έλεγχο ακολουθία συγκρίνεται με τις ακολουθίες της βάσης με κάποιον από τους γνωστούς αλγορίθμους αναζήτησης ομοιοτήτων:***

Ομοιότητες με σημαντικά μεγάλες βαθμολογίες θα προκύψουν μόνο εάν η ακολουθία που εξετάζουμε περιέχει περιοχές ακραίας αμινοξικής σύστασης. Η ακολουθία της βάσης με την οποία βρίσκεται η ομοιότητα ταυτοποιεί άμεσα τον τύπο του κυρίαρχου καταλοίπου και η περιοχή της ομοιότητας μας δίνει τα όρια της περιοχής χαμηλής πολυπλοκότητας.
Ξεκινώντας από την ιδέα της αναζήτησης ομοιότητας (με ήδη υπάρχον λογισμικό) έναντι της βάσης εκφυλισμένων ακολουθιών που μόλις περιγράφηκε, μπορούμε να διατυπώσουμε με πιο κομψό τρόπο ένα αλγόριθμο ο οποίος θα εντοπίζει (με μια επαναληπτική διαδικασία) όλες τις περιοχές χαμηλής πολυπλοκότητας σε μια ακολουθία. Η διατύπωση η οποία παρουσιάζεται στα αμέσως επόμενα

χρησιμοποιεί μια διαδικασία δυναμικού προγραμματισμού η οποία ακολουθεί τα ίδια βήματα με τα περισσότερα προγράμματα αναζήτησης ομοιοτήτων.

Στην ειδική περίπτωση που εξετάζουμε, μπορούμε να τροποποιήσουμε τη διαδικασία αναζήτησης τοπικών ομοιοτήτων που βασίζεται στον κλασικό αλγόριθμο Smith-Waterman (SW, Smith and Waterman, 1981). Αυτό είναι δυνατό να συμβεί αφού όπως ήδη τονίσθηκε η θέση που κάθε φορά μελετάμε στην ακολουθία (η οποία σχετίζεται με τη ροή του αλγορίθμου SW) στην περίπτωση του ομοπολυμερούς είναι ασήμαντη. Επομένως, για κάθε τύπο καταλοίπου *a* απαιτείται ένα απλό πέρασμα κατά μήκος της ακολουθίας για να εντοπισθούν περιοχές με μέγιστη βαθμολογία.

Ας συμβολίσουμε μια βιολογική ακολουθία *r* μήκους *n* ως:

$$r \equiv r_1 r_2 r_3 \ldots r_i \ldots r_n$$

με $1 \le i \le n$ και $i, n \in \mathbb{N}^*$.

Για τον έλεγχο της ύπαρξης στην ακολουθία *r* περιοχής με προτίμηση στον τύπο καταλοίπου *a* υπολογίζουμε μια βαθμολογία για κάθε θέση της ακολουθίας με τελικό σκοπό να εντοπίσουμε περιοχές που μεγιστοποιούν τη βαθμολογία. Ο υπολογισμός της βαθμολογίας s_i^α στη θέση *i* πραγματοποιείται προσθέτοντας την τιμή της βαθμολογίας για την προηγούμενη θέση (s_{i-1}^α) με την τιμή $\mathbf{M}_{\alpha, r_i}$ που προβλέπει ο πίνακας αντικατάστασης **M** για ταίριασμα του τύπου του καταλοίπου r_i με ένα κατάλοιπο τύπου *a*.

$$s_i^\alpha = \mathbf{M}_{\alpha, r_i} + \begin{cases} s_{i-1}^\alpha, & s_{i-1}^\alpha \ge 0 \\ 0, & s_{i-1}^\alpha < 0 \end{cases} \quad (7)$$

Οι περιοχές που επιτυγχάνουν μέγιστη βαθμολογία προσδιορίζονται ως συνεχείς περιοχές για τις οποίες οι τιμές της βαθμολογίας για κάποιο τύπο καταλοίπου α είναι διαρκώς θετικές και ξεκινούν από το *i* που αντιστοιχεί στην πρώτη θετική τιμή μέχρι τη θέση που προκύπτει το μέγιστο. Εφαρμόζοντας αυτό το βήμα και για τους είκοσι τύπους καταλοίπων εντοπίζουμε όλες τις πιθανές περιοχές με ακραία αμινοξική σύσταση. Οι τιμές των μέγιστων βαθμολογιών δίνουν ένα ποσοτικό κριτήριο του 'εκφυλισμού' στην περιοχή εκείνη της ακολουθίας και εξαρτάται (για δεδομένη ακολουθία) από τον πίνακα αντικατάστασης φυσικά τον τύπο του επικρατούς καταλοίπου. Αυτή η υλοποίηση του αλγορίθμου CAST, σε αντίθεση με τον κλασικό αλγόριθμο Smith-Waterman, δεν απαιτεί ειδικούς χειρισμούς για την εισαγωγή κενών κατά σύγκριση των ακολουθιών, ουσιαστικά αποκλείοντας κάτι τέτοιο. Διαφορετικά ο αλγόριθμος θα κατέληγε να γίνει αδικαιολόγητα πολύπλοκος με την αναίτια εισαγωγή εξωτερικών παραμέτρων για τον έλεγχο της εισαγωγής κενών. Στην πραγματικότητα, με βάση την τελική διατύπωση του αλγορίθμου, είναι προφανής η ισοδυναμία του με πολλαπλά περάσματα μιας σύγκρισης Smith-Waterman μεταξύ της ακολουθίας που μελετάμε και ενός συνόλου ομοπολυμερών ίσου με αυτήν μήκους, με παραμέτρους ποινής για την εισαγωγή και επέκταση κενών οι οποίες τείνουν στο άπειρο.

Φιλτράρισμα (masking) ακολουθιών για αναζητήσεις ομοιότητας

Πριν τη διαδικασία αναζήτησης ομοιοτήτων οι ακολουθίες προετοιμάζονται κατάλληλα για την εξάλειψη των ομοιοτήτων που θα οφείλονται σε ταίριασμα περιοχών με ακραία αμινοξική σύσταση. Όπως αναφέρθηκε, οι μέθοδοι που αναπτύχθηκαν αρχικά (Claverie and States, 1993; Wootton and Federhen, 1993) πραγματοποιούν φιλτράρισμα των περιοχών που εντοπίζουν με το χαρακτήρα X (Εικόνα 10). Κατά τον τρόπο αυτό σταθμίζονται αρνητικά ομοιότητες

σε αυτές τις περιοχές και προβάλλονται περισσότερο οι ομοιότητες στο υπόλοιπο μήκος της ακολουθίας.

```
>test_sequence
MWFHSTLLLAILVAVSADTCPAGFTALSTSKKCVKLITDVAKHSDATANCSSYGGHLISV
QNAIDNNAYLQLAAVSVTPYWLGIKCSLSGNPASCQWDDQSGNAGGYNGFAPGYPLVEVG
NCVYVPTSGSFAGKWLSGDCNTMSLNFICETAPTSPITDTCSFQYNGNCYYPSLSALPKQ
DAQFSCQQACGNLVSIHSIEENNYVQSLFTTNAPTYIRIGAVANNQNSNSWIDGTSWNYD
NIGYSNINLGMCWSMALSNDIVSTGKWISSSSSSSLPFVCKRKVGTQCGTTSGPTQTPGQ
CTSPMFMDNSGRFYSPSWPYSYIGEQNPCNYILDTPVGSLVQIRFPVMNLDSQASISIYS
RIEDTTPLVVLQGNSASNQWYTSTTNTMKVVFRPCIANCPNDGVNYRWEADFKPSTDVTQ
PPVTVVVVVVVVVVPSGVVVVAPGSISTPNYPNYYPNFLLCMYHLSTTGGYRINLDFGAI
DTEQCCDIIEVHDGPLLGSPKLGIVSGTWPAHAKTYQSSSNSMLVTFSTDSSGQGSGFSA
NFWAL
```

CAST REPORT

```
>test_sequence
V-rich region from 423 to 442 corrected with score 53
S-rich region from 127 to 317 corrected with score 40
MWFHSTLLLAILVAVSADTCPAGFTALSTSKKCVKLITDVAKHSDATANCSSYGGHLISV
QNAIDNNAYLQLAAVSVTPYWLGIKCSLSGNPASCQWDDQSGNAGGYNGFAPGYPLVEVG
NCVYVPTXGXFAGKWLXGDCNTMXLNFICETAPTXPITDTCXFQYNGNCYYPXLXALPKQ
DAQFXCQQACGNLVXIHXIEENNYVQXLFTTNAPTYIRIGAVANNQNXNXWIDGTXWNYD
NIGYXNINLGMCWXMALXNDIVXTGKWIXXXXXXXLPFVCKRKVGTQCGTTXGPTQTPGQ
CTXPMFMDNXGRFYXPXWPYSYIGEQNPCNYILDTPVGSLVQIRFPVMNLDSQASISIYS
RIEDTTPLVVLQGNSASNQWYTSTTNTMKVVFRPCIANCPNDGVNYRWEADFKPSTDVTQ
PPXTXXXXXXXXXXPSGXXXXAPGSISTPNYPNYYPNFLLCMYHLSTTGGYRINLDFGAI
DTEQCCDIIEVHDGPLLGSPKLGIVSGTWPAHAKTYQSSSNSMLVTFSTDSSGQGSGFSA
NFWAL
```

Εικόνα 11: Φιλτράρισμα (masking) με τον Αλγόριθμο CAST.

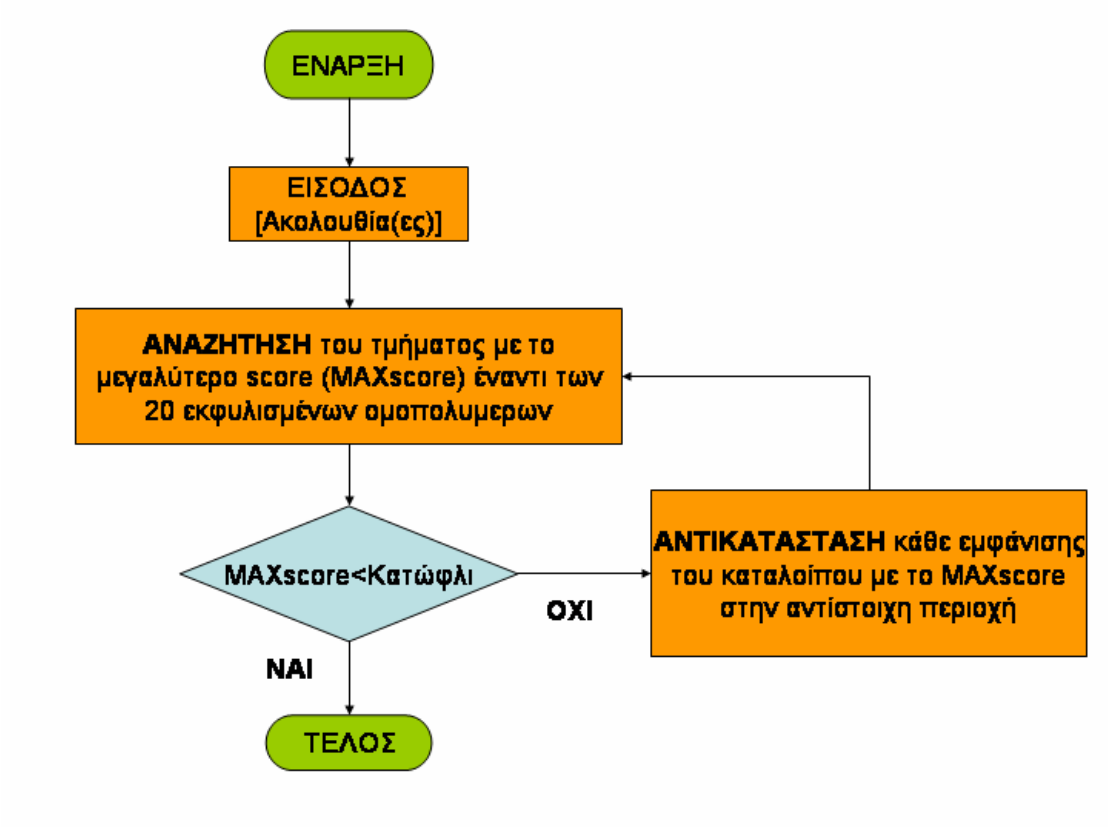
Απεικονίζεται η ακολουθία που δίνεται στην είσοδο (πλαίσιο στην κορυφή της εικόνας) και η πλήρης (verbose) έξοδος του προγράμματος. Έχουν υπογραμμιστεί και χρωματιστεί οι περιοχές στις οποίες εντοπίσθηκαν περιοχές ακραίας σύστασης σε βαλίνη και σερίνη. Στην έξοδο του προγράμματος γίνεται προφανής ο 'χειρουργικός' τρόπος με τον οποίο πραγματοποιείται το φιλτράρισμα.

Ο αλγόριθμος CAST δεν εντοπίζει απλά τις περιοχές χαμηλής πολυπλοκότητας αλλά, παράλληλα, συσχετίζει με κάθε τέτοια περιοχή και τον τύπο *α* ενός καταλοίπου που προκαλεί το φαινόμενο.

Έχοντας διαθέσιμη αυτήν την επιπλέον πληροφορία, η διαδικασία του φιλτραρίσματος είναι δυνατόν να πραγματοποιηθεί με ένα έξυπνο, σχεδόν 'χειρουργικό', τρόπο με μεγαλύτερη ακρίβεια. Για κάθε περιοχή που εντοπίζεται στο προηγούμενο στάδιο του αλγορίθμου αντικαθιστούμε με X μόνο εκείνον τον τύπο καταλοίπου που εντοπίστηκε κατά την αναζήτηση, στο οποίο θα αναφερόμαστε στο εξής ως τον 'τύπο προτίμησης'.

Τα υπόλοιπα κατάλοιπα παραμένουν ανεπηρέαστα, οπότε κάποιο ποσό πληροφορίας της ακολουθίας παραμένει ανέπαφο ακόμα και μετά την εφαρμογή αυτής της διόρθωσης και είναι δυνατό να συνεισφέρει θετικά στη διαδικασία αναζήτησης ομοιότητας (Εικόνα 11).

Η όλη διαδικασία πραγματοποιείται με τη χρήση ενός πίνακα αντικατάστασης, ο οποίος, ως γνωστόν, είναι δυνατόν να δίνει θετική βαθμολογία για το ταίριασμα ακόμη και μη ταυτόσημων καταλοίπων. Η υψηλή σύσταση μιας περιοχής σε ένα αμινοξικό τύπο (π.χ. αργινίνη) μπορεί να οδηγήσει σε μεγάλη βαθμολογία για ένα άλλο κατάλοιπο με παρόμοιες φυσικοχημικές ιδιότητες (π.χ. λυσίνη). Για να αποφύγουμε τέτοια ανεπιθύμητα φαινόμενα, όπου θα φιλτράραμε αδικαιολόγητα πολλά κατάλοιπα (τείνοντας η μέθοδος να φιλτράρει συγκρίσιμα με τις SEG, XNU), η διαδικασία του φιλτραρίσματος πραγματοποιείται με επαναληπτικό τρόπο. Σε κάθε επανάληψη, μετά τον εντοπισμό των υποψήφιων περιοχών για φιλτράρισμα, φιλτράρουμε επιλεκτικά για τον 'τύπο προτίμησης' μόνο την περιοχή εκείνη η οποία έχει εντοπιστεί με τη μεγαλύτερη βαθμολογία, εφόσον αυτή υπερβαίνει μια εμπειρική τιμή κατωφλίου. Αυτός ο έλεγχος, πέρα από την κατά σειρά επιλογή των τμημάτων που χρίζουν φιλτραρίσματος, εξασφαλίζει επιπλέον τον τερματισμό της επαναληπτικής διαδικασίας, όταν πλέον για κανένα τύπο καταλοίπου δεν είναι δυνατόν να βρεθεί τμήμα της ακολουθίας που να εμφανίζει βαθμολογία μεγαλύτερη του κατωφλίου. Στην Εικόνα 12 παρουσιάζεται το διάγραμμα ροής του αλγορίθμου CAST.



Εικόνα 12: Διάγραμμα ροής του Αλγορίθμου CAST.

Υλοποίηση του αλγορίθμου CAST

Ο αλγόριθμος CAST, όπως περιγράφηκε παραπάνω, υλοποιήθηκε σε πρόγραμμα για ηλεκτρονικό υπολογιστή το οποίο στην είσοδό του δέχεται μια αμινοξική ακολουθία στην οποία εντοπίζει και φιλτράρει περιοχές χαμηλής πολυπλοκότητας η βαθμολογία των οποίων υπερβαίνει μια καθορισμένη τιμή κατωφλίου, η οποία είναι δυνατόν να αλλαχθεί από το χρήστη. Η προκαθορισμένη τιμή της παραμέτρου κατωφλίου είναι 40 half-bits, μια τιμή η οποία έχει αναφερθεί ότι είναι βέλτιστη για αναζητήσεις ομοιότητας με το πακέτο λογισμικού BLAST. Ως προεπιλεγμένος πίνακας αντικατάστασης χρησιμοποιείται μια κατάλληλα τροποποιημένη εκδοχή του BLOSUM62 (Henikoff and

Henikoff, 1992): με βάση τον BLOSUM62 επανυπολογίζονται οι βαθμολογίες για ταίριασμα με τα X ως η μέση τιμή κάθε γραμμής ή στήλης, σύμφωνα με την πρόταση του Altschul και των συνεργατών του (Altschul et al., 1994). Αυτή η τροποποίηση γίνεται με σκοπό την εξαφάνιση της επίδρασης των 'ουδέτερων' χαρακτήρων (X), που χρησιμοποιούνται κατά τη διαδικασία του φιλτραρίσματος, στην επόμενη επανάληψη της διαδικασίας προσδιορισμού περιοχών ακραίας αμινοξικής σύστασης. Η εμπειρία στη χρήση των συγκεκριμένων παραμέτρων (κατωφλίου και πίνακα αντικατάστασης) δείχνει ότι οδηγούν σε ικανοποιητική απόδοση του αλγορίθμου όσον αφορά την ταχύτητα αλλά και την ευαισθησία. Παρόλα αυτά, είναι δυνατόν να χρησιμοποιηθούν διαφορετικές τιμές κατωφλίου ή/και πίνακες αντικατάστασης, καθώς θεωρούνται από το πρόγραμμα ως παράμετροι οι οποίες ορίζονται από το χρήστη κατά την κλήση του προγράμματος.

3 Και φυσικά … Ερωτήσεις

(Παραδίδονται **ΓΡΑΠΤΑ**, Ημερομηνία Παράδοσης: **Θα καθοριστεί στο Μάθημα**)

Ερώτηση 1

Α. Πόσοι (προσεγγιστικά) υπολογισμοί απαιτούνται για τον προσδιορισμό της βέλτιστης στοίχισης κατά ζεύγη δύο ακολουθιών ίσου μήκους Μ;
Β. Πόσοι (προσεγγιστικά) υπολογισμοί απαιτούνται για τον προσδιορισμό της στοίχισης των παραπάνω ακολουθιών αν περιοριστούμε στις k-το-πλήθος διαγώνιες εκατέρωθεν της κυρίας διαγωνίου του πίνακα δυναμικού προγραμματισμού;
Γ. Ποιο είναι το επί τοις εκατό (%) «κέρδος» σε υπολογισμούς ακολουθώντας τη δεύτερη προσέγγιση;
Δ. Να υπολογίσετε την τιμή που προκύπτει για το υποερώτημα Γ στις ακόλουθες περιπτώσεις:

1. M=100, k = 5
2. M=100, k =10
3. M=100, k = 50

Ερώτηση 2

Να δημιουργήσετε ευρετήρια των k-tuples για k=1 και k=2 για τις ακολουθίες GSVLAVFDYRLI και AVLAGFPVLI.
Εντοπίστε τις «καλές» διαγωνίους πάνω στους πίνακες δυναμικού προγραμματισμού και στο Dot-Plot της ερώτησης 3 του προηγούμενου φυλλαδίου.
Διατυπώστε τις παρατηρήσεις-σχόλιά σας.

4 Συμπληρωματικό Υλικό

4.1 Χρήσιμες Πηγές στο Διαδίκτυο (και όχι μόνο ...)

1. Ένα εξαιρετικό άρθρο ανασκόπησης για τις μεθόδους Δυναμικού Προγραμματισμού (Sankoff, 2000) και τις χρήσεις τους και σε άλλες εφαρμογές Βιοπληροφορικής. Το συγκεκριμένο τεύχος του περιοδικού Bioinformatics (ελεύθερη πρόσβαση από Η/Υ του ΕΚΠΑ στο URL: http://bioinformatics.oupjournals.org/content/vol16/issue1/index.shtml) περιέχει ιστορικές αναφορές στις πρώτες ημέρες του πεδίου της Βιοπληροφορικής. Από τον ίδιο συγγραφέα, επίσης, το βιβλίο «Time warps, string edits, and macromolecules : the theory and practice of sequence comparison» (Sankoff and Kruskal, 1983) στο οποίο η πληθώρα εφαρμογών που αναλύεται (από τη σύγκριση ακολουθιών μέχρι την ανάλυση του κελαηδήματος των πουλιών!!!) αποδεικνύει την ευρύτητα σκέψης που πρέπει να επιδεικνύει όποιος θέλει να ασχολείται σοβαρά με το πεδίο της Βιοπληροφορικής.

2. Το κεφάλαιο 3 από το βιβλίο «Introduction to Computational Molecular Biology» (Setubal and Meidanis, 1997). Στο κεφάλαιο αυτό, παρέχεται σε σχετικά μικρή έκταση η μαθηματική θεμελίωση για διάφορους από τους αλγορίθμους που συζητήσαμε στο μάθημα (αλλά και παραλλαγές τους).

4.2 Βιβλιογραφία

Allison, L., L. Stern, T. Edgoose and T. I. Dix (2000). Sequence complexity for biological sequence analysis. *Comput Chem*, **24**(1): 43-55.

Altman, R. B. and J. M. Dugan, Eds. (2003). *In* Defining Bioinformatics and Structural Bioinformatics. Structural Bioinformatics. Hoboken, NJ, John Wiley & Sons.

Altschul, S. F., M. S. Boguski, W. Gish and J. C. Wootton (1994). Issues in searching molecular sequence databases. *Nat Genet*, **6**(2): 119-29.

Altschul, S. F. and W. Gish (1996). Local alignment statistics. *Methods Enzymol*, **266**: 460-80.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). Basic local alignment search tool. *J Mol Biol*, **215**(3): 403-10.

Altschul, S. F., T. L. Madden, A. A. Schèaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research.*, **25(17)**: 3389-402.

Andrade, M. A., N. P. Brown, C. Leroy, S. Hoersch, A. de Daruvar, C. Reich, A. Franchini, J. Tamames, A. Valencia, C. Ouzounis and C. Sander (1999). Automated genome sequence analysis and annotation. *Bioinformatics (Oxford, England)*, **15(5)**: 391-412.

Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell and D. L. Wheeler (2003). GenBank. *Nucleic acids research.*, **31(1)**: 23-7.

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne (2000). The Protein Data Bank. *Nucleic acids research.*, **28(1)**: 235-42.

Bhatia, U., K. Robison and W. Gilbert (1997). Dealing with database explosion: a cautionary note. *Science*, **276**(5319): 1724-5.

Boeckmann, B., A. Bairoch, R. Apweiler, M. C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout and M. Schneider (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic acids research.*, **31(1)**: 365-70.

Bork, P. and E. V. Koonin (1998). Predicting functions from protein sequences--where are the bottlenecks? *Nat Genet*, **18**(4): 313-8.

Claverie, J.-M. and D. J. States (1993). Information enhancement methods for large scale sequence analysis. *Computers & Chemistry*, **17**(2): 191-201.

Etzold, T., A. Ulyanov and P. Argos (1996). SRS: information retrieval system for molecular biology data banks. *Methods Enzymol*, **266**: 114-28.

Galperin, M. Y. and E. V. Koonin (1998). Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. *In Silico Biol*, **1**(1): 55-67.

Gilks, W. R., B. Audit, D. De Angelis, S. Tsoka and C. A. Ouzounis (2002). Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics*, **18**(12): 1641-9.

Gumbel, E. J. (1958). Statistics of extremes. New York,, Columbia University Press.

Henikoff, S. and J. G. Henikoff (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, **89**(22): 10915-9.

Iliopoulos, I., S. Tsoka, M. A. Andrade, A. J. Enright, M. Carroll, P. Poullet, V. Promponas, T. Liakopoulos, G. Palaios, C. Pasquier, S. Hamodrakas, J. Tamames, A. T. Yagnik, A. Tramontano, D. Devos, C. Blaschke, A. Valencia, D. Brett, D. Martin, C. Leroy, I. Rigoutsos, C. Sander and C. A. Ouzounis (2003). Evaluation of annotation strategies using an entire genome sequence. *Bioinformatics*, **19**(6): 717-26.

Iliopoulos, I., S. Tsoka, M. A. Andrade, P. Janssen, B. Audit, A. Tramontano, A. Valencia, C. Leroy, C. Sander and C. A. Ouzounis (2001). Genome sequences and great expectations. *Genome biology*, **2(1)**.

Karlin, S. and S. F. Altschul (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences of the United States of America.*, **87(6)**: 2264-8.

Milosavljevic, A., Ed. (1999). *In* Discovering patterns in DNA sequences by the Algorithmic Significance Method. Pattern discovery in biomolecular data. Tools, techniques, and applications. Oxford, Oxford University Press.

Miyazaki, S., H. Sugawara, T. Gojobori and Y. Tateno (2003). DNA Data Bank of Japan (DDBJ) in XML. *Nucleic acids research.*, **31(1)**: 13-6.

Pearson, W. R. (1990). Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods in enzymology.*, **183**.

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.

Promponas, V. J., A. J. Enright, S. Tsoka, D. P. Kreil, C. Leroy, S. Hamodrakas, C. Sander and C. A. Ouzounis (2000). CAST: an iterative algorithm for the complexity analysis of sequence tracts. *Bioinformatics (Oxford, England)*, **16(10)**: 915-22.

Robison, K., W. Gilbert and G. M. Church (1994). Large scale bacterial gene discovery by similarity search. *Nature genetics.*, **7(2)**: 205-14.

Sankoff, D. (2000). The early introduction of dynamic programming into computational biology. *Bioinformatics*, **16**(1): 41-7.

Sankoff, D. and J. B. Kruskal (1983). Time warps, string edits, and macromolecules : the theory and practice of sequence comparison. Reading, Mass., Addison-Wesley Pub. Co. Advanced Book Program.

Setubal, J. C. and J. Meidanis (1997). Introduction to computational molecular biology. Boston, PWS Pub.

Smith, T. F. and M. S. Waterman (1981). Identification of common molecular subsequences. *Journal of molecular biology.*, **147(1)**: 195-7.

Stoesser, G., W. Baker, A. van den Broek, M. Garcia-Pastor, C. Kanz, T. Kulikova, R. Leinonen, Q. Lin, V. Lombard, R. Lopez, R. Mancuso, F. Nardone, P. Stoehr, M. A. Tuli, K. Tzouvara and R. Vaughan (2003). The EMBL Nucleotide Sequence Database: major new developments. *Nucleic acids research.*, **31(1)**: 17-22.

Tsoka, S., V. Promponas and C. A. Ouzounis (1999). Reproducibility in genome sequence annotation: the *Plasmodium falciparum* chromosome 2 case. *FEBS Lett*, **451**(3): 354-5.

Wise, M. J. (2001). 0j.py: a software tool for low complexity proteins and protein domains. *Bioinformatics*, **17**(90001): 288S-295.

Wootton, J. C. (1994). Non-globular domains in protein sequences: automated segmentation using complexity measures. *Computers & chemistry.*, **18(3)**: 269-85.

Wootton, J. C. and S. Federhen (1993). Statistics of local complexity in amino acid sequences and sequence databases. *Computers & Chemistry*, **17**(2): 149-163.

Wu, C. H., L. S. Yeh, H. Huang, L. Arminski, J. Castro-Alvear, Y. Chen, Z. Hu, P. Kourtesis, R. S. Ledley, B. E. Suzek, C. R. Vinayaka, J. Zhang and

W. C. Barker (2003). The Protein Information Resource. *Nucleic acids research.*, **31(1)**: 345-7.