ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΑΚΟΛΟΥΘΙΑΚΗ ΑΝΑΛΥΣΗ ΑΛΥΣΙΔΩΝ DNA

Φραντζέσκα Δ. Παπά

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς
Μάρτιος 2007

Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΑΚΟΛΟΥΘΙΑΚΗ ΑΝΑΛΥΣΗ ΑΛΥΣΙΔΩΝ DNA

Φραντζέσκα Δ. Παπά

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς
Μάρτιος 2007

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. ……. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- **Κούτρας Μάρκος (Επιβλέπων)**
- **Κατέρη Μαρία**
- **Αντζουλάκος Δημήτριος**

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

DEPARTMENT OF STATISTICS
AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

SEQUENCIAL ANALYSIS OF DNA STRINGS

By

Frantzeska D. Papa

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece
March 2007

Στους γονείς μου
Δημήτρη και Άννα,
και στην αδελφή μου
Βασιλική

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον καθηγητή μου και επιβλέποντα κύριο Μάρκο Κούτρα για την πολύτιμη βοήθεια, συμπαράσταση και υπομονή που έδειξε καθ' όλη τη διάρκεια της συγγραφής της παρούσας εργασίας. Εκφράζω τις ευχαριστίες μου στην κυρία Μαρία Κατέρη και στον κύριο Δημήτριο Αντζουλάκο για τη συμμετοχή τους στην Τριμελή Εξεταστική Επιτροπή.

Θέλω να ευχαριστήσω τον διδάκτορα Σωτήριο Μπερσίμη για το χρόνο που διέθεσε, τις επισημάνσεις του και τη σημαντική βοήθεια που μου προσέφερε. Ακόμα, ευχαριστώ τον Παναγιώτη Σταθάκη καθώς επίσης και τις συμφοιτήτριές μου Μαργέντη Παπακωνσταντίνου, Ελένη Μπάστα και Δέσποινα Καρατίσογλου για την ηθική υποστήριξη και συμπαράστασή τους.

Ευχαριστώ επίσης, την οικογένειά μου και τους φίλους μου για την υπομονή και κατανόηση που έδειξαν ώστε να ολοκληρώσω τη φοίτησή μου στο Μεταπτυχιακό Πρόγραμμα του Πανεπιστημίου Πειραιά.

Φραντζέσκα Παπά

Πειραιάς, Μάρτιος 2007

Περίληψη

Η έλικα του DNA αποτελεί ένα σημαντικό κομμάτι μελέτης στο χώρο της Μοριακής Βιολογίας. Όλες οι λειτουργίες και τα κληρονομικά γνωρίσματα των οργανισμών αποκαλύπτονται μέσα από τη μελέτη των δομικών λίθων του νουκλεϊνικού οξέος DNA. Για το λόγο αυτό έχουν αναπτυχθεί διάφορες Βάσεις Δεδομένων, μια εκ των οποίων είναι η GenBank, οι οποίες διαχειρίζονται τον μεγάλο αριθμό ακολουθιακών δομών που συλλέγουν καθημερινά. Στη παρούσα εργασία γίνεται παρουσίαση των μαθηματικών και των στατιστικών εργαλείων για την διαχείριση και την ανάλυση αυτών των δεδομένων.

Συγκεκριμένα, γίνεται σύγκριση δύο αλυσίδων DNA αντιστοιχίζοντας την ακολουθιακή τους δομή προς εύρεση κοινών χαρακτηριστικών ή προέλευσης των οργανισμών. Η σύγκριση αυτή επιτυγχάνεται με τη χρήση αλγορίθμων Δυναμικού Προγραμματισμού για την εύρεση των βέλτιστων μεταξύ τους αντιστοιχίσεων. Οι αντιστοιχίες μεταξύ των αλυσίδων επιτρέπουν την εμφάνιση αντικαταστάσεων, προσθηκών ή διαγραφών βάσεων, ακόμα και αναστροφών που αφορούν την αντιστοίχιση με το αντίστροφο συμπλήρωμα μιας ακολουθίας. Ακόμα περιγράφεται ένας σημαντικός αλγόριθμος εύρεσης επαναλαμβανόμενων σχηματισμών (*tandem repeats*) σε μια ακολουθία DNA που η συχνότητα εμφάνισής τους διαπιστώνει την παρουσία σημαντικών γενετικών προβλημάτων.

Με αφορμή τη σημαντικότητα της εμφάνισης των επαναλαμβανόμενων σχηματισμών γίνεται παρουσίαση ενός μοντέλου πιθανοτήτων. Πλέον, οι υπό μελέτη ακολουθίες DNA είναι ίσου μήκους και οδηγούν σε μια δίτιμη ακολουθία επιτυχιών και αποτυχιών για την οποία μελετάται η ακριβής κατανομή του συνολικού αριθμού επιτυχιών στις ροές ενός τουλάχιστον επιθυμητού αριθμού επιτυχιών. Στην παρούσα εργασία περιγράφονται ορισμένα πρόσφατα δημοσιευμένα αποτελέσματα που αφορούν την ακριβή κατανομή της προαναφερθείσας στατιστικής συνάρτησης.

Abstract

The helix of DNA has become an important area of study in the field of Molecular Biology. The performance and the heredity of several organisms may be revealed by the study of small molecules that the DNA consists of, the nucleotides. For this reason, lots of databases have been developed, such as GenBank, where all publicly available DNA sequences are collected on a daily basis. In the present study, mathematical and statistical models that have been developed for the DNA sequences analysis are presented.

The comparison of two DNA molecules is usually carried out by aligning the single strands of the molecules one on top of the other. Dynamic programming algorithms are a useful framework where the sequence matching problems can be embedded in order to find the optimal alignments. In that case, substitutions, insertions or deletions, even inversions are allowed in the alignment process. An important algorithm is presented for finding tandem repeats which have been considered that offer a strong evidence for the occurrence of several genetic diseases.

Given the prominent significance of tandem repeats in biological studies, an interesting statistic closely associated to them is described in some detail. In that case, the sequences under study are taken to have the same length, they are aligned one on top of the other and are converted into a sequence of successes/matches. In the new bistate sequence what we are looking at is the total number of successes in success runs of length k or longer. In the present thesis we present several recently published results on its exact distribution evaluation.

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Σχημάτων

Κατάλογος Συντομογραφιών

σ.π.π.	συνάρτηση πυκνότητας πιθανότητας
τ.μ.	τυχαία μεταβλητή

Κ Ε Φ Α Λ Α Ι Ο 1

Εισαγωγή

1.1 Παρουσίαση της Αλυσίδας DNA

1.1.1 Ορισμός του DNA

Το *DNA* (δεσοξυριβονουκλεϊνικό οξύ) είναι ένα είδος νουκλεϊνικού οξέος, που διακρίνεται για την ικανότητά του να κατευθύνει τη σύνθεση των πρωτεϊνών στα κύτταρα και να ελέγχει όλες τις λειτουργίες και τα κληρονομικά γνωρίσματα των οργανισμών. Εντοπίζεται στον πυρήνα των ευκαρυωτικών κυττάρων (κύτταρα που διαθέτουν πυρήνα) ως συστατικό των χρωμοσωμάτων αλλά ένα μικρό ποσοστό μπορεί να βρεθεί στα μιτοχόνδρια και στους χλωροπλάστες. Σε κάθε άνθρωπο, στον πυρήνα των κυττάρων τους συναντάται το ίδιο DNA.

Όλη η πληροφορία του νουκλεϊνικού οξέος DNA αποθηκεύεται κωδικοποιημένη στους δομικούς του λίθους, τα νουκλεοτίδια (*nucleotides*). Τα νουκλεοτίδια προέρχονται από τη σύνδεση, με ομοιοπολικό δεσμό, τριών διαφορετικών μορίων: Μιας πεντόζης (σάκχαρο με πέντε άτομα άνθρακα) που ονομάζεται δεσοξυριβόζη, ενός μορίου φωσφορικού οξέος και μιας οργανικής αζωτούχας βάσης (*base*). Οι αζωτούχες βάσεις που συναντώνται στο μόριο του DNA είναι η αδενίνη (*A*), η γουανίνη (*G*), η θυμίνη (*T*) και η κυτοσίνη (*C*). Οι βάσεις αδενίνη (*A*) και γουανίνη (*G*) ανήκουν στις πουρίνες (*purines*) και οι βάσεις θυμίνη (*T*) και κυτοσίνη (*C*) στις πυριμιδίνες (*pyrimidines*).

Τα μονοφωσφορικά νουκλεοτίδια ενώνονται μεταξύ τους με ομοιοπολικό δεσμό δημιουργώντας ένα πολυνουκλεοτίδιο, δηλαδή έναν απεριόριστο αριθμό διαφορετικών αλληλουχιών νουκλεοτιδίων. Έτσι, εμφανίζονται μεγάλου μήκους αλυσίδες νουκλεϊνικών οξέων που δικαιολογούν τη μοναδική ιδιότητα του DNA να είναι ο φορέας όλων των πληροφοριών που χρειάζεται ένας οργανισμός για να οικοδομηθεί και να λειτουργήσει. Το βακτήριο του εντέρου *Escherichia coli* (E. coli) είναι ένας οργανισμός που περιέχει 4.6×10^6 ζεύγη βάσεων ενώ το ανθρώπινο DNA συνίσταται από 3×10^9 ζεύγη βάσεων. Περισσότερο

από 99% των βάσεων είναι τα ίδια σε όλους τους ανθρώπους, Clote & Backofen (2000).

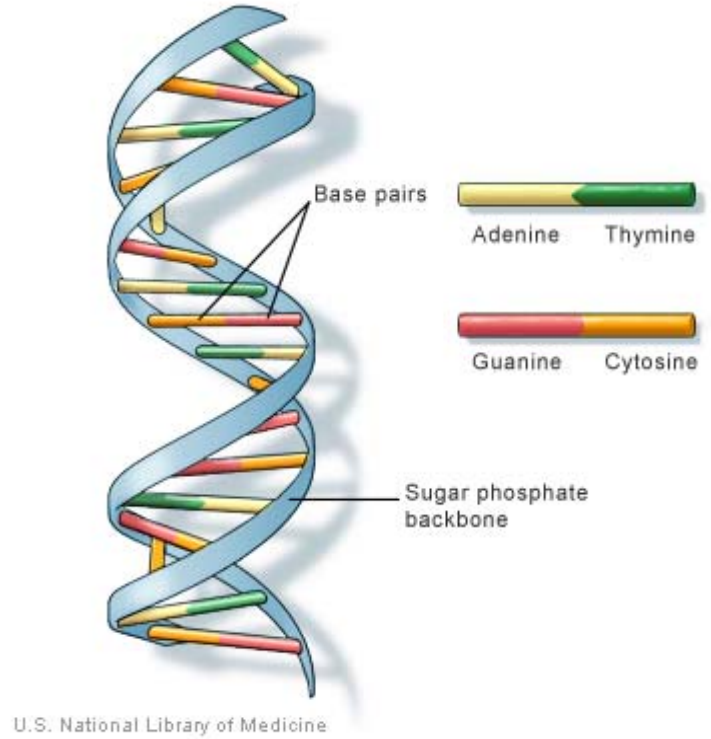
1.1.2 Δομή του DNA

Τα πολυνουκλεοτίδια διατάσσονται στο χώρο σύμφωνα με το μοντέλο της διπλής έλικας του DNA (*double-stranded DNA*) που προτάθηκε το 1953 από τους Watson και Crick. Το μοντέλο παρουσιάζεται στο *Σχήμα 1.1* και έχει τα εξής χαρακτηριστικά:

Αποτελείται από δύο πολυνουκλεοτιδικές αλυσίδες, τους κλώνους (*strands*), που σχηματίζουν διπλή έλικα. Σε κάθε κλώνο οι αζωτούχες βάσεις είναι κάθετες στον κύριο άξονα του μορίου και προεξέχουν προς το εσωτερικό του, ενώ οι δύο κλώνοι συγκρατούνται μεταξύ τους με δεσμούς υδρογόνου που σχηματίζονται μεταξύ των βάσεων.

Σχήμα 1.1

Τμήμα μορίου DNA



Τα ζεύγη των βάσεων (*base pairs*) είναι καθορισμένα και έτσι συναντάμε την αδενίνη (*A*) με τη θυμίνη (*T*) και τη γουανίνη (*G*) με τη κυτοσίνη (*C*) σύμφωνα με τους κανόνες των Watson-Crick (1953). Οι βάσεις αυτές που συνδέονται μεταξύ τους χαρακτηρίζονται ως συμπληρωματικές (*complementary bases*) και η ιδέα αυτής της συμπληρωματικότητας παρουσιάζεται με το παρακάτω παράδειγμα του Waterman (1995).

Θεωρούμε το τμήμα της αλυσίδας DNA

$$5'ACCTGAC3'$$

όπου τα σύμβολα 5′ και 3′ χρησιμοποιούνται για να δηλώσουν την καθορισμένη «κατεύθυνση» του τμήματος του μορίου DNA.

Τότε το τμήμα αυτό συνδέεται με το συμπληρωματικό του σύμφωνα με το παρακάτω σχήμα.

$$\begin{array}{ccccccccc} 5' & A & C & C & T & G & A & C & 3' \\ & | & | & | & | & | & | & | & \\ 3' & T & G & G & A & C & T & G & 5' \end{array}$$

1.2 Ειδικοί Σχηματισμοί και η Σημασία τους

Η μελέτη της αλυσίδας DNA εστιάζεται στον έναν από τους δύο κλώνους του μορίου ο οποίος αντιμετωπίζεται ως μια ακολουθία γραμμάτων από το σύνολο $\mathcal{B} = \{A,C,G,T\}$. Σε μια ακολουθία αναζητούνται συγκεκριμένοι σχηματισμοί νουκλεοτιδίων (*pattern of nucleotides*) οι οποίοι συνήθως σχηματίζονται από 2-5 βάσεις. Στο χώρο της Βιολογίας, το ενδιαφέρον των επιστημόνων έχει στραφεί κυρίως στην αναζήτηση συγκεκριμένων σχηματισμών με 3 βάσεις που επαναλαμβάνονται περισσότερο από 2 φορές ο ένας μετά τον άλλο (*tandem repeats*). Η επαναλαμβανόμενη εμφάνιση των τρινουκλεοτιδίων σε φυσιολογικά πλαίσια αποτελεί το 10% του ανθρώπινου γονιδίου και παίζει πολύ σημαντικό ρόλο στην ανάπτυξη του ανοσοποιητικού συστήματος των κυττάρων. Όμως, τα τελευταία χρόνια η έρευνα για τον προσδιορισμό των επαναλαμβανόμενων σχηματισμών οδήγησε στην ανακάλυψη της

συσχέτισης των διαδοχικών τρινουκλεοτιδίων με την εμφάνιση κάποιων ασθενειών. Η έρευνα αυτή εφιστάται κυρίως σε 5 ασθένειες που προσβάλουν τον ανθρώπινο οργανισμό και προκύπτουν από τον επαναλαμβανόμενο αριθμό εμφάνισης των σχηματισμών αυτών στις ακολουθίες DNA, Benson (1999).

Οι Verkerk et al. (1991) αναφέρουν ότι το σύνδρομο Χ, μια μορφή διανοητικής καθυστέρησης που εμφανίζεται στο χρωμόσωμα Χ (*fragile-X mental retardation*), οφείλεται στη συνεχόμενη επανάληψη του προτύπου CGG στην αλυσίδα DNA περισσότερες από 200 φορές. Μια παρόμοια διαρκής επανάληψη ενός προτύπου 3 βάσεων έχει εντοπιστεί και σε άλλες ασθένειες.

Η μυοτονική δυστροφία (*myotonic dystrophy*) είναι μια κληρονομική ασθένεια που προσβάλει τους μύες και προκύπτει από τη διαδοχική εμφάνιση των προτύπων AGC/CTG. Στα άτομα που δεν έχουν προσβληθεί από την ασθένεια αυτή, τα συγκεκριμένα πρότυπα εμφανίζονται από 5 έως 27 φορές. Οι πάσχοντες της συγκεκριμένης ασθένειας έχουν το λιγότερο 50 επαναλήψεις των προτύπων, ενώ μπορεί να εμφανίζονται έως και μερικές χιλιάδες πρότυπα, Fu et al. (1992).

Η «αταξία» του Friedreich (*Friedreich's ataxia*), που οφείλει το όνομά της στο Γερμανό γιατρό Nicolaus Friedreich, είναι μια κληρονομική ασθένεια που προσβάλει και τα δύο φύλα. Οι Campuzano et al. (1996) παρατήρησαν ότι προέρχεται από τη συνεχή επανάληψη του προτύπου GAA στο DNA των μιτοχονδρίων του ανθρώπου περισσότερες από 1000 φορές, ενώ ένας υγιής άνθρωπος διαθέτει από 8-30 διαδοχικές επαναλήψεις του προτύπου.

Ακόμα μελετήθηκαν δύο ασθένειες που προκύπτουν από τη διαδοχική επανάληψη στην αλυσίδα DNA του προτύπου CAG. Σύμφωνα με έρευνα της ομάδας Huntington's Disease Collaborative Research Group (1993), άτομα που νοσούν με την ασθένεια του Huntington (*Huntington's Disease*) εμφανίζουν τον συγκεκριμένο σχηματισμό κατ' επανάληψη από 40 έως 100 φορές, ενώ σε φυσιολογικά πλαίσια ο σχηματισμός επαναλαμβάνεται ως 26 φορές. Στην περίπτωση της ασθένειας της νωτιαίας και βολβικής μυϊκής ατροφίας ή ασθένεια του Kennedy (*spinal and bulbar muscular atrophy or Kennedy's disease*), η οποία προσβάλει μόνο τους άντρες, οι La Spada et al. (1991) παρατήρησαν την επανάληψη του σχηματισμού CAG πάνω από 35 φορές στο γονίδιο AR[(1)] (*androgen receptor*).

Μεγάλο ενδιαφέρον παρουσιάζει επίσης η αναζήτηση ομοιοτήτων μεταξύ των ακολουθιών δύο διαφορετικών οργανισμών που μπορεί να συνεπάγεται τη συσχέτιση μεταξύ

[(1)] Το γονίδιο AR είναι ένα τμήμα DNA που εντοπίζεται στο χρωμόσωμα Χ και παρέχει την πληροφορία για την παραγωγή της πρωτεΐνης που ονομάζεται υποδοχέας ανδρογόνων.

των λειτουργιών τους ή της προέλευσης τους. Κατά τη διαδικασία σύγκρισης των ακολουθιών, αναζητούνται κάποια τμήματα (υπακολουθίες) που ταυτίζονται. Συνήθως οι ακολουθίες έχουν διαφορετικό μήκος και η ανεύρεση των κοινών υπακολουθιών αντιμετωπίζεται με δύο διαφορετικούς τρόπους. Οι ακολουθίες είτε στοιχίζονται η μία κάτω από την άλλη (*aligned without shifts*), είτε ελέγχονται όλες οι διαδοχικές δυνατές στοιχίσεις (*the shifting case*) αναζητώντας την καλύτερη ή σχεδόν την καλύτερη αντιστοίχιση μεταξύ αυτών, γεγονός που αποτελεί πιο πολύπλοκο πρόβλημα. Και στις δύο περιπτώσεις χρησιμοποιείται μια δίτιμη μεταβλητή που λαμβάνει την τιμή 1, στην περίπτωση που οι βάσεις που βρίσκονται η μία κάτω από την άλλη είναι οι ίδιες (*matching case*) και την τιμή 0 όταν οι βάσεις είναι διαφορετικές.

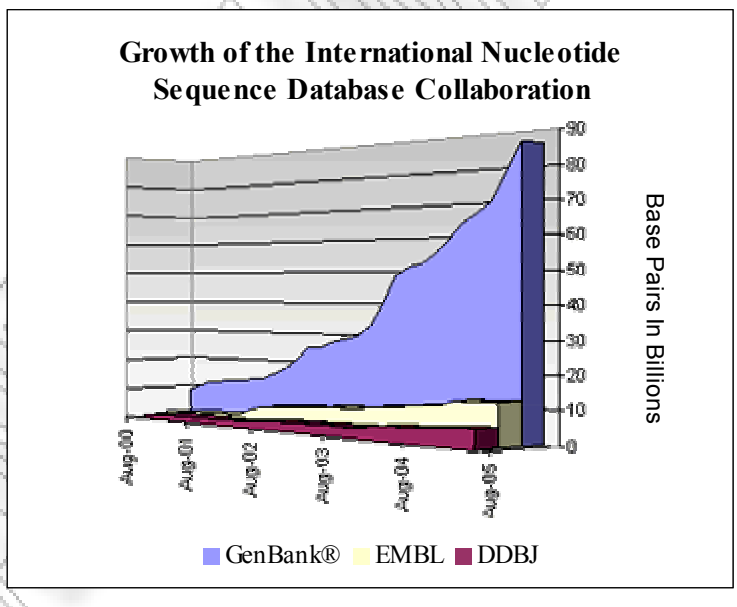
1.3 Προβλήματα Εντοπισμού Ειδικών Σχηματισμών

Προηγουμένως αναφερθήκαμε στη σημασία των ειδικών σχηματισμών στις ακολουθίες DNA για την εξέλιξη του κλάδου της Μοριακής Βιολογίας. Ωστόσο, η αναγνώρισή τους στις αλυσίδες DNA περιορίζεται σημαντικά εξαιτίας του μεγάλου μήκους των αλυσίδων. Όταν αναζητούνται συγκεκριμένοι επαναλαμβανόμενοι σχηματισμοί σε μια ακολουθία, ο εντοπισμός τους δεν είναι εύκολος, αφού ο έλικας του DNA είναι ένα βιολογικό μόριο που υπόκειται σε τυχαίες μεταλλαγές στο χρόνο. Έτσι, οι επαναλαμβανόμενοι σχηματισμοί συνίστανται από «κατά προσέγγιση» αντίγραφα (*approximate copies*), στα οποία εμφανίζονται διαφορετικές βάσεις κατά την επανάληψη ή παρεμβάλλονται βάσεις μεταξύ των αντιγράφων. Το θέμα αυτό απασχόλησε τον Benson (1999) που εισήγαγε ένα λογισμικό, που ονομάζεται *Tandem Repeat Finder* (*TRF*), για την εύρεση των «κατά προσέγγιση» επαναλαμβανόμενων σχηματισμών. Ακόμα, η σύγκριση δύο ακολουθιών οι οποίες έχουν μεγάλα μήκη οδηγεί σε χρονοβόρες διαδικασίες όταν μάλιστα ελέγχονται όλες οι δυνατές διαδοχικές στοιχίσεις (*the shifting case*). Έτσι, έχουν δοθεί λύσεις με χρήση ασυμπτωτικών αποτελεσμάτων από τους Gordon et al. (1986) για το πρόβλημα της αντιστοίχισης (*sequence alignment*) και από τους Arratia et al. (1986, 1990a) για την εύρεση κοινών προτύπων μεταξύ δύο ακολουθιών (*perfect matches*) ή κοινών προτύπων όπου επιτρέπονται και μη κοινά γράμματα (*mismatches*).

1.4 Βάσεις Δεδομένων DNA

Ο μεγάλος αριθμός δεδομένων των ακολουθιών έχει οδηγήσει στην ανάπτυξη των βάσεων δεδομένων DNA που συλλέγουν και διαχειρίζονται όλες τις διαθέσιμες ακολουθίες DNA που έχουν δημοσιευθεί σε διάφορες εργασίες. Συγκεκριμένα μεταξύ αυτών συγκαταλέγονται τρεις μεγάλες βάσεις δεδομένων. Στην Ευρώπη βρίσκεται ο οργανισμός European Molecular Biology Laboratory (EMBL), στην Ιαπωνία η τράπεζα δεδομένων DNA DataBank of Japan (DDBJ) και στην Αμερική η GenBank στο Διεθνές Κέντρο Πληροφόρησης Βιοτεχνολογίας (*International Centre for Biotechnology Information*). Αυτοί οι οργανισμοί αποτελούν μέρος της Διεθνούς Συνεργασίας Βάσεων Δεδομένων Ακολουθιών Νουκλεοτιδίων (*International Nucleotide Sequence Database Collaboration*) και συνεργάζονται καθημερινά για την ανταλλαγή δεδομένων.

Σχήμα 1.2



Από τις παραπάνω Βάσεις ξεχωρίζει ως προς τον όγκο των δεδομένων η GenBank, που εμπλουτίζει το υλικό που διαθέτει κάθε δύο μήνες. Από πληροφορίες μέσω του Διαδικτύου και συγκεκριμένα της διευθύνσεως http://www.ncbi.nlm.nih.gov/Genbank οι βάσεις δεδομένων DNA της GenBank ξεπερνούν σήμερα τα 100 giga ενώ από το παραπάνω

διάγραμμα (*Σχήμα 1.2*), είναι εμφανής η υπεροχή της σε σχέση με τις άλλες δύο Βιβλιοθήκες Δεδομένων. Μάλιστα σύμφωνα με την έκδοση του Οκτωβρίου του 2005 (GenBank® Release 150) περιέχει περισσότερες από 46 εκατομμύρια καταχωρήσεις ακολουθιών και συνολικά περισσότερα από 51 δισεκατομμύρια ζεύγη βάσεων.

Οι ακολουθίες που διαθέτει η GenBank μπορούν να αναζητηθούν μέσω της *National Centre for Biotechnology Information* (NCBI). Κάθε καταχωρημένη ακολουθία συνοδεύεται από μια περιεκτική περιγραφή της. Δίνεται ο επιστημονικός ορισμός που συχνά χρησιμοποιείται στη βιβλιογραφία ή το κοινό όνομα του οργανισμού τον οποίο αφορά η ακολουθία. Παρέχεται ένας αριθμός πρόσβασης, δηλαδή ένας κωδικός που περιγράφει την ακολουθία και κάποιες λέξεις-κλειδιά που περιγράφουν προϊόντα γονιδίων (*gene products*) της καταχώρησης. Επίσης, παρέχονται στοιχεία που αφορούν το πλήθος των νουκλεοτιδικών βάσεων που έχουν χρησιμοποιηθεί σε διάφορα άρθρα, όπως και ο τίτλος των άρθρων, οι συγγραφείς, τα ονόματα των δημοσιευμάτων, κάποια σχόλια και χαρακτηριστικά και το τμήμα της αλυσίδας που έχει αποκωδικοποιηθεί.

Ένα χαρακτηριστικό υπόδειγμα αυτής της εγγραφής της GenBank που βρέθηκε στην Διεύθυνση http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&val=198054 παρουσιάζεται παρακάτω.

```
LOCUS       MUSIGVCD 153 bp DNA linear
            ROD 27-APR-1993
DEFINITION  Mouse Ig germline H-chain D region, 5' flank.
ACCESSION   M60958
VERSION     M60958.1  GI:198054
KEYWORDS    D-region; germline; immunoglobulin heavy chain.
SOURCE      Mus musculus (house mouse)
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
            Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
            Sciurognathi; Muroidea; Muridae; Murinae; Mus.
REFERENCE   1  (bases 1 to 153)
  AUTHORS   Gu,H., Kitamura,D. and Rajewsky,K.
```

```
  TITLE     B cell development regulated by gene rearrangement: arrest of
            maturation by membrane-bound D mu protein and selection of DH
            element reading frames
  JOURNAL   Cell 65 (1), 47-54 (1991)
   PUBMED   2013094
COMMENT     Original source text: Mouse, DNA.
FEATURES             Location/Qualifiers
     source          1..153
                     /organism="Mus musculus"
                     /mol_type="genomic DNA"
                     /strain="CB20"
                     /db_xref="taxon:10090"
ORIGIN
        1 gagtaaaaat gctggatgtc tcttaaggat ggcccctgac actctgcact gctacctctg
       61 gccccaccag acaatgttcc tgcagaacct gttaccttac ttggcaggga ttttgtcaa
      121 gggatctatt actgtgtcta ctatggtaac tac
//
```

Κ Ε Φ Α Λ Α Ι Ο 2

Αριθμητικοί Αλγόριθμοι Αντιστοίχισης Δυο Ακολουθιών DNA

2.1 Εισαγωγή

Η εξέλιξη της Μοριακής Βιολογίας εμφανίζεται από τα μέσα του 1960 με τη σύγκριση των μέχρι τότε διαθέσιμων πρωτεϊνικών ακολουθιών προς αναζήτηση ομοιοτήτων μεταξύ των οργανισμών. Αρχίζουν να μελετώνται οι αλλαγές που υπεισέρχονται σε μια ακολουθία DNA κατά την εξέλιξη των οργανισμών στο πέρασμα του χρόνου. Έτσι, προκύπτει για τη σύγκριση των ακολουθιών, η εισαγωγή των αριθμητικών αλγορίθμων στο χώρο της Βιολογίας, επιδιώκοντας την άριστη αντιστοίχιση (*optimal alignment*) μεταξύ δύο ακολουθιών κατά την τοποθέτηση της μιας κάτω από την άλλη. Οι αλγόριθμοι, που χρησιμοποιούνται στο συγκεκριμένο κεφάλαιο, εντάσσονται στην περιοχή του Δυναμικού Προγραμματισμού (*Dynamic Programming*) που πρωτοεισήχθηκε από τους Needleman & Wunsch (1970).

Το ενδιαφέρον για την εξέλιξη των μορίων εστιάζεται στη μελέτη και σύγκριση ακολουθιών DNA για τον έλεγχο ύπαρξης κοινού προγόνου μεταξύ των οργανισμών. Κατά τη διαδικασία σύγκρισης δύο ακολουθιών, οι αλλαγές που παρατηρούνται μεταξύ τους προέρχονται από αντικαταστάσεις νουκλεϊνικών βάσεων και οφείλονται σε προσθήκες ή αφαιρέσεις βάσεων. Θεωρώντας την ακολουθία $\mathrm{a} = ACTGC$ παρατηρείται ότι η αντικατάσταση της βάσης $a_3 = T$ από τη βάση C οδηγεί στην ακολουθία $\mathrm{b} = ACCGC$. Τότε η στοίχιση μεταξύ τους είναι η παρακάτω

$$\begin{array}{ll} \mathbf{a:} & A \quad C \quad T \quad G \quad C \\ \mathbf{b:} & A \quad C \quad C \quad G \quad C \end{array}$$

Αν επιπλέον παρατηρηθεί αφαίρεση της βάσης $a_2 = C$, θα προκύψει η ακολουθία

c $= A - CGC$ όπου η βάση a_2 έχει αντικατασταθεί από ένα μηδενικό στοιχείο και η μεταξύ τους στοίχιση τότε είναι η εξής

$$\begin{array}{lccccc} \textbf{a:} & A & C & T & G & C \\ \textbf{c:} & A & - & C & G & C \end{array}$$

Στη συνέχεια η εισαγωγή της βάσης T μεταξύ των βάσεων $c_3 = G$ και $c_4 = C$ οδηγεί σε μια νέα ακολουθία d $= ACGTC$ και η στοίχιση με την ακολουθία a είναι η εξής

$$\begin{array}{lccccc} \textbf{a:} & A & C & T & G & - & C \\ \textbf{d:} & A & - & C & G & T & C \end{array}$$

Ακολουθώντας διαδοχικά τις αλλαγές που υφίσταται η ακολουθία a $= ACTGC$ ώστε να προκύψει η ακολουθία d $= ACGTC$, η στοίχιση που προκύπτει μεταξύ τους είναι η ακόλουθη:

$$\begin{array}{cccccc} A & C & T & G & - & C \\ | & & & | & & | \\ A & - & C & G & T & C \end{array} \qquad \textbf{(2.1)}$$

όπου παρατηρούνται 3 βάσεις της ακολουθίας a να είναι ταυτόσημες (*identities*) με 3 βάσεις της ακολουθίας d κατά την τοποθέτηση της μιας ακολουθίας κάτω από την άλλη. Η υποκατάσταση της βάσης T από τη βάση C οδηγεί στην αντιστοίχιση δύο μη-ομοίων βάσεων, που ονομάζεται αντικατάσταση (*substitution* ή *mismatch*), ενώ ακόμα η αφαίρεση (*deletion*) της κυτοσίνης της ακολουθίας a και η προσθήκη (*insertion*) της θυμίνης στην ακολουθία d εμφανίζονται με τη μορφή αντιστοίχισης βάσης με το μηδενικό στοιχείο "-" (*gap*).

Η αντιστοίχιση μεταξύ των παραπάνω ακολουθιών θα μπορούσε να είναι διαφορετική, αν δεν ήταν γνωστές οι αλλαγές που υφίσταται η ακολουθία a ώστε να προκύψει η d . Για παράδειγμα θα μπορούσαμε να θεωρήσουμε την αντιστοίχιση

$$
\begin{array}{cccccc}
A & C & T & G & - & C \\
| & | & & | & & | \\
A & C & - & G & T & C
\end{array}.
$$

Έτσι, παρατηρούνται 4 ομοιότητες (*identities*) και 2 αντιστοιχίσεις βάσεων με το μηδενικό στοιχείο "-". Εδώ δε γίνεται διαχωρισμός μεταξύ προσθήκης ή αφαίρεσης βάσης διότι δεν είναι γνωστά τα ενδιάμεσα στάδια που αποφέρουν την αντιστοίχιση. Θα μπορούσε να θεωρηθεί ότι κατά την εξέλιξη των οργανισμών δύο μόνο βάσεις είτε προστέθηκαν είτε αφαιρέθηκαν, ενώ οι υπόλοιπες βάσεις παρέμειναν οι ίδιες. Έχει επικρατήσει οι έννοιες της προσθήκης και της αφαίρεσης βάσεων να θεωρούνται πανομοιότυπες και σύμφωνα με τον Kruskal (1983) χαρακτηρίζονται ως *indels* από τα πρώτα γράμματα των λέξεων *insertions-deletions*. Όπως είναι φανερό από το παραπάνω παράδειγμα, η στοίχιση μεταξύ δύο ακολουθιών μπορεί να επιτευχθεί με πολλούς διαφορετικούς τρόπους, η χρήση όμως αριθμητικών αλγορίθμων μας επιτρέπει την εύρεση της άριστης στοίχισης, Waterman (1995).

Πριν παρουσιαστούν οι αλγόριθμοι, είναι απαραίτητη η εισαγωγή των *scores* της αντιστοίχισης, μια υπολογιστική εκτίμηση της ομοιότητας δύο ακολουθιών. Έστω p η πιθανότητα ομοιότητας δύο βάσεων κατά τη στοίχιση, q η πιθανότητα υποκατάστασης και r η πιθανότητα εμφάνισης ενός *indel*. Τότε η πιθανότητα της στοίχισης (2.1) ισούται με $\Pr = p^3 q r^2$ και το *score* της στοίχισης S προσδιορίζεται, αφού πρώτα λογαριθμίσουμε την πιθανότητα, ως εξής:

$$
\begin{aligned}
S &= \log \Pr - 5(\log s), \qquad \text{όπου } 5(\log s) = \text{σταθερή ποσότητα} \\
&= 3(\log p) + \log q + 2(\log r) - 5(\log s) \\
&= 3(\log p - \log s) + (\log q - \log s) + 2(\log r - \log \sqrt{s}) \\
&= 3(\log(p/s)) - (\log(s/q)) - 2(\log(\sqrt{s}/r)) \\
&= 3 - \mu - 2\delta
\end{aligned}
$$

όπου θεωρούμε το s τέτοιο ώστε να ισχύει $\log(p/s) = 1$ και θέτουμε $\mu = \log(s/q)$ και $\delta = \log(\sqrt{s}/r)$. Ο συντελεστής του $\log s$, όπου στην περίπτωση μας ισούται με την τιμή 5, επιλέγεται ίσος με

$$(\#identities) + (\#substitutions) + (1/2)(\#indels)$$

ώστε να οδηγεί σε *score* της μορφής

$$S = (\#identities) - \mu \times (\#substitutions) - \delta \times (\#indels).$$

Γενικά, το *score* ορίζεται από την παρακάτω σχέση:

$$S = \max\{(\#identities) - \mu \times (\#substitutions) - \delta \times (\#indels)\},$$

όπου το μέγιστο λαμβάνεται ως προς όλες τις δυνατές αντιστοιχίσεις μεταξύ δυο ακολουθιών.

2.2 Το πλήθος των Αντιστοιχίσεων

Θεωρούμε τις ακολουθίες a=$a_1 a_2 ... a_n$ και b=$b_1 b_2 ... b_m$ μεγέθους n και m αντίστοιχα, όπου τα στοιχεία $a_i, i = 1,2,...,n$ και $b_j, j = 1,2,...,m$ προέρχονται από το σύνολο $\mathcal{B} = \{A, C, G, T\}$. Η αντιστοίχιση μπορεί να επιτευχθεί αυξάνοντας τα μήκη των δύο ακολουθιών σε L με την προσθήκη μηδενικών στοιχείων "-" (*gaps*). Έτσι, η ακολουθία a μετατρέπεται στην ακολουθία $a^* = a_1^* a_2^* ... a_L^*$ και η ακολουθία b στην $b^* = b_1^* b_2^* ... b_L^*$ ενώ η στοίχιση μεταξύ τους είναι η ακόλουθη

$$\begin{matrix} a_1^* & a_2^* ..... & a_L^* \\ b_1^* & b_2^* ..... & b_L^*. \end{matrix}$$

Το μήκος L των δύο νέων ακολουθιών a^* και b^* κυμαίνεται μεταξύ του μέγιστου μήκους της μίας εκ των δύο ακολουθιών και του αθροίσματος των μηκών τους, δηλαδή $\max(n,m) \le L \le n+m$. Η περίπτωση $L = n+m$ προκύπτει από την αντιστοίχιση των βάσεων των δύο ακολουθιών με μηδενικά στοιχεία

$$\begin{matrix} a_1 & ... & a_n & - & ... & - \\ - & ... & - & b_1 & ... & b_m. \end{matrix}$$

Η αντιστοίχιση δύο μηδενικών στοιχείων $\begin{pmatrix} - \\ - \end{pmatrix}$ δεν είναι αποδεκτή, καθώς δεν έχει νόημα το «ταίριασμα» (*matching*) μεταξύ δύο αφαιρέσεων.

Το πλήθος όλων των δυνατών αντιστοιχίσεων (*number of sequence alignments*) δύο ακολουθιών a και b μήκους n και m αντίστοιχα είναι σύμφωνα με τον Laquer (1981) συνυφασμένο με τους αριθμούς των Stanton & Cowan (1970).

Ορίζουμε $f(n,m)$ το πλήθος των αντιστοιχίσεων μιας ακολουθίας μήκους n με μια m μήκους ακολουθία. Τότε ισχύουν:

(i) $$f(n,m) = f(n-1,m) + f(n-1,m-1) + f(n,m-1)$$

και

(ii) $$f(n,n) \sim 2^{5/4}(n\pi)^{-1/2}(1+\sqrt{2})^{2n+1},$$

όπου η ισοδυναμία $c(n) \sim d(n)$ δηλώνει ότι $\lim \frac{c(n)}{d(n)} = 1$, καθώς $n \to \infty$.

Η σχέση ***(i)*** με την αρχική συνθήκη

$$f(i,0) = f(0,j) = 1, \text{ για κάθε } i = 1,...,n \text{ και } j = 1,...,m$$

προκύπτει από την παρατήρηση του Waterman (1984b) ότι οι αντιστοιχίες μεταξύ δύο ακολουθιών μπορεί να καταλήγουν σε μία από τις 3 περιπτώσεις:

$$\begin{matrix} \ldots & a_n \\ \ldots & - \end{matrix} \qquad \begin{matrix} \ldots & a_n \\ \ldots & b_m \end{matrix} \qquad \begin{matrix} \ldots & - \\ \ldots & b_m \end{matrix}$$

Η περίπτωση $\begin{pmatrix} a_n \\ - \end{pmatrix}$ αντιστοιχεί σε αφαίρεση της βάσης a_n, η περίπτωση $\begin{pmatrix} a_n \\ b_m \end{pmatrix}$ αντιστοιχεί σε ομοιότητα ή αντικατάσταση βάσης με άλλη βάση και η περίπτωση $\begin{pmatrix} - \\ b_m \end{pmatrix}$ σε προσθήκη της βάσης b_m.

Η σχέση ***(ii)*** δίνει το πλήθος των δυνατών αντιστοιχίσεων δύο ακολουθιών ίσου μήκους n.

Ωστόσο η εξίσωση $f(n,m) = f(n-1,m) + f(n-1,m-1) + f(n,m-1)$ υπερεκτιμά το πλήθος των δυνατών αντιστοιχίσεων αφού οι περιπτώσεις αντιστοίχισης

$$\begin{matrix} a_i & - \\ - & b_j \end{matrix} \quad \text{και} \quad \begin{matrix} - & a_i \\ b_j & - \end{matrix}$$

είναι ταυτόσημες στο χώρο της Βιολογίας. Γι' αυτό γίνεται χρήση της συνάρτησης $g(n,m)$ που τις παραπάνω περιπτώσεις τις αντιμετωπίζει σαν μία.

Αν μια αντιστοίχιση καταλήγει σε $\begin{pmatrix} a_n \\ - \end{pmatrix}$, τότε εμφανίζονται 3 πιθανές περιπτώσεις:

$$\begin{matrix} \dots & a_{n-1} & a_n \\ \dots & b_m & - \end{matrix} \qquad \begin{matrix} \dots & a_{n-1} & a_n \\ \dots & - & - \end{matrix} \qquad \begin{matrix} \dots & - & a_n \\ \dots & b_m & - \end{matrix}$$

ενώ στην περίπτωση $\begin{pmatrix} - \\ b_m \end{pmatrix}$, οι πιθανότητες είναι:

$$\begin{matrix} \dots & a_n & - \\ \dots & b_{m-1} & b_m \end{matrix} \qquad \begin{matrix} \dots & - & - \\ \dots & b_{m-1} & b_m \end{matrix} \qquad \begin{matrix} \dots & a_n & - \\ \dots & - & b_m \end{matrix}.$$

Το νέο πλήθος αντιστοιχίσεων $g(n,m)$ για το οποίο ισχύει $g(n,m) \le f(n,m)$ ικανοποιεί την αναδρομική σχέση

$$\begin{aligned} g(n,m) &= g(n-1,m) + g(n,m-1) + g(n-1,m-1) - g(n-1,m-1) \\ &= g(n-1,m) + g(n,m-1). \end{aligned}$$

Η νέα συνάρτηση ικανοποιεί την αρχική συνθήκη $g(0,0) = g(0,1) = g(1,0) = 1$ και ισούται με το Διωνυμικό Συντελεστή (*binomial coefficient*)

$$g(n,m)=\binom{n+m}{n}=\binom{n+m}{m}=\left[\begin{matrix}m+1\\ n\end{matrix}\right],$$

όπου το σύμβολο $\left[\begin{matrix}\cdot\\ \cdot\end{matrix}\right]$ παριστάνει τον αριθμό των επαναληπτικών συνδυασμών των $m+1$ στοιχείων ανά n.

Στην περίπτωση ακολουθιών ίσου μήκους εφαρμόζεται ο τύπος του Stirling

$$n!\sim\sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n}$$

και προκύπτει η σχέση

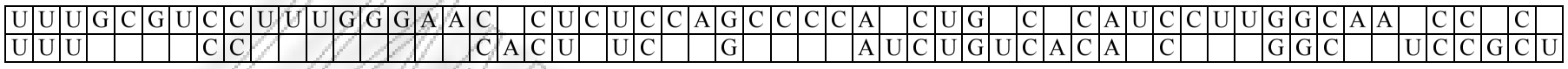
$$\begin{aligned}g(n,n)=\binom{2n}{n}=\frac{(2n)!}{(n!)^2}&\sim\frac{\sqrt{2\pi}(2n)^{2n+\frac{1}{2}}e^{-2n}}{\left(\sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n}\right)^2}=\\ &=\frac{\sqrt{2\pi}2^{2n+\frac{1}{2}}n^{2n+\frac{1}{2}}e^{-2n}}{\left(\sqrt{2\pi}\right)^2n^{2n+1}e^{-2n}}\\ &=\frac{2^{2n}2^{\frac{1}{2}}n^{-\frac{1}{2}}}{2^{\frac{1}{2}}\pi^{\frac{1}{2}}}\\ &=2^{2n}\left(\sqrt{n\pi}\right)^{-1},\ \text{καθώς } n\to\infty.\end{aligned}\qquad(2.2)$$

Συχνά οι Βιολόγοι ενδιαφέρονται για αντιστοιχίσεις μεταξύ ακολουθιών στις οποίες οι ομοιότητες βάσεων ή οι αντιστοιχίες μεταξύ διαφορετικών βάσεων εμφανίζονται σε *blocks* μεγέθους τουλάχιστον b χωρίς την παρουσία *indels*. Με αυτό το κίνητρο οι Griggs et al. (1986) παρουσίασαν μια καινούργια συνάρτηση για το πλήθος των αντιστοιχίσεων.

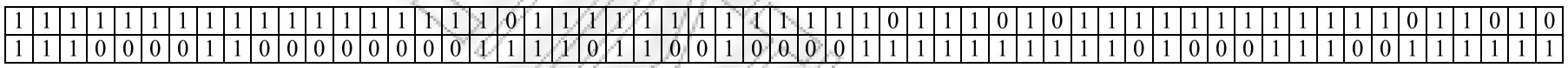
Η νέα συνάρτηση $g(b,n)$ εκφράζει το πλήθος των αντιστοιχίσεων δύο ακολουθιών ίσου μήκους n επιτρέποντας οι αντιστοιχίες μεταξύ ομοίων ή ανόμοιων βάσεων (*identities & mismatches*) να συμβαίνουν σε *blocks* μεγέθους τουλάχιστον b. Οι βάσεις ακολουθιών στην αντιστοίχιση αντικαθίστανται από το 0 και το 1, όπου το 0 δηλώνει την παρουσία μηδενικού

Σχήμα 2.1

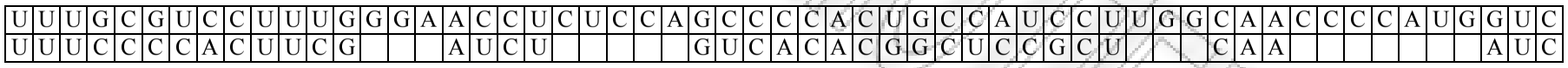
(a)



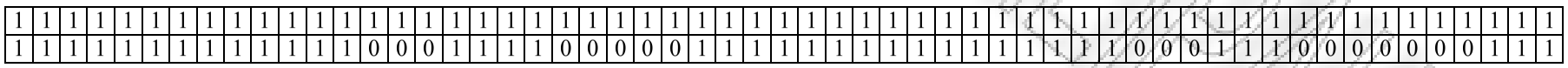
(b)



(c)



(d)



στοιχείου "-" και το 1 παραπέμπει σε παρουσία βάσης. Η συνάρτηση $g(b,n)$ εκφράζει το πλήθος εμφάνισης των πινάκων με στοιχεία το 0 και το 1 διαστάσεων $2\times L$, έτσι ώστε κάθε γραμμή να περιέχει ακριβώς n άσσους και κάθε στήλη να περιέχει τουλάχιστον 1 άσσο ενώ στήλες με 2 άσσους να παρουσιάζονται συνεχόμενες σε μέγεθος τουλάχιστον b.

Ένα παράδειγμα δύο διαφορετικών αντιστοιχίσεων δύο ακολουθιών παρουσιάζεται στο *Σχήμα 2.1*. Πρόκειται για ακολουθίες αγγελιοφόρου RNA (*mRNA*). Το RNA είναι το δεύτερο είδος νουκλεϊνικού οξέος και ο βιολογικός ρόλος του συγκεκριμένου τύπου είναι η μεταφορά της γενετικής πληροφορίας από το DNA στα ριβοσώματα για τη σύνθεση μιας πολυπεπτιδικής αλυσίδας (οι άλλοι δύο τύποι του μορίου RNA είναι το μεταφορικό RNA (*tRNA*) και το ριβοσωμικό RNA (*rRNA*)). Στο μόριο του RNA συναντώνται οι αζωτούχες βάσεις: αδενίνη (*A*), γουανίνη (*G*), ουρακίλη (*U*) και κυτοσίνη (*C*). Η πρώτη ακολουθία που χρησιμοποιείται είναι *β*-hemoglobin *mRNA* όρνιθας όπου διατίθενται τα νουκλεοτίδια 115-171 και η δεύτερη ακολουθία είναι *α*-hemoglobin *mRNA* όρνιθας με νουκλεοτίδια 118-156. Οι δύο mRNA ακολουθίες είναι γνωστό ότι προέρχονται από έναν κοινό πρόγονο, Fitch & Smith (1983).

Οι περιπτώσεις (a) και (c) αποτελούν δύο διαφορετικούς τρόπους αντιστοίχισης μεταξύ των ακολουθιών και προκύπτουν για διαφορετικές τιμές των παραμέτρων α,β μιας συνάρτησης $g(k)=\alpha+\beta(k-1)$, που ανατίθεται στην αντιστοίχιση για την επιβάρυνση εμφάνισης k συνεχόμενων *indels* κατά την εφαρμογή των αλγορίθμων. Πιο λεπτομερής αναφορά στη συνάρτηση $g(k)$ θα γίνει στην Παράγραφο 2.3.1. Στην περίπτωση (a) οι παράμετροι έχουν τεθεί ίσοι με το 0. Η περίπτωση αντιστοίχισης (c) αποτελεί μια άριστη αντιστοίχιση για συγκεκριμένες τιμές των παραμέτρων α,β και ανήκει σε ένα σύνολο άριστων αντιστοιχίσεων για τις ίδιες τιμές των παραμέτρων που ονομάζεται «σύνολο λύσης» (*solution set*). Οι Fitch & Smith (1983) παρουσίασαν άλλες ακόμα 6 άριστες αντιστοιχίσεις από 6 διαφορετικά «σύνολα λύσης» με διαφορετικές τιμές των παραμέτρων το καθένα. Στην αντιστοίχιση (c) οι Griggs et al. (1986) παρατήρησαν τις αντιστοιχίες μεταξύ ομοίων και ανόμοιων βάσεων να συμβαίνουν σε *blocks* μεγέθους τουλάχιστον 3. Έτσι, η αντιστοίχιση αυτή μπορεί να θεωρηθεί ότι ανήκει στην κατηγορία για την οποία ισχύει $b\le 3$, ενώ για την περίπτωση αντιστοίχισης (a) ισχύει $b=1$. Οι περιπτώσεις (b) και (d) είναι αντίστοιχα η αναγωγή των περιπτώσεων (a) και (c) σε 0 και 1.

Μας ενδιαφέρει η ασυμπτωτική συμπεριφορά της συνάρτησης $g(b,n)$ για ορισμένες τιμές

του b καθώς $n \to \infty$. Αντιστοιχίσεις μεταξύ δύο ακολουθιών στις οποίες, μετά την αναγωγή τους σε 0 και 1, το άθροισμα κάθε στήλης ισούται με 1, δηλαδή αντιστοιχίσεις βάσης με μηδενικό στοιχείο, παραπέμπουν σε μεταθέσεις ακολουθιών μήκους n. Έτσι, εμφανίζονται n στήλες με 1 στην 1^{η} γραμμή και n στήλες με 1 στη 2^{η} γραμμή δημιουργώντας αντιστοιχίσεις μήκους $L = 2n$. Αυτή η περίπτωση είναι ικανοποιητική για οποιαδήποτε τιμή της παραμέτρου b και έτσι προκύπτει η ισχύς της σχέσης

$$g(b,n) \geq \binom{2n}{n}.$$

Εφαρμόζοντας στο συνδυασμό $\binom{2n}{n}$ τον τύπο του Stirling για σταθερή τιμή του b καθώς το $n \to \infty$, προκύπτει ότι (βλέπε Σχέση (2.2))

$$g(b,n) \geq (n\pi)^{-1/2}\left(4^n + o(1)\right),$$

όπου $o(1)$ παριστάνει μια μηδενική ακολουθία για $n \to \infty$.

Μια προσέγγιση για τη συνάρτηση $g(b,n)$ για $b \geq 1$ δίνεται μέσω του παρακάτω θεωρήματος των Griggs et al. (1986).

Θεώρημα 2.1

Έστω η συνάρτηση $g(b,n)$ που εκφράζει τον αριθμό των αντιστοιχίσεων δύο ακολουθιών ίσου μήκους n, όπου οι αντιστοιχίες μεταξύ των βάσεων εμφανίζονται σε *blocks* μεγέθους τουλάχιστον $b \geq 1$. Ορίζουμε τη συνάρτηση

$$h(x) = (1-x)^2 - 4x(x^b - x + 1)^2$$

και έστω $\rho = \min\{x : h(x) = 0\}$, η μικρότερη θετική ρίζα της εξίσωσης $h(x) = 0$. Τότε ισχύει

$$g(b,n) \sim \left(\gamma_b n^{-1/2}\right)\rho^{-n}, \text{ καθώς } n \to \infty$$

όπου $\gamma_b = (\rho^b - \rho + 1)(-\pi\rho h'(\rho))^{-1/2}$.

Παρατηρείται για τη συνάρτηση $h(x) = 1 - 6x + 9x^2 - 4x^3 - 8x^{b+1} + 8x^{b+2} - 4x^{2b+1}$ ότι $h(0) = 1$ και $h(\frac{1}{4}) = \left(\frac{3}{4}\right)^2 - \left(\left(\frac{1}{4}\right)^b + \frac{3}{4}\right)^2 < 0$. Έτσι, η συνάρτηση h σύμφωνα με το Θεώρημα Bolzano έχει τουλάχιστον μια πραγματική ρίζα στο διάστημα $(0, \frac{1}{4})$.

Από τον ορισμό της συνάρτησης $g(b,n)$ προκύπτει, για $b=1$, η ισότητα $g(1,n) = f(n,n)$, αφού τότε η συνάρτηση $g(1,n)$ υπολογίζει όλες τις δυνατές αντιστοιχίσεις μεταξύ 2 ακολουθιών. Η ισότητα αποδεικνύεται ακόμα με την εφαρμογή του *Θεωρήματος 2.1* για $b=1$. Η συνάρτηση $h(x)$ παίρνει τη μορφή: $h(x) = (1-x)^2 - 4x = x^2 - 6x + 1$ και μηδενίζεται για τις τιμές $x = 3 + 2\sqrt{2}$ ή $x = 3 - 2\sqrt{2}$. Η μικρότερη θετική ρίζα της εξίσωσης είναι η $\rho = 3 - 2\sqrt{2}$ και με αντικατάσταση στη σχέση $\gamma_b = (\rho^b - \rho + 1)(-\pi\rho h'(\rho))^{-1/2}$, όπου $h'(x) = 2x - 6$, προκύπτει η τιμή $\gamma_1 = 0.5727$. Το πλήθος των αντιστοιχίσεων 2 ακολουθιών μήκους n για $b=1$ προκύπτει ισοδύναμο με $g(1,n) \sim (0.5727)n^{-1/2}(5.828)^n$. Όμως ισχύει $f(n,n) \sim 2^{5/4}(n\pi)^{-1/2}(1+\sqrt{2})^{2n+1}$, όπου $(1+\sqrt{2})^{2n} = (5.828)^n$. Έτσι καταλήγουμε στην ισοδυναμία $g(1,n) = f(n,n)$.

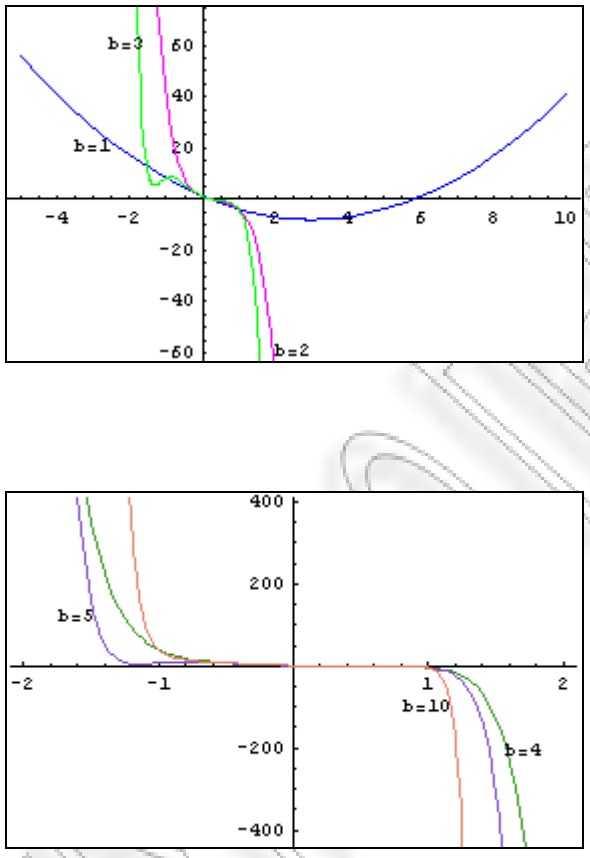
Λήμμα 2.1

Η μοναδική πραγματική και θετική ρίζα (*smallest modulus*) της συνάρτησης $h(x)$ είναι η ρ και μάλιστα η ρ είναι απλή ρίζα της $h(x)$.

Για διάφορες τιμές της παραμέτρου b, η συνάρτηση $h(x)$ παίρνει τις παρακάτω μορφές όπως φαίνεται από τις επόμενες γραφικές παραστάσεις μέσω *Mathematica*.

Σχήμα 2.2

Γραφικές παραστάσεις της συνάρτησης $h(x)$ για διάφορες τιμές της παραμέτρου b



Στον πίνακα που ακολουθεί εμφανίζονται οι τιμές των ρ και γ_b για διάφορες τιμές της παραμέτρου b.

Πίνακας 2.1

b	ρ	γ_b
1	0.1716	0.57268
2	0.2213	0.53206
3	0.2410	0.54290
4	0.2475	0.55520
5	0.2493	0.56109
10	0.2499	0.564183

Οι Griggs et al. (1986) από τις τιμές του Πίνακα 2.1 παρατήρησαν ότι καθώς το $b \to \infty$, η μικρότερη ρίζα ρ της $h(x)$ αυξάνεται και προσεγγίζει την τιμή $\frac{1}{4}$ και έτσι οδηγήθηκαν στο εξής πόρισμα.

Πόρισμα 2.1

Καθώς $b \to \infty$ προκύπτουν ότι $\rho \to \frac{1}{4}$ και $\gamma_b \to \pi^{-1/2}$.

Ο παρακάτω πίνακας περιέχει τις τιμές της συνάρτησης $g(b,n)$ για διάφορες τιμές των παραμέτρων b και n.

Πίνακας 2.2

Τιμές της συνάρτησης $g(b,n)$

n \ b	*1*	*5*	*9*	*10*
100	2.056×10^{75}	1.165×10^{59}	9.075×10^{58}	9.068×10^{58}
200	5.220×10^{151}	1.712×10^{119}	1.032×10^{119}	1.031×10^{119}
300	1.530×10^{228}	2.903×10^{179}	1.355×10^{179}	1.353×10^{179}
400	4.758×10^{304}	5.223×10^{239}	1.888×10^{239}	1.883×10^{239}
500	1.528×10^{381}	9.705×10^{299}	2.717×10^{299}	2.707×10^{299}
600	5.008×10^{457}	1.840×10^{360}	3.990×10^{359}	3.972×10^{359}
700	1.665×10^{534}	3.539×10^{420}	5.942×10^{419}	5.911×10^{419}
800	5.591×10^{610}	6.877×10^{480}	8.941×10^{479}	8.887×10^{479}
900	1.892×10^{687}	1.347×10^{541}	1.356×10^{540}	1.347×10^{540}
1000	6.445×10^{763}	6.654×10^{601}	2.069×10^{600}	2.054×10^{600}

Από τις τιμές του παραπάνω πίνακα παρατηρείται μείωση των τιμών της συνάρτησης $g(b,n)$ για σταθερές τιμές του μήκους n των ακολουθιών καθώς αυξάνει το μέγεθος των

blocks. Ακόμα παρατηρούνται οι τιμές της συνάρτησης $g(b,n)$ για $b=5$, $b=9$ και $b=10$ να πλησιάζουν μεταξύ τους για σταθερές τιμές του n. Έτσι, καθώς η τιμή του b αυξάνεται ο λόγος των τιμών της συνάρτησης $g(b,n)$ τείνει στη μονάδα.

2.3 Συνολική Στοίχιση με τη Μέθοδο της Απόστασης

Στις αρχές του 1970 αρκετοί Μαθηματικοί και ανάμεσα τους ο Stan Ulam (1972) εισήγαγαν την έννοια της απόστασης για την εύρεση της καλύτερης στοίχισης μεταξύ δύο ακολουθιών a και b. Στη Μαθηματική Ανάλυση, η απόσταση $D(x,y)$ ορίζεται ως μια συνάρτηση σε ένα μετρικό χώρο με τις εξής ιδιότητες:

1. $D(x,y)=0$ αν και μόνο αν $x=y$
2. $D(x,y)=D(y,x)$ (***Ιδιότητα της Συμμετρίας***)
3. $D(x,y)\le D(x,z)+D(z,y)$ για οποιαδήποτε x,y,z (***Τριγωνική Ανισότητα***).

Θεωρώντας τις δύο ακολουθίες a=$a_1a_2...a_n$ και b=$b_1b_2...b_m$, ορίζουμε ως $d(a,b)$ τη μετρική της απόστασης μεταξύ των βάσεων a και b από το σύνολο $\mathcal{B}=\{A,C,G,T\}$. Δεδομένης της $1^{ης}$ ιδιότητας της μετρικής, στην περίπτωση ομοίων βάσεων ισχύει $d(a,a)=0$, ενώ για ανόμοιες βάσεις ισχύει $d(a,b)>0$, όπου συνήθως επιλέγεται η τιμή $d(a,b)=1$. Στο χώρο των ακολουθιών, η απόσταση d αντιπροσωπεύει το κόστος μεταβολής της βάσης a στη βάση b κατά την αντιστοίχισή τους. Στην περίπτωση αντιστοίχισης βάσης με μηδενικό στοιχείο χρησιμοποιείται η συνάρτηση $g(a)$ που εκφράζει το θετικό κόστος προσθήκης ή αφαίρεσης μιας βάσης a σύμφωνα με τη σχέση $g(a)=d(a,-)=d(-,a)$.

Τότε η απόσταση μεταξύ των ακολουθιών a και b, λαμβάνοντας υπόψη όλες τις αντιστοιχίσεις μεταξύ των βάσεων, ορίζεται ως η ελάχιστη απόσταση στοίχισης με το μικρότερο σταθμικό άθροισμα (*weighted sum*) των αντικαταστάσεων (*mismatches*), των προσθηκών (*insertions*) και των αφαιρέσεων (*deletions*) βάσεων και προκύπτει από τη σχέση:

$$D(\mathbf{a},\mathbf{b})=\min\sum_{i=1}^{L}d(a_i^*,b_i^*).$$

Τα στοιχεία a_i^*, b_i^* για $i = 1,...,L$ αποτελούν τα στοιχεία των ακολουθιών a* και b* κατά την αντιστοίχιση

$$\begin{matrix} a_1^* & a_2^* ..... & a_L^* \\ b_1^* & b_2^* ..... & b_L^* \end{matrix}.$$

Αν η $d(a,b)$ είναι μια μετρική στο σύνολο $\mathcal{B} = \{A,C,G,T\}$, τότε και η απόσταση D(a,b) θα είναι μια μετρική στο χώρο των ακολουθιών και χαρακτηρίζεται ως απόσταση του Levenshtein (*Levenshtein distance*) ικανοποιώντας όλες τις παραπάνω ιδιότητες, Levenshtein (1965).

Το θεώρημα που ακολουθεί οφείλεται στον Sellers (1974) και δίνει έναν τρόπο υπολογισμού της ελάχιστης ολικής απόστασης.

Θεώρημα 2.2

Έστω οι ακολουθίες a$=a_1a_2...a_n$ και b$=b_1b_2...b_m$. Θέτουμε τις αρχικές συνθήκες

$$D_{0,0} = 0 \quad , \quad D_{0,j} = \sum_{k=1}^{j} d(-,b_k) \quad , \quad D_{i,0} = \sum_{k=1}^{i} d(a_k,-) .$$

Αν η απόσταση μεταξύ των τμημάτων $a_1,...,a_i$ και $b_1,...,b_j$ των ακολουθιών a και b αντίστοιχα ορίζεται από τη σχέση

$$D_{i,j} = D(a_1a_2...a_i, b_1b_2...b_j) \text{, για } 0 \le i \le n \text{ και } 0 \le j \le m$$

τότε υπολογίζεται από τον αναδρομικό τύπο

$$D_{i,j} = \min\begin{Bmatrix} D_{i-1,j} + d(a_i,-), \\ D_{i-1,j-1} + d(a_i,b_j), \\ D_{i,j-1} + d(-,b_j) \end{Bmatrix}.$$

Η ελάχιστη ολική απόσταση (*minimum global distance*) στοίχισης μεταξύ των δύο ακολουθιών είναι ίση με $D_{n,m}$.

Απόδειξη

Η στοίχιση των τμημάτων $a_1a_2...a_i$ και $b_1b_2...b_j$ δύο ακολουθιών a και b θα καταλήγει σε μία από τις τρεις περιπτώσεις

$$\begin{matrix} ...a_i & ...a_i & ...- \\ ...- & ...b_j & ...b_j \end{matrix}.$$

Αν η βέλτιστη στοίχιση καταλήγει σε $\begin{pmatrix} a_i \\ - \end{pmatrix}$, τότε το κόστος στοίχισης είναι $D_{i-1,j} + d(a_i,-)$ μιας και η στοίχιση για τα τμήματα $a_1a_2...a_{i-1}$ και $b_1b_2...b_j$ θα είναι βέλτιστη.

Αν η βέλτιστη στοίχιση καταλήγει σε $\begin{pmatrix} a_i \\ b_j \end{pmatrix}$, τότε το κόστος στοίχισης είναι $D_{i-1,j-1} + d(a_i,b_j)$ μιας και η στοίχιση για τα τμήματα $a_1a_2...a_{i-1}$ και $b_1b_2...b_{j-1}$ θα είναι βέλτιστη.

Αν η βέλτιστη στοίχιση καταλήγει σε $\begin{pmatrix} - \\ b_j \end{pmatrix}$, τότε το κόστος στοίχισης είναι $D_{i,j-1} + d(-,b_j)$ μιας και η στοίχιση για τα τμήματα $a_1a_2...a_i$ και $b_1b_2...b_{j-1}$ θα είναι βέλτιστη.

Για να επιτύχουμε την καλύτερη δυνατή στοίχιση (*optimal alignment*) χρειάζεται να βρούμε το μικρότερο κόστος των τριών περιπτώσεων και έτσι αποδεικνύεται το θεώρημα του Sellers (1974). □

Ο Αλγόριθμος 2.1 είναι η διατύπωση του θεωρήματος με σταθερό κόστος αντιστοίχισης των βάσεων των ακολουθιών με μηδενικά στοιχεία ίσο με $\delta = d(a_i,-) = d(-,b_j)$.

Αλγόριθμος 2.1

Βήμα 1 : Θέτουμε $D_{1,1} = 0$,

$D_{i,1} = (i-1)\delta$ για $1 \le i \le n+1$ και

$D_{1,j} = (j-1)\delta$ για $1 \le j \le m+1$.

Βήμα 2 : Υπολογίζουμε κάθε κελί (i,j) ($i \ge 2$ & $j \ge 2$) του πίνακα από τη σχέση

$$D_{i,j} = \min\{D_{i-1,j} + \delta, D_{i-1,j-1} + d(a_i,b_j), D_{i,j-1} + \delta\}.$$

Βήμα 3 : Επαναλαμβάνουμε το *Βήμα 2* για $2 \le i \le n+1$ και $2 \le j \le m+1$.

Η εύρεση της άριστης στοίχισης μεταξύ δύο ακολουθιών επιτυγχάνεται με χρήση του παραπάνω αλγορίθμου του οποίου ο χρόνος περάτωσης είναι της τάξης $O(nm)$ ή $O(n^2)$ αν $n = m$. Στον αλγόριθμο ορίζουμε τις δύο ακολουθίες a και b, την απόσταση $d(a,b)$ και το κόστος αντιστοίχισης των βάσεων των ακολουθιών με μηδενικά στοιχεία ίσο με δ υπολογίζοντας έτσι τον πίνακα της απόστασης D.

Μια υλοποίηση του παραπάνω αλγορίθμου στο *Mathematica* για τις ακολουθίες $\mathrm{a} = \mathrm{GCTGATATAGCT}$ και $\mathrm{b} = \mathrm{GGGTGATTAGCT}$ με τιμές των παραμέτρων $\delta = 1$, για την εμφάνιση *indel*, και $\mu = 1$, για την αντιστοίχιση μη-ομοίων βάσεων, είναι η ακόλουθη:

```
a = {"G", "C", "T", "G", "A", "T", "A", "T", "A", "G", "C", "T"};
b = {"G", "G", "G", "T", "G", "A", "T", "T", "A", "G", "C", "T"};
aa = Prepend[a, "-"];
bb = Prepend[b, "-"];
n = Length[a];
m = Length[b];
nn = Length[aa];
mm = Length[bb];
D1 = Table[0, {i, 1, nn}, {j, 1, mm}];
d = Table[0, {i, 1, nn}, {j, 1, mm}];
delta = 1;
Do[
  If[aa[[i]] == bb[[j]], x = 0, x = 1]; d[[i, j]] = x, {i, 1, nn}, {j, 1, mm}];
Do[D1[[1, j]] = (j - 1) * delta, {j, 2, mm}];
Do[D1[[i, 1]] = (i - 1) * delta, {i, 2, nn}];
Do[
  Do[
   D1[[i, j]] = Min[D1[[i - 1, j]] + delta, D1[[i - 1, j - 1]] + d[[i, j]],
     D1[[i, j - 1]] + delta],
   {j, 2, mm}], {i, 2, nn}];
D = TableForm[D1, TableHeadings → {aa, bb}]
```

Στο *Σχήμα 2.3* που ακολουθεί, δίνεται ο πίνακας D με ελάχιστη ολική απόσταση στοίχισης ίση με $D_{n,m} = 3$. Οι υπογραμμισμένοι αριθμοί παραπέμπουν στα ζεύγη βάσεων που αντιστοιχίζονται για την εύρεση της άριστης αντιστοίχισης (*optimal alignment*) των δύο ακολουθιών. Η άριστη στοίχιση, όπως φαίνεται στο *Σχήμα 2.4*, επιτυγχάνεται με μια διαδικασία που θα περιγραφεί αναλυτικά στην Παράγραφο 2.8. Η διαδικασία ονομάζεται

tracebacks και αποτυπώνει κάθε φορά με βήματα προς τα πίσω το ένα εκ των 3 ζευγών βάσεων που περιλήφθηκε στον υπολογισμό της τιμής $D_{i,j}$. Ξεκινώντας από το τελευταίο ζεύγος $(n+1, m+1)$ του πίνακα, τα 3 ζεύγη που συμμετείχαν στον υπολογισμό της τιμής $D_{n+1,m+1}$ είναι τα $(n, m+1)$, $(n+1, m)$ και (n, m). Όμως τελικά η τιμή $D_{n+1,m+1}$ προκύπτει από τη συμμετοχή του ζεύγους (n, m) και έτσι προκύπτει η αντιστοίχιση του ζεύγους $\binom{C}{C}$. Η τιμή κάθε κελιού (i, j) προκύπτει από την επιλογή του ενός από τα κελιά $(i-1, j)$, $(i, j-1)$ και $(i-1, j-1)$. Σε περίπτωση που υπάρξουν περισσότερες από μια επιλογές, που δίνουν την ίδια τιμή, τότε οι πιθανές θέσεις τοποθετούνται σε έναν σωρό (*stack*) και ακολουθείται η αρχή *last in-first out* για την αποτύπωση των άριστων αντιστοιχίσεων.

Σχήμα 2.3

Στοίχιση με τη μέθοδο της απόστασης

		G	G	G	T	G	A	T	T	A	G	C	T
	0	**1**	2	3	4	5	6	7	8	9	10	11	12
G	1	0	**1**	2	3	4	5	6	7	8	9	10	11
C	2	1	1	**2**	3	4	5	6	7	8	9	9	10
T	3	2	2	2	**2**	3	4	5	6	7	8	9	9
G	4	3	2	2	3	**2**	3	4	5	6	7	8	9
A	5	4	3	3	3	3	**2**	3	4	5	6	7	8
T	6	5	4	4	3	4	3	**2**	3	4	5	6	7
A	7	6	5	5	4	4	4	3	**3**	3	4	5	6
T	8	7	6	6	5	5	5	4	3	4	4	5	5
A	9	8	7	7	6	6	5	5	4	**3**	4	5	6
G	10	9	8	7	7	6	6	6	5	4	**3**	4	5
C	11	10	9	8	8	7	7	7	6	5	4	**3**	4
T	12	11	10	9	8	8	8	7	7	6	5	4	**3**

Σχήμα 2.4

Άριστη στοίχιση

–	G	C	T	G	A	T	A	T	A	G	C	T
	|		|	|	|	|		|	|	|	|	|
G	G	G	T	G	A	T	–	T	A	G	C	T

2.3.1 Συναρτήσεις που αφορούν στοίχιση Βάσης-Μηδενικού Στοιχείου

Συχνά μετά τη στοίχιση δύο ακολουθιών μπορεί *n* βάσεις που βρίσκονται σε σειρά να αντιστοιχίζονται με μηδενικά στοιχεία "-" (*gaps*) σχηματίζοντας 1 *indel* μεγέθους *n*. Σύμφωνα με τον αλγόριθμο του Sellers (1974), η αφαίρεση (ή προσθήκη) βάσεων είναι αποτέλεσμα αρκετών ξεχωριστών αφαιρέσεων (ή προσθηκών). Στη Βιολογία όμως, η αφαίρεση (ή προσθήκη) πολλαπλών βάσεων κατά την εξέλιξη ενός οργανισμού μπορεί να αποτελεί ένα γεγονός και να μην είναι άθροισμα ξεχωριστών αφαιρέσεων (ή προσθηκών). Έτσι, κρίνεται απαραίτητος ο ορισμός συναρτήσεων που εξαρτώνται από το μέγεθος των *indels*. Συγκεκριμένα, ορίζουμε τη συνάρτηση $g(k): N \to \Re$ που εκφράζει τη στάθμη στοίχισης k συνεχόμενων βάσεων από το σύνολο $\mathcal{B} = \{A,C,G,T\}$ με το μηδενικό στοιχείο. Για τη συνάρτηση αυτή ισχύει η υποπροσθετική συνθήκη (*subadditivity condition*) $g(k) \le kg(1)$, έτσι η προσθήκη ή αφαίρεση ενός *block* βάσεων είναι πιο πιθανή από την αθροιστική προσθήκη ή αφαίρεση ξεχωριστών βάσεων του *block*, Lange (2002). Μια γενίκευση του *Θεωρήματος 2.2* αποτελεί ο Αλγόριθμος WSB μέσω του επόμενου θεωρήματος, Waterman, Smith & Beyer (1976).

Θεώρημα 2.3

Έστω $g(k): \mathrm{N} \to \Re$ ο συντελεστής (βάρος) εμφάνισης k συνεχόμενων *indels* και $d(a,b)$ το κόστος μεταβολής της βάσης a της ακολουθίας $\mathrm{a} = a_1 ... a_n$ στη βάση b της ακολουθίας $\mathrm{b} = b_1 ... b_m$ κατά την διαδικασία της στοίχισης. Τότε για δύο ακολουθίες ορίζουμε την απόσταση

$$D_{i,j} = D(a_1 ... a_i, b_1 .. b_j)$$

για την οποία ισχύουν τα εξής

$$D_{0,0} = 0$$
$$D_{0,j} = g(j)$$
$$D_{i,0} = g(i)$$

$$D_{i,j} = \min\begin{Bmatrix} \min\{D_{i,j-k} + g(k) : 1 \le k \le j\} \\ D_{i-1,j-1} + d(a_i, b_j) \\ \min\{D_{i-k,j} + g(k) : 1 \le k \le i\} \end{Bmatrix}.$$

Ο χρόνος υπολογισμού του παραπάνω αλγορίθμου είναι τάξης

$$\sum_{i=1}^{n}\sum_{j=1}^{m}(i+j) = \sum_{i=1}^{n}\binom{m+1}{2} + \sum_{j=1}^{m}\binom{n+1}{2} = \frac{n(m+1)m}{2} + \frac{m(n+1)n}{2} = O(n^2 m + nm^2)$$

και $O(n^3)$ σε περίπτωση ακολουθιών ίσου μήκους ($n = m$), Lange (2002).

Ο χρόνος αυτός μπορεί να μειωθεί για διαφορετικές μορφές της συνάρτησης $g(k)$. Σύμφωνα με τον Gotoh (1982), η εφαρμογή του αλγορίθμου για γραμμικές συναρτήσεις (*linear functions*) της μορφής $g(k) = \alpha + \beta(k-1)$, όπου α, β : σταθεροί αριθμοί, μειώνει το χρόνο υπολογισμού σε $O(nm)$ ή $O(n^2)$ για $n = m$. Η εισαγωγή του νέου αλγορίθμου προϋποθέτει τη χρήση τριών πινάκων αντί του ενός και εκφράζεται μέσω του θεωρήματος που ακολουθεί.

Θεώρημα 2.4

Έστω $g(k) = \alpha + \beta(k-1)$ η γραμμική συνάρτηση βάρους εμφάνισης k συνεχόμενων *indels (affine gap penalty)* και $d(a,b)$ το κόστος μεταβολής της βάσης a της ακολουθίας $\mathrm{a} = a_1 \ldots a_n$ στη βάση b της ακολουθίας $\mathrm{b} = b_1 \ldots b_n$. Ορίζουμε τους πίνακες $(E_{i,j})$, $(F_{i,j})$ και $(D_{i,j})$ για τους οποίους ισχύουν τα εξής:

$$\begin{aligned} &D_{0,0} = 0 \\ &D_{0,j} = g(j), \forall j \\ &D_{i,0} = g(i), \forall i \end{aligned} \qquad \begin{aligned} &F_{0,j} = \infty, \forall j \\ &E_{i,0} = \infty, \forall i \end{aligned} \qquad \begin{aligned} &F_{i,j} = \min\{D_{i-1,j} + \alpha, F_{i-1,j} + \beta\} \\ &E_{i,j} = \min\{D_{i,j-1} + \alpha, E_{i,j-1} + \beta\}. \end{aligned}$$

Τότε $D_{i,j} = \min\begin{Bmatrix} F_{i,j} \\ D_{i-1,j-1} + d(a_i, b_j) \\ E_{i,j} \end{Bmatrix}$.

Απόδειξη

Αρκεί να αποδειχτούν οι ισότητες που αφορούν τους πίνακες $(E_{i,j})$ και $(F_{i,j})$ για $1 \le i \le n$ και $1 \le j \le m$.

Σύμφωνα με το *Θεώρημα 2.3* για τους δύο πίνακες ισχύουν

$$E_{i,j} = \min\{D_{i,j-k} + g(k) : 1 \le k \le j\}$$
$$F_{i,j} = \min\{D_{i-l,j} + g(l) : 1 \le l \le i\}.$$

Για την πρώτη ισότητα έχουμε, κάνοντας χρήση της ιδιότητας $g(k+1) = g(k) + \beta$ που ισχύει για τη γραμμική συνάρτηση βάρους που χρησιμοποιούμε:

$$\begin{aligned}
E_{i,j} &= \min\{D_{i,j-k} + g(k) : 1 \le k \le j\} \\
&= \min\{D_{i,j-1} + g(1), \min\{D_{i,j-k} + g(k) : 2 \le k \le j\}\} \\
&= \min\{D_{i,j-1} + \alpha, \min\{D_{i,j-k} + g(k) : 1 \le k-1 \le j-1\}\} \\
&= \min\{D_{i,j-1} + \alpha, \min\{D_{i,j-l-1} + g(l+1) : 1 \le l \le j-1\}\} \\
&= \min\{D_{i,j-1} + \alpha, \min\{D_{i,(j-1)-l} + g(l) + \beta : 1 \le l \le j-1\}\} \\
&= \min\{D_{i,j-1} + \alpha, \min\{D_{i,(j-1)-l} + g(l) : 1 \le l \le j-1\} + \beta\} \\
&= \min\{D_{i,j-1} + \alpha, E_{i,j-1} + \beta\}.
\end{aligned}$$

Όμοια αποδεικνύεται και η δεύτερη ισότητα. □

Η συνάρτηση $g(k)$ μπορεί ακόμη να είναι κοίλη (*concave*) και τότε ισχύει η συνθήκη $g(m+k+l) - g(m+k) \le g(k+l) - g(k)$ για $m,k,l \ge 0$. Μια μορφή αυτής της συνάρτησης μπορεί να είναι η εξής $g(k) = \alpha + \beta \log(k)$ με την οποία ασχολήθηκαν οι Waterman (1984a) και Myers & Miller (1988a) που παρουσίασαν έναν πολύπλοκο αλγόριθμο αλλά μείωσαν το χρόνο υπολογισμού σε $O(n^2 \log(n))$. Οι Galil & Giancarlo (1989) επεκτάθηκαν και σε αλγορίθμους στην περίπτωση κυρτής συνάρτησης (*convex*) $g(k)$. Για αυτής της μορφής τις συναρτήσεις ισχύει η αντίστροφη συνθήκη $g(m+k+l) - g(m+k) \ge g(k+l) - g(k)$ για $m,k,l \ge 0$. Ωστόσο, η χρήση των κυρτών συναρτήσεων δεν αρμόζει στη Βιολογία για τη σύγκριση των ακολουθιών, Waterman (1995).

Στον Αλγόριθμο του Sellers η τιμή κάθε κελιού (i, j) προκύπτει από τον έλεγχο των τιμών

3 κελιών, $(i-1, j-1)$, $(i-1, j)$ και $(i, j-1)$, ενώ στην περίπτωση εισαγωγής πολλαπλών *indels* με χρήση της συνάρτησης $g(k)$ συνολικά ελέγχονται οι τιμές $(i+j+1)$ κελιών, της $i-$γραμμής, της $j-$στήλης και η τιμή του κελιού $(i-1, j-1)$. Με τους τρόπους αυτούς, η άριστη αντιστοίχιση μεταξύ των ακολουθιών μπορεί να θεωρηθεί ως άθροισμα πολλών διαδοχικών βημάτων μέχρι την κατάληξη στο ζεύγος βάσεων (i, j). Υπάρχει ένας ακόμη τρόπος που επιτυγχάνει τη συντομότερη πορεία (*shortest path*) με τη βοήθεια μιας εναλλακτικής συνάρτησης $\hat{g}$ που δίνεται από τη σχέση

$$\hat{g}(k) = \min\left\{\sum_{i=1}^{k} g(l_i) : \sum l_i = k, 0 \le l_i \le k\right\}.$$

Τότε η τιμή κάθε κελιού (i, j) προκύπτει από τον έλεγχο των τιμών μιας ορθογώνιας περιοχής κελιών $(i+1)(j+1)-1$, με εξαίρεση το κελί (i, j).
Για τη συνάρτηση αυτή ισχύει η υποπροσθετική ανισότητα (*subadditivity condition*)

$$\hat{g}(k+l) \le \hat{g}(k) + \hat{g}(l) \text{ , για } k, l \ge 0 .$$

Αν μας ενδιαφέρει η γραμμική συνάρτηση $g(k) = \alpha + \beta(k-1)$ με $0 \le \alpha \le \beta$, τότε η συνάρτηση που προτιμάται να χρησιμοποιηθεί είναι η $\hat{g}(k) = \alpha k$, ενώ αν ισχύει $\alpha > \beta$, τότε $\hat{g}(k) = g(k)$. Η συνάρτηση $\hat{g}$ δεν μπορεί να θεωρηθεί ότι είναι μονότονη ή κοίλη αφού όταν $g(k) = k$, με $k \ne 3$ και $g(3) = 0$ τότε προκύπτει $\hat{g}(k) = k \bmod 3$, Waterman (1995).

Για την εύρεση των τιμών της $\hat{g}$ χρησιμοποιείται ο επόμενος αλγόριθμος:

Αλγόριθμος 2.2

Βήμα 1 : Ορίζουμε τη συνάρτηση $g(k) = \alpha + \beta(k-1)$ και τα μήκη των ακολουθιών, *n* και *m*.

Βήμα 2 : Αναθέτουμε την τιμή της συνάρτησης g στην $\hat{g}$, $\hat{g}(k) = g(k)$, ξεκινώντας από την τιμή $k = 1$.

Βήμα 3 : Η τιμή της $\hat{g}$ προκύπτει από τη σχέση $\hat{g}(k) = \min\{\hat{g}(i) + g(k-i), \hat{g}(k)\}$, για $i = 1, \ldots, k-1$.

Βήμα 4 : Επαναλαμβάνουμε τα *Βήματα 2,3* για $1 \le k \le \max\{n, m\}$.

Μια υλοποίηση του παραπάνω αλγορίθμου στο *Mathematica* για τη συνάρτηση $g(k) = k$ με $k \neq 3$ και $g(3) = 0$, επιλέγοντας για τα μήκη τις τιμές $n = 13$ και $m = 14$, οδηγεί στις τιμές της συνάρτησης $\hat{g}(k) = k \bmod 3$.

```
a = 1; b = 1;
g[k_] = a + b * (k - 1);
k ≠ 3;
g[3] = 0;
n = 13;
m = 14;
M = Max[n, m];
t = {};
Do[gnew[k] = g[k];
  Do[gnew[k] = Min[gnew[i] + g[k - i], gnew[k]], {i, 1, k - 1}];
  AppendTo[t, gnew[k]], {k, 1, M}];
Print[t]
```

2.3.2 Βάρη εξαρτώμενα από τις θέσεις των ζευγών Βάσεων των Ακολουθιών

Τα βάρη (*weights*) που αναθέτουμε στην αντιστοίχιση μη-ομοίων βάσεων μεταξύ δύο ακολουθιών μπορεί να εξαρτώνται από τις θέσεις των βάσεων στις ακολουθίες. Μια τροποποίηση του Θεωρήματος του Gotoh για γραμμικές συναρτήσεις εμφάνισης *gaps* είναι η εξής:

$$F_{i,j} = \min\{D_{i-1,j} + \alpha_i, F_{i-1,j} + \beta_i\}$$
$$E_{i,j} = \min\{D_{i,j-1} + \gamma_j, E_{i,j-1} + \delta_j\}$$

και τότε η απόσταση μεταξύ των ακολουθιών ορίζεται ως εξής

$$D_{i,j} = \min\{D_{i-1,j-1} + s_{i,j}(a_i, b_j), F_{i,j}, E_{i,j}\}.$$

Εδώ η συνάρτηση $s_{i,j}(a_i, b_j)$ αντιπροσωπεύει το κόστος αντιστοίχισης δύο διαφορετικών βάσεων a_i, b_j και εξαρτάται από τις θέσεις τους (i, j) στις ακολουθίες a και b αντίστοιχα. Στη θέση i της ακολουθίας a, το κόστος είναι α_i, αν η αφαίρεση της βάσης a_i προκαλεί τη

στοίχιση βάσης με μηδενικό στοιχείο και β_i , αν επεκτείνεται η αντιστοιχία με μηδενικό στοιχείο. Ομοίως, ορίζονται τα βάρη με την αφαίρεση της βάσης b_j της ακολουθίας b.

Σε αυτήν την περίπτωση μεταβλητού βάρους είναι προφανής η ισχύς της σχέσης $D(a_1...a_n, b_1...b_m) \neq D(a_n...a_1, b_m...b_1)$ για τις ακολουθίες a και b και τις αντίστροφες τους. Αυτό το πρόβλημα διορθώνεται αν ισχύουν $\alpha_i = \beta_i$ και $\gamma_j = \delta_j$ για όλα τα i, j ή αν όλες οι παράμετροι $\alpha_i, \beta_i, \gamma_j, \delta_j$ είναι σταθερές.

Εφαρμόζοντας μεταβλητά βάρη κατά την στοίχιση δύο ακολουθιών προκύπτουν αλγόριθμοι τάξης $O(n^2)$, Gribskov et al. (1987).

2.4 Συνολική Στοίχιση με τη Μέθοδο της Ομοιότητας

Ο πρώτος αλγόριθμος αντιστοίχισης δύο ακολουθιών παρουσιάστηκε από τους Needleman & Wunsch (1970). Ο αλγόριθμος αυτός αποτέλεσε έμπνευση για τον Sellers (1974) που εισήγαγε τη Μέθοδο της Απόστασης όπως περιγράφηκε στην Παράγραφο 2.2. Στη συγκεκριμένη μέθοδο χρησιμοποιείται μια συνάρτηση $s(a,b)$ που ονομάζεται συνάρτηση ομοιότητας (*similarity function*). Η συνάρτηση αυτή μετρά την ομοιότητα μεταξύ των βάσεων κατά τη στοίχιση και παίρνει θετικές τιμές στην περίπτωση ταυτόσημων βάσεων, $s(a,a) > 0$, και αρνητικές τιμές όταν οι βάσεις είναι διαφορετικές, $s(a,b) < 0$ όταν $a \neq b$. Η αντιστοίχιση βάσης με μηδενικό στοιχείο σταθμίζεται αρνητικά, έτσι ώστε $s(a,-) = s(-,a) = -h(a)$.

Τότε η ομοιότητα (*similarity*) μεταξύ των ακολουθιών a και b ορίζεται από τη σχέση:

$$S(\mathrm{a},\mathrm{b}) = \max \sum_{i=1}^{L} s(a_i^*, b_i^*),$$

όπου το μέγιστο λαμβάνεται για όλες τις δυνατές αντιστοιχίσεις μεταξύ των ακολουθιών.

Ο αλγόριθμος των Needleman-Wunsch για την εύρεση της βέλτιστης αντιστοίχισης (*optimal alignment*) των δύο ακολουθιών εκφράζεται μέσω του *Θεωρήματος 2.5*.

Θεώρημα 2.5

Έστω οι ακολουθίες a=$a_1a_2...a_n$ και b=$b_1b_2...b_m$. Θέτουμε τις αρχικές συνθήκες

$$S_{0,0} = 0 \quad , \quad S_{0,j} = \sum_{k=1}^{j} s(-, b_k) \quad \text{και} \quad S_{i,0} = \sum_{k=1}^{i} s(a_k, -) .$$

Αν η ομοιότητα μεταξύ των τμημάτων $a_1,...,a_i$ και $b_1,...,b_j$ των ακολουθιών a και b αντίστοιχα ορίζεται από τη σχέση

$$S_{i,j} = S(a_1a_2...a_i, b_1b_2...b_j) \text{, για } 0 \le i \le n \text{ και } 0 \le j \le m$$

τότε υπολογίζεται από τον αναδρομικό τύπο

$$S_{i,j} = \max\begin{Bmatrix} S_{i-1,j} + s(a_i, -), \\ S_{i-1,j-1} + s(a_i, b_j), \\ S_{i,j-1} + s(-, b_j) \end{Bmatrix}.$$

Η μέγιστη ολική ομοιότητα (*maximum global similarity*) στοίχισης μεταξύ των δύο ακολουθιών είναι ίση με $S_{n,m}$.

Ο *Αλγόριθμος 2.3* είναι η διατύπωση του θεωρήματος με σταθερή τιμή κόστους αντιστοίχισης των βάσεων των ακολουθιών με μηδενικά στοιχεία ίση με $s(a_i, -) = s(-, b_j) = -\hat{\delta}$.

Αλγόριθμος 2.3

Βήμα 1 : Θέτουμε $S_{1,1} = 0$,

$S_{i,1} = -(i-1)\hat{\delta}$ για $1 \le i \le n+1$ και

$S_{1,j} = -(j-1)\hat{\delta}$ για $1 \le j \le m+1$.

Βήμα 2 : Υπολογίζουμε κάθε κελί (i, j) ($i \geq 2$ & $j \geq 2$) του πίνακα από τη σχέση

$$S_{i,j} = \max\left\{S_{i-1,j} - \hat{\delta}, S_{i-1,j-1} + s(a_i, b_j), S_{i,j-1} - \hat{\delta}\right\}.$$

Βήμα 3 : Επαναλαμβάνουμε το *Βήμα 2* για $2 \leq i \leq n+1$ και $2 \leq j \leq m+1$.

Μια υλοποίηση του *Αλγορίθμου 2.3* μέσω *Mathematica* στις ακολουθίες a = GCTGATATAGCT και b = GGGTGATTAGCT με τιμές της συνάρτησης ομοιότητας $s(a,a) = 1$, $s(a,b) = -1$ για $a \neq b$ και τιμή της παραμέτρου $\hat{\delta} = 2$ δίνεται παρακάτω με πίνακα ομοιότητας S που παρουσιάζεται στο *Σχήμα 2.5*.

```
a = {"G", "C", "T", "G", "A", "T", "A", "T", "A", "G", "C", "T"};
b = {"G", "G", "G", "T", "G", "A", "T", "T", "A", "G", "C", "T"};
aa = Prepend[a, "-"]
bb = Prepend[b, "-"]
n = Length[a];
m = Length[b];
nn = Length[aa];
mm = Length[bb];
delta = 2;
S1 = Table[0, {i, 1, nn}, {j, 1, mm}];
s = Table[0, {i, 1, nn}, {j, 1, mm}];
Do[
  If[aa[[i]] == bb[[j]], x = 1, x = -1]; s[[i, j]] = x, {i, 1, nn}, {j, 1, mm}];
Do[S1[[1, j]] = -(j - 1) * delta, {j, 2, mm}];
Do[S1[[i, 1]] = -(i - 1) * delta, {i, 2, nn}];
Do[
  Do[
   S1[[i, j]] = Max[S1[[i - 1, j]] - delta, S1[[i - 1, j - 1]] + s[[i, j]],
     S1[[i, j - 1]] - delta],
   {j, 2, mm}], {i, 2, nn}];
S = TableForm[S1, TableHeadings → {aa, bb}]
```

Σχήμα 2.5

Στοίχιση με τη μέθοδο της ομοιότητας

	-	*G*	*G*	*G*	*T*	*G*	*A*	*T*	*T*	*A*	*G*	*C*	*T*
-	0	**-2**	-4	-6	-8	-10	-12	-14	-16	-18	-20	-22	-24
G	-2	1	**-1**	-3	-5	-7	-9	-11	-13	-15	-17	-19	-21
C	-4	-1	0	**-2**	-4	-6	-8	-10	-12	-14	-16	-16	-18
T	-6	-3	-2	-1	**-1**	-3	-5	-7	-9	-11	-13	-15	-15
G	-8	-5	-2	-1	-2	**0**	-2	-4	-6	-8	-10	-12	-14
A	-10	-7	-4	-3	-2	-2	**1**	-1	-3	-5	-7	-9	-11
T	-12	-9	-6	-5	-2	-3	-1	**2**	0	-2	-4	-6	-8
A	-14	-11	-8	-7	-4	-3	-2	0	**1**	1	-1	-3	-5
T	-16	-13	-10	-9	-6	-5	-4	-1	**1**	0	0	-2	-2
A	-18	-15	-12	-11	-8	-7	-4	-3	-1	**2**	0	-1	-3
G	-20	-17	-14	-11	-10	-7	-6	-5	-3	0	**3**	1	-1
C	-22	-19	-16	-13	-12	-9	-8	-7	-5	-2	1	**4**	2
T	-24	-21	-18	-15	-12	-11	-10	-7	-6	-4	-1	2	**5**

Η άριστη στοίχιση προκύπτει με τον ίδιο τρόπο που περιγράφηκε στη Μέθοδο Απόστασης και δίνεται στο σχήμα που ακολουθεί, όπως παρουσιάστηκε στο *Σχήμα 2.4*.

Σχήμα 2.6

Άριστη στοίχιση

–	*G*	*C*	*T*	*G*	*A*	*T*	*A*	*T*	*A*	*G*	*C*	*T*
	|		|	|	|	|		|	|	|	|	|
G	*G*	*G*	*T*	*G*	*A*	*T*	–	*T*	*A*	*G*	*C*	*T*

Όμοια με τη Μέθοδο της Απόστασης, στην περίπτωση εμφάνισης πολλαπλών *indels* κατά την αντιστοίχιση ακολουθιών, ισχύει το επόμενο θεώρημα.

Θεώρημα 2.6

Έστω $h(k): \mathrm{N} \to \Re$ ο συντελεστής (βάρος) εμφάνισης k συνεχόμενων *indels* και $s(a,b)$ το μέτρο ομοιότητας της βάσης a της ακολουθίας $\mathrm{a} = a_1 \ldots a_n$ με τη βάση b της ακολουθίας $\mathrm{b} = b_1 \ldots b_m$. Τότε για δύο ακολουθίες ορίζουμε την ομοιότητα

$$S_{i,j} = S(a_1 ... a_i, b_1 .. b_j)$$

για την οποία ισχύουν οι σχέσεις

$$\begin{aligned}
&S_{0,0} = 0 \\
&S_{0,j} = h(j) \\
&S_{i,0} = h(i) \\
&S_{i,j} = \max\begin{Bmatrix} \max\{S_{i,j-k} + h(k) : 1 \le k \le j\} \\ S_{i-1,j-1} + s(a_i, b_j) \\ \max\{S_{i-k,j} + h(k) : 1 \le k \le i\} \end{Bmatrix}.
\end{aligned}$$

Ο χρόνος υπολογισμού του αλγορίθμου μειώνεται για γραμμικές συναρτήσεις της μορφής $g(k) = \alpha + \beta(k-1)$, όπου α, β σταθεροί αριθμοί. Όμοια με τη Μέθοδο της Απόστασης ο νέος αλγόριθμος εκφράζεται με το θεώρημα που ακολουθεί.

Θεώρημα 2.7

Έστω $g(k) = \alpha + \beta(k-1)$ η γραμμική συνάρτηση βάρους εμφάνισης k συνεχόμενων *indels (affine gap penalty)* και $s(a,b)$ το μέτρο ομοιότητας μιας βάσης a της πρώτης ακολουθίας με μια βάση b της δεύτερης ακολουθίας. Ορίζουμε τους πίνακες $(E_{i,j})$, $(F_{i,j})$ και $(S_{i,j})$ για τους οποίους ισχύουν τα εξής:

$$\begin{aligned}
&S_{0,0} = 0 \\
&S_{0,j} = -h(j), \forall j \\
&S_{i,0} = -h(i), \forall i
\end{aligned}
\qquad
\begin{aligned}
&F_{0,j} = -\infty, \forall j \\
&E_{i,0} = -\infty, \forall i
\end{aligned}
\qquad
\begin{aligned}
&F_{i,j} = \max\{S_{i-1,j} - \alpha, F_{i-1,j} - \beta\} \\
&E_{i,j} = \max\{S_{i,j-1} - \alpha, E_{i,j-1} - \beta\}.
\end{aligned}$$

Τότε η ομοιότητα των ακολουθιών ισούται με

$$S_{i,j} = \max\begin{Bmatrix} F_{i,j} \\ S_{i-1,j-1} + s(a_i, b_j) \\ E_{i,j} \end{Bmatrix}.$$

Οι δύο μέθοδοι αντιστοίχισης μεταξύ δύο ακολουθιών, που περιγράφηκαν με τους αλγορίθμους Needleman-Wunsch και Sellers είναι ισοδύναμοι. Η ισοδυναμία αυτή αποδεικνύεται με το θεώρημα που ακολουθεί και οφείλεται στους Smith, Waterman & Fitch (1981).

Θεώρημα 2.8

Έστω $d(a,b)$ η μετρική που εκφράζει την απόσταση μεταξύ δύο βάσεων a και b με συνάρτηση βάρους εμφάνισης k συνεχόμενων *indels* $g(k)$ και $s(a,b)$ το μέτρο ομοιότητας με *score* εμφάνισης *indels* $h(k)$. Υποθέτουμε ότι υπάρχει μια σταθερά c, τέτοια ώστε να ισχύουν οι σχέσεις

$$s(a,b) = c - d(a,b) \text{ και } h(k) = g(k) - \frac{kc}{2}.$$

Τότε μια αντιστοίχιση μεταξύ δύο ακολουθιών είναι βέλτιστη με τη Μέθοδο της Απόστασης αν και μόνο αν είναι βέλτιστη με τη Μέθοδο της Ομοιότητας και ισχύει

$$D(\mathrm{a},\mathrm{b}) + S(\mathrm{a},\mathrm{b}) = c\frac{(n+m)}{2}, \text{ όπου } 0 \le c \le \max_{a,b} d(a,b).$$

Απόδειξη

Θεωρούμε τις ακολουθίες $\mathrm{a} = a_1 \ldots a_n$ και $\mathrm{b} = b_1 \ldots b_m$. Τότε κατά τη διαδικασία αντιστοίχισής τους ισχύει για το άθροισμα των μεγεθών τους

$$n + m = 2(\#matches) + \sum_k k\Delta_k,$$

όπου με *#matches* εκφράζεται το πλήθος των ζευγών βάσεων που αντιστοιχίζονται, είτε είναι όμοια είτε όχι, και με Δ_k ο αριθμός εμφάνισης *indels* μεγέθους *k*. Από τον ορισμό της απόστασης ισχύει

$$D(\mathrm{a},\mathrm{b}) = \min\left\{ \sum_{matches} d(a,b) + \sum_k g(k)\Delta_k \right\}$$

$$= \min\left\{ \sum_{matches} \left(c - s(a,b)\right) + \sum_{k} \left(h(k) + \frac{kc}{2} \right) \Delta_k \right\}$$

$$= \min\left\{ \sum_{matches} c + \sum_{k} k \frac{c}{2} \Delta_k - \sum_{matches} s(a,b) + \sum_{k} h(k) \Delta_k \right\}$$

$$= \min\left\{ \frac{c}{2} \left(2(\#matches) + \sum_{k} k\Delta_k \right) - \sum_{matches} s(a,b) + \sum_{k} h(k) \Delta_k \right\}$$

$$= c\left(\frac{n+m}{2} \right) - \max\left\{ \sum_{matches} s(a,b) - \sum_{k} h(k) \Delta_k \right\}$$

$$= c\left(\frac{n+m}{2} \right) - S(\mathrm{a}, \mathrm{b}). \qquad \square$$

Σύμφωνα με αυτό το θεώρημα η χρήση οποιασδήποτε από τις δύο μεθόδους αποφέρει τα ίδια συμπεράσματα.

2.5 Προσαρμογή μιας Ακολουθίας Μικρού Μήκους σε Ακολουθία Μεγάλου Μήκους

Ένα θέμα που ενδιαφέρει τη Βιολογία είναι η εύρεση της κατά προσέγγιση βέλτιστης προσαρμογής μιας ακολουθίας μικρού μήκους σε μια ακολουθία μεγάλου μήκους. Το γεγονός αυτό βρίσκει εφαρμογή στην αναζήτηση προτύπων που έχουν λειτουργική σημασία σε μια ακολουθία DNA, όπως το πρότυπο *TATAAT* στην ακολουθία του βακτηρίου *Escherichia coli*, Waterman (1995).

Για τη διατύπωση του προβλήματος θεωρούμε μια ακολουθία a μήκους n και μια ακολουθία b μήκους m, έτσι ώστε $n << m$. Τότε υπάρχει ένας αλγόριθμος τάξης

$$\sum_{k=1}^{m} \sum_{l=k}^{m} n(l-k) = O(nm^3)$$

για την εύρεση της βέλτιστης προσαρμογής της ακολουθίας a στην b μέσω της σχέσης

$$T(\mathrm{a}, \mathrm{b}) = \max\left\{ S(\mathrm{a}, b_k b_{k+1} \ldots b_{l-1} b_l) : 1 \le k \le l \le m \right\}.$$

Η σχέση αυτή υπολογίζει τη μέγιστη ομοιότητα της ακολουθίας a με ένα τμήμα της ακολουθίας b μήκους (*l-k+1*) με τη Μέθοδο της Ομοιότητας.

Ο χρόνος υλοποίησης του αλγορίθμου μπορεί να μειωθεί σε *O(nm)* μέσω μιας διαφορετικής προσέγγισης που αναγράφεται στο βιβλίο του Waterman (1995).

Θεώρημα 2.9

Έστω οι ακολουθίες $\mathrm{a} = a_1 \ldots a_n$ και $\mathrm{b} = b_1 \ldots b_m$ με $n << m$. Ορίζουμε την ποσότητα

$$T_{i,j} = \max\{S(a_1 a_2 \ldots a_i, b_k b_{k+1} \ldots b_j) : 1 \le k \le j\}$$

που εκφράζει τη μέγιστη ομοιότητα μεταξύ του τμήματος $a_1 a_2 \ldots a_i$ μιας ακολουθίας a και ενός τμήματος $b_k b_{k+1} \ldots b_j$ της ακολουθίας b, που λαμβάνεται ως προς k, σε κάθε θέση (*i,j*).

(i) Θέτουμε τις αρχικές συνθήκες

$$T_{0,j} = 0, \text{ για } 0 \le j \le m \qquad \text{και} \qquad T_{i,0} = \sum_{k=1}^{i} s(a_k, -), \text{ για } 1 \le i \le n.$$

Τότε κάθε τιμή του πίνακα T μεγέθους $(n+1)\times(m+1)$ προκύπτει από τη σχέση

$$T_{i,j} = \max\begin{Bmatrix} T_{i-1,j} + s(a_i, -) \\ T_{i-1,j-1} + s(a_i, b_j) \\ T_{i,j-1} + s(-, b_j) \end{Bmatrix}.$$

(ii) Η εύρεση του καλύτερου προτύπου βρίσκεται αναζητώντας στην τελευταία γραμμή του πίνακα τη μέγιστη τιμή, μέσω της σχέσης

$$T_{\mathrm{a,b}} = \max\{T_{n,j} : 1 \le j \le n\}.$$

Ο παρακάτω αλγόριθμος είναι η διατύπωση του θεωρήματος με σταθερό κόστος αντιστοίχισης των βάσεων των ακολουθιών με μηδενικά στοιχεία ίσο με $s(a_i,-) = s(-,b_j) = -\hat{\delta}$.

Αλγόριθμος 2.4

Βήμα 1 : Θέτουμε $T_{1,j} = 0$ για $1 \le j \le m+1$ και

$$T_{i,1} = -(i-1)\hat{\delta} \text{ για } 2 \le i \le n+1.$$

Βήμα 2 : Υπολογίζουμε κάθε κελί (i,j) ($i \ge 2$ & $j \ge 2$) του πίνακα από τη σχέση

$$T_{i,j} = \max\left\{T_{i-1,j} - \hat{\delta}, T_{i-1,j-1} + s(a_i,b_j), T_{i,j-1} - \hat{\delta}\right\}.$$

Βήμα 3 : Επαναλαμβάνουμε το *Βήμα 2* για $2 \le i \le n+1$ και $2 \le j \le m+1$.

Η συνάρτηση $s(a,b)$, όπως έχει περιγραφεί, εκφράζει το κόστος αντιστοίχισης δύο στοιχείων από το σύνολο $\mathcal{D} = \{A,C,G,T,-\}$. Θα εφαρμόσουμε τον παραπάνω αλγόριθμο για τις ακολουθίες $\text{a} = TATAAT$ και $\text{b} = GACACCATCGAATGGCGCAAAACCTTTCG$ $CGGTATGGCATGATAGCGCCCGGAAGAGAGT$, με στόχο την εύρεση της καλύτερης προσαρμογής του προτύπου a στην ακολουθία του βακτηρίου *E. coli*. Επιλέγουμε για τη συνάρτηση $s(a,b)$ τις τιμές που ακολουθούν

$$s(a,b) = \begin{cases} 1 & , \text{αν } a = b \\ -1 & , \text{αν } a \ne b \\ 2 & , \text{αν } a = - \text{ ή } b = -. \end{cases}$$

Παρακάτω δίνεται ένα πρόγραμμα με εντολές του *Mathematica* για την υλοποίηση του *Αλγορίθμου 2.4*.

```
a = {"T", "A", "T", "A", "A", "T"};
b = {"G", "A", "C", "A", "C", "C", "A", "T", "C", "G", "A", "A", "T", "G", "G",
   "C", "G", "C", "A", "A", "A", "A", "C", "C", "T", "T", "T", "C", "G", "C",
   "G", "G", "T", "A", "T", "G", "G", "C", "A", "T", "G", "A", "T", "A", "G",
   "C", "G", "C", "C", "C", "G", "G", "A", "A", "G", "A", "G", "A", "G", "T"};
aa = Prepend[a, "-"];
bb = Prepend[b, "-"];
n = Length[a];
m = Length[b];
nn = Length[aa];
mm = Length[bb];
delta = 2;
T1 = Table[0, {i, 1, nn}, {j, 1, mm}];
s = Table[0, {i, 1, nn}, {j, 1, mm}];
Do[
  If[aa[[i]] == bb[[j]], x = 1, x = -1]; s[[i, j]] = x, {i, 2, nn}, {j, 2, mm}];
Do[T1[[i, 1]] = -(i - 1) * delta, {i, 2, nn}];
Do[
  Do[
   T1[[i, j]] = Max[T1[[i - 1, j]] - delta, T1[[i - 1, j - 1]] + s[[i, j]],
     T1[[i, j - 1]] - delta],
   {j, 2, mm}], {i, 2, nn}];
T3 = Transpose[Rest[Transpose[Rest[T1]]]];
T = TableForm[T3, TableHeadings → {a, b}]
```

Με εφαρμογή του προγράμματος αυτού προκύπτει ο πίνακας στο *Σχήμα 2.7*.

Από την τελευταία γραμμή αναζητούμε τη μέγιστη τιμή : $\max\{T(6,j): 1 \le j \le 60\} = 2$ και βρίσκουμε δύο λύσεις στις θέσεις (6,13) και (6,43). Οι λύσεις αυτές αντιστοιχούν στις δύο στοιχίσεις

$$\begin{array}{cccccc} T & A & T & A & A & T \\ | & & & | & | & | \\ T & C & G & A & A & T \end{array} \quad \text{και} \quad \begin{array}{cccccc} T & A & T & A & A & T \\ & | & | & & | & | \\ C & A & T & G & A & T \end{array}.$$

Παρατηρείται μια καλή προσαρμογή, αφού και στις δύο περιπτώσεις αντιστοίχισης εμφανίζονται 4 ίσα ζεύγη βάσεων, ενώ 2 βάσεις του προτύπου *TATAAT* δεν είναι όμοιες με τις βάσεις του τμήματος της ακολουθίας a.

Σχήμα 2.7

-	G	A	C	A	C	C	A	T	C	G	A
T	-1	-1	-1	-1	-1	-1	-1	**1**	-1	-1	-1
A	-3	0	-2	0	-2	-2	0	-1	**0**	-2	0
T	-5	-2	-1	-2	-1	-3	-2	1	-1	**-1**	-2
A	-7	-4	-3	0	-2	-2	-2	-1	0	-2	**0**
A	-9	-6	-5	-2	-1	-3	-1	-3	-2	-1	-1
T	-11	-8	-7	-4	-3	-2	-3	0	-2	-3	-2

A	T	G	G	C	G	C	A	A	A	A	C
-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0	-1	0	-2	-2	-2	-2	0	0	0	0	-2
-1	1	-1	-1	-3	-3	-3	-2	-1	-1	-1	-1
-1	-1	0	-2	-2	-4	-4	-2	-1	0	0	-2
1	-1	-2	-1	-3	-3	-5	-3	-1	0	1	-1
-1	**2**	0	-2	-2	-4	-4	-5	-3	-2	-1	0

C	T	T	T	C	G	C	G	G	T	A	T
-1	1	1	1	-1	-1	-1	-1	-1	1	-1	1
-2	-1	0	0	0	-2	-2	-2	-2	-1	2	0
-3	-1	0	1	-1	-1	-3	-3	-3	-1	0	3
-2	-3	-2	-1	0	-2	-2	-4	-4	-3	0	1
-3	-3	-4	-3	-2	-1	-3	-3	-5	-5	-2	-1
-2	-2	-2	-3	-4	-3	-2	-4	-4	-4	-4	-1

G	G	C	A	T	G	A	T	A	G	C	G
-1	-1	**-1**	-1	1	-1	-1	1	-1	-1	-1	-1
0	-2	-2	**0**	-1	0	0	-1	2	0	-2	-2
1	-1	-3	-2	**1**	-1	-1	1	0	1	-1	-3
2	0	-2	-2	-1	**0**	0	-1	2	0	0	-2
0	1	-1	-1	-3	-2	**1**	-1	0	1	-1	-1
-2	-1	0	-2	0	-2	-1	**2**	0	-1	0	-2

C	C	C	G	G	A	A	G	A	G	A	G	T
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1
-2	-2	-2	-2	-2	0	0	-2	0	-2	0	-2	-1
-3	-3	-3	-3	-3	-2	-1	-1	-2	-1	-2	-1	-1
-4	-4	-4	-4	-4	-2	-1	-2	0	-2	0	-2	-2
-3	-5	-5	-5	-5	-3	-1	-2	-1	-1	-1	-1	-3
-2	-4	-6	-6	-6	-5	-3	-2	-3	-2	-2	-2	0

Η προσαρμογή μιας ακολουθίας μικρού μήκους σε μια ακολουθία μεγάλου μήκους προκύπτει με ανάλογη εφαρμογή της Μεθόδου της Απόστασης και εκφράζεται με το θεώρημα που ακολουθεί, Sellers (1980).

Θεώρημα 2.10

Έστω οι ακολουθίες $\mathrm{a} = a_1 \ldots a_n$ και $\mathrm{b} = b_1 \ldots b_m$ με $n << m$. Ορίζουμε την ποσότητα

$$D_{i,j} = \min\{D(a_1 a_2 \ldots a_i, b_k b_{k+1} \ldots b_j) : 1 \le k \le j\}$$

που εκφράζει την ελάχιστη απόσταση μιας ακολουθίας a και ενός τμήματος της ακολουθίας b σε κάθε θέση (i,j).

(i) Θέτουμε τις αρχικές συνθήκες

$$D_{0,j} = 0, \ \text{για} \ 0 \le j \le m \qquad \text{και} \qquad D_{i,0} = \sum_{k=1}^{i} d(a_k, -), \ \text{για} \ 1 \le i \le n.$$

Τότε κάθε τιμή του πίνακα D που κατασκευάζουμε, μεγέθους $(n+1)\times(m+1)$ προκύπτει από τη σχέση

$$D_{i,j} = \min\begin{Bmatrix} D_{i-1,j} + d(a_i, -) \\ D_{i-1,j-1} + d(a_i, b_j) \\ D_{i,j-1} + d(-, b_j) \end{Bmatrix}.$$

(ii) Η εύρεση του καλύτερου προτύπου βρίσκεται αναζητώντας στην τελευταία γραμμή του πίνακα την ελάχιστη τιμή, μέσω της σχέσης

$$D_{\mathrm{a,b}} = \min\{D_{n,j} : 1 \le j \le n\}.$$

2.6 Τοπική Στοίχιση και Groups

Ένα ακόμα θέμα που παρουσιάζει μεγάλο ενδιαφέρον είναι η αντιστοίχιση μεταξύ τμημάτων (*segments*) ακολουθιών, που ονομάζεται «τοπική» αντιστοίχιση (*local alignment*). Το πρόβλημα αυτό πρωτοαπασχόλησε τον Sellers (1979, 1980) που βασίστηκε στη χρήση της μετρικής της απόστασης για τον υπολογισμό των "*forward*" και "*backward*" πινάκων. Μια όμως πιο εύχρηστη μέθοδος που στηρίζεται σε συναρτήσεις ομοιότητας, περιγράφεται από τους Smith & Waterman (1981a,b).

Η ιδέα της μεθόδου είναι η εύρεση της μέγιστης ομοιότητας μεταξύ τμημάτων των δύο ακολουθιών $\mathrm{a} = a_1 ... a_n$ και $\mathrm{b} = b_1 .. b_m$ με τη βοήθεια ενός πίνακα H μεγέθους $(n+1)\times(m+1)$. Ο αλγόριθμος περιλαμβάνει τη συνάρτηση ομοιότητας $s(a,b)$ μεταξύ δυο βάσεων a & b και τη συνάρτηση $g(k)$ που εκφράζει το συντελεστή εμφάνισης *indel* μεγέθους *k*. Ο υπολογισμός της «τοπικής» αντιστοίχισης γίνεται με χρήση του θεωρήματος που ακολουθεί.

Θεώρημα 2.11

Έστω οι ακολουθίες $\mathrm{a} = a_1 ... a_n$ και $\mathrm{b} = b_1 .. b_m$. Ορίζουμε την ποσότητα

$$H_{i,j} = \max\left\{0; S(a_x a_{x+1} ... a_i, b_y b_{y+1} ... b_j) : 1 \le x \le i, 1 \le y \le j\right\}$$

που εκφράζει τη μέγιστη ομοιότητα για δύο τμήματα των ακολουθιών a και b που καταλήγουν στις βάσεις a_i και b_j.

(i) Θέτουμε τις αρχικές συνθήκες

$$H_{i,0} = H_{0,j} = 0, \quad \text{για} \quad 0 \le i \le n \quad \text{και} \quad 0 \le j \le m .$$

Τότε κάθε στοιχείο του πίνακα H μεγέθους $(n+1)\times(m+1)$ υπολογίζεται από τη σχέση

$$H_{i,j} = \max\begin{Bmatrix} 0 \\ H_{i-1,j-1} + s(a_i, b_j) \\ \max\{H_{i-k,j} - g(k) : 1 \le k \le i\} \\ \max\{H_{i,j-l} - g(l) : 1 \le l \le j\} \end{Bmatrix}.$$

(ii) Το ζεύγος των τμημάτων με τη μέγιστη ομοιότητα εντοπίζεται με την εύρεση της μέγιστης τιμής του πίνακα H

$$H_{\text{a,b}} = \max\{H_{k,l} : 1 \le k \le n, 1 \le l \le m\}.$$

Στη σχέση υπολογισμού των στοιχείων του πίνακα H συμπεριλαμβάνεται η τιμή του μηδενός προς αποφυγή εμφάνισης αρνητικής τιμής, που υποδεικνύει τη μη-ομοιότητα μεταξύ των βάσεων a_i και b_j.

Η περίπτωση της γραμμικής μορφής της συνάρτησης $g(k)$ περιγράφεται στο επόμενο θεώρημα.

Θεώρημα 2.12

Έστω $g(k) = \alpha + \beta(k-1)$ η γραμμική συνάρτηση βάρους εμφάνισης k συνεχόμενων *indels (affine gap penalty)* και $s(a,b)$ το μέτρο ομοιότητας της βάσης a της πρώτης ακολουθίας με τη βάση b της δεύτερης ακολουθίας. Ορίζουμε τους πίνακες $(E_{i,j})$, $(F_{i,j})$ και $(H_{i,j})$ για τους οποίους ισχύουν τα εξής:

$$E_{i,j} = F_{i,j} = H_{i,j} = 0 \text{, αν } i \cdot j = 0$$
$$E_{i,j} = \max\{H_{i,j-1} - \alpha, E_{i,j-1} - \beta\}$$
$$F_{i,j} = \max\{H_{i-1,j} - \alpha, F_{i-1,j} - \beta\}$$

Τότε

$$H_{i,j} = \max\begin{Bmatrix} 0 \\ E_{i,j} \\ F_{i,j} \\ H_{i-1,j-1} + s(a_i, b_j) \end{Bmatrix}.$$

Κατά τη διαδικασία εύρεσης του ζεύγους τμημάτων με την άριστη «τοπική» αντιστοίχιση μπορεί να βρεθούν περισσότερες από μια θέσεις στον πίνακα *H* που εμφανίζουν τη μέγιστη τιμή $H_{\text{a,b}} = \max\{H_{k,l} : 1 \le k \le n, 1 \le l \le m\}$. Απ' αυτές τις θέσεις θα επιλεγεί η θέση (*i,j*), που ικανοποιεί ένα από τα επόμενα κριτήρια:

(*i*) $H_{i,j} = H_{k,l}$ και $i + j < k + l$ ή

(*ii*) $H_{i,j} = H_{k,l}$, $i + j = k + l$ και $i < k$.

Στη συνέχεια είναι απαραίτητη η εύρεση της θέσης (*p,q*) που αντιστοιχεί στο τελικό ζεύγος βάσεων σύμφωνα με τη διαδικασία των *tracebacks*. Σε περίπτωση εύρεσης δύο υποψήφιων θέσεων (*p,q*) και (*r,s*), έχει καθιερωθεί η επιλογή της θέσης που αντιστοιχεί σε μικρότερου μήκους τοπική στοίχιση. Έτσι, επιλέγεται η θέση (*p,q*), αν ισχύει $r + s < p + q$ ή $p + q = r + s$ και $p > r$, Waterman (1984).

Με τη μέθοδο που περιγράφηκε, βρίσκεται μόνο ένα ζεύγος τμημάτων άριστης αντιστοίχισης, ενώ στο χώρο της Βιολογίας ενδιαφέρονται για την εύρεση περισσότερων «τοπικών» αντιστοιχίσεων που δεν έχουν μεταξύ τους κανένα κοινό ζεύγος αντιστοίχισης (*non-intersecting alignments*). Οι Waterman & Eggert (1987) περιγράφουν μια μέθοδο επαναϋπολογισμού του πίνακα *H* και χρησιμοποιούν την πρόταση των Kruskal & Sankoff (1983) για αντικατάσταση των τιμών των θέσεων (*i,j*), που εμπλέκονται στην προηγούμενη άριστη «τοπική» αντιστοίχιση, με 0. Ο επαναϋπολογισμός του πίνακα *H* επηρεάζει τις τιμές του πίνακα που βρίσκονται δίπλα σε εκείνες που αντιστοιχούν στην αρχική «τοπική» στοίχιση και συγκεκριμένα τις τιμές προς τα δεξιά και προς τα κάτω. Η μέθοδος που θα περιγραφεί αφορά την περίπτωση μεμονωμένης αντιστοίχισης βάσης με μηδενικό στοιχείο, όπου η συνάρτηση $g(k)$ δίνεται από τη σχέση $g(k) = k\delta$. Ο τύπος αυτός προκύπτει από τη γραμμική συνάρτηση $g(k) = \alpha + \beta(k-1)$ για $\alpha = \beta = \delta$.

Έστω (i,j) η θέση που αντιστοιχεί στο πρώτο ζεύγος βάσεων της αρχικής «τοπικής» αντιστοίχισης. Τότε ο νέος πίνακας H^* ικανοποιεί τη σχέση $H^*_{k,l} = H_{k,l}$ για $k < i$ ή $l < j$. Για τη θέση (i,j) ορίζουμε την τιμή του πίνακα ως εξής : $H^*_{i,j} = \max\{0, H^*_{i-1,j} - \delta, H^*_{i,j-1} - \delta\}$. Στη συνέχεια θεωρούμε τη γραμμή (i,l) με $l > j$ και επαναϋπολογίζουμε για $l = j+1, j+2,...$ τις τιμές του πίνακα μέχρι η νέα τιμή $H^*_{i,l}$ να ισούται με την τιμή $H_{i,l}$ του προηγούμενου πίνακα. Τότε για τις υπόλοιπες τιμές της i - γραμμής θα ισχύει $H^*_{i,l} = H_{i,l}$. Ανάλογα, θεωρούμε τη στήλη (k,j) με $k > i$ και επαναϋπολογίζουμε για $k = i+1, i+2,...$ μέχρι η νέα τιμή $H^*_{k,j}$ να ισούται με την προηγούμενη $H_{k,j}$. Η διαδικασία συνεχίζεται και υπολογίζουμε το νέο πίνακα H^*.

Η μέθοδος αυτή μπορεί να τροποποιηθεί κατάλληλα για την περίπτωση της γραμμικής μορφής της συνάρτησης $g(k) = \alpha + \beta(k-1)$. Πλέον υπολογίζονται τρεις πίνακες H^* , E^* και F^* όπου η διαδικασία υπολογισμού των στοιχείων των πινάκων προχωρά κατά γραμμή και στήλη μέχρι την ισχύ των ισοτήτων: $H_{i,j} = H^*_{i,j}$, $E_{i,j} = E^*_{i,j}$ και $F_{i,j} = F^*_{i,j}$.

Ο αλγόριθμος που περιγράφει το *Θεώρημα 2.11* για την περίπτωση της απλής συνάρτησης $g(k) = k\delta$ είναι ο ακόλουθος:

Αλγόριθμος 2.5

Βήμα 1 : Θέτουμε $H_{i,1} = H_{1,j} = 0$ για $1 \le i \le n+1$ και $1 \le j \le m+1$.

Βήμα 2 : Υπολογίζουμε το κελί (i,j) ($i \ge 2$ & $j \ge 2$) του πίνακα από τη σχέση

$$H_{i,j} = \max\{0, H_{i-1,j} - \delta, H_{i-1,j-1} + s(a_i, b_j), H_{i,j-1} - \delta\}.$$

Βήμα 3 : Επαναλαμβάνουμε το *Βήμα 2* για $2 \le i \le n+1$ και $2 \le j \le m+1$.

Η υλοποίηση του αλγορίθμου για την τοπική αντιστοίχιση παρουσιάζεται παρακάτω με το παράδειγμα των Waterman & Eggert (1987).

```
a = {"C", "C", "A", "A", "T", "C", "T", "A", "C", "T", "A", "C", "T", "G", "C",
"T", "T", "G", "C", "A", "G", "T", "A", "C"};
b = {"A", "G", "T", "C", "C", "G", "A", "G", "G", "G", "C", "T", "A", "C", "T",
"C", "T", "A", "C", "T", "G", "A", "A", "C"};
aa = Prepend[a, "-"];
bb = Prepend[b, "-"];
n = Length[a];
m = Length[b];
nn = Length[aa];
mm = Length[bb];
delta = 20;
H1 = Table[0, {i, 1, nn}, {j, 1, mm}];
s = Table[0, {i, 1, nn}, {j, 1, mm}];
Do[
  If[aa[[i]] == bb[[j]], x = 10, x = -9]; s[[i, j]] = x, {i, 1, nn}, {j, 1, mm}];
Do[H1[[1, j]] = 0, {j, 2, mm}];
Do[H1[[i, 1]] = 0, {i, 2, nn}];
Do[
  Do[
   H1[[i, j]] = Max[0, H1[[i - 1, j]] - delta, H1[[i - 1, j - 1]] + s[[i, j]],
     H1[[i, j - 1]] - delta], {j, 2, mm}], {i, 2, nn}];
H = TableForm[H1, TableHeadings → {aa, bb}]
```

Η ακολουθία $\mathrm{a} = \mathrm{CCAATCTACTACTGCTTGCAGTAC}$ συγκρίνεται με την ακολουθία $\mathrm{b} = \mathrm{AGTCCGAGGGCTACTCTACTGAAC}$ με τιμές $s(a,a) = 10$, $s(a,b) = -9$ αν $a \neq b$ και $g(k) = 20k$ (δηλαδή $\delta = 20$). Από τον Αλγόριθμο 2.5 προκύπτει ο αρχικός πίνακας H (*Σχήμα 2.8*), όπου βρίσκοντας τη μέγιστη τιμή $H_{\mathrm{a,b}} = 62$ στη θέση (11,21) με τη διαδικασία των *tracebacks* προκύπτει η «τοπική» αντιστοίχιση

$$\begin{array}{cccccccccc} C & T & A & C & T & C & T & A & C & T \\ | & & | & & | & | & | & | & | & | \\ C & C & A & A & T & C & T & A & C & T \end{array}.$$

Η επόμενη καλύτερη «τοπική» αντιστοίχιση προκύπτει με τη βοήθεια του *Αλγορίθμου 2.6* για $H^{*}_{a,b} = 61$ που αντιστοιχεί στη θέση (17,21) και είναι η εξής

$$\begin{array}{cccccccccc} C & T & A & C & T & - & C & T & A & C & T \\ | & | & | & | & | & & | & | & & | & | \\ C & T & A & C & T & A & C & T & G & C & T \end{array}.$$

Αλγόριθμος 2.6

Βήμα 1 : Έστω (k,l) το $1^{ο}$ ζεύγος βάσεων της τοπικής αντιστοίχισης τότε θέτουμε $H^*_{i,j} = H_{i,j}$ για $i < k$ ή $j < l$.

Βήμα 2 : Θέτουμε $H^*_{k.l} = \max\{0, H^*_{k-1,l} - \delta, H^*_{k,l-1} - \delta\}$.

Βήμα 3 : Υπολογίζουμε τις τιμές της *k*-γραμμής για $j > l$ από τη σχέση

$H^*_{k,j} = \max\{0, H^*_{k-1,j} - \delta, H^*_{k-1,j-1} + s(a_k, b_j), H^*_{k,j-1} - \delta\}$ μέχρι $H^*_{k,j} = H_{k,j}$. Οι υπόλοιπες τιμές της *k*-γραμμής θα είναι ίσες με τις τιμές του πίνακα *H*.

Βήμα 4 : Υπολογίζουμε τις τιμές της *l*-στήλης για $i > k$ από τη σχέση

$H^*_{i,l} = \max\{0, H^*_{i-1,l} - \delta, H^*_{i-1,l-1} + s(a_i, b_l), H^*_{i,l-1} - \delta\}$ μέχρι $H^*_{i,l} = H_{i,l}$. Οι υπόλοιπες τιμές της *l*-στήλης θα είναι ίσες με τις τιμές του πίνακα *H*.

Βήμα 5 : Επαναλαμβάνουμε τα *Βήματα 2,3 & 4* για τα υπόλοιπα ζεύγη αντιστοίχισης.

Βήμα 6 : Τα υπόλοιπα κελιά υπολογίζονται από τη σχέση

$H_{i,j} = \max\{0, H_{i-1,j} - \delta, H_{i-1,j-1} + s(a_i, b_j), H_{i,j-1} - \delta\}$.

Σχήμα 2.8

Πρώτη «τοπική» στοίχιση για τις ακολουθίες a και b

	-	*A*	*G*	*T*	*C*	*C*	*G*	*A*	*G*	*G*	*G*	*C*	*T*	*A*	*C*	*T*	*C*	*T*	*A*	*C*	*T*	*G*	*A*	*A*	*C*
-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	10	10	0	0	0	0	0	***10***	0	0	10	0	10	0	0	10	0	0	0	0	10
C	0	0	0	0	10	20	1	0	0	0	0	10	***1***	0	10	1	10	1	0	10	1	0	0	0	10
A	0	10	0	0	0	1	11	11	0	0	0	0	1	***11***	0	1	0	1	11	0	1	0	10	10	0
A	0	10	1	0	0	0	0	21	2	0	0	0	0	11	***2***	0	0	0	11	2	0	0	10	20	1
T	0	0	1	11	0	0	0	1	12	0	0	0	10	0	2	***12***	0	10	0	2	12	0	0	1	11
C	0	0	0	0	21	10	0	0	0	3	0	10	0	1	10	0	***22***	2	1	10	0	3	0	0	11
T	0	0	0	10	1	12	1	0	0	0	0	0	20	0	0	20	2	***32***	12	0	20	0	0	0	0
A	0	10	0	0	1	0	3	11	0	0	0	0	0	30	10	0	11	12	***42***	22	2	11	10	10	0
C	0	0	1	0	10	11	0	0	2	0	0	10	0	10	40	20	10	2	22	***52***	32	12	2	1	20
T	0	0	0	11	0	1	2	0	0	0	0	0	20	0	20	50	30	20	2	32	***62***	42	22	2	0
A	0	10	0	0	2	0	0	12	0	0	0	0	0	30	10	30	41	21	30	12	42	53	52	32	12
C	0	0	1	0	10	12	0	0	3	0	0	10	0	10	40	20	40	32	12	40	22	33	44	43	42
T	0	0	0	11	0	1	3	0	0	0	0	0	20	0	20	50	30	50	30	20	50	30	24	35	34
G	0	0	10	0	2	0	11	0	10	10	10	0	0	11	0	30	41	30	41	21	30	60	40	20	26
C	0	0	0	1	10	12	0	2	0	1	1	20	0	0	21	10	40	32	21	51	31	40	51	31	30
T	0	0	0	10	0	1	3	0	0	0	0	0	30	10	1	31	20	50	30	31	61	41	31	42	22
T	0	0	0	10	1	0	0	0	0	0	0	0	10	21	1	11	22	30	41	21	41	52	32	22	33
G	0	0	10	0	1	0	10	0	10	10	10	0	0	1	12	0	2	13	21	32	21	51	43	23	13
C	0	0	0	1	10	11	0	1	0	1	1	20	0	0	11	3	10	0	4	31	23	31	42	34	33
A	0	10	0	0	0	1	2	10	0	0	0	0	11	10	0	2	0	1	10	11	22	14	41	52	32
G	0	0	20	0	0	0	11	0	20	10	10	0	0	2	1	0	0	0	0	1	2	32	21	32	43
T	0	0	0	30	10	0	0	2	0	11	1	1	10	0	0	11	0	10	0	0	11	12	23	12	23
A	0	10	0	10	21	1	0	10	0	0	2	0	0	20	0	0	2	0	20	0	0	2	22	33	13
C	0	0	1	0	20	31	1	0	1	0	0	12	0	0	30	10	10	0	0	30	10	0	2	13	43

Σχήμα 2.9

Δεύτερη «τοπική» στοίχιση για τις ακολουθίες a και b

	-	*A*	*G*	*T*	*C*	*C*	*G*	*A*	*G*	*G*	*G*	*C*	*T*	*A*	*C*	*T*	*C*	*T*	*A*	*C*	*T*	*G*	*A*	*A*	*C*
-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	10	10	0	0	0	0	0	*0*	*0*	0	10	0	10	0	0	10	0	0	0	0	10
C	0	0	0	0	10	20	1	0	0	0	0	*10*	*0*	*0*	10	1	10	1	0	10	1	0	0	0	10
A	0	10	0	0	0	1	11	11	0	0	0	0	*1*	*0*	*0*	1	0	1	11	0	1	0	10	10	0
A	0	10	1	0	0	0	0	21	2	0	0	0	0	*11*	*0*	*0*	0	0	11	2	0	0	10	20	1
T	0	0	1	11	0	0	0	1	12	0	0	0	10	0	*2*	*0*	*0*	10	0	2	12	0	0	1	11
C	0	0	0	0	21	10	0	0	0	3	0	***10***	0	1	10	*0*	*0*	*0*	*1*	10	0	3	0	0	11
T	0	0	0	10	1	12	1	0	0	0	0	0	***20***	0	0	20	*0*	*0*	*0*	*0*	20	0	0	0	0
A	0	10	0	0	1	0	3	11	0	0	0	0	0	***30***	10	0	*11*	*0*	*0*	*0*	*0*	*11*	10	10	0
C	0	0	1	0	10	11	0	0	2	0	0	10	0	10	***40***	20	10	*2*	*0*	*0*	*0*	*0*	*2*	1	20
T	0	0	0	11	0	1	2	0	0	0	0	0	20	0	20	***50***	30	20	*0*	*0*	*0*	*0*	*0*	*0*	*0*
A	0	10	0	0	2	0	0	12	0	0	0	0	0	30	10	***30***	41	21	*30*	*10*	*0*	*0*	*10*	*10*	*0*
C	0	0	1	0	10	12	0	0	3	0	0	10	0	10	40	20	***40***	32	12	*40*	*20*	*0*	*0*	*1*	*20*
T	0	0	0	11	0	1	3	0	0	0	0	0	20	0	20	50	30	***50***	30	20	*50*	*30*	*10*	*0*	*0*
G	0	0	10	0	2	0	11	0	10	10	10	0	0	11	0	30	41	30	***41***	21	30	60	*40*	*20*	*0*
C	0	0	0	1	10	12	0	2	0	1	1	20	0	0	21	10	40	32	21	***51***	31	40	51	31	*30*
T	0	0	0	10	0	1	3	0	0	0	0	0	30	10	1	31	20	50	30	31	***61***	41	31	42	22
T	0	0	0	10	1	0	0	0	0	0	0	0	10	21	1	11	22	30	41	21	41	52	32	22	33
G	0	0	10	0	1	0	10	0	10	10	10	0	0	1	12	0	2	13	21	32	21	51	43	23	13
C	0	0	0	1	10	11	0	1	0	1	1	20	0	0	11	3	10	0	4	31	23	31	42	34	33
A	0	10	0	0	0	1	2	10	0	0	0	0	11	10	0	2	0	1	10	11	22	14	41	52	32
G	0	0	20	0	0	0	11	0	20	10	10	0	0	2	1	0	0	0	0	1	2	32	21	32	43
T	0	0	0	30	10	0	0	2	0	11	1	1	10	0	0	11	0	10	0	0	11	12	23	12	23
A	0	10	0	10	21	1	0	10	0	0	2	0	0	20	0	0	2	0	20	0	0	2	22	33	13
C	0	0	1	0	20	31	1	0	1	0	0	12	0	0	30	10	10	0	0	30	10	0	2	13	43

Η υλοποίηση του αλγορίθμου, όπου στη λίστα *lst1* περιέχονται τα ζεύγη της πρώτης τοπικής αντιστοίχισης, είναι η επόμενη:

```
lst1 = {{2, 12}, {3, 13}, {4, 14}, {5, 15}, {6, 16}, {7, 17}, {8, 18}, {9, 19},
   {10, 20}, {11, 21}};
Hnew = Table[0, {i, 1, nn}, {j, 1, mm}];
Do[If[i < lst1[[1, 1]]  || j < lst1[[1, 2]], Hnew[[i, j]] = H1[[i, j]], 0],
 {i, 1, nn}, {j, 1, mm}]
Do[
  Hnew[[lst1[[z, 1]], lst1[[z, 2]]]] =
   Max[0, Hnew[[lst1[[z, 1]] - 1, lst1[[z, 2]]]] - delta,
    Hnew[[lst1[[z, 1]], lst1[[z, 2]] - 1]] - delta];
  Do[
   Hnew[[i, j]] = Max[0, Hnew[[i - 1, j]] - delta, Hnew[[i - 1, j - 1]] + s[[i, j]],
     Hnew[[i, j - 1]] - delta],
   {j, lst1[[z, 2]] + 1, mm}, {i, lst1[[z, 1]], lst1[[z, 1]]}];
  Do[
   Hnew[[i, j]] = Max[0, Hnew[[i - 1, j]] - delta, Hnew[[i - 1, j - 1]] + s[[i, j]],
     Hnew[[i, j - 1]] - delta],
   {i, lst1[[z, 1]] + 1, nn}, {j, lst1[[z, 2]], lst1[[z, 2]]}],
  {z, 1, len}];
Do[
  Do[
   Hnew[[i, j]] = Max[0, Hnew[[i - 1, j]] - delta, Hnew[[i - 1, j - 1]] + s[[i, j]],
     Hnew[[i, j - 1]] - delta], {j, Last[lst1][[2]] + 1, mm}], {i, Last[lst1][[1]] + 1, nn}];
Hnew2 = TableForm[Hnew, TableHeadings → {aa, bb}]
```

Ο νέος πίνακας H^* που προκύπτει δίνεται στο *Σχήμα 2.9* και υπολογίζεται με τη βοήθεια του *Αλγορίθμου 2.6*. Οι αριθμοί με πλάγια έντονη γραφή αντιστοιχούν στα ζεύγη της $2^{ης}$ τοπικής αντιστοίχισης ενώ οι αριθμοί με πλάγια και υπογραμμισμένη γραφή στις νέες τιμές του πίνακα που προκύπτουν από τη μεθοδολογία των Waterman & Eggert (1987).

2.6.1 Επαναλαμβανόμενα Πρότυπα (Tandem Repeats)

Η εμφάνιση επαναλαμβανόμενων προτύπων (*tandem repeats*) σε μια ακολουθία DNA είναι σημαντική για τον καθορισμό των λειτουργιών του οργανισμού. Η αύξηση της συχνότητάς τους έχει συνδεθεί με την παρουσία κάποιων ασθενειών και έτσι κρίνεται

απαραίτητη η εύρεση αυτών των ειδικών σχηματισμών με τη βοήθεια αριθμητικών αλγορίθμων. Έχουν προταθεί διάφοροι αλγόριθμοι για την εύρεση αυτών των σχηματισμών που αυτοί είτε δίνονται εκ των προτέρων και ελέγχεται η συχνότητα εμφάνισής τους κατά τη διαδικασία της στοίχισης (Myers & Miller (1989) και Fischetti et al. (1992)) είτε αναζητούνται άγνωστου μήκους επαναλαμβανόμενοι σχηματισμοί. Η περίπτωση εύρεσης ενός γνωστού σχηματισμού μήκους *k* σε μια ακολουθία μήκους *n*, που θα μας απασχολήσει στη συνέχεια, εντάσσεται στην περιοχή του Wraparound Δυναμικού Προγραμματισμού (*Wraparound Dynamic Programming*) ο οποίος οφείλει την ονομασία του στον τρόπο σύνθεσης του αλγορίθμου κατά την αντιστοίχιση της ακολουθίας με τον συγκεκριμένο σχηματισμό. Η τεχνική αυτής της μεθόδου περιλαμβάνει μια συνάρτηση ομοιότητας και διαφοροποιείται από τους αλγορίθμους που έχουμε συναντήσει μέχρι τώρα διότι οι τιμές των κελιών κάθε γραμμής αναθεωρούνται κατά το δεύτερο υπολογισμό σύμφωνα με τον Benson (1997).

Ο Αλγόριθμος, που θα περιγραφεί, εισήχθη πρώτα από τους Myers & Miller (1988b) και στη συνέχεια βελτιώθηκε από τους Fischetti et al. (1992) και βασίζεται σε δύο παρατηρήσεις:

1. Θεωρούμε την ακολουθία a μήκους *n* και την ακολουθία b μήκους *k* που επαναλαμβάνεται τουλάχιστον *n* φορές. Τότε ο πίνακας που δημιουργείται κατά τη διαδικασία της αντιστοίχισης θα αποτελείται από *n* γραμμές και *nk* στήλες. Είναι σαφής η ανεξαρτησία της τιμής στο κελί (*i,j*) με την τιμή του κελιού (*i,j-k*) της *i* γραμμής, αφού η τιμή κάθε κελιού επηρεάζεται μόνο από γειτονικά κελιά. Το κελί (*i,j-k*) θα πάρει κάποια τιμή *score* αλλά εξαιτίας της επαναληπτικής φύσης του προτύπου b, την ίδια τιμή *score* θα παρουσιάζει και το κελί (*i,j*). Όμως, αν ληφθεί υπ' όψιν ο συνυπολογισμός της αρνητικής τιμής που οφείλεται στις *k* αφαιρέσεις στην τιμή του *score*, τότε το κελί (*i,j-k*) θα έχει μεγαλύτερο *score* από το κελί (*i,j*). Έτσι, καταλήγουμε σε αντίφαση αφού προηγουμένως δεχτήκαμε την ισότητα των *scores* των δύο κελιών.

Ορίζουμε $R_{i,j}$ την τιμή του βέλτιστου *score* του κελιού (*i,j+1*) για την αντιστοίχιση μιας ακολουθίας $a_1 a_2 ... a_i$ με το πρότυπο $b_1 b_2 ... b_{j+1}$, όπου $1 \le j \le k-1$. Υποθέτουμε ότι η τιμή $R_{i-1,j}$, για $1 \le j \le k-1$ είναι γνωστή και υπολογίζουμε την τιμή $R^*_{i,j}$ από τη σχέση

$$R^*_{i,j} = \max\{0, R_{i,j-1} - \delta, R_{i-1,j-1} + s(a_i, b_{j+1}), R_{i-1,j} - \delta\}, \text{ όπου } 0 \le i \le n \text{ και } 1 \le j \le k-1.$$

Για $j=1$ από τη δοσμένη σχέση προκύπτει ότι χρειάζεται να γνωρίζουμε τις τιμές $R_{i,0}$ και $R_{i-1,0}$ που όμως δεν είναι γνωστές εκ των προτέρων αλλά λόγω περιοδικότητας θα ισχύουν οι σχέσεις $R_{i,0}=R_{i,k-1}$ και $R_{i-1,0}=R_{i-1,k-1}$. Έτσι, θέτουμε εξαρχής $R_{i,0}=0$ για να υπολογίσουμε την τιμή $R^*_{i,1}$. Επομένως, θα ισχύει $R^*_{i,1}=R_{i,1}$, εκτός αν η σωστή αντιστοίχιση προέρχεται από την αντιστοίχιση που καταλήγει σε a_i και b_{k-1}. Καθώς ο αλγόριθμος συνεχίζεται, καταλήγουμε στον υπολογισμό της τιμής $R^*_{i,k-1}$ από την ισότητα $R_{i,k-1}=R^*_{i,k-1}$ εκτός και αν η αντιστοίχιση προέρχεται από την τιμή $R_{i,0}$. Αυτό όμως δεν μπορεί να ισχύει οπότε $R_{i,k-1}=R^*_{i,k-1}$.

2. Ο επαναϋπολογισμός της *i*-γραμμής με χρήση της σωστής τιμής $R_{i,0}=R_{i,k-1}$ δίνει τις τιμές $R_{i,j}$ για $1\le j\le k-1$.

Έτσι, προκύπτει ένας αλγόριθμος της τάξης $O(nk)$ που υπολογίζει τις τιμές των *scores* ενός πίνακα R διαστάσεων $(n+1)\times k$ και περιγράφεται με τα παρακάτω βήματα.

Αλγόριθμος 2.7 (Wraparound Dynamic Programming)

Βήμα 1 : Θέτουμε $R_{1,j}=0$, για $1\le j\le k$ και

$R_{i,1}=0$, για $1\le i\le n+1$.

Βήμα 2 : Υπολογίζουμε για την *i*- γραμμή του πίνακα, ξεκινώντας από $i=2$, τις τιμές

$R_{i,j}=\max\{0,R_{i,j-1}-\delta,R_{i-1,j-1}+s(a_i,b_j),R_{i-1,j}-\delta\}$ για $j=2,...,k$.

Βήμα 3 : Θέτουμε για $i=2$ την τιμή $R_{i,1}=\max\{0,R_{i,k}-\delta,R_{i-1,k}+s(a_i,b_1),R_{i-1,1}-\delta\}$.

Βήμα 4 : Επαναϋπολογίζουμε τις τιμές της *i*- γραμμής για $i=2$ από τη σχέση

$R_{i,j}=\max\{0,R_{i,j-1}-\delta,R_{i-1,j-1}+s(a_i,b_j),R_{i-1,j}-\delta\}$ για $j=2,...,k$.

Βήμα 5 : Επαναλαμβάνουμε τα *Βήματα 2,3* και *4* για $i=3$ μέχρι $n+1$.

Βήμα 6 : Βρίσκουμε την καλύτερη τιμή του *score* από τη σχέση

$R(\mathrm{a},\mathrm{b})=\max\{R_{i,j}:2\le i\le n+1,1\le j\le k\}$.

Η υλοποίηση του Αλγορίθμου *Wraparound Dynamic Programming* παρουσιάζεται μέσω του *Mathematica* για την αντιστοίχιση του επαναλαμβανόμενου σχηματισμού *CGG* και

ενός τμήματος της ακολουθίας DNA του γονιδίου X-FMR-1 (*fragile-X mental retardation*) a=*CGTGCGGCAG CGCGG* , Benson & Waterman (1994).

```
a = {"C", "G", "T", "G", "C", "G", "G", "C", "A", "G", "C", "G", "C", "G", "G"};
b = {"C", "G", "G"};
aa = Prepend[a, "-"];
n = Length[a]; k = Length[b];
nn = Length[aa]; delta = 2;
R1 = Table[0, {i, 1, nn}, {j, 1, k}];
s = Table[0, {i, 1, nn}, {j, 1, k}];
Do[If[aa[[i]] == b[[j]], x = 2, x = -1]; s[[i, j]] = x, {i, 1, nn}, {j, 1, k}];
Do[R1[[1, j]] = 0, {j, 1, k}];
Do[R1[[i, 1]] = 0, {i, 1, nn}];
Do[Do[R1[[i, j]] = Max[0, R1[[i, j - 1]] - delta,
     R1[[i - 1, j - 1]] + s[[i, j]], R1[[i - 1, j]] - delta], {j, 2, k}];
  R1[[i, 1]] = Max[0, R1[[i, k]] - delta, R1[[i - 1, k]] + s[[i, 1]],
    R1[[i - 1, 1]] - delta];
  Do[R1[[i, j]] = Max[0, R1[[i, j - 1]] - delta,
     R1[[i - 1, j - 1]] + s[[i, j]], R1[[i - 1, j]] - delta],
   {j, 2, k}], {i, 2, nn}];
R = TableForm[R1, TableHeadings → {aa, b}]
```

Σχήμα 2.10

Αντιστοίχιση ακολουθίας με τον επαναλαμβανόμενο σχηματισμό *CGG*

	C	*G*	*G*
	0	0	0
C	**2**	0	0
G	0	**4**	2
T	1	**2**	3
G	2	3	**4**
C	**6**	4	2
G	4	**8**	6
G	8	6	**10**
C	**12**	10	8
A	10	**11**	9
G	11	12	**13**
C	**15**	13	11
G	13	**17**	**15**
C	**17**	15	16
G	15	**19**	17
G	19	17	**21**

Ο *Αλγόριθμος 2.7* για τις τιμές των παραμέτρων $s(a,a)=2$ στην περίπτωση αντιστοίχισης δύο ίδιων βάσεων, $s(a,b)=-1$ κατά την αντικατάσταση βάσης από βάση και $\delta=2$ στην περίπτωση εμφάνισης *indel* δίνει τον πίνακα τιμών των *scores* $R(i,j)$ με μέγιστη τιμή $R(\mathrm{a},\mathrm{b})=21$ στη θέση (16,3).

Η αντιστοίχιση μεταξύ της ακολουθίας και του επαναλαμβανόμενου προτύπου υλοποιείται στο *Mathematica* με τροποποίηση της μεθόδου των *tracebacks* ως εξής:

```
a = Max[R1];
b = Position[R1, a];
i = b[[1]][[1]];
j = b[[1]][[2]];
l = {{i, j}};
While[R1[[i, j]] > 0,
  R1[[i, j]];
  If[j ≠ 1,
   If[R1[[i, j]] == R1[[i - 1, j - 1]] + s[[i, j]], AppendTo[l, {i = i - 1, j = j - 1}],
    If[
     R1[[i, j]] == R1[[i, j - 1]] - delta, AppendTo[l, {i = i, j = j - 1}],
     If[
      R1[[i, j]] == R1[[i - 1, j]] - delta, AppendTo[l, {i = i - 1, j = j}]]]],
   i = i - 1;
   j = j + k - 1;
   AppendTo[l, {i, j}]]];
Print[Drop[l, -1]]
```

Η τροποποιημένη μέθοδος των *tracebacks* δίνει σε μια λίστα *l* τα ζεύγη βάσεων που αντιστοιχίζονται και η αντιστοίχιση που προκύπτει δίνεται παρακάτω.

C	G	T	G		C	G	G		C	A	G		C	–	G		C	G	G	
|	|		|		|	|	|		|		|		|		|	|	|	
C	G	–	G		C	G	G		C	G	G		C	G	G		C	G	G	

Οι Benson & Waterman (1994), βασισμένοι στον Αλγόριθμο *Wraparound Dynamic Programming*, δημιούργησαν ένα πρόγραμμα στη Γλώσσα Προγραμματισμού *C* σύμφωνα με το οποίο ανιχνεύονται σε μια ακολουθία DNA όλα τα «ύποπτα» πρότυπα μεγέθους μέχρι 32 βάσεων που μπορεί να επαναλαμβάνονται έως και 27 φορές. Για κάθε πρότυπο που

εντοπίζει το Πρόγραμμα, εφαρμόζεται ο Αλγόριθμος *Wraparound Dynamic Programming* και αποτυπώνονται οι αντιστοιχίσεις μεταξύ της ακολουθίας και του προτύπου. Ένα παράδειγμα εύρεσης ενός προτύπου μεγέθους 7 που επαναλαμβάνεται ακριβώς (*exact matching*) πάνω από 7 φορές στο ανθρώπινο γονίδιο *carbonic anhydrase* II (*CA*II) παρουσιάζεται παρακάτω:

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

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

2.7 Αλγόριθμοι Γραμμικού Χώρου

Η εύρεση της κοινής υπακολουθίας δύο μεγάλου μήκους ακολουθιών κατά τη διαδικασία της ολικής (*global*) ή της τοπικής (*local*) αντιστοίχισης εκτός από αρκετό χρόνο υπολογισμού απαιτεί και αρκετό χώρο στη μνήμη ενός Η/Υ. Έτσι, κρίνεται απαραίτητη η εισαγωγή Αλγορίθμων Δυναμικού Προγραμματισμού που αποσκοπούν στη μείωση του χώρου που καταλαμβάνουν κατά την πραγματοποίησή τους. Η μείωση σε χώρο ανάλογο του αθροίσματος των μηκών των ακολουθιών στη διαδικασία της ολικής αντιστοίχισης οφείλεται στον Hirschberg (1975). Έτσι, επιτυγχάνεται η εύρεση της βέλτιστης ολικής αντιστοίχισης σε γραμμικό χώρο (*linear space*). Οι αλγόριθμοι που χρησιμοποιούνται για την απλή περίπτωση της συνάρτησης $g(k) = k\delta$ πραγματοποιούνται με μεθόδους ομοιότητας (*similarity methods*).

Θεωρούμε τον παρακάτω αλγόριθμο που αποτελεί εφαρμογή του *Θεωρήματος 2.5* στην περίπτωση της συνάρτησης $g(k) = k\delta$.

Αλγόριθμος 2.8 (A)

Βήμα 0 : Δηλώνουμε τις δύο ακολουθίες a και b, την τιμή του δ και υπολογίζουμε τις τιμές της συνάρτησης $s(a_i, b_j)$ για $i = 1,...,n+1$ και $j = 1,...,m+1$.

Βήμα 1 : Αναθέτουμε στην 1^{η} στήλη τις τιμές $S_{i,1} \leftarrow -(i-1)\delta$ για $i = 1,...,n+1$ και στην 1^{η}

γραμμή τις τιμές $S_{1,j} \leftarrow -(j-1)\delta$ για $j = 1,...,m+1$.

Βήμα 2 : Υπολογίζουμε τις τιμές του υπόλοιπου $n \times m$ πίνακα από τη σχέση

$S_{i,j} = \max\{S_{i,j-1} - \delta, S_{i-1,j-1} + s(a_i, b_j), S_{i-1,j} - \delta\}$ για $i = 2,...,n+1$

και $j = 2,...,m+1$.

Βήμα 3 : Αναθέτουμε την τιμή $S \leftarrow S_{n+1,m+1}$.

Ο χρόνος περάτωσης του αλγορίθμου αυτού είναι τάξης $O(nm)$ και ο χώρος που καταλαμβάνει στη μνήμη του Η/Υ είναι επίσης τάξης $O(nm)$. Ο Hirschberg (1975) για την εισαγωγή των αλγορίθμων του, βασίστηκε στην παρατήρηση της εξάρτησης του υπολογισμού των τιμών της $i-$γραμμής από τις τιμές της $(i-1)-$γραμμής και παρουσίασε τον παρακάτω αλγόριθμο με χρόνο υπολογισμού τάξης $O(nm)$ και γραμμικό χώρο τάξης $O(n+m)$. Είναι αξιοσημείωτο ότι δεν υπολογίζει τις τιμές ενός $(n+1)\times(m+1)$ πίνακα αλλά κάθε φορά αναθεωρεί τις τιμές μιας λίστας $(m+1)$ στοιχείων με χρήση ενός πίνακα διαστάσεων $2\times(m+1)$. Το τελευταίο στοιχείο της λίστας που προκύπτει αντιστοιχεί στο στοιχείο $S_{n+1,m+1}$ που δίνει ο προηγούμενος αλγόριθμος.

Αλγόριθμος 2.9 (Β)

Βήμα 0 : Δηλώνουμε τις δύο ακολουθίες a και b, την τιμή του δ και υπολογίζουμε τις τιμές της συνάρτησης $s(a_i, b_j)$ για $i = 1,...,n+1$ και $j = 1,...,m+1$.

Βήμα 1 : Αναθέτουμε στη 2^{η} γραμμή του $2\times(m+1)$ πίνακα τις τιμές $T_{2,j} \leftarrow -(j-1)\delta$ για $j = 1,...,m+1$.

Βήμα 2 : Αναθέτουμε στην 1^{η} γραμμή τις τιμές $T_{1,j} \leftarrow T_{2,j}$ για $j = 1,...,m+1$.

Βήμα 3 : Αναθέτουμε την τιμή $T_{2,1} \leftarrow -(i-1)\delta$, όπου $i \geq 2$.

Βήμα 4 : Υπολογίζουμε τις τιμές της $2^{ης}$ γραμμής από τη σχέση

$T_{2,j} = \max\{T_{1,j} - \delta, T_{1,j-1} + s(a_i, b_j), T_{2,j-1} - \delta\}$ για $j = 2,...,m+1$.

Βήμα 5 : Επαναλαμβάνουμε τα *Βήματα 2,3* και *4* για $2 \leq i \leq n+1$.

Βήμα 6 : Αναθέτουμε $S_{n,j} \leftarrow T_{2,j}$ για $j = 1,...,m+1$.

Μια υλοποίηση του αλγορίθμου στο *Mathematica* στις ακολουθίες $\mathrm{a}=\mathrm{GCTGATATAGCT}$ και $\mathrm{b}=\mathrm{GGGTGATTAGCT}$ με $s(a,a)=1$, $s(a,b)=-1$ για $a\neq b$ και $\delta=1$, παρουσιάζεται ακολούθως και δίνει τις ίδιες τιμές στις λίστες με τις τιμές στις γραμμές του πίνακα S.

```
a = {"G", "C", "T", "G", "A", "T", "A", "T", "A", "G", "C", "T"};
b = {"G", "G", "G", "T", "G", "A", "T", "T", "A", "G", "C", "T"};
aa = Prepend[a, "-"];
bb = Prepend[b, "-"];
n = Length[a];
m = Length[b];
nn = Length[aa];
mm = Length[bb];
T = Table[0, {i, 1, 2}, {j, 1, mm}];
s = Table[0, {i, 1, nn}, {j, 1, mm}];
Do[If[aa[[i]] == bb[[j]], x = 1, x = -1]; s[[i, j]] = x, {i, 1, nn}, {j, 1, mm}];
delta = 2;
Do[T[[2, j]] = -(j - 1) * delta, {j, 1, mm}];
Do[
  Do[T[[1, j]] = T[[2, j]], {j, 1, mm}];
  T[[2, 1]] = -(i - 1) * delta;
  Do[T[[2, j]] = Max[T[[1, j]] - delta, T[[1, j - 1]] + s[[i, j]],
     T[[2, j - 1]] - delta], {j, 2, mm}],
  {i, 2, nn}];
S = T[[2]]
```

Ωστόσο η εφαρμογή του αλγορίθμου κρίνεται ανεπαρκής για την εύρεση της άριστης αντιστοίχισης. Όμως, μπορεί να χρησιμοποιηθεί για κατάλληλες υπακολουθίες των ακολουθιών a και b δίνοντας την καλύτερη αντιστοίχιση καταλαμβάνοντας γραμμικό χώρο. Συγκεκριμένα, επιλέγοντας μια υπακολουθία της ακολουθίας a μεγέθους ίσου με $i=\left\lceil \frac{n}{2} \right\rceil$, όπου $n=$ μήκος της ακολουθίας a, μπορούμε να εφαρμόσουμε τον <u>*Αλγόριθμο 2.9 (Β)*</u> για τις ακολουθίες:

$$\mathrm{a}_{1,i}=a_1a_2\ldots a_i \qquad \text{και} \qquad \mathrm{b}_{1,m}=b_1b_2\ldots b_m$$

και για τις ακολουθίες

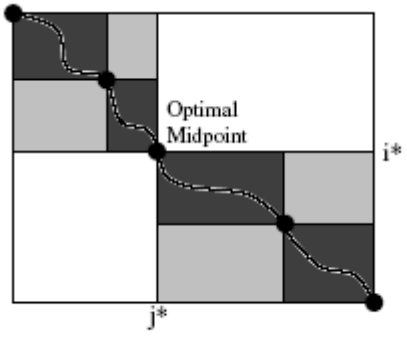
$$\hat{\mathrm{a}}_{n,i+1}=a_na_{n-1}\ldots a_{i+1} \qquad \text{και} \qquad \hat{\mathrm{b}}_{m,1}=b_mb_{m-1}\ldots b_1$$

όπου με $\hat{a}_{n,i+1}$ και $\hat{b}_{m,1}$ δηλώνονται οι αντίστροφες ακολουθίες.

Στην πρώτη εφαρμογή του <u>*Αλγόριθμου Β*</u> τυπώνονται οι τιμές των *scores* $S(a_{1,i},\emptyset)$, $S(a_{1,i},b_1)$, $S(a_{1,i},b_1b_2)$, … , $S(a_{1,i},b_1b_2...b_m)$ και αντίστοιχα στη δεύτερη οι τιμές $S(\hat{a}_{n,i+1},\emptyset)$, $S(\hat{a}_{n,i+1},b_m)$, $S(\hat{a}_{n,i+1},b_mb_{m-1})$, …, $S(\hat{a}_{n,i+1},b_mb_{m-1}...b_1)$, οι οποίες τοποθετούνται σε δύο λίστες. Στη συνέχεια επιλέγεται η τιμή M που μεγιστοποιεί το άθροισμα ενός στοιχείου της πρώτης λίστας που βρίσκεται στη θέση j με το στοιχείο της θέσης $m-j+1$ της άλλης λίστας, για $j=1,...,m+1$ και τίθεται ως k η ελάχιστη τιμή του j σε περίπτωση εμφάνισης της μέγιστης τιμής περισσότερες από μια φορές. Η τιμή του k επιλέγεται για την εφαρμογή του νέου <u>*Αλγορίθμου 2.10 (C)*</u> στις ακολουθίες $a_{1,i}$ & $b_{1,k}$ και στην εφαρμογή του στις ακολουθίες $a_{i+1,n}$ & $b_{k+1,m}$, όπου το ζεύγος τιμών $(\left\lceil \frac{n}{2} \right\rceil, k)$ χαρακτηρίζεται ως βέλτιστο μεσαίο σημείο της αντιστοίχισης (*optimal midpoint of alignment*). Στην ουσία ο <u>*Αλγόριθμος C*</u> επιτυγχάνει διαχωρισμό του προβλήματος σε υποπροβλήματα σύμφωνα με το παρακάτω σχεδιάγραμμα , όπου $i^* = \left\lceil \frac{n}{2} \right\rceil$ και $j^* = k$. Η καμπύλη δείχνει την ενδεχόμενη βέλτιστη αντιστοίχιση, Myers & Miller (1988b).

Σχήμα 2.11

Απεικόνιση εφαρμογής του <u>*Αλγορίθμου C*</u>



Η διατύπωση του αλγορίθμου σε μορφή βημάτων είναι η εξής:

Αλγόριθμος C

Βήμα 0 : Δηλώνουμε τις δύο ακολουθίες a και b, την τιμή του δ και υπολογίζουμε τις τιμές της συνάρτησης $s(a_i, b_j)$ για $i = 1,...,n+1$ και $j = 1,...,m+1$.

Βήμα 1 : Αν ισχύει $m = 0$ τότε αναθέτουμε $c \leftarrow$"–" διαφορετικά

αν $n = 1$ και $\exists j : s(a_1, b_j) = \max\{s(a_1, b_k), 1 \le k \le m+1\}$

τότε αναθέτουμε $c \leftarrow (1, j)$

διαφορετικά $c \leftarrow$"–".

Βήμα 2 : Θέτουμε $i \leftarrow \lfloor n/2 \rfloor$ και εφαρμόζουμε τον *Αλγόριθμο B* ως εξής

$Αλγόριθμο\, B(i, m, \mathrm{a}_{1,\mathrm{i}}, \mathrm{b}_{1,\mathrm{m}}, S1)$

$Αλγόριθμο\, B(n-i, m, \hat{\mathrm{a}}_{\mathrm{n,i+1}}, \hat{\mathrm{b}}_{\mathrm{m,1}}, S2)$

Βήμα 3 : Αναθέτουμε τις τιμές

$M \leftarrow \max\{S1(j) + S2(m-j+1) : 1 \le j \le m+1\}$

$k \leftarrow \min\{j : S1(j) + S2(m-j+1) = M\}$

Βήμα 4 : Εφαρμόζουμε από την αρχή τον *Αλγόριθμο C* ως εξής

$Αλγόριθμο\, C(i, k, \mathrm{a}_{1,\mathrm{i}}, \mathrm{b}_{1,\mathrm{k}}, c_1)$

$Αλγόριθμο\, C(n-i, m-k, \mathrm{a}_{\mathrm{i+1,n}}, \mathrm{b}_{\mathrm{k+1,m}}, c_2)$

Βήμα 5 : Αναθέτουμε $c \leftarrow c_1 \,||\, c_2$

Στο τελευταίο Βήμα με $c_1 \,||\, c_2$ δηλώνεται η αλληλουχία των δύο λιστών c_1 και c_2 που περιέχουν τις θέσεις των βάσεων που αντιστοιχίζονται.

Ο *Αλγόριθμος C* μπορεί να τροποποιηθεί κατάλληλα για την περίπτωση της μορφής της συνάρτησης $g(k) = \alpha + \beta(k-1)$, Myers & Miller (1988b). Με την ελάττωση του χώρου για την πραγματοποίηση τοπικής αντιστοίχισης ασχολήθηκαν οι Huang & Miller (1991). Συγκεκριμένα υπολόγισαν k βέλτιστες «μη - τεμνόμενες» (*non - intersecting*) τοπικές αντιστοιχίσεις καταλαμβάνοντας χώρο τάξης $O(n+m+k)$. Οι αντιστοιχίσεις θεωρούνται «μη – τεμνόμενες» όταν δεν έχουν μεταξύ τους κοινά ζεύγη βάσεων.

Ένας ακόμα γρήγορος αλγόριθμος που καταλαμβάνει χώρο τάξης $O(D_{n,m} \min\{n, m\})$ προσδιορίζει τη βέλτιστη αντιστοίχιση με μεθόδους απόστασης. Αν δεν επιδιώκεται η

αντιστοίχιση, τότε ο χώρος που καταλαμβάνει είναι τάξης $O(D_{n,m})$, Ukkonen (1983, 1984). Ο Ukkonen (1983) παρουσιάζει έναν αλγόριθμο υπολογίζοντας την απόσταση των ακολουθιών a και b σύμφωνα με τη σχέση

$$D_{i,j} = \min\left\{D_{i-1,j-1} + d(a_i, b_j), \min_{k\geq 1}\left\{D_{i,j-k} + g(k)\right\}, \min_{k\geq 1}\left\{D_{i-k,j} + g(k)\right\}\right\}$$

και θεωρεί ότι στην περίπτωση αντιστοίχισης διαφορετικών βάσεων ισχύει η ισότητα $d(a,b) = 1$. Το Λήμμα στο οποίο βασίστηκε για την εφαρμογή του αλγορίθμου του, ισχύει και στη γενική περίπτωση εμφάνισης πολλαπλών *indels*.

Λήμμα 2.2

Έστω δύο ακολουθίες a και b μεγέθους *n* και *m* αντίστοιχα. Τότε για την κατασκευή του πίνακα τιμών $D_{i,j}$ για $1 \leq i \leq n+1$ και $1 \leq j \leq m+1$ ισχύει η σχέση

$$D_{i,j} - 1 \leq D_{i-1,j-1} \leq D_{i,j}.$$

Απόδειξη

Η απόδειξη θα γίνει επαγωγικά ως προς το άθροισμα $i + j$. Η σχέση $D_{i,j} - 1 \leq D_{i-1,j-1}$ είναι προφανής από τον ορισμό της Απόστασης. Για την απόδειξη της σχέσης $D_{i-1,j-1} \leq D_{i,j}$ παρατηρούνται τα εξής: Σε περίπτωση που για τη θέση (i, j) ισχύει η ισότητα $D_{i,j} = D_{i-1,j-1} + d(a_i, b_j)$, ανάλογα με την αντιστοίχιση των βάσεων a_i και b_j θα ισχύει είτε $D_{i,j} = D_{i-1,j-1} + 1$ είτε $D_{i,j} = D_{i-1,j-1}$. Άρα και για τις δυο εκδοχές ισχύει $D_{i,j} \geq D_{i-1,j-1}$. Διαφορετικά έστω ότι ισχύει $D_{i,j} = D_{i-k,j} + g(k)$. Επαγωγικά, αφού $i + j - k \leq i + j$ θα ισχύει $D_{i-k,j} \geq D_{i-(k+1),j-1}$ και έτσι προκύπτει $D_{i,j} = D_{i-k,j} + g(k) \geq D_{i-1-k,j-1} + g(k)$. Αφού όμως $D_{i-1-k,j-1} + g(k) \geq D_{i-1,j-1}$ τελικά προκύπτει το ζητούμενο, $D_{i,j} \geq D_{i-1,j-1}$. □

Η σημαντικότητα του Λήμματος είναι η απόδειξη της ύπαρξης μιας μη-φθίνουσας συνάρτησης $D_{i,i+c}$ ως προς i σύμφωνα με την οποία οι τιμές του πίνακα είναι κατανεμημένες από τη μικρότερη, $D_{1,1} = 0$, που αντιστοιχεί στην αντιστοίχιση των δύο

μηδενικών στοιχείων, μέχρι τη μεγαλύτερη $D_{n+1,m+1}$. Θεωρώντας $g(k)=k$, η βασική ιδέα του αλγορίθμου του Ukkonen (1983) είναι η μετακίνηση κατά μήκος της κύριας διαγωνίου από τη θέση (1,1), όπου $D_{1,1}=0$, μέχρι τη θέση (*i*,*i*) ώστε να προκύψει $D_{i,i}=1$. Γενικά στόχος του αλγορίθμου είναι η εστίαση στις θέσεις αλλαγών των τιμών των κελιών, όταν η τιμή $D_{i,i+c}=k$ αλλάζει σε $D_{i+1,i+1+c}=k+1$. Στη συνέχεια ελέγχουμε όλες τις θέσεις για τις οποίες ισχύει $j-i=c$ έως ότου $D_{i,j}=k+1$. Η επέκταση στην τιμή $k+1$ καθορίζεται από τις προηγούμενες τιμές $k,k-1,...$ και από τον έλεγχο της ισότητας των βάσεων. Η διαδικασία συνεχίζεται μέχρι την εύρεση της τιμής $D_{n+1,m+1}$.

2.8 Παραγωγή Αντιστοιχίσεων με χρήση Tracebacks

Η αντιστοίχιση μεταξύ δύο ακολουθιών πραγματοποιείται μετά την εύρεση του πίνακα των *scores*. Υπάρχουν δύο μέθοδοι που καταλήγουν στην εύρεση όλων των άριστων αντιστοιχίσεων και στηρίζονται στην προς τα πίσω αποτύπωση των ζευγών βάσεων από την μέγιστη τιμή του *score* μέχρι τη μικρότερη (*traceback*). Η πρώτη μέθοδος επιδιώκει την αποθήκευση σε μια λίστα των δεικτών (*saving pointers*) που συμμετέχουν στον υπολογισμό κάθε τιμής $D_{i,j}$ κατά τη διάρκεια εξέλιξης του αλγορίθμου. Στην απλή περίπτωση της συνάρτησης $g(k)=k\delta$, κάθε τιμή στη θέση (i,j) μπορεί να προκύψει από τη συμβολή των τιμών των γειτονικών θέσεων $(i-1,j)$, $(i-1,j-1)$ και $(i,j-1)$ σύμφωνα με τη σχέση

$$D_{i,j}=\min\{D_{i-1,j}+\delta, D_{i-1,j-1}+s(a_i,b_j), D_{i,j-1}+\delta\}.$$

Έτσι, αφού βρεθεί η μέγιστη τιμή $D_{n+1,m+1}$ οι δείκτες «ακολουθούνται» προς τα πίσω για την παραγωγή της άριστης στοίχισης.

Η δεύτερη μέθοδος επιτυγχάνεται μετά τον υπολογισμό και την αποθήκευση του πίνακα $D_{i,j}$, επαναϋπολογίζοντας (*recomputation*) τις τιμές $D_{i-1,j}+\delta$, $D_{i-1,j-1}+s(a_i,b_j)$ και $D_{i,j-1}+\delta$ ώστε να βρεθεί εκείνη που συμβάλλει στον καθορισμό της τιμής $D_{i,j}$. Οι Wagner

& Fischer (1974) πρότειναν τον αλγόριθμο που ακολουθεί, με χρόνο εκτέλεσης τάξης $O(n+m)$.

Αλγόριθμος 2.11 (Wagner & Fischer)

Βήμα 0 : Αφού βρεθεί ο πίνακας $D_{i,j}$ θέτουμε $i = n+1$ και $j = m+1$.

Βήμα 1 : Αν ισχύει $D_{i,j} = D_{i-1,j} + \delta$, τότε θέτουμε $i = i-1$.

Βήμα 2 : Αν ισχύει $D_{i,j} = D_{i,j-1} + \delta$, τότε θέτουμε $j = j-1$.

Βήμα 3 : Αποτυπώνεται το ζεύγος (i, j).

Βήμα 4 : Θέτουμε $i = i-1$ και $j = j-1$.

Βήμα 5 : Τα *Βήματα 1, 2, 3 & 4* επαναλαμβάνονται καθώς $i \neq 1$ και $j \neq 1$.

Η υλοποίηση του αλγορίθμου στο *Mathematica* για τις ακολουθίες $\mathrm{a} = \mathrm{GCTGATATAGCT}$ και $\mathrm{b} = \mathrm{GGGTGATTAGCT}$ με $\delta = 1$ και $\mu = 1$, αφού πρώτα εφαρμοστεί ο *Αλγόριθμος 2.1*, παρουσιάζεται παρακάτω.

```
i = Length[aa]
j = Length[bb]
l = {{i, j}};
While[i ≠ 1 && j ≠ 1,
  Do[
   If[D1[[i, j]] == D1[[i - 1, j]] + delta, i = i - 1; j = j,
    If[D1[[i, j]] == D1[[i, j - 1]] + delta, i = i; j = j - 1,
     i = i - 1;
     j = j - 1]];
   AppendTo[l, {i, j}];
  ]];
Print[l]
```

Τα ζεύγη βάσεων που αντιστοιχίζονται αποτυπώνονται σε μια λίστα *l* και η άριστη αντιστοίχιση παρουσιάζεται στο *Σχήμα 2.12*.

Σχήμα 2.12

Άριστη στοίχιση

$$\begin{array}{cccccccccccc} G & C & - & T & G & A & T & A & T & A & G & C & T \\ | & & & | & | & | & | & & | & | & | & | & | \\ G & G & G & T & G & A & T & - & T & A & G & C & T \end{array}$$

Ο Αλγόριθμος μπορεί να εφαρμοστεί και στην περίπτωση αντιστοίχισης με τη μέθοδο ομοιότητας όπου υπολογίζεται ο πίνακας $S_{i,j}$ σύμφωνα με τη σχέση

$$S_{i,j} = \max\left\{S_{i-1,j} - \delta, S_{i-1,j-1} + s(a_i, b_j), S_{i,j-1} - \delta\right\}.$$

Πολλές φορές από τη βέλτιστη τιμή του *score* με τη διαδικασία των *tracebacks* μπορεί να προκύψουν περισσότερες από μία άριστες αντιστοιχίσεις μεταξύ των ακολουθιών. Αυτό συμβαίνει στην περίπτωση προέλευσης της τιμής του κελιού (i, j) από περισσότερες από μία εκ των τιμών $D_{i-1,j} + \delta$, $D_{i-1,j-1} + s(a_i, b_j)$ και $D_{i,j-1} + \delta$, όταν αυτές προκύψουν ίσες. Τότε οι θέσεις των πιθανών επιλογών τοποθετούνται σε έναν σωρό (*stack*) και σύμφωνα με την αρχή *last in – first out* αποτυπώνεται η άριστη αντιστοίχιση. Μετά την αποτύπωση μιας άριστης αντιστοίχισης, επιστρέφουμε στο σωρό για την αποτύπωση των υπολοίπων άριστων αντιστοιχίσεων.

Ο αλγόριθμος επιδέχεται κατάλληλες τροποποιήσεις για την παραγωγή των τοπικών αντιστοιχίσεων. Οι δυνατότητες επιλογής της τιμής της θέσης (i, j) είναι 4 συμπεριλαμβανομένου του μηδενός, πιο συγκεκριμένα $H_{i-1,j} - \delta$, $H_{i-1,j-1} + s(a_i, b_j)$, $H_{i,j-1} + \delta$ και 0. Πλέον η μέγιστη τιμή του *score* αναζητείται μέσα από όλες τις τιμές του πίνακα και όχι από την τιμή της θέσης $(n+1, m+1)$ ενώ η επιλογή του 0 κατά τη διαδικασία των *tracebacks* δηλώνει τη λήξη της τοπικής αντιστοίχισης.

Ο αλγόριθμος που χρησιμοποιείται στην περίπτωση της τοπικής αντιστοίχισης είναι ο ακόλουθος:

Αλγόριθμος 2.12

Βήμα 0 : Αφού βρεθεί ο πίνακας H, βρίσκουμε τη θέση (i, j) για την οποία ισχύει

$$H_{i,j} = \max\{H_{k,l} : 1 \le k \le n+1, 1 \le l \le m+1\}.$$

Βήμα 1 : Αν ισχύει $H_{i,j} = H_{i-1,j} - \delta$, τότε θέτουμε $i = i-1$.

Βήμα 2 : Αν ισχύει $H_{i,j} = H_{i,j-1} - \delta$, τότε θέτουμε $j = j-1$.

Βήμα 3 : Αποτυπώνεται το ζεύγος (i, j).

Βήμα 4 : Διαφορετικά θέτουμε $i = i-1$ και $j = j-1$.

Βήμα 5 : Τα *Βήματα 1, 2, 3 & 4* επαναλαμβάνονται καθώς $H_{i,j} > 0$.

Η υλοποίηση του αλγορίθμου στο *Mathematica* για την εύρεση της «πρώτης» τοπικής αντιστοίχισης των δυο ακολουθιών, που παρουσιάζεται στο *Σχήμα 2.8*, είναι η ακόλουθη:

```
a = Max[H1];
b = Position[H1, a];
i = b[[1]][[1]];
j = b[[1]][[2]];
delta = 20;
l = {{i, j}};
H1[[i, j]];
While[H1[[i, j]] > 0,
  H1[[i, j]];
  Which[H1[[i, j]] == H1[[i - 1, j - 1]] + s[[i, j]], AppendTo[l, {i = i - 1, j = j - 1}],
   H1[[i, j]] == H1[[i, j - 1]] - delta, AppendTo[l, {i = i, j = j - 1}],
   H1[[i, j]] == H1[[i - 1, j]] - delta, AppendTo[l, {i = i - 1, j = j}]];
 ];
lst = Drop[l, -1];
len = Length[lst];
lst1 = Reverse[lst]
Do[AppendTo[d, lst1[[i, 1]]] && AppendTo[e, lst1[[i, 2]]], {i, 1, len}]
```

Η περίπτωση εμφάνισης πολλαπλών *indels* προϋποθέτει την εφαρμογή του αλγορίθμου ελέγχοντας τις τιμές 3 πινάκων, Clote & Backofen (2000).

2.9 Σχεδόν-Άριστες Αντιστοιχίσεις

Η άριστη αντιστοίχιση μεταξύ δύο ακολουθιών που προκύπτει από τις μεθόδους που μέχρι τώρα έχουν περιγραφεί δεν ανταποκρίνεται πολλές φορές στην «αληθή» αντιστοίχιση, που προκαλείται από τις μεταβολές που έχει υποστεί μια ακολουθία ώστε να προκύψει μια άλλη. Αυτό κυρίως οφείλεται στην αυθαιρεσία επιλογής των παραμέτρων στην πραγματοποίηση των αριθμητικών αλγορίθμων και σε κάποιους άγνωστης φύσης περιορισμούς στις ακολουθίες. Για την καταπολέμηση αυτών των δυσκολιών έχει προταθεί ένας αλγόριθμος για την παραγωγή όλων των αντιστοιχίσεων με τιμή του *score* μέσα στα όρια μιας διευκρινισμένης απόστασης από το μέγιστο. Ο αλγόριθμος αυτός επιτυγχάνει την εύρεση όλων των σχεδόν άριστων αντιστοιχίσεων (*near optimal alignments*) και εφαρμόζεται στην περίπτωση του πίνακα των *scores* απόστασης ή ομοιότητας.

Αφού υπολογιστεί ο πίνακας $D_{i,j}$ στην απλή περίπτωση της συνάρτησης $g(k) = k\delta$ που δίνεται από τη σχέση

$$D_{i,j} = \min\left\{D_{i-1,j} + \delta, D_{i-1,j-1} + d(a_i, b_j), D_{i,j-1} + \delta\right\}, \text{ για } i = 1,...,n+1 \text{ και } j = 1,...,m+1$$

και βρεθεί η μέγιστη τιμή του *score* $D_{n+1,m+1}$, το πρόβλημα που τίθεται είναι η εύρεση όλων των αντιστοιχίσεων με *score* μικρότερο ή ίσο μιας τιμής $D_{n+1,m+1} + e$, όπου $e \geq 0$. Στην ουσία προσεγγίζεται μια τροποποίηση της διαδικασίας των *tracebacks*. Έστω $T_{i,j}$ το *score* αντιστοίχισης κατά την πραγματοποίηση της διαδικασίας *traceback* από τη θέση $(n+1, m+1)$ στη θέση (i, j). Η τιμή $T_{i,j}$ ορίζεται ως το άθροισμα των τιμών για τα βάρη που αναλογούν στις αντιστοιχίες των βάσεων σε όλα τα βήματα που ακολουθούνται από τη θέση $(n+1, m+1)$ μέχρι την εξαιρετέα τιμή της θέσης (i, j). Η εύρεση της επόμενης θέσης στη διαδικασία των *tracebacks* εξαρτάται από την ισχύ των παρακάτω υποθέσεων:

1. Αν ισχύει $T_{i,j} + (D_{i-1,j} + \delta) \leq D_{n+1,m+1} + e$ επιλέγεται η θέση $(i-1, j)$ με $T_{i-1,j} = T_{i,j} + \delta$
2. Αν ισχύει $T_{i,j} + (D_{i-1,j-1} + d(a_i, b_j)) \leq D_{n+1,m+1} + e$ επιλέγεται η θέση $(i-1, j-1)$ με $T_{i-1,j-1} = T_{i,j} + d(a_i, b_j)$

3. Αν ισχύει $T_{i,j} + (D_{i,j-1} + \delta) \le D_{n+1,m+1} + e$ επιλέγεται η θέση $(i, j-1)$ με $T_{i,j-1} = T_{i,j} + \delta$

Κρίνεται απαραίτητη η εξέταση της περίπτωσης επιλογής περισσότερων από μιας θέσης. Γι' αυτό επιδιώκεται η αποθήκευση σε έναν σωρό των πιθανών δεικτών και η αποτύπωση όλων των αντιστοιχίσεων σύμφωνα με την αρχή *last in – first out*, Waterman (1983).

Οι Vingron & Argos (1990) αναφέρουν μια διαφορετική προσέγγιση του προβλήματος χωρίς την απαίτηση εύρεσης όλων των σχεδόν άριστων αντιστοιχίσεων. Στην περίπτωση υπολογισμού του πίνακα ομοιότητας $S_{i,j}$ ελέγχεται η συμμετοχή της θέσης (i, j) σε οποιαδήποτε σχεδόν άριστη αντιστοίχιση με τη βοήθεια της σχέσης

$$S(a_1 \ldots a_{i-1}, b_1 \ldots b_{j-1}) + s(a_i, b_j) + S(a_{i+1} \ldots a_n, b_{j+1} \ldots b_m) \ge S_{n+1,m+1} - e\,.$$

Έτσι για την εφαρμογή της μεθοδολογίας απαιτείται ο υπολογισμός δύο πινάκων ομοιότητας για τις ακολουθίες $\mathrm{a} = a_1 a_2 \ldots a_n$ - $\mathrm{b} = b_1 b_2 \ldots b_m$ και τις αντίστροφες τους $\hat{\mathrm{a}} = a_n a_{n-1} \ldots a_1$ - $\hat{\mathrm{b}} = b_m b_{m-1} \ldots b_1$.

2.10 Αναστροφές

Η αναστροφή (*inversion*) μιας ακολουθίας DNA ορίζεται ως το αντίστροφο συμπλήρωμα (*reverse complement*) της ακολουθίας. Ο Wagner (1975) ασχολήθηκε με την αντιστοίχιση μεταξύ δύο ακολουθιών επιτρέποντας αντικαταστάσεις (*substitutions*), προσθήκες (*insertions*), αφαιρέσεις (*deletions*) και αναστροφές (*inversions*) βάσεων. Όμως, η εισαγωγή των αναστροφών απαιτούσε αρκετό χρόνο αριθμητικού υπολογισμού και θεωρήθηκε υπολογιστικά μη εφαρμόσιμη. Η επόμενη απόπειρα εύρεσης άριστης αντιστοίχισης επιτρέποντας τις αναστροφές πραγματοποιήθηκε μέσω μιας διαφορετικής προσέγγισης.

Οι Schöniger & Waterman (1992) περιγράφουν δύο αλγορίθμους για την εύρεση της βέλτιστης τοπικής αντιστοίχισης μεταξύ δύο ακολουθιών DNA επιτρέποντας την εμφάνιση αναστροφών με χρήση της γραμμικής συνάρτησης $g(k)$ ανάθεσης *indels*, όπου οι αναστροφές δεν επιτρέπεται να τέμνονται μεταξύ τους. Ο λόγος της συμπληρωματικότητας του αντίστροφου τμήματος της ακολουθίας αφορά τη διατήρηση της πολικότητας (*polarity*)

της ακολουθίας DNA.

Η τοπική αντιστοίχιση δύο ακολουθιών a και b επιτυγχάνεται με χρήση της συνάρτησης

$$Z(g,h;i,j) = H(a_g a_{g+1}...a_i, \overline{b}_j \overline{b}_{j-1}...\overline{b}_h)$$

όπου με $Z(g,h;i,j)$ ορίζεται η αντιστοίχιση του τμήματος $a_g a_{g+1}...a_i$ της ακολουθίας a με το τμήμα $b_h b_{h+1}...b_j$ της ακολουθίας b μετά την αναστροφή του με αρχικό ζεύγος αντιστοίχισης με δείκτες (g,h) και τελικό ζεύγος με δείκτες (i,j). Η συνάρτηση H έχει οριστεί μέσω του *Θεωρήματος 2.12* με συναρτήσεις $s_1(a,b)$ και $g_1(k) = \alpha_1 + \beta_1(k-1)$. Η εφαρμογή της αναστροφής επιβαρύνει με κόστος γ την πραγματοποίηση ενός δεύτερου αλγορίθμου για τις αρχικές ακολουθίες a και b με αντίστοιχες συναρτήσεις $s_2(a,b)$ και $g_2(k) = \alpha_2 + \beta_2(k-1)$.

Ο αλγόριθμος που ακολουθεί υπολογίζει όλες τις τοπικές αντιστοιχίσεις που μπορούν να πραγματοποιηθούν για δύο ακολουθίες DNA επιτρέποντας την ύπαρξη αναστροφών.

Αλγόριθμος 2.13 (All Inversions)

Βήμα 0 : Ορίζουμε τις ακολουθίες a και b με την εισαγωγή ενός μηδενικού στοιχείου σε κάθε μία.

Βήμα 1 : Για $i = 1$ ή $j = 1$ θέτουμε $U(i,j) = V(i,j) = W(i,j) = 0$.

Βήμα 2 : Ξεκινώντας από $i = 2$ και $j = 2$ υπολογίζουμε τις σχέσεις

$$U(i,j) = \max\{U(i-1,j) + \beta_2, W(i-1,j) + \alpha_2\}$$

$$V(i,j) = \max\{V(i,j-1) + \beta_2, W(i,j-1) + \alpha_2\}.$$

Βήμα 3 : Για $g = 2,...,i$ και $h = 2,...,j$ υπολογίζεται η συνάρτηση $Z(g,h;i,j)$.

Βήμα 4 : Υπολογίζεται η ποσότητα

$$W(i,j) = \max\left\{\begin{matrix} \max\limits_{\substack{2\le g\le i \\ 2\le h\le j}}\{W(g-1,h-1) + Z(g,h;i,j)\} + \gamma, \\ W(i-1,j-1) + s_2(a_i,b_j), U(i,j), V(i,j), 0 \end{matrix}\right\}.$$

Βήμα 5 : Τα *Βήματα 2, 3 & 4* επαναλαμβάνονται για $i = 3,...,n+1$ και $j = 3,...,m+1$.

Βήμα 6 : Το *score* της καλύτερης αναστροφής βρίσκεται μέσω της σχέσης

$$\max\{W(i,j) : 2 \le i \le n+1,\ \ 2 \le j \le m+1\}.$$

Ο αλγόριθμος αυτός υπολογίζει όλες τις βέλτιστες τοπικές αντιστοιχίσεις των ακολουθιών σε υπολογιστικό χρόνο τάξης $O(n^6)$ για $n = m$ (*Βήμα 3*). Όμως, στη Βιολογία ενδιαφέρονται μόνο για τις τοπικές αντιστοιχίσεις, όπου οι αντιστοιχίσεις μεταξύ της ακολουθίας a και της ανάστροφης της b είναι μεγάλου μήκους, σύμφωνα με τους Howe et al. (1988) και Zhou et al. (1988), οπότε ο αλγόριθμος χρειάζεται να τροποποιηθεί ώστε να ικανοποιήσει αυτό το αίτημα στο μικρότερο δυνατό υπολογιστικό χρόνο.

Ένας αποδοτικός αλγόριθμος πραγματοποιείται σε δύο μέρη μειώνοντας αρκετά το χρόνο εύρεσης της βέλτιστης αναστροφής. Πρώτα εφαρμόζεται ο αλγόριθμος τοπικής αντιστοίχισης για τις ακολουθίες $\mathbf{a} = a_1 ... a_n$ και $\mathbf{b}^{(inv)} = \overline{b}_m ... \overline{b}_1$ που περιγράφεται μέσω του *Θεωρήματος 2.12* με συναρτήσεις $s_1(a,b)$ και $g_1(k) = \alpha_1 + \beta_1(k-1)$. Χρησιμοποιώντας τη μέθοδο επαναϋπολογισμού του πίνακα H των Waterman & Eggert (1987), προκύπτουν οι K βέλτιστες τοπικές αντιστοιχίσεις γι' αυτές τις ακολουθίες. Έτσι, για την παραγωγή της λίστας $\mathcal{L}$ των βέλτιστων αντιστοιχίσεων ο απαιτούμενος χρόνος υπολογισμού είναι τάξης $O(nm + \sum_{i=1}^{K} L_i^2)$, όπου το L_i συμβολίζει το μήκος της i αντιστοίχισης. Η επιλογή του K είναι αυθαίρετη αλλά ικανή να μειώσει το χρόνο υπολογισμού εύρεσης των τοπικών αντιστοιχίσεων. Ο χρόνος υπολογισμού μπορεί ακόμα να μειωθεί επιλέγοντας κατάλληλη τιμή C_1, ώστε να υπάρχει μικρή πιθανότητα εύρεσης K-οστής αντιστοίχισης με τιμή του *score* μεγαλύτερη από αυτήν.

Αλγόριθμος 2.14 (Best Inversions)

Βήμα 0 : Ορίζουμε τις ακολουθίες a και b με την εισαγωγή ενός μηδενικού στοιχείου σε κάθε μία.

(I)

Βήμα 1 : Εφαρμόζουμε τη μέθοδο των Waterman - Eggert στις ακολουθίες **a** και $\mathbf{b}^{(inv)}$ και προκύπτει η λίστα $\mathcal{L} = \{Z(g,h;i,j),(g,h),(i,j) : K \quad \beta\acute{\varepsilon}\lambda\tau\iota\sigma\tau\varepsilon\varsigma\}$.

(II)

Βήμα 2 : Για $i = 1$ ή $j = 1$ θέτουμε $U(i,j) = V(i,j) = W(i,j) = 0$.

Βήμα 3 : Ξεκινώντας από $i = 2$ και $j = 2$ υπολογίζουμε τις ποσότητες

$$U(i,j) = \max\{W(i-1,j) + \alpha_2, U(i-1,j) + \beta_2\}$$

$$V(i,j) = \max\{W(i,j-1) + \alpha_2, V(i,j-1) + \beta_2\}$$

$$W(i,j) = \max\begin{Bmatrix} \max\limits_{L}\{W(g-1,h-1) + Z(g,h;i,j)\} + \gamma, \\ W(i-1,j-1) + s_2(a_i,b_j), U(i,j), V(i,j), 0 \end{Bmatrix}.$$

Βήμα 4 : Τα *Βήματα 2 & 3* επαναλαμβάνονται για $i = 3,...,n+1$ και $j = 3,...,m+1$.

Βήμα 6 : Το *score* της καλύτερης αναστροφής βρίσκεται μέσω της σχέσης

$$\max\{W(i,j) : 2 \le i \le n+1,\ \ 2 \le j \le m+1\}.$$

Η εφαρμογή του παραπάνω αλγορίθμου προϋποθέτει την εισαγωγή δύο συναρτήσεων για τον υπολογισμό της ομοιότητας των βάσεων, $s_1(a,b)$ και $s_2(a,b)$, και δύο συναρτήσεων για την ανάθεση *indels*, $g_1(k)$ και $g_2(k)$. Οι συναρτήσεις αυτές μπορεί να έχουν ίδιες τιμές ή διαφορετικές. Η επιλογή των τιμών των παραμέτρων εξαρτάται από την εμπειρία εφαρμογής των αλγορίθμων και τον τρόπο λειτουργίας της αντιστοίχισης κατά την αναστροφή της ακολουθίας b σε σχέση με την αντιστοίχιση των ακολουθιών a και b. Αν οι δύο αντιστοιχίσεις εκτυλίσσονται διαφορετικά, τότε οι τιμές των παραμέτρων των συναρτήσεων $s_1(a,b)$ - $s_2(a,b)$ και $g_1(k)$ - $g_2(k)$ είναι διαφορετικές. Η παράμετρος γ ικανοποιεί τη σχέση $\gamma \ge g(1)$ ενώ η επιλογή της τιμής του K είναι θέμα εμπειρίας και καθορίζεται κατά την εξέλιξη του αλγορίθμου της τοπικής αντιστοίχισης.

Μια υλοποίηση του Μέρους *(I)* του *Αλγορίθμου 2.14* στο *Mathematica* για τις ακολουθίες a = CCAATCTACTACT και b = GCCACTCTCGCTGTACTGTG είναι η ακόλουθη:

```
a = {"C", "C", "A", "A", "T", "C", "T", "A", "C", "T", "A", "C", "T", "G",
   "C", "T", "T", "G", "C", "A"};
b = {"G", "C", "C", "A", "C", "T", "C", "T", "C", "G", "C", "T", "G", "T",
   "A", "C", "T", "G", "T", "G"};
m = Length[b];
binv = {};
Do[If[b[[m + 1 - i]] == "G", AppendTo[binv, "C"],
   If[b[[m + 1 - i]] == "T", AppendTo[binv, "A"],
    If[b[[m + 1 - i]] == "C", AppendTo[binv, "G"],
     If[b[[m + 1 - i]] == "A", AppendTo[binv, "T"]]]]], {i, 1, m}];
aa = Prepend[a, "-"];
bb = Prepend[binv, "-"];
nn = Length[aa];
mm = Length[bb];
a = -20;
b = -5;
H = Table[0, {i, 1, nn}, {j, 1, mm}];
Em = Table[0, {i, 1, nn}, {j, 1, mm}];
F = Table[0, {i, 1, nn}, {j, 1, mm}];
s = Table[0, {i, 1, nn}, {j, 1, mm}];
Do[
  If[aa[[i]] == bb[[j]], x = 10, x = -11]; s[[i, j]] = x, {i, 1, nn}, {j, 1, mm}];
Do[If[i == 1 || j == 1,
   H[[i, j]] = Em[[i, j]] = F[[i, j]] = 0], {i, 1, nn}, {j, 1, mm}];
Do[
  Do[
   Em[[i, j]] = Max[H[[i, j - 1]] + a, Em[[i, j - 1]] + b];
   F[[i, j]] = Max[H[[i - 1, j]] + a, F[[i - 1, j]] + b];
   H[[i, j]] = Max[0, H[[i - 1, j - 1]] + s[[i, j]], Em[[i, j]], F[[i, j]]],
   {j, 2, mm}], {i, 2, nn}];
Print[TableForm[H, TableHeadings → {aa, bb}]]
```

Με τις παραπάνω εντολές αρχικά τυπώνεται ο πίνακας των *scores* H από τον οποίο βρίσκουμε την $1^{η}$ βέλτιστη τοπική αντιστοίχιση για τις ακολουθίες **a** και $\mathbf{b}^{(\mathbf{inv})}$, δηλαδή την $1^{η}$ αναστροφή (*inversion*). Έχουν επιλεγεί η συνάρτηση ομοιότητας να είναι ίση με:

$$s_1(a,b) = \begin{cases} 10, & a = b \\ -11, & a \neq b \end{cases}$$

και η συνάρτηση ανάθεσης *indels* η εξής:

$$g_1(k) = -20 - 5(k-1).$$

Με την εφαρμογή της μεθόδου των *tracebacks* για την εύρεση της $1^{ης}$ αναστροφής βρίσκουμε τη μέγιστη τιμή του *score* ίση με $Z(g,h;i,j) = 39$, όπου $g = 10$, $i = 15$, $h = 10$ και $j = 15$, αφού έχουμε αγνοήσει την αρχική εισαγωγή των μηδενικών στοιχείων (*gaps*) στις δύο ακολουθίες . Μια υλοποίηση της μεθόδου στο *Mathematica* είναι η εξής:

```
p = Position[H, Max[H]];
i = p[[1]][[1]];
j = p[[1]][[2]];
l = {{i, j}};
H[[i, j]];
While[H[[i, j]] > 0,
  H[[i, j]];
  Which[H[[i, j]] == H[[i - 1, j - 1]] + s[[i, j]], AppendTo[l, {i = i - 1, j = j - 1}],
   H[[i, j]] == Em[[i, j]], AppendTo[l, {i = i, j = j - 1}],
   H[[i, j]] == F[[i, j]], AppendTo[l, {i = i, j = j - 1}]]];
lst = Drop[l, -1];
len = Length[lst];
lst1 = Reverse[lst]
```

από την οποία αποτυπώνεται η λίστα *lst1* με τα εξής ζεύγη βάσεων:

```
{{11, 7}, {12, 8}, {13, 9}, {14, 10}, {15, 11}, {16, 12}}
```

Στη συνέχεια, επιλέγοντας την τιμή $K = 2$, εφαρμόζουμε τον αλγόριθμο των *Waterman-Eggert* για την εύρεση της $2^{ης}$ αναστροφής και η υλοποίηση του αλγορίθμου είναι η επόμενη.

```
lst1 = {{11, 7}, {12, 8}, {13, 9}, {14, 10}, {15, 11}, {16, 12}};
d = {};
e = {};
Do[AppendTo[d, lst1[[i, 1]]] && AppendTo[e, lst1[[i, 2]]], {i, 1, len}]
f = d[[1]];
g = e[[1]];
Hnew = Table[0, {i, 1, nn}, {j, 1, mm}];
Emn = Table[0, {i, 1, nn}, {j, 1, mm}];
Fne = Table[0, {i, 1, nn}, {j, 1, mm}];
Do[If[i < f  || j < g, Hnew[[i, j]] = H[[i, j]], 0],
 {i, 1, nn}, {j, 1, mm}]
Do[If[i < f  || j < g, Emn[[i, j]] = Em[[i, j]], 0],
 {i, 1, nn}, {j, 1, mm}]
Do[If[i < f  || j < g, Fne[[i, j]] = F[[i, j]], 0],
 {i, 1, nn}, {j, 1, mm}]
Do[
  Emn[[lst1[[z, 1]], lst1[[z, 2]]]] = Max[H[[lst1[[z, 1]], lst1[[z, 2]] - 1]] + a,
    Em[[lst1[[z, 1]], lst1[[z, 2]] - 1]] + b]
   Fne[[lst1[[z, 1]], lst1[[z, 2]]]] = Max[H[[lst1[[z, 1]] - 1, lst1[[z, 2]]]] + a,
     F[[lst1[[z, 1]] - 1, lst1[[z, 2]]]] + b]
    Hnew[[lst1[[z, 1]], lst1[[z, 2]]]] = Max[0,
      Emn[[lst1[[z, 1]], lst1[[z, 2]]]], F[[lst1[[z, 1]], lst1[[z, 2]]]]];
  Do[
   Emn[[i, j]] = Max[Hnew[[i, j - 1]] + a, Emn[[i, j - 1]] + b];
   Fne[[i, j]] = Max[Hnew[[i - 1, j]] + a, Fne[[i - 1, j]] + b];
   Hnew[[i, j]] = Max[0, Hnew[[i - 1, j - 1]] + s[[i, j]], Emn[[i, j]], Fne[[i, j]]],
   {j, lst1[[z, 2]] + 1, mm}, {i, lst1[[z, 1]], lst1[[z, 1]]}];
  Do[
   Emn[[i, j]] = Max[Hnew[[i, j - 1]] + a, Emn[[i, j - 1]] + b];
   Fne[[i, j]] = Max[Hnew[[i - 1, j]] + a, Fne[[i - 1, j]] + b];
   Hnew[[i, j]] = Max[0, Hnew[[i - 1, j - 1]] + s[[i, j]], Emn[[i, j]], Fne[[i, j]]],
   {i, lst1[[z, 1]] + 1, nn}, {j, lst1[[z, 2]], lst1[[z, 2]]}],
  {z, 1, len}];
Do[
  Do[
   Emn[[i, j]] = Max[Hnew[[i, j - 1]] + a, Emn[[i, j - 1]] + b];
   Fne[[i, j]] = Max[Hnew[[i - 1, j]] + a, Fne[[i - 1, j]] + b];
   Hnew[[i, j]] = Max[0, Hnew[[i - 1, j - 1]] + s[[i, j]], Emn[[i, j]], Fne[[i, j]]],
   {j, Last[e] + 1, mm}], {i, Last[d] + 1, nn}];
Hnew2 = TableForm[Hnew, TableHeadings → {aa, bb}]
```

Εφαρμόζουμε πάλι τη μέθοδο των *tracebacks* για την εύρεση της $2^{ης}$ αναστροφής και βρίσκουμε τη μέγιστη τιμή του *score* ίση με $Z(g,h;i,j)=30$, όπου $g=7$, $i=9$, $h=13$ και

Σχήμα 2.13

Πρώτη «τοπική» στοίχιση για τις ακολουθίες **a** και $\mathbf{b}^{(inv)}$

		C	A	C	A	G	T	A	C	A	G	C	G	A	G	A	G	T	G	G	C
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	10	0	10	0	0	0	0	10	0	0	10	0	0	0	0	0	0	0	0	10
C	0	10	0	10	0	0	0	0	10	0	0	10	0	0	0	0	0	0	0	0	10
A	0	0	20	0	20	0	0	10	0	20	0	0	0	10	0	10	0	0	0	0	0
A	0	0	10	9	10	9	0	10	0	10	9	0	0	10	0	10	0	0	0	0	0
T	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	10	0	0	0
C	0	10	0	10	0	0	0	8	10	0	0	10	0	0	0	0	0	0	0	0	10
T	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	10	0	0	0
A	0	0	10	0	10	0	0	20	0	10	0	0	0	10	0	10	0	0	0	0	0
C	0	10	0	20	0	0	0	0	30	10	5	10	0	0	0	0	0	0	0	0	10
T	0	0	0	0	9	0	**10**	0	10	19	0	0	0	0	0	0	0	10	0	0	0
A	0	0	10	0	10	0	0	**20**	5	20	8	0	0	10	0	10	0	0	0	0	0
C	0	10	0	20	0	0	0	0	**30**	10	9	18	0	0	0	0	0	0	0	0	10
T	0	0	0	0	9	0	10	0	10	**19**	0	0	7	0	0	0	0	10	0	0	0
G	0	0	0	0	0	19	0	0	5	0	**29**	9	10	0	10	0	10	0	20	10	0
C	0	10	0	10	0	0	8	0	10	0	9	**39**	19	14	9	4	0	0	0	9	20
T	0	0	0	0	0	0	10	0	0	0	4	19	28	8	3	0	0	10	0	0	0
T	0	0	0	0	0	0	10	0	0	0	0	14	8	17	0	0	0	10	0	0	0
G	0	0	0	0	0	10	0	0	0	0	10	9	24	4	27	7	10	0	20	10	0
C	0	10	0	10	0	0	0	0	10	0	0	20	4	13	7	16	0	0	0	9	20
A	0	0	20	0	20	0	0	10	0	20	0	0	9	14	2	17	5	0	0	0	0

Σχήμα 2.14

Δεύτερη «τοπική» στοίχιση για τις ακολουθίες **a** και $\mathbf{b}^{(inv)}$

		C	A	C	A	G	T	A	C	A	G	C	G	A	G	A	G	T	G	G	C
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	10	0	10	0	0	0	0	10	0	0	10	0	0	0	0	0	0	0	0	10
C	0	10	0	10	0	0	0	0	10	0	0	10	0	0	0	0	0	0	0	0	10
A	0	0	20	0	20	0	0	10	0	20	0	0	0	10	0	10	0	0	0	0	0
A	0	0	10	9	10	9	0	10	0	10	9	0	0	10	0	10	0	0	0	0	0
T	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	10	0	0	0
C	0	10	0	10	0	0	0	8	10	0	0	10	0	0	0	0	0	0	0	0	10
T	0	0	0	0	0	0	**10**	0	0	0	0	0	0	0	0	0	0	10	0	0	0
A	0	0	10	0	10	0	0	**20**	0	10	0	0	0	10	0	10	0	0	0	0	0
C	0	10	0	20	0	0	0	0	**30**	10	5	10	0	0	0	0	0	0	0	0	10
T	0	0	0	0	9	0	0	0	10	19	0	0	0	0	0	0	0	10	0	0	0
A	0	0	10	0	10	0	0	0	5	20	8	0	0	10	0	10	0	0	0	0	0
C	0	10	0	20	0	0	0	0	0	0	9	18	0	0	0	0	0	0	0	0	10
T	0	0	0	0	9	0	10	0	0	0	0	0	7	0	0	0	0	10	0	0	0
G	0	0	0	0	0	19	0	0	0	0	0	0	10	0	10	0	10	0	20	10	0
C	0	10	0	10	0	0	8	0	10	0	0	0	0	0	0	0	0	0	0	9	20
T	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	10	0	0	0
T	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	10	0	0	0
G	0	0	0	0	0	10	0	0	0	0	10	0	10	0	10	0	10	0	20	10	0
C	0	10	0	10	0	0	0	0	10	0	0	20	0	0	0	0	0	0	0	9	20
A	0	0	20	0	20	0	0	10	0	20	0	0	9	10	0	10	0	0	0	0	0

$j=15$, αφού έχουμε αγνοήσει την αρχική εισαγωγή των μηδενικών στοιχείων (*gaps*) στις δύο ακολουθίες . Έτσι, αποτυπώνεται η λίστα *lst2* με τα εξής ζεύγη βάσεων:

```
{{8, 7}, {9, 8}, {10, 9}}
```

Οι δύο πίνακες που βρέθηκαν με τους παραπάνω αλγορίθμους, παρουσιάζονται αντίστοιχα στα *Σχήματα 2.13* & *2.14* στα οποία εμφανίζονται οι παρακάτω βέλτιστες τοπικές αντιστοιχίσεις:

$$\begin{array}{cccccc} T & A & C & T & G & C \\ | & | & | & & | & | \\ T & A & C & A & G & C \end{array}$$

και

$$\begin{array}{ccc} T & A & C \\ | & | & | \\ T & A & C \end{array}.$$

Η υλοποίηση του Μέρους *(II)* του <u>*Αλγορίθμου 2.14*</u> στο *Mathematica* με συνάρτηση ομοιότητας $s_2(a,b)=s_1(a,b)$, συνάρτηση ανάθεσης *indels* $g_2(k)=g_1(k)$ και κόστος αναστροφής ίσο με $\gamma=-2$ δίνεται παρακάτω απ' όπου προκύπτει η άριστη τοπική αντιστοίχιση με χρήση της $1^{ης}$ αναστροφής.

$$\begin{array}{cccccccccccccccc} C & C & A & A & T & C & T & A & * & * & * & * & * & * & * & T & T & G \\ | & | & | & & | & | & | & & & & & & & & & & | & | \\ C & C & A & C & T & C & T & C & * & * & * & * & * & * & * & C & T & G \end{array}$$

Στη βέλτιστη τοπική αντιστοίχιση, με αστερίσκο έχουν αντικατασταθεί τα ζεύγη βάσεων που έχουν αντιστοιχηθεί κατά την εφαρμογή της $1^{ης}$ αναστροφής. Ο τελικός πίνακας W και η άριστη αντιστοίχιση παρουσιάζονται στο *Σχήμα 2.15*. Η έντονη γραφή και υπογράμμιση δηλώνει τα ζεύγη βάσεων της άριστης τοπικής αντιστοίχισης, ενώ με πλάγια γραφή και υπογράμμιση δηλώνεται η άριστη τοπική αντιστοίχιση με χρήση της $1^{ης}$ αναστροφής.

```
a = {"C", "C", "A", "A", "T", "C", "T", "A", "C", "T", "A", "C", "T", "G",
"C", "T", "T", "G", "C", "A"};
b = {"G", "C", "C", "A", "C", "T", "C", "T", "C", "G", "C", "T", "G", "T",
   "A", "C", "T", "G", "T", "G"};
n = Length[a];
m = Length[b];
aa = Prepend[a, "-"];
bb = Prepend[b, "-"];
nn = Length[aa];
mm = Length[bb];
a = -20;
b = -5;
gama = -2;
W = Table[0, {i, 1, nn}, {j, 1, mm}];
U = Table[0, {i, 1, nn}, {j, 1, mm}];
V = Table[0, {i, 1, nn}, {j, 1, mm}];
s = Table[0, {i, 1, nn}, {j, 1, mm}];
L = {{39, {11, 11}, {16, 16}}, {30, {8, 14}, {10, 16}}};
Do[
  If[aa[[i]] == bb[[j]], x = 10, x = -11]; s[[i, j]] = x, {i, 1, nn}, {j, 1, mm}];
Do[If[i == 1 || j == 1,
   W[[i, j]] = U[[i, j]] = V[[i, j]] = 0], {i, 1, nn}, {j, 1, mm}];
Do[
  If[i == L[[1]][[3]][[1]] && j == L[[1]][[3]][[2]],
   Do[
    V[[i, j]] = Max[W[[i, j - 1]] + a, V[[i, j - 1]] + b];
    U[[i, j]] = Max[W[[i - 1, j]] + a, U[[i - 1, j]] + b];
    W[[i, j]] = Max[W[[L[[1]][[2]][[1]] - 1, L[[1]][[2]][[2]] - 1]] + L[[1]][[1]] + gama,
      0, W[[i - 1, j - 1]] + s[[i, j]], U[[i, j]], V[[i, j]]]],
   If[i == L[[2]][[3]][[1]] && j == L[[2]][[3]][[2]],
    Do[
     V[[i, j]] = Max[W[[i, j - 1]] + a, V[[i, j - 1]] + b];
     U[[i, j]] = Max[W[[i - 1, j]] + a, U[[i - 1, j]] + b];
      W[[i, j]] = Max[W[[L[[2]][[2]][[1]] - 1, L[[2]][[2]][[2]] - 1]] + L[[2]][[1]] + gama,
        0, W[[i - 1, j - 1]] + s[[i, j]], U[[i, j]], V[[i, j]]]],
    Do[
     V[[i, j]] = Max[W[[i, j - 1]] + a, V[[i, j - 1]] + b];
     U[[i, j]] = Max[W[[i - 1, j]] + a, U[[i - 1, j]] + b];
     W[[i, j]] = Max[0, W[[i - 1, j - 1]] + s[[i, j]], U[[i, j]], V[[i, j]]]]]],
  {i, 2, nn}, {j, 2, mm}];
Print[TableForm[W, TableHeadings → {aa, bb}]]
```

Σχήμα 2.15

Άριστη «τοπική» στοίχιση για τις ακολουθίες **a** και **b** επιτρέποντας αναστροφές

		G	C	C	A	C	T	C	T	C	G	C	T	G	T	A	C	T	G	T	G
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	**10**	10	0	10	0	10	0	10	0	10	0	0	0	0	10	0	0	0	0
C	0	0	10	**20**	0	10	0	10	0	10	0	10	0	0	0	0	10	0	0	0	0
A	0	0	0	0	**30**	10	5	0	0	0	0	0	0	0	0	10	0	0	0	0	0
A	0	0	0	0	10	**19**	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0
T	0	0	0	0	5	0	**29**	9	10	0	0	0	10	0	10	0	0	10	0	10	0
C	0	0	10	10	0	15	9	**39**	19	20	9	10	0	0	0	0	10	0	0	0	0
T	0	0	0	0	0	0	25	19	**49**	29	24	19	20	9	10	0	0	20	0	10	0
A	0	0	0	0	10	0	5	14	29	38	18	13	8	9	0	20	0	0	9	0	0
C	0	0	10	10	0	20	0	15	24	**39**	27	28	9	4	0	28	30	10	5	0	0
T	0	0	0	0	0	0	30	10	25	19	*28*	*16*	*38*	*18*	*14*	*8*	17	40	20	15	10
A	0	0	0	0	10	0	10	19	14	14	*8*	*17*	*18*	*27*	*7*	*24*	5	20	29	9	4
C	0	0	10	10	0	20	5	20	9	24	*4*	*18*	*13*	*7*	*16*	*4*	34	15	9	18	0
T	0	0	0	0	0	0	30	10	30	10	*13*	*0*	*28*	*8*	*17*	*5*	14	44	24	19	14
G	0	10	0	0	0	0	10	19	10	19	*20*	*2*	*8*	*38*	*18*	*13*	9	24	54	34	29
C	0	0	20	10	0	10	5	20	8	20	*8*	*30*	*10*	*18*	*27*	*76*	56	51	46	43	36
T	0	0	0	9	0	0	20	0	30	10	9	10	40	20	28	56	**65**	66	46	56	36
T	0	0	0	0	0	0	10	9	10	19	0	5	20	29	30	51	45	**75**	55	56	45
G	0	10	0	0	0	0	0	0	5	0	29	9	15	30	18	46	40	55	**85**	65	66
C	0	0	20	10	0	10	0	10	0	15	9	39	19	14	19	41	56	50	65	74	54
A	0	0	0	9	20	0	0	0	0	0	4	19	28	8	3	36	36	45	60	54	63

Κ Ε Φ Α Λ Α Ι Ο 3

Ένα Μοντέλο Πιθανοτήτων για τη Μελέτη Ακολουθιών DNA

3.1 Εισαγωγή

Στην παράγραφο αυτή, εξετάζουμε το πρόβλημα της αντιστοίχισης δύο αλυσίδων DNA ίσου μήκους, όπου η μία ακολουθία συγκρίνεται με την άλλη, αντιστοιχίζοντας την *i*-οστή βάση της μιας με την *i*-οστή βάση της άλλης, για $i = 1, 2, \ldots$. Στην περίπτωση αυτή, παραλείπεται το μηδενικό στοιχείο "-" και δεν χρησιμοποιούνται οι έννοιες της προσθήκης ή αφαίρεσης βάσεων (*indels*) κατά την τοποθέτηση της μιας ακολουθίας κάτω από την άλλη. Όταν, κατά την αντιστοίχιση, βρίσκουμε ταυτόσημες βάσεις λέμε ότι έχουμε σύμπτωση των στοιχείων (*match*) ενώ όταν εμφανίζονται διαφορετικές μεταξύ τους βάσεις, έχουμε μη σύμπτωση (*mismatch*). Οι δύο αυτές έννοιες στο χώρο των πιθανοτήτων μεταφράζονται ως επιτυχία (*success*) και αποτυχία (*failure*) αντίστοιχα ή συμβολικά χρησιμοποιούνται η μονάδα «1» και το μηδέν «0». Έτσι, όλη η διαδικασία αντιστοίχισης δύο ακολουθιών DNA ανάγεται σε μια ακολουθία δύο αποτελεσμάτων, όπως φαίνεται στο παρακάτω σχήμα που αντιστοιχίζονται δύο τμήματα ακολουθιών μήκους $n = 12$.

G	*A*	*C*	*T*	*T*	*G*	*A*	*T*	*G*	*G*	*T*	*C*
G	*G*	*C*	*T*	*A*	*T*	*A*	*T*	*G*	*A*	*T*	*C*
1	*0*	*1*	*1*	*0*	*0*	*1*	*1*	*1*	*0*	*1*	*1*

Η ακολουθία δίτιμων αποτελεσμάτων μελετάται ως προς την εμφάνιση ροών επιτυχιών (*success runs*), γενικευμένων ροών επιτυχιών ή συναρτήσεων σάρωσης (*scans*) και ειδικών σχηματισμών (*patterns*). Οι ροές επιτυχιών αποτελούν μια αδιάκοπη σειρά επιτυχιών που εμφανίζονται μεταξύ δυο αποτυχιών ενώ οι σαρώσεις, ως μια γενίκευση των ροών, αποτελούν τμήματα ακολουθιών στα οποία μεταξύ των επιτυχιών επιτρέπεται η παρουσία

ενός μεγίστου αριθμού αποτυχιών. Έτσι, λέγοντας r-γενικευμένη ροή επιτυχιών μήκους k, εννοούμε τον αριθμό των τμημάτων (*windows*) μήκους k στα οποία περιέχονται τουλάχιστον r επιτυχίες σε μια ακολουθία n αποτελεσμάτων. Οι ροές επιτυχιών και οι συναρτήσεις σάρωσης αποτελούν ειδικές περιπτώσεις των προτύπων ή των σχηματισμών που βρίσκουν μεγάλη εφαρμογή στο πεδίο των ακολουθιών DNA για τη μελέτη ειδικών επαναλαμβανόμενων σχηματισμών (*tandem repeats*).

Όσον αφορά τις ροές που σχετίζονται με δίτιμες δοκιμές *Bernoulli*, γίνεται χρήση διάφορων στατιστικών συναρτήσεων. Ωστόσο, στην παρούσα διπλωματική θα περιοριστούμε στη μελέτη της ακριβούς κατανομής της στατιστικής συνάρτησης $S_{n,k}^{(2)}$ (*k-tuple statistic*), που εκφράζει το συνολικό αριθμό επιτυχιών στις ροές επιτυχίας μεγέθους τουλάχιστον k (*total number of successes in success runs of length at least* k).

Για αυτήν τη στατιστική συνάρτηση έχει μελετηθεί εκτενώς η ασυμπτωτική της κατανομή και έχουν δοθεί προσεγγίσεις και φράγματα. Ακόμα, με χρήση τεχνικών συνδυαστικής ανάλυσης (*combinatory analysis*) έχει μελετηθεί η ακριβής κατανομή της, που όμως είναι αρκετά δύσκολο να υπολογιστεί. Αρκετά πρόσφατα, με τις μελέτες των Fu (1986) και Fu & Koutras (1994), η ακριβής κατανομή της αντιμετωπίζεται με έναν διαφορετικό τρόπο με τη χρήση της τεχνικής της περιγραφής της τυχαίας μεταβλητής, που μας ενδιαφέρει, με μια πεπερασμένη Μαρκοβιανή αλυσίδα (*finite Markov chain imbedding* ή *FMCI*). Σύμφωνα με αυτήν την τεχνική, αρκεί να οριστούν ένας κατάλληλος χώρος καταστάσεων (*state space*), μια διαμέριση (*partition*) και οι πίνακες πιθανοτήτων μετάβασης (*transition probability matrices*) της αλυσίδας. Η προσέγγιση αυτή μπορεί να εφαρμοστεί σε ακολουθίες δίτιμων αποτελεσμάτων (*bistate trials*) ή πολλαπλών αποτελεσμάτων (*multistate trials*) που είναι ανεξάρτητα και ισόνομα κατανεμημένα (*independent and identically distributed* ή *i.i.d.*) ή σε ακολουθίες με Μαρκοβιανά εξαρτημένες δίτιμες δοκιμές με πίνακα πιθανοτήτων μετάβασης

$$\mathcal{A} = \begin{bmatrix} p_{FF} & p_{FS} \\ p_{SF} & p_{SS} \end{bmatrix}$$

ή ακόμα και σε ακολουθία ανεξάρτητων αλλά όχι ισόνομων δίτιμων αποτελεσμάτων.

---

(2) Ο συμβολισμός που χρησιμοποιείται είναι σύμφωνος με την εργασία της Lou (2003). Στη βιβλιογραφία χρησιμοποιούνται επίσης και οι συμβολισμοί S_n^k ή R_n^k.

3.2 Θεωρία των Ροών και των Σχηματισμών (Theory of Runs and Patterns)

Η στοχαστική ανάλυση των αλυσίδων DNA βασίζεται στη θεωρία των ροών και των σχηματισμών (*theory of runs and patterns*) για δίτιμες (*bistate*) δοκιμές *Bernoulli*. Όπως έχουμε αναφέρει, η ανίχνευση των επαναλαμβανόμενων ειδικών σχηματισμών κρίνεται απαραίτητη για τον κλάδο της Μοριακής Βιολογίας. Με αφορμή τη σημαντικότητα της ανίχνευσης των σχηματισμών αυτών σε ακολουθίες DNA, η στατιστική συνάρτηση $S_{n,k}$ βρίσκει σημαντική εφαρμογή. Παρακάτω περιγράφεται η διαδικασία εμφάνισης επαναλαμβανόμενων σχηματισμών σε ακολουθίες DNA.

Το μόριο του DNA υφίσταται τυχαίες μεταβολές (*random mutations*) καθώς οι γενετικές πληροφορίες μεταβιβάζονται από γενιά σε γενιά. Για το λόγο αυτό, οι επαναλαμβανόμενοι σχηματισμοί αποτελούνται από «κατά προσέγγιση» αντίγραφα (*approximate copies*), σημειώνοντας διαφορές μεταξύ τους. Αυτό έχει ως αποτέλεσμα να επιτρέπεται η εμφάνιση αντικαταστάσεων και προσθηκών ή αφαιρέσεων βάσεων. Ένα πραγματικό παράδειγμα εμφάνισης ενός «κατά προσέγγιση» επαναλαμβανόμενου σχηματισμού, που αποτελείται από 39 βάσεις, σε μια ακολουθία DNA εμφανίζεται παρακάτω όπου παρουσιάζονται μόνο 2 από τα 8 αντίγραφα, Benson & Su (1998). Η πραγματική ακολουθία εμφανίζεται στην πρώτη γραμμή και ο επαναλαμβανόμενος σχηματισμός στη δεύτερη γραμμή.

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

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

Στην αντιστοίχιση του τμήματος της ακολουθίας με τον ειδικό σχηματισμό, ο συμβολισμός $*$ δηλώνει την αντικατάσταση των βάσεων και ο συμβολισμός $-$ την εμφάνιση *indel*.

Προκειμένου να ανιχνευθούν τέτοιοι σχηματισμοί σε μια ακολουθία, αναζητούνται ζεύγη ομοίων βάσεων μεγέθους τουλάχιστον k (*matching k-tuples*), όπου η τιμή του k επιλέγεται να είναι ίση τουλάχιστον με 2 για περισσότερη υπολογιστική αποδοτικότητα. Για την

κατανόηση της εφαρμογής του στατιστικού $S_{n,k}$ στις ακολουθίες DNA περιγράφεται το εξής παράδειγμα που έχει δοθεί από την Lou (2003). Σε ένα τμήμα μιας ακολουθίας DNA θεωρούμε ότι εμφανίζεται ένα «κατά προσέγγιση» επαναλαμβανόμενο αντίγραφο ενός ειδικού σχηματισμού 2 φορές. Στο σχήμα που ακολουθεί, όπως προαναφέρθηκε, στην πρώτη γραμμή εμφανίζεται το τμήμα της ακολουθίας DNA και στη δεύτερη ο επαναλαμβανόμενος σχηματισμός ως εξής:

```
*               *                 *     *
C A A G T G T G G G T C   G A A G T G A G G A G C
G A A G T G T G G A T C   G A A G T G T G G A T C
```

Στην αντιστοίχιση της ακολουθίας με τον ειδικό σχηματισμό, το σύμβολο $*$ δηλώνει την αναπλήρωση των βάσεων. Αν αντιστοιχίσουμε τα δύο αυτά τμήματα της ακολουθίας, μπορούμε να δημιουργήσουμε μια ακολουθία δίτιμων δοκιμών που αποτελείται από τις τιμές 0 και 1. Έτσι, η ακολουθία DNA μετά την αντιστοίχιση του ενός τμήματος κάτω από το άλλο μετατρέπεται σε ακολουθία που αποτελείται από τις τιμές 0 και 1 ως εξής:

```
C A A G T G T G G G T C
G A A G T G A G G A G C
0 1 1 1 1 1 0 1 1 0 0 1
  \_____/     \___/
    k=5         k=2
```

Για τη δίτιμη ακολουθία (*binary sequence*) που σχηματίστηκε, θεωρώντας την τιμή του k για τη ροή επιτυχίας ίση με 2, $k = 2$, παρατηρούμε ότι υπάρχουν δύο ροές επιτυχιών μεγέθους τουλάχιστον 2, μια που αποτελείται από μια ροή επιτυχίας μεγέθους 5 και μια ροή μεγέθους ακριβώς ίσου με $k = 2$. Επομένως προκύπτει $S_{12,2} = 7$.

Αφού βρεθεί η τιμή της στατιστικής συνάρτησης $S_{n,k}$, κρίνεται στη συνέχεια απαραίτητη η εύρεση της στατιστικής σημαντικότητας εμφάνισης συνολικών επιτυχιών σε ροές επιτυχίας μεγέθους τουλάχιστον k. Ο υπολογισμός της στατιστικής σημαντικότητας απαιτεί γνώση της κατανομής του στατιστικού $S_{n,k}$ και η επιτυχία της προσέγγισης της μεθόδου αυτής

εξαρτάται από την ακρίβεια υπολογισμού της κατανομής του.

Εξαιτίας της σημαντικότητας των επαναλαμβανόμενων σχηματισμών στο χώρο της Μοριακής Βιολογίας, αρκετοί ήταν αυτοί που ασχολήθηκαν με την εύρεση αλγορίθμων εντοπισμού τους. Ανάμεσα τους, αναδεικνύεται η προσφορά του Benson (1999) με την ανάπτυξη του αλγορίθμου «*Tandem-Repeats-Finder*» (*TRF*). Στη φάση εντοπισμού των σχηματισμών, οι υποψήφιες ακολουθίες υπόκεινται σε έλεγχο υποθέσεων ενός σχηματισμού μήκους n με χρήση της στατιστικής συνάρτησης $S_{n,k}$ υπό την υπόθεση ότι η δίτιμη ακολουθία είναι ανεξάρτητη. Λόγω της πολυπλοκότητας υπολογισμού της ακριβούς κατανομής της στατιστικής συνάρτησης, οι Benson & Su (1998) χρησιμοποίησαν την Κανονική Προσέγγιση υπολογίζοντας τη μέση τιμή και τη διακύμανση με χρήση υπολογιστικών αλγορίθμων.

Οι Goldstein & Waterman (1992) μελέτησαν την κατανομή του στατιστικού $S_{n,k}$ για ακολουθία δοκιμών με πιθανότητα επιτυχίας ίση με $p = 0.25$, χρησιμοποιώντας τη σύνθετη κατανομή *Poisson* (*compound Poisson distribution*) για να προσεγγίσουν την πιθανότητα $P(S_{n,k} = x)$. Οι Fu et al. (2002), για τη μελέτη της ασυμπτωτικής κατανομής του $S_{n,k}$, εφάρμοσαν την τεχνική των Fu & Koutras (1994) σε ακολουθίες Μαρκοβιανών εξαρτημένων δίτιμων δοκιμών (*homogeneous Markov-dependent two-state trials*). Αργότερα, η Lou (2003) χρησιμοποίησε την ίδια μέθοδο για τον υπολογισμό της ακριβούς κατανομής του στατιστικού για ανεξάρτητες και ισόνομες τ.μ. *Bernoulli* με πιθανότητα επιτυχίας ίση με p. Η τεχνική FMCI τροποποιείται από τους Antzoulakos, Bersimis and Koutras (2003) και η ακριβής κατανομή του στατιστικού εξετάζεται υπό την έννοια της τυχαίας μεταβλητής εμφυτεύσιμης σε Μαρκοβιανή αλυσίδα πολυωνυμικού τύπου (MVP), προς υπολογιστική διευκόλυνση. Σε μια τελευταία εργασία, χρησιμοποιείται η τεχνική FMCI από τον Martin (2005) και επεκτείνονται τα αποτελέσματα σε δίτιμες Μαρκοβιανές αλυσίδες μεγαλύτερης τάξης ίσης με *m* (*m-th order Markovian sequences*).

Στη συνέχεια, αναπτύσσονται, με τη χρήση της τεχνικής *FMCI*, η μελέτη της ακριβούς κατανομής του στατιστικού $S_{n,k}$ που αφορά ακολουθίες ανεξάρτητων και ισόνομων δίτιμων δοκιμών (*independently and identically distributed* ή *i.i.d.*) *Bernoulli* με πιθανότητα επιτυχίας ίση με p. Επίσης, για την *i.i.d.* περίπτωση, θα αντιμετωπιστεί η συνάρτηση $S_{n,k}$ ως μια τυχαία μεταβλητή εμφυτεύσιμη σε Μαρκοβιανή αλυσίδα πολυωνυμικού τύπου.

3.3 Ακριβής κατανομή με χρήση της Τεχνικής Finite Markov Chain Imbedding (FMCI)

3.3.1 Μεταβλητή εμφυτεύσιμη σε Μαρκοβιανή Αλυσίδα (Markov chain embeddable variable)

Ο συνολικός αριθμός επιτυχιών σε ροές επιτυχίας μεγέθους τουλάχιστον k μαθηματικά ορίζεται από τη σχέση

$$S_{n,k} = \sum_{i=k}^{n} iR_{n,i} \ , \tag{3.1}$$

όπου με $R_{n,i}$ συμβολίζεται ο αριθμός των ροών επιτυχίας μεγέθους ακριβώς i σε μια ακολουθία n δοκιμών. Οι τιμές της $R_{n,i}$ για κάθε $i = k, k+1, ..., n$ υπολογίζονται στην περίπτωση που η ροή επιτυχίας λαμβάνεται μεταξύ δυο αποτυχιών (*non-overlap counting*), ενώ δεν υπολογίζονται οι επικαλυπτόμενες ροές μεγέθους $\geq k$, που εμφανίζονται σε μια ροή επιτυχίας μεγαλύτερου μήκους (*overlap counting*). Έτσι, για παράδειγμα αν θεωρήσουμε την δίτιμη ακολουθία

1 1 0 1 0 1 1 1

με $n = 8$ και $k = 2$, έχουμε μια ροή μεγέθους ακριβώς ίσου με 2, οπότε $R_{8,2} = 1$, και μια ροή μεγέθους ακριβώς 3, οπότε $R_{8,3} = 1$. Επομένως, ο συνολικός αριθμός επιτυχιών μεγέθους τουλάχιστον 2 θα προκύπτει από τον τύπο (3.1) ως $S_{8,2} = 2 \times 1 + 3 \times 1 = 5$.

Η ακριβής κατανομή του στατιστικού, $S_{n,k}$, προκύπτει σύμφωνα με την αρχή FMCI που αναπτύχθηκε από τους Fu (1986) και Fu & Koutras (1994). Η αρχή αυτή καθιστά εφικτή την προσαρμογή της στατιστικής συνάρτησης $S_{n,k}$ στα πλαίσια μιας πεπερασμένης Μαρκοβιανής αλυσίδας της οποίας η ακριβής κατανομή μπορεί να περιγραφεί με τη βοήθεια πινάκων πιθανοτήτων μετάβασης της προσαρμοσμένης Μαρκοβιανής αλυσίδας.

Σύμφωνα με την αρχή αυτή, η συνάρτηση πιθανότητας (*probability mass function*) μιας πεπερασμένης μη-αρνητικής ακέραιας τ.μ. (*non-negative finite integer-valued random variable*) X_n μπορεί να εκφραστεί με τη βοήθεια μιας Μαρκοβιανής αλυσίδας, αν ικανοποιεί τις προϋποθέσεις του παρακάτω ορισμού.

Ορισμός 3.1 **(Fu & Koutras (1994))**

Μια τυχαία μεταβλητή X_n καλείται μεταβλητή εμφυτεύσιμη σε Μαρκοβιανή αλυσίδα εάν, ικανοποιούνται τα εξής:

(i) Ορίζεται μια πεπερασμένη Μαρκοβιανή αλυσίδα $\{Y_t : t = 0,1,...,n\}$ στον πεπερασμένο χώρο καταστάσεων (*finite state space*) $\{\Omega_t\}$ με πίνακες πιθανοτήτων μετάβασης $\boldsymbol{M_t}$, $t = 1,...,n$ και αρχικό διάνυσμα πιθανότητας (*initial probability vector*) $\boldsymbol{\xi}_0$,

(ii) Υπάρχει μια πεπερασμένη διαμέριση (*partition*) $\{C_x : x = 0,1,...,l\}$ στο χώρο Ω_n, έτσι ώστε για κάθε $x = 0,1,...,l$ η συνάρτηση πιθανότητας ισούται με: $P(X_n = x) = P(Y_n \in C_x \mid \boldsymbol{\xi}_0)$.

Έτσι, αν η $S_{n,k}$ μπορεί να εκφραστεί με τη βοήθεια μιας Μαρκοβιανής αλυσίδας, τότε μπορεί να επιτευχθεί η εύρεση της ακριβούς κατανομής της με την κατάλληλη κατασκευή τριών αναγκαίων συνιστωσών: ενός κατάλληλου χώρου καταστάσεων Ω_n, μιας κατάλληλης διαμέρισης $\{C_x\}$ στο χώρο Ω_n για κάθε ένα από τα $x = 0,1,...,l$ και μιας Μαρκοβιανής αλυσίδας με τους αντίστοιχους πίνακες πιθανοτήτων μετάβασης $\boldsymbol{M}_t$, $t = 1,...,n$.

Η Lou (2003) στην εργασία της, για μια ακολουθία δίτιμων δοκιμών $\{Z_i\}$ με $i = 1,2,...,n$, προσαρμόζει τη στατιστική συνάρτηση $S_{n,k}$ σε Μαρκοβιανή αλυσίδα με την προσθήκη μιας βοηθητικής μεταβλητής E_t, ορίζοντας τα εξής:

(a) Μια Μαρκοβιανή αλυσίδα Y_t, στο σύνολο $w=\{z_1,z_2,...,z_n\}$, είναι ίση με $Y_t(w)=(S_{t,k},E_t)$, όπου με $S_{t,k}$ παριστάνεται ο αριθμός των επιτυχιών μήκους τουλάχιστον k στις πρώτες t δοκιμές και με E_t παριστάνεται η κατάληξη (*ending block*) των πρώτων t δοκιμών. Η κατάληξη E_t καταγράφει τον αριθμό επιτυχιών στην τρέχουσα ροή t και ισούται με 0 κατά την εμφάνιση αποτυχίας, $Z_t=0$, ενώ μπορεί να πάρει τις τιμές από 1 έως $k-1$ όταν εμφανίζονται στη σειρά επιτυχίες με μήκος μικρότερο του k. Χρησιμοποιείται επίσης μια τιμή k^* η οποία συμβολίζει τη ροή επιτυχιών μήκους μεγαλύτερου ή ίσου του k και χρησιμεύει προκειμένου να μειωθεί το μέγεθος του χώρου καταστάσεων. Έτσι, ορίζεται ένα σύνολο $\mathscr{E}=\{0,1,...,k-1,k^*\}$, που αποτελεί τη συλλογή όλων των πιθανών τιμών της βοηθητικής μεταβλητής E_t για κάθε δοκιμή.

Για να γίνει πιο κατανοητή η παραπάνω διαδικασία ορισμού της Μαρκοβιανής αλυσίδας θεωρούμε την ακολουθία $w=\{11010\}$ μεγέθους $n=5$ και επιλέγουμε την τιμή $k=2$. Τότε οι τιμές που παίρνει η ποσότητα $Y_t(w)$, για $t=1,...,5$ είναι οι εξής:

$$\{Y_1=(0,1),\ Y_2=(2,2^*),\ Y_3=(2,0),\ Y_4=(2,1),\ Y_5=(2,0)\}.$$

(b) Το χώρο καταστάσεων $\Omega_t=\{(S_{t,k},E_t)\}$, για $t=1,...,n$, όπου το μέγεθος του Ω_n δίνεται από τη σχέση $d=k+\dfrac{(l+1)(l+2)}{2}+(k+1)(n-k-l)$ με $l=\min\{k,n-k\}$.

(c) Μια διαμέριση που διαμορφώνεται ως εξής

$$C_0=\{(0,0)\},\ C_k=\{(k,0),\ldots,(k,k^*)\},\ldots,\ C_x=\{(x,i):x=0,k,...,n\ \&\ i\in\mathscr{E}\},$$

όπου $\mathscr{E}=\{0,1,...,k-1,k^*\}$.

Τότε για τον υπολογισμό της συνάρτησης πιθανότητας της στατιστικής συνάρτησης $S_{n,k}$

ισχύει το επόμενο θεώρημα που στηρίζεται στο θεώρημα των Fu & Koutras (1994) και αποτελεί άμεση συνέπεια των εξισώσεων Chapman-Kolmogorov.

Θεώρημα 3.4 **(Lou (2003))**

Αν η στατιστική συνάρτηση $S_{n,k}$ μπορεί να εμφυτευτεί σε μια Μαρκοβιανή αλυσίδα, τότε ισχύει

$$P(S_{n,k} = x) = \boldsymbol{\xi}_0 \left(\prod_{t=1}^{n} \boldsymbol{M}_t \right) \boldsymbol{U}'(C_x), \quad x = 0, k, ..., n \tag{3.2}$$

όπου $\boldsymbol{\xi}_0$ είναι το αρχικό διάνυσμα πιθανοτήτων, $\boldsymbol{U}'(C_x) = (0, \ldots, 0, 1, \ldots, 1, 0, \ldots, 0)'$ είναι ένα διάνυσμα στήλη διάστασης $d \times 1$ (d ο πληθικός αριθμός του συνόλου Ω_n), στο οποίο οι μονάδες αντιστοιχούν στις θέσεις που συνδέονται με τις καταστάσεις της διαμέρισης C_x, και $\boldsymbol{M}_t = \left(p_{(u,v)(x,y)} \right), \quad t = 1, ..., n$, είναι οι πίνακες πιθανοτήτων μετάβασης της Μαρκοβιανής αλυσίδας. Θεωρώντας $p = P(Z_i = 1)$ και $q = P(Z_i = 0)$, οι πίνακες πιθανοτήτων μετάβασης ορίζονται από την ακόλουθη σχέση:

$$p_{(u,v)(x,y)} = P\{Y_t = (u,v) \mid Y_{t-1} = (x,y)\} = \begin{cases} q, & u = x, v = 0, \text{ για } \textit{όλες} \text{ τις τιμές του } y \\ p, & u = x, v = y+1, \text{ για } y < k-1 \\ p, & u = x+k, v = y+1, \text{ για } y = k-1 \\ p, & u = x+1, v = y+1, \text{ για } y = k^* \\ 0, & \text{διαφορετικά} \end{cases} \tag{3.3}$$

Στην περίπτωση που ισχύει $y \geq k-1$, προκύπτουν οι εξής ισοδυναμίες $(x+k, y+1) \equiv (x+k, k^*)$ και $(x+1, y+1) \equiv (x+1, k^*)$.

Απόδειξη

Οι πιθανότητες μετάβασης $p_{(u,v)(x,y)}$ προκύπτουν άμεσα από τον τρόπο ορισμού της αλυσίδας $Y_t = (S_{t,k}, E_t)$. Έτσι, ο συνολικός αριθμός επιτυχιών σε ροές επιτυχίας μεγέθους τουλάχιστον k, $S_{n,k}$, μπορεί να εκφραστεί μέσω μιας Μαρκοβιανής αλυσίδας και η εξίσωση

υπολογισμού των πιθανοτήτων του, προκύπτει άμεσα από το Θεώρημα των Fu-Koutras (1994). □

Οι ροπές *r*-τάξης του στατιστικού $S_{n,k}$, $E(S_{n,k}^r)$, δίνονται από τη σχέση

$$E(S_{n,k}^r) = \boldsymbol{\xi}_0 \left(\prod_{t=1}^{n} \boldsymbol{M}_t \right) \boldsymbol{V}_r , \tag{3.4}$$

όπου $\boldsymbol{V}_r = \sum_x x^r \boldsymbol{U}'(C_x)$ και για $r = 1$ προκύπτει η επόμενη σχέση για τη μέση τιμή

$$E(S_{n,k}) = \boldsymbol{\xi}_0 \left(\prod_{t=1}^{n} \boldsymbol{M}_t \right) \sum_x x \boldsymbol{U}'(C_x) .$$

Ο υπολογισμός της διακύμανσης επιτυγχάνεται με τη βοήθεια της σχέσης $Var(S_{n,k}) = \left(E(S_{n,k})\right)^2 - E(S_{n,k}^2)$, όπου η ποσότητα $E(S_{n,k}^2)$ προκύπτει από τη σχέση (3.4) για $r = 2$.

Αξίζει να σημειωθεί ότι η Μαρκοβιανή αλυσίδα $\{Y_t\}$ είναι ομογενής (*homogeneous*) και οι πιθανότητες μετάβασής της είναι ανεξάρτητες της παραμέτρου t, δηλαδή ισχύει $M_t = M$ για όλες τις τιμές του $t = 1, ..., n$. Έτσι, η εξίσωση (3.2) μπορεί να γραφεί στη μορφή

$$P(S_{n,k} = x) = \boldsymbol{\xi}_0 \boldsymbol{M}^n \boldsymbol{U}'(C_x), \quad x = 0, k, ..., n . \tag{3.5}$$

Η μεθοδολογία αυτή μπορεί να εφαρμοστεί για την εύρεση της ακριβούς κατανομής της $S_{n,k}$. Ωστόσο, η κατανομή της στατιστικής συνάρτησης μπορεί να προσεγγιστεί από την κανονική κατανομή για μεγάλες τιμές του n, τιμή του p κοντά στο 1 και αρκετά μικρή τιμή του k/n, Lou (2003).

3.3.2 Μεταβλητή εμφυτεύσιμη σε Μαρκοβιανή αλυσίδα Πολυωνυμικού Τύπου (Markov chain embeddable variable of polynomial type, MVP)

Στην πράξη, πολλές φορές το μήκος των προς σύγκριση ακολουθιών είναι αρκετά μεγάλο και η εφαρμογή της μεθόδου της Μαρκοβιανής εμφύτευσης όπως ορίστηκε από τους Fu & Koutras (1994) απαιτεί τη χρήση πινάκων μεγάλης διάστασης. Κατά συνέπεια, καθιστά μη υπολογιστικά εφαρμόσιμη την παραπάνω μεθοδολογία για ακολουθίες μεγάλου μήκους, αφού στις πράξεις υπεισέρχονται πολλαπλασιασμοί πινάκων μεγάλης διάστασης.

Για την αντιμετώπιση αυτής της δυσκολίας, οι Koutras & Alexandrou (1995) εισήγαγαν την έννοια της τυχαίας μεταβλητής Διωνυμικού Τύπου εμφυτεύσιμης σε Μαρκοβιανή αλυσίδα (*Markov chain embeddable variable of Binomial Type, MVB*). Η ιδέα αυτής της τροποποιημένης τεχνικής της Μαρκοβιανής εμφύτευσης στηρίχτηκε στην παρατήρηση ότι, κάθε πίνακας πιθανοτήτων μετάβασης $\boldsymbol{M}_t$, για $t = 1,\ldots,n$, μπορεί να εμφανιστεί σε μια μορφή διαμερισμένου πίνακα (*blocked matrix*) με μη-μηδενικά στοιχεία μόνο στα διαγώνια τμήματα (υποπίνακες) και στα τμήματα που βρίσκονται ακριβώς δεξιά αυτών (σχηματίζοντας επίσης μια μη κύρια διαγώνιο). Με αυτόν το διαχωρισμό των πινάκων, η ακριβής κατανομή μιας στατιστικής συνάρτησης μπορεί να υπολογιστεί θεωρώντας κατάλληλα διανύσματα πιθανοτήτων (*probability vectors*) που περιγράφουν τη συνολική κατάσταση μιας Μαρκοβιανής διαδικασίας σε χρόνο t.

Μια ευρύτερη κατηγορία μεταβλητών που εμφυτεύονται σε Μαρκοβιανές αλυσίδες είναι οι λεγόμενες μεταβλητές Πολυωνυμικού Τύπου (*Markov chain embeddable variables of polynomial type*, *MVP*) που εισήχθησαν από τους Antzoulakos, Bersimis and Koutras (2003). Για αυτήν την κατηγορία μεταβλητών οι διαμερίσεις C_x, για $x = 0,1,\ldots$, αποτελούνται από το ίδιο πεπερασμένο πλήθος στοιχείων (*cardinality*) που συμβολίζεται με $s = |C_x|$. Η στατιστική συνάρτηση $S_{n,k}$ ανήκει στην κατηγορία των *MVP* μεταβλητών και στη συνέχεια της παραγράφου αυτής θα παραθέσουμε τον ορισμό και τη μεθοδολογία που σχετίζεται με τις μεταβλητές τύπου *MVP* για τον υπολογισμό της συνάρτησης πιθανότητας, της μονής (*generating function*) και της διπλής γεννήτριας συνάρτησης (*double generating function*), καθώς και της ροπογεννήτριας συνάρτησης.

Γενικά, μια πεπερασμένη μη-αρνητική ακέραια τ.μ. X_n ονομάζεται *MVP*, αν ικανοποιεί τις τρεις προϋποθέσεις του επόμενου ορισμού.

Ορισμός 3.2 **(τ.μ. *MVP*)**

Μια τυχαία μεταβλητή X_n καλείται *MVP* (*Markov Chain embeddable variable of polynomial type (MVP)*), αν ικανοποιούνται τα εξής:

(i) Ορίζεται μια Μαρκοβιανή αλυσίδα $\{Y_t : t = 0,1,...,n\}$ στο διακριτό χώρο καταστάσεων (*discrete state space*) Ω που αποτελείται από τις διαμερίσεις $C_x = \{c_{x,0}, c_{x,1}, \ldots, c_{x,s-1}\}$, όπου $c_{x,j} = (x, j)$, έτσι ώστε να ισχύει $\Omega = \bigcup_{x \geq 0} C_x$,

(ii) Υπάρχει ένας θετικός ακέραιος αριθμός $m > 1$ έτσι ώστε για $t \geq 1$ να ισχύει $P(Y_t \in C_y \mid Y_{t-1} \in C_x) = 0$ για όλες τις τιμές του $y \neq x, x+1, \ldots, x+m$ και

(iii) Για κάθε $x \geq 0$ και $n \geq 0$ ως προς τις πιθανότητες ισχύει: $P(X_n = x) = P(Y_n \in C_x)$.

Στην περίπτωση που για το θετικό ακέραιο αριθμό m ισχύει $m = 1$, η τυχαία μεταβλητή X_n θα ανήκει στην ειδικότερη κατηγορία των εμφυτεύσιμων μεταβλητών Μαρκοβιανής αλυσίδας Διωνυμικού Τύπου, Koutras & Alexandrou (1995).

Η χαρακτηριστική ιδιότητα μιας μεταβλητής *MVP* είναι ότι οι διαμερίσεις C_x, $x \geq 0$, ταξινομούνται κατά τέτοιον τρόπο έτσι ώστε όταν η αλυσίδα βρίσκεται σε μια κατάσταση της διαμέρισης C_x μπορεί να μεταβεί σε μια κατάσταση που ανήκει σε μία εκ των διαμερίσεων $C_x, C_{x+1}, \ldots, C_{x+m}$. Οι $m+1$ πίνακες πιθανοτήτων μετάβασης

$$A_{t,i}(x) = \left(P(Y_t = c_{x+i,j'} \mid Y_{t-1} = c_{x,j})\right)_{s \times s}, \text{ για } 0 \leq i \leq m,\ t \geq 1 \text{ και } x \geq 0$$

ικανοποιούν την προϋπόθεση *(ii)* του *Ορισμού 3.2* έτσι ώστε να καθιστούν τον πίνακα $\sum_{i=0}^{m} A_{t,i}(x)$ στοχαστικό.

Επιπλέον, από τα διανύσματα πιθανοτήτων

$$\boldsymbol{f}_t(x) = \left(P(Y_t = c_{x,0}), P(Y_t = c_{x,1}), \ldots, P(Y_t = c_{x,s-1})\right), \text{ για } t \geq 0 \text{ και } x \geq 0$$

λαμβάνοντας υπόψη την προϋπόθεση *(iii)* του *Ορισμού 3.2*, προκύπτει η συνάρτηση πιθανότητας της τ.μ. X_n από την ακόλουθη σχέση

$$P(X_n = x) = \boldsymbol{f}_n(x)(1, 1, \ldots, 1)' = \boldsymbol{f}_n(x)\mathbf{1}', \text{ για } n \geq 0 \text{ και } x \geq 0,$$

όπου με **1** συμβολίζεται το διάνυσμα στήλη διάστασης $1 \times s$ με όλα τα στοιχεία του ίσα με 1.

Για $n = 0$ και $x = 0$ ισχύει η συνθήκη $P(X_0 = 0) = 1$ που οδηγεί στα συμπεράσματα

$$\boldsymbol{\xi}_0\mathbf{1}' = \boldsymbol{f}_0(0)\mathbf{1}' = \left(P(Y_0 = c_{0,0}), P(Y_0 = c_{0,1}), \ldots, P(Y_0 = c_{0,s-1})\right)\mathbf{1}' = 1$$
$$\boldsymbol{\xi}_x\mathbf{1}' = \boldsymbol{f}_0(x)\mathbf{1}' = 0, \text{ για } x \geq 1.$$

Για την εύρεση της συνάρτησης πιθανότητας μιας *MVP* τ.μ. X_n, αρκεί να υπολογιστεί η ακολουθία διανυσμάτων $\boldsymbol{f}_t(x)$. Το επόμενο θεώρημα δίνει μια επαναληπτική σχέση για τον υπολογισμό της.

Θεώρημα 3.5

Η ακολουθία διανυσμάτων (*sequence of vectors*) $\boldsymbol{f}_t(x)$ ικανοποιεί την επαναληπτική σχέση

$$\boldsymbol{f}_t(x) = \sum_{i=0}^{\min(x,m)} \boldsymbol{f}_{t-1}(x-i)A_{t,i}(x-i), \text{ για } t \geq 1 \text{ και } x \geq 0.$$

Απόδειξη

Έστω, $t \geq 1$, $x \geq 0$ και $0 \leq j \leq s-1$. Κάνοντας χρήση του Θεωρήματος ολικής πιθανότητας (*Total Probability Theorem*) προκύπτει ότι

$$P(Y_t = c_{x,j}) = \sum_{i=0}^{\min(x,m)} \sum_{r=0}^{s-1} P(Y_t = c_{x,j} \mid Y_{t-1} = c_{x-i,r})P(Y_{t-1} = c_{x-i,r})$$

$$= \sum_{i=0}^{\min(x,m)} \sum_{r=0}^{s-1} \boldsymbol{e}_{r+1} A_{t,i}(x-i) \boldsymbol{e}'_{j+1} P(Y_{t-1} = c_{x-i,r})$$

$$= \sum_{i=0}^{\min(x,m)} \sum_{r=0}^{s-1} \boldsymbol{f}_{t-1}(x-i) A_{t,i}(x-i) \boldsymbol{e}'_{j+1}$$

$$= \sum_{i=0}^{\min(x,m)} \boldsymbol{f}_{t-1}(x-i) A_{t,i}(x-i) \boldsymbol{e}'_{j+1} ,$$

όπου τα διανύσματα $\boldsymbol{e}_i$ συμβολίζουν τα μοναδιαία διανύσματα-γραμμές του συνόλου $\Re^s$. □

Στη συνέχεια θεωρούμε τη γεννήτρια συνάρτηση, $\varphi_t(z)$, και τη διπλή γεννήτρια συνάρτηση, $\Phi(z,w)$, οι οποίες ορίζονται αντίστοιχα από τις σχέσεις

$$\varphi_t(z) = \sum_{x=0}^{\infty} P(X_t = x) z^x$$

$$\Phi(z,w) = \sum_{t=0}^{\infty} \varphi_t(z) w^t ,$$

ενώ ορίζουμε με $\boldsymbol{\varphi}_t(z)$ και $\boldsymbol{\Phi}(z,w)$ τις αντίστοιχες διανυσματικές γεννήτριες συναρτήσεις της ακολουθίας διανυσμάτων $\boldsymbol{f}_t(x)$ που δίνονται από τις σχέσεις

$$\boldsymbol{\varphi}_t(z) = \sum_{x=0}^{\infty} \boldsymbol{f}_t(x) z^x \text{ , για } t \geq 0$$

$$\boldsymbol{\Phi}(z,w) = \sum_{t=0}^{\infty} \boldsymbol{\varphi}_t(z) w^t .$$

Για αυτές τις συναρτήσεις θα ισχύουν τα εξής

$$\boldsymbol{\varphi}_0(z) = \boldsymbol{\xi}_0 ,$$

$$\varphi_t(z) = \boldsymbol{\varphi}_t(z) \mathbf{1}' \text{, για } t \geq 1 \text{ και}$$

$$\Phi(z,w) = \boldsymbol{\Phi}(z,w)\mathbf{1}'.$$

Συνήθως, οι πίνακες πιθανοτήτων μετάβασης $A_{t,i}(x)$ δεν εξαρτώνται από το x δηλαδή ισχύει $A_{t,i}(x) = A_{t,i}$, για $t \geq 1$ και $x \geq 0$. Αυτό έχει σαν αποτέλεσμα την έκφραση της γεννήτριας συνάρτησης υπό μορφή γινομένου, όπως δίνεται στο παρακάτω θεώρημα.

Θεώρημα 3.6

Αν ισχύει $A_{t,i}(x) = A_{t,i}$, για $t \geq 1$ και $x \geq 0$, το διάνυσμα της γεννήτριας συνάρτησης της τ.μ. X_t προκύπτει από τη σχέση

$$\boldsymbol{\varphi}_t(z) = \boldsymbol{\xi}_0 \prod_{r=1}^{t} \left(\sum_{i=0}^{m} A_{r,i} z^i \right), \text{ για } t \geq 1.$$

Απόδειξη

Για $t \geq 1$, χρησιμοποιώντας το *Θεώρημα 3.5*, η διανυσματική γεννήτρια συνάρτηση προκύπτει ως εξής

$$\boldsymbol{\varphi}_t(z) = \sum_{x=0}^{\infty} \boldsymbol{f}_t(x) z^x = \sum_{x=0}^{m} \sum_{i=0}^{x} \boldsymbol{f}_{t-1}(x-i) A_{t,i} z^x + \sum_{x=m+1}^{\infty} \sum_{i=0}^{m} \boldsymbol{f}_{t-1}(x-i) A_{t,i} z^x$$

$$= \sum_{i=0}^{m} z^i \left(\sum_{x=i}^{m} \boldsymbol{f}_{t-1}(x-i) z^{x-i} \right) A_{t,i} + \sum_{i=0}^{m} z^i \left(\sum_{x=m+1}^{\infty} \boldsymbol{f}_{t-1}(x-i) z^{x-i} \right) A_{t,i}$$

$$= \sum_{i=0}^{m} z^i \left(\sum_{y=0}^{\infty} \boldsymbol{f}_{t-1}(y) z^y \right) A_{t,i} = \boldsymbol{\varphi}_{t-1}(z) \left(\sum_{i=0}^{m} A_{t,i} z^i \right).$$

Επαναλαμβάνοντας την ίδια διαδικασία για τον υπολογισμό των διανυσματικών συναρτήσεων $\boldsymbol{\varphi}_{t-1}(z)$, $\boldsymbol{\varphi}_{t-2}(z)$, …, $\boldsymbol{\varphi}_1(z)$ προκύπτει η προς απόδειξη σχέση. □

Στην περίπτωση που η τ.μ. X_t απαριθμεί σχηματισμούς σε μια ακολουθία ανεξάρτητων

και ισόνομων δοκιμών, οι πίνακες $A_{t,i}(x)$ είναι ανεξάρτητοι από τις παραμέτρους x και t (ομογενής περίπτωση). Τότε, για τη διανυσματική γεννήτρια συνάρτηση $\boldsymbol{\Phi}(z,w)$ προκύπτει το παρακάτω θεώρημα.

Θεώρημα 3.7

Αν ισχύει $A_{t,i}(x) = A_i$ για $t \geq 1$ και $x \geq 0$, τότε η διπλή διανυσματική γεννήτρια συνάρτηση της τ.μ. X_t δίνεται από τη σχέση

$$\boldsymbol{\Phi}(z,w) = \boldsymbol{\xi}_0 \left(I - w \sum_{i=0}^{m} A_i z^i \right)^{-1},$$

όπου ο πίνακας I είναι ο μοναδιαίος πίνακας.

Απόδειξη

Προκύπτει άμεσα με τη βοήθεια του *Θεωρήματος 3.6* ως εξής

$$\boldsymbol{\Phi}(z,w) = \sum_{t=0}^{\infty} \boldsymbol{\varphi}_t(z) w^t = \boldsymbol{\xi}_0 \sum_{t=0}^{\infty} \left(w \sum_{i=0}^{m} A_i z^i \right)^t = \boldsymbol{\xi}_0 \left(I - w \sum_{i=0}^{m} A_i z^i \right)^{-1},$$

όπου η τελευταία ισότητα ισχύει με την προϋπόθεση ότι περιοριζόμαστε σε μια κατάλληλη περιοχή γύρω από το μηδέν για την παράμετρο w. □

Επίσης, για την ομογενή *MVP* τ.μ. X_t το παρακάτω θεώρημα μπορεί να χρησιμοποιηθεί για την εύρεση της μέσης τιμής $\mu_t = E(X_t)$ για $t \geq 1$ και της γεννήτριας συνάρτησης

$$M(w) = \sum_{t=1}^{\infty} \mu_t w^t$$

κάνοντας χρήση των πινάκων πιθανοτήτων μετάβασης A_i, για $i = 0,1,\ldots,m$.

Θεώρημα 3.8

Αν $A_{t,i}(x) = A_i$ για $t \geq 1$ και $x \geq 0$ τότε ισχύουν οι ακόλουθες σχέσεις για τη μέση τιμή και τη γεννήτρια συνάρτηση των μέσων τιμών της τ.μ. X_t

$$\mu_t = E(X_t) = \boldsymbol{\xi}_0 \left(\sum_{r=1}^{t} \left(\sum_{i=0}^{m} A_i \right)^{r-1} \right) \left(\sum_{i=1}^{m} iA_i \right) I'$$

$$M(w) = \sum_{t=1}^{\infty} \mu_t w^t = \frac{w}{1-w} \boldsymbol{\xi}_0 \left(I - w \sum_{i=0}^{m} A_i \right)^{-1} \left(\sum_{i=1}^{m} iA_i \right) \mathbf{1}'.$$

Απόδειξη

Χρησιμοποιώντας τη σχέση

$$\frac{d}{dz} \left(\sum_{i=0}^{m} A_i z^i \right)^t = \sum_{r=1}^{t} \left[\left(\sum_{i=0}^{m} A_i z^i \right)^{r-1} \left(\sum_{i=1}^{m} iA_i z^{i-1} \right) \left(\sum_{i=0}^{m} A_i z^i \right)^{t-r} \right]$$

προκύπτει για τη μέση τιμή ο εξής τύπος

$$\mu_t = \frac{d}{dz} [\boldsymbol{\varphi}_t(z) \mathbf{1}'] \Big|_{z=1} = \boldsymbol{\xi}_0 \sum_{r=1}^{t} \left[\left(\sum_{i=0}^{m} A_i \right)^{r-1} \left(\sum_{i=1}^{m} iA_i \right) \left(\sum_{i=0}^{m} A_i \right)^{t-r} \right] \mathbf{1}',$$

όπου χρησιμοποιείται το γεγονός ότι ο πίνακας $\sum_{i=0}^{m} A_i$ είναι στοχαστικός.

Η γεννήτρια συνάρτηση των μέσων τιμών γράφεται στη μορφή

$$M(w) = \boldsymbol{\xi}_0 \sum_{t=1}^{\infty} \sum_{r=1}^{t} \left[\left(\sum_{i=0}^{m} A_i \right)^{r-1} \left(\sum_{i=1}^{m} iA_i \right) \right] w^t \mathbf{1}'$$

$$= \boldsymbol{\xi}_0 w \sum_{r=1}^{\infty} \left(\sum_{i=0}^{m} A_i \right)^{r-1} w^{r-1} \sum_{t=r}^{\infty} w^{t-r} \left(\sum_{i=1}^{m} iA_i \right) \mathbf{1}'.$$

Όμως, γνωρίζουμε ότι για μια γεωμετρική σειρά $\sum_{r=1}^{\infty} a\omega^{r-1}$ με $a, \omega \in \Re$ και $\omega \neq 0$ ισχύει

$$\lim(\sum_{r=1}^{\infty} a\omega^{r-1}) = \frac{a}{1-\omega} \text{ για } |\omega| < 1,$$

οπότε ανάλογα προκύπτει η ισότητα

$$\sum_{r=1}^{\infty}\left(\sum_{i=0}^{m} A_i\right)^{r-1} w^{r-1} = \left(I - w\sum_{i=0}^{m} A_i\right)^{-1}$$

και με αντικατάσταση προκύπτει η προς απόδειξη σχέση. □

Όλες οι προηγούμενες σχέσεις για $m=1$ δίνουν τα αποτελέσματα που πρότειναν οι Koutras & Alexandrou (1997) για μια τ.μ. *MVB*.

3.3.2.1 Η κατανομή της στατιστικής συνάρτησης $S_{n,k}$ με χρήση της μεθόδου MVP

Όλα τα παραπάνω θεωρήματα εφαρμόζονται για την περίπτωση της αντιστοίχισης δύο ακολουθιών DNA. Θεωρούμε μια ακολουθία $Z_1, Z_2, \ldots, Z_n$ ανεξάρτητων και ισόνομων δίτιμων δοκιμών *Bernoulli*, που έχει προκύψει μετά την αντιστοίχιση δύο ακολουθιών DNA ίσου μήκους n, με πιθανότητα επιτυχίας ίση με $p = P(Z_i = 1)$ και πιθανότητα αποτυχίας ίση με $q = P(Z_i = 0) = 1 - p$, για $i = 1, \ldots, n$. Επίσης, έχουμε αντικαταστήσει την παράμετρο m με το θετικό ακέραιο αριθμό k, όπου $k \leq n$ και για $k \leq i \leq n$ και $l \geq 0$ ορίζουμε τη μεταβλητή

$$U_i = \begin{cases} k+l, & \text{αν} \quad Z_{i-k-l+1} = Z_{i-k-l+2} = \ldots = Z_i = 1 \quad \text{και} \quad Z_{i-k-l} = Z_{i+1} = 0 \\ 0, & \text{διαφορετικά} \end{cases},$$

με αρχική συνθήκη $Z_0 = Z_{n+1} = 0$.

Τότε ο συνολικός αριθμός επιτυχιών σε ροές επιτυχίας μεγέθους τουλάχιστον k σε μια ακολουθία μήκους n εκφράζεται μέσω της σχέσης

$$S_{n,k} = \sum_{i=k}^{n} U_i \ . \tag{3.6}$$

Οι τιμές που μπορεί να πάρει η στατιστική συνάρτηση $S_{n,k}$ ανήκουν στο σύνολο $\{0, k, k+1, \ldots, n\}$, ενώ για $n \le k$ θέτουμε $S_{n,k} = 0$ και το σύνολο τιμών της αποτελείται μόνο από το μηδενικό στοιχείο.

Το στατιστικό $S_{n,k}$ μπορεί να μελετηθεί ως *MVP* εισάγοντας τις διαμερίσεις $C_x = \{c_{x,0}, c_{x,1}, \ldots, c_{x,k}\}$, όπου $c_{x,j} = (x, j)$ με $0 \le j \le k$, $x \ge 0$, και ορίζοντας μια Μαρκοβιανή αλυσίδα $\{Y_t : t \ge 0\}$ στο σύνολο $\Omega = \bigcup_{x \ge 0} C_x$ ως ακολούθως:

$$Y_t = c_{x,j} = (x, j)$$

αν στις πρώτες t δοκιμές, έστω $1001\ldots0\underbrace{11\ldots1}_{r}$, ο συνολικός αριθμός επιτυχιών σε ροές επιτυχίας μεγέθους τουλάχιστον k ισούται με x και

$$j = \begin{cases} r, & \text{αν} \quad r = 0, 1, \ldots, k-1 \\ k, & \text{αν} \quad r \ge k. \end{cases}$$

Είναι προφανές ότι, αν η αλυσίδα βρίσκεται σε μια κατάσταση της διαμέρισης C_x, τότε μπορεί να μεταβεί σε μια κατάσταση που ανήκει σε μία εκ των διαμερίσεων C_x, C_{x+1} ή C_{x+k}. Ως εκ τούτου, η στατιστική συνάρτηση $S_{n,k}$ ανήκει στην κατηγορία *MVP*. Οι πίνακες πιθανοτήτων μετάβασης $A_{t,i} = A_i$, για $i = 0, 1, \ldots, k$, αφού αναφερόμαστε στην *i.i.d.* περίπτωση, προκύπτουν από την παρατήρηση ότι αν ισχύει $Y_t = c_{x,k}$, στην αμέσως επόμενη θέση, $(t+1)$, είτε θα εμφανιστεί μια επιτυχία, $Z_{t+1} = 1$, και η αλυσίδα θα μεταβεί στην

κατάσταση $c_{x+1,k}$ είτε θα εμφανιστεί αποτυχία, $Z_{t+1}=0$, και η αλυσίδα θα μεταβεί στην κατάσταση $c_{x,0}$. Επομένως, η γενική μορφή των πινάκων πιθανοτήτων μετάβασης διαστάσεων $(k+1)\times(k+1)$ παρουσιάζεται παρακάτω και για $i=0$ είναι ως εξής

$$A_0=\begin{bmatrix} q & p & 0 & \dots & 0 & 0 & 0 \\ q & 0 & p & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ q & 0 & 0 & \dots & 0 & p & 0 \\ q & 0 & 0 & \dots & 0 & 0 & 0 \\ q & 0 & 0 & \dots & 0 & 0 & 0 \end{bmatrix}_{(k+1)\times(k+1)}.$$

Για $i=1$, ο πίνακας αποτελείται από μηδενικά στοιχεία εκτός από το στοιχείο της θέσης $(k+1,k+1)$ που ισούται με p

$$A_1=\begin{bmatrix} 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 & p \end{bmatrix}_{(k+1)\times(k+1)},$$

ενώ για $i=2,\dots,k-1$ οι πίνακες πιθανοτήτων μετάβασης $A_2,\dots,A_{k-1}$ διαστάσεων $(k+1)\times(k+1)$ αποτελούνται από μηδενικά στοιχεία. Τέλος, ο πίνακας A_k έχει ένα μοναδικό μη-μηδενικό στοιχείο p στη θέση $(k,k+1)$, δηλαδή έχει τη μορφή

$$A_k = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 & p \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 \end{bmatrix}_{(k+1)\times(k+1)}.$$

Το αρχικό διάνυσμα πιθανοτήτων ισούται με $\boldsymbol{\xi}_0 = (1, 0, 0, \dots, 0)$.

Για παράδειγμα, αν θεωρήσουμε μια ακολουθία δίτιμων δοκιμών μήκους $n = 3$ με πιθανότητα επιτυχίας p και μέγεθος ροών επιτυχιών τουλάχιστον $k = 2$, τότε ο πίνακας μετάβασης $\boldsymbol{M}$ και οι πίνακες A_i, για $i = 0, 1, 2$ είναι οι εξής:

$$\boldsymbol{M} = \begin{bmatrix} q & p & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & & \\ q & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p & & & \\ q & 0 & 0 & 0 & 0 & p & 0 & 0 & 0 & & & \\ & & & q & p & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & q & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p \\ & & & q & 0 & 0 & 0 & 0 & p & 0 & 0 & 0 \\ & & & & & & q & p & 0 & 0 & 0 & 0 \\ & & & & & & q & 0 & 0 & 0 & 0 & 0 \\ & & & & & & q & 0 & 0 & 0 & 0 & p \\ & & & & & & & & & q & p & 0 \\ & & & & & & & & & q & 0 & 0 \\ & & & & & & & & & q & 0 & 1 \end{bmatrix},$$

$$\boldsymbol{A}_0 = \begin{bmatrix} q & p & 0 \\ q & 0 & 0 \\ q & 0 & 0 \end{bmatrix},$$

$$\boldsymbol{A}_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & p \end{bmatrix},$$

$$\boldsymbol{A}_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & p \\ 0 & 0 & 0 \end{bmatrix}.$$

Με τη βοήθεια του *Θεωρήματος 3.5* βρίσκεται η συνάρτηση πιθανότητας της $S_{n,k}$ ενώ μέσω του *Θεωρήματος 3.6* μπορεί να βρεθεί η πιθανογεννήτρια $\varphi_n(z)$ από τον τύπο

$$\varphi_n(z) = \sum_{x=0}^{\infty} P(S_{n,k} = x) z^x = \boldsymbol{\xi}_o \prod_{r=1}^{n} (A_0 + zA_1 + z^k A_k)\mathbf{1}' .$$

Η διπλή διανυσματική γεννήτρια συνάρτηση προκύπτει άμεσα με εφαρμογή του *Θεωρήματος 3.7* και ισούται με το λόγο πολυωνύμων

$$\boldsymbol{\Phi}(z,w) = \sum_{n=0}^{\infty} \varphi_n(z) w^n = \boldsymbol{\xi}_o \left(I - w(A_0 + zA_1 + z^k A_k)\right)^{-1} \mathbf{1}' = \frac{P_1(z,w)}{P_2(z,w)} , \qquad (3.7)$$

όπου

$$P_1(z,w) = 1 - wpz - (wp)^k (1 - z^k) - (wp)^{k+1} (z^k - z),$$
$$P_2(z,w) = 1 - w(1 + pz) - w^2 pz + w^{k+1} qp^k (1 - z^k) + w^{k+2} qp^{k+1} (z^k - z)$$

(με I συμβολίζεται ο μοναδιαίος πίνακας διαστάσεων $(k+1, k+1)$).

Επίσης, η διπλή γεννήτρια συνάρτηση μπορεί να γραφεί και στην μορφή

$$\Phi(z,w) = \sum_{n=0}^{\infty} \varphi_n(z) w^n = \frac{1 - wa_1(z) - w^k a_2(z) - w^{k+1} a_3(z)}{1 - \left[wb_1(z) + w^2 b_2(z) + w^{k+1} b_3(z) + w^{k+2} b_4(z)\right]},$$

όπου οι $a_i(z)$, για $i=1,2,3$ και $b_i(z)$, για $i=1,2,3,4$ είναι κατάλληλες συναρτήσεις ως προς z.

Χρησιμοποιώντας τη σχέση (3.7) προκύπτει ένας εύχρηστος επαναληπτικός τύπος για την πιθανογεννήτρια συνάρτηση $\varphi_n(z)$, όπως δείχνεται στο θεώρημα που ακολουθεί.

Θεώρημα 3.9

Αν $Z_1, Z_2, \ldots, Z_n$ είναι μια ακολουθία ανεξάρτητων και ισόνομων τ.μ. *Bernoulli*, τότε η γεννήτρια συνάρτηση $\varphi_n(z)$ της τ.μ. $S_{n,k}$ ικανοποιεί τον επαναληπτικό τύπο

$$\begin{aligned}\varphi_n(z) &= (1+pz)\varphi_{n-1}(z) - pz\varphi_{n-2}(z) \\ &\quad - qp^k(1-z^k)\varphi_{n-k-1}(z) - qp^{k+1}(z^k - z)\varphi_{n-k-2}(z), \qquad n \geq k+2\end{aligned}$$

με αρχικές συνθήκες

$$\varphi_n(z) = \begin{cases} 1, & 0 \leq n < k \\ 1 - p^k + (pz)^k, & n = k \\ 1 - p^k(1+q) + 2q(pz)^k + (pz)^{k+1}, & n = k+1. \end{cases}$$

Απόδειξη

Η προς απόδειξη σχέση προκύπτει αφού γράψουμε τη σχέση (3.7) στη μορφή

$$P_2(z,w)\sum_{n=0}^{\infty}\varphi_n(z)w^n = P_1(z,w).$$

Στη συνέχεια, εκτελούμε τους πολλαπλασιασμούς στο αριστερό μέλος της εξίσωσης και αντιστοιχίζουμε μεταξύ τους, τους συντελεστές του w^n, για $n=0,1,2,\ldots$, στη δυναμοσειρά που προκύπτει. □

Η συνάρτηση πιθανότητας της $S_{n,k}$, $g_n(x) = P(S_{n,k} = x)$, για $x \geq 0$, μπορεί να υπολογιστεί με εφαρμογή του *Θεωρήματος 3.5*, με επαναληπτικούς υπολογισμούς διανυσμάτων πιθανότητας χρησιμοποιώντας πίνακες πιθανοτήτων μετάβασης. Συγκεκριμένα μπορεί να χρησιμοποιηθεί η έκφραση

$$P(S_{n,k} = x) = \boldsymbol{f}_n(x)\mathbf{1}', \text{ για } n \geq 0 \text{ και } x = 0,1,2,\ldots,n,$$

με $\boldsymbol{f}_n(x) = \boldsymbol{0}$ για $x = 1,2,\ldots,k-1$ και αναδρομικό τύπο

$$\boldsymbol{f}_n(x) = \sum_{i=0}^{\min(x,k)} \boldsymbol{f}_{n-1}(x-i)A_i \text{ , για } n \geq 1 \text{ και } x \geq 0 .$$

Ωστόσο, στην περίπτωση που αναφερόμαστε σε *i.i.d.* μεταβλητές Z_i, για $i = 1,\ldots,n$, ο υπολογισμός της συνάρτησης πιθανότητας της στατιστικής συνάρτησης $S_{n,k}$ βρίσκεται ευκολότερα αποφεύγοντας τη χρήση διανυσμάτων. Πιο συγκεκριμένα ισχύει το επόμενο θεώρημα.

Θεώρημα 3.10

Αν $Z_1, Z_2, \ldots, Z_n$ είναι μια ακολουθία ανεξάρτητων και ισόνομων τ.μ. *Bernoulli*, τότε η συνάρτηση πιθανότητας $g_n(x) = P(S_{n,k} = x)$ της τ.μ. $S_{n,k}$ ικανοποιεί την αναδρομική σχέση

$$\begin{aligned} g_n(x) = {} & g_{n-1}(x) + pg_{n-1}(x-1) - pg_{n-2}(x-1) \\ & - qp^k(g_{n-k-1}(x) - g_{n-k-1}(x-k)) \\ & - qp^{k+1}(g_{n-k-2}(x-k) - g_{n-k-2}(x-1)), \qquad n \geq k+2,\ x \geq 0 \end{aligned}$$

με αρχικές συνθήκες

$$g_n(x) = 0, \text{ για } x < 0 \text{ ή } x > n$$

$$g_n(x) = \begin{cases} 1, & x = 0 \\ 0, & x > 0 \end{cases} \qquad \text{για } 0 \leq n < k$$

$$g_n(x)=\begin{cases}1-p^k, & x=0\\ p^k, & x=k\\ 0, & 1\le x\le k-1\end{cases}\qquad \gamma\iota\alpha\quad n=k$$

$$g_n(x)=\begin{cases}1-p^k(1+q), & x=0\\ 2qp^k, & x=k\\ p^{k+1}, & x=k+1\\ 0, & 1\le x\le k-1\end{cases}\qquad \gamma\iota\alpha\quad n=k+1.$$

Απόδειξη

Αρκεί να αντικατασταθεί η γεννήτρια συνάρτηση $\varphi_n(z)$, στο προηγούμενο θεώρημα, με τη δυναμοσειρά

$$\varphi_n(z)=\sum_{x=0}^{\infty} g_n(x)z^x$$

και εν συνεχεία να αντιστοιχηθούν στην ισότητα που θα προκύψει οι μεταξύ του z^x συντελεστές του. □

Για τον υπολογισμό της ακριβούς κατανομής της $S_{n,k}$ μπορούμε να προσφύγουμε στην εφαρμογή του *Θεωρήματος 3.10*. Ωστόσο, για ακολουθίες μεγάλου μήκους, n, και μεγάλες τιμές της παραμέτρου k, οι υπολογισμοί είναι αρκετά πολύπλοκοι. Σε αυτές τις περιπτώσεις είναι προτιμότερο να καταφύγει κανείς στην ασυμπτωτική κατανομή της στατιστικής συνάρτησης $S_{n,k}$ η οποία έχει μελετηθεί με εφαρμογή της μεθόδου των Chen-Stein. Στα πλαίσια της παρούσας διπλωματικής εργασίας δεν θα ασχοληθούμε με αποτελέσματα που αφορούν την ασυμπτωτική κατανομή της $S_{n,k}$.

Το *Θεώρημα 3.9* είναι επίσης χρήσιμο για τον υπολογισμό των ροπών (*raw moments*) της στατιστικής συνάρτησης $S_{n,k}$ γύρω από το μηδέν. Αρκεί να παρατηρηθεί ότι, η ροπογεννήτρια συνάρτηση της $S_{n,k}$ είναι ίση με $E\left[\exp(zS_{n,k})\right]=\varphi_n(e^z)$ και να βρεθεί μια αναδρομική σχέση γι' αυτή με αντικατάσταση της μεταβλητής z από την e^z στην επαναληπτική σχέση του *Θεωρήματος 3.9*.

Για τις ροπές είναι γνωστή η ακόλουθη ισότητα

$$\mu_{n,r} = E\left[(S_{n,k})^r\right] = \begin{cases} \dfrac{d^r}{dz^r} E\left[\exp(zS_{n,k})\right]\Big|_{z=0}, & r \geq 1 \\ 1, & r = 0 \end{cases}$$

και κάνοντας χρήση του τύπου

$$\frac{d^r}{dz^r}\left(e^{kz} E\left[\exp(zS_{n,k})\right]\right)\Big|_{z=0} = \sum_{i=0}^{r} \binom{r}{i} k^{r-i} \mu_{n,i}$$

καταλήγουμε στο επόμενο θεώρημα.

Θεώρημα 3.11

Οι ροπές γύρω από το μηδέν $\mu_{n,r}$, για $r \geq 1$, της τ.μ. $S_{n,k}$ ικανοποιούν τον αναδρομικό τύπο

$$\begin{aligned} \mu_{n,r} = \mu_{n-1,r} &+ p\sum_{i=0}^{r} \binom{r}{i} \left(\mu_{n-1,i} - \mu_{n-2,i}\right) - qp^k \mu_{n-k-1,r} \\ &+ qp^k \sum_{i=0}^{r} \binom{r}{i} \left(k^{r-i}(\mu_{n-k-1,i} - p\mu_{n-k-2,i}) + p\mu_{n-k-2,i}\right), \qquad n \geq k+2 \end{aligned}$$

με αρχικές συνθήκες

$$\mu_{n,r} = \begin{cases} 0, & 0 \leq n < k \\ k^r p^k, & n = k \\ 2k^r qp^k + (k+1)^r p^{k+1}, & n = k+1. \end{cases}$$

Αντικαθιστώντας στην παραπάνω σχέση όπου r την τιμή 1 προκύπτει η ακόλουθη εξίσωση διαφορών δεύτερης τάξης

$$\mu_s - \mu_{s-1} = E(S_{s,k}) - E(S_{s-1,k}) = p(\mu_{s-1} - \mu_{s-2}) + qp^k(kq + p), \text{ για } s \geq k+2.$$

Για τις διάφορες τιμές της παραμέτρου s, $s = k+2, k+3, \ldots, n$, υπολογίζουμε τις ροπές και αθροίζοντας κατά μέλη, καταλήγουμε στην επόμενη σχέση

$$\mu_n - \mu_{k+1} = E(S_{n,k}) - E(S_{k+1,k}) = p(\mu_{n-1} - \mu_k) + (n-k-1)qp^k(kq+p), \text{ για } n \geq k+2.$$

Στη σχέση αυτή, αν αντικαταστήσουμε τις απλές ροπές $(k)-$τάξεως και $(k+1)-$τάξεως αντίστοιχα, μ_k και μ_{k+1}, σύμφωνα με τις σχέσεις του παραπάνω θεωρήματος, προκύπτει η ακόλουθη εξίσωση διαφορών πρώτης τάξης για τη μέση τιμή της $S_{n,k}$

$$\mu_n = E(S_{n,k}) = p\mu_{n-1} + kqp^k + (kq+p)p^k(p+(n-kq)), \text{ για } n \geq k+1.$$

Επίσης, λαμβάνοντας υπόψη και τις αρχικές συνθήκες για τη μέση τιμή της $S_{n,k}$

$$\mu_n = 0, \text{ για } 0 \leq n < k$$

$$\mu_k = kp^k$$

καταλήγουμε στο ακόλουθο θεώρημα.

Θεώρημα 3.12

Η μέση τιμή της τ.μ. $S_{n,k}$ δίνεται από τη σχέση

$$\mu_n = E(S_{n,k}) = p^k(k+(n-k)(kq+p)), \text{ για } n \geq k.$$

Στην ίδια σχέση θα μπορούσαμε να καταλήξουμε χρησιμοποιώντας τη γεννήτρια συνάρτηση των μέσων τιμών

$$M(w) = \sum_{n=0}^{\infty} E(S_{n,k})w^n = \frac{(wp)^k(k-wp(k-1))}{(1-w)^2},$$

που προκύπτει εύκολα από το *Θεώρημα 3.8*.

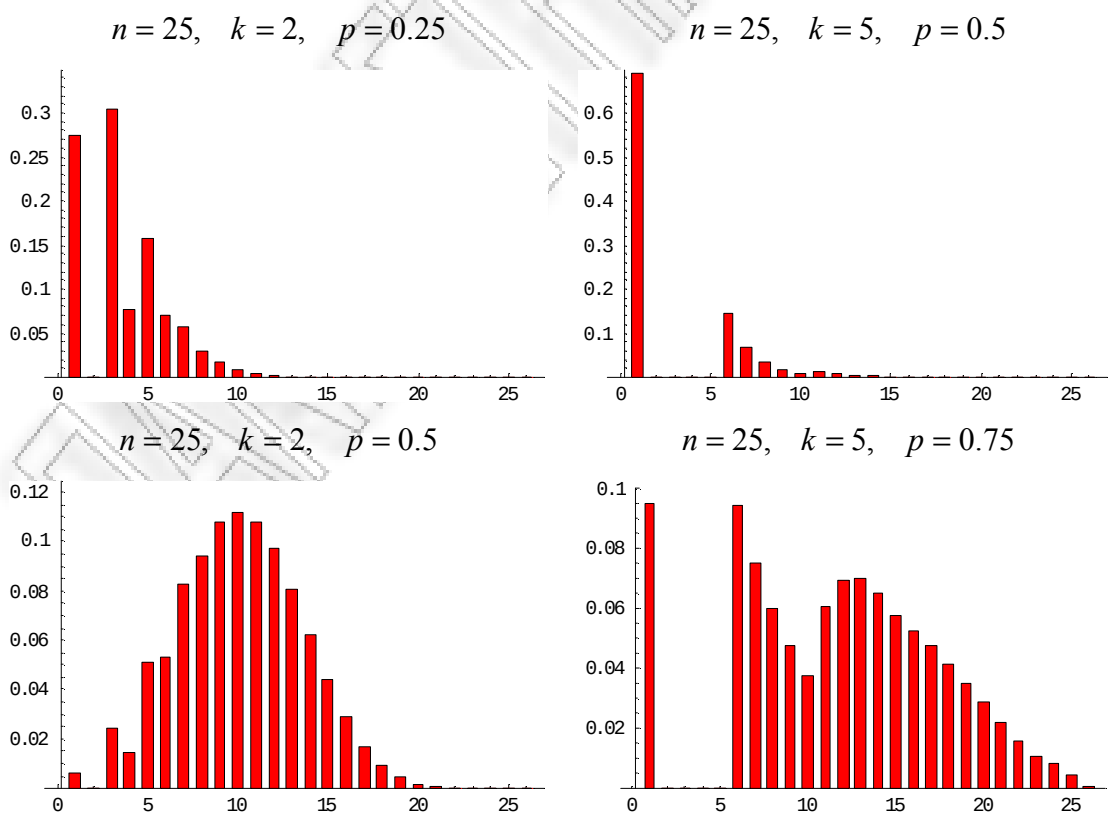
Τέλος, το *Θεώρημα 3.9* μπορεί να εφαρμοστεί επίσης και για τον υπολογισμό των παραγοντικών ροπών r τάξεως (*factorial moments*), αφού ως γνωστόν ισχύει η ισότητα

$$E(S_{n,k}(S_{n,k}-1)\ldots(S_{n,k}-r+1)) = \frac{d^r}{dz^r}\varphi_n(z)\Big|_{z=1}.$$

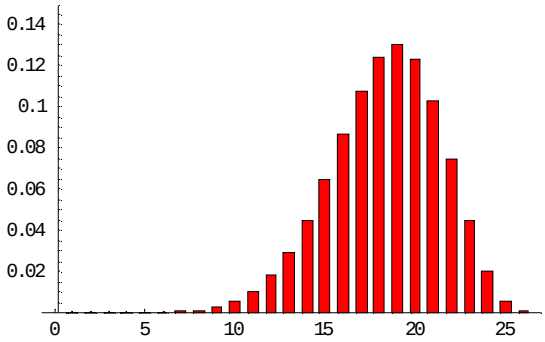
Παρακάτω δίνεται η γραφική παράσταση της ακριβούς κατανομής της τ.μ. $S_{n,k}$, η οποία υπολογίστηκε με χρήση της αναδρομικής σχέσης του *Θεωρήματος 3.10*, για διάφορες τιμές των παραμέτρων n, k και p. Για το μήκος των ακολουθιών έχουν επιλεγεί οι τιμές από $n = 20$ έως 100 ενώ οι τιμές της παραμέτρου k είναι έως 10, καθώς είναι οι πιο συνηθισμένες στην ανάλυση των επαναλαμβανόμενων σχηματισμών.

Σχήμα 3.1

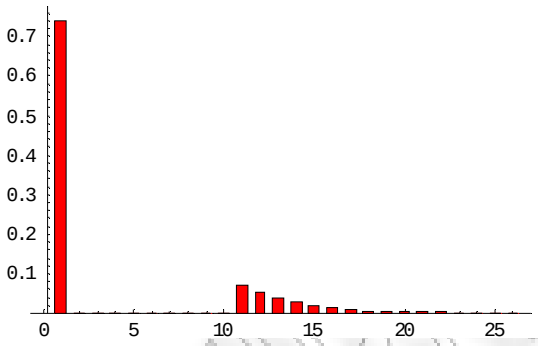
Συνάρτηση πιθανότητας της $S_{n,k}$ για διάφορες τιμές των παραμέτρων n, k και p



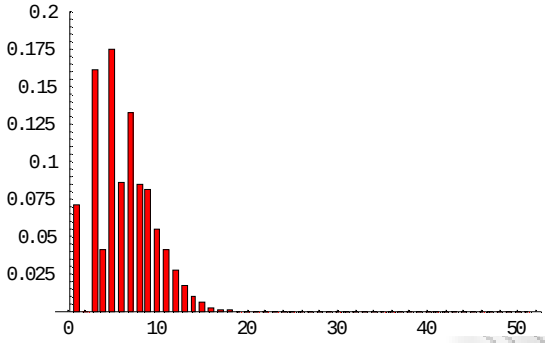
$n = 25, \quad k = 2, \quad p = 0.75$



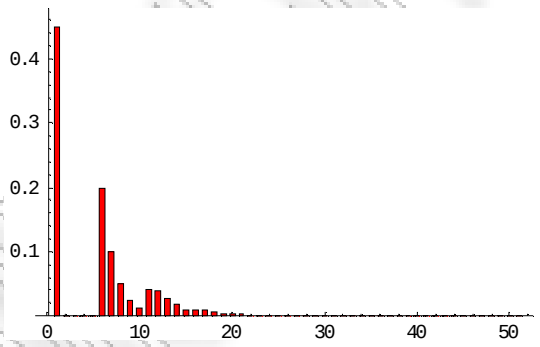
$n = 25, \quad k = 10, \quad p = 0.75$



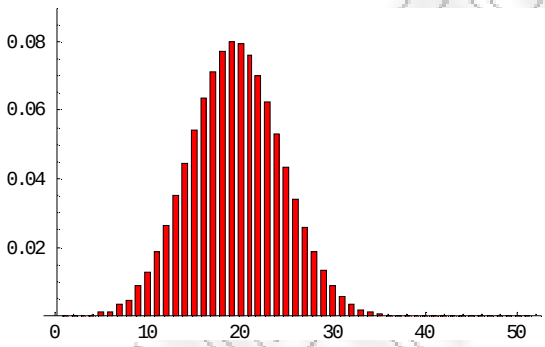
$n = 50, \quad k = 2, \quad p = 0.25$



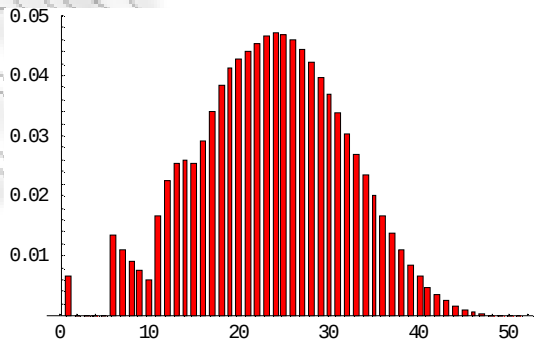
$n = 50, \quad k = 5, \quad p = 0.5$



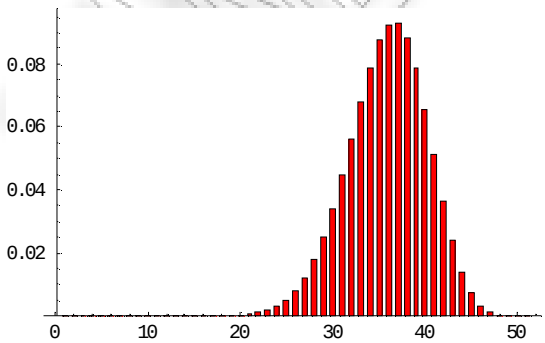
$n = 50, \quad k = 2, \quad p = 0.5$



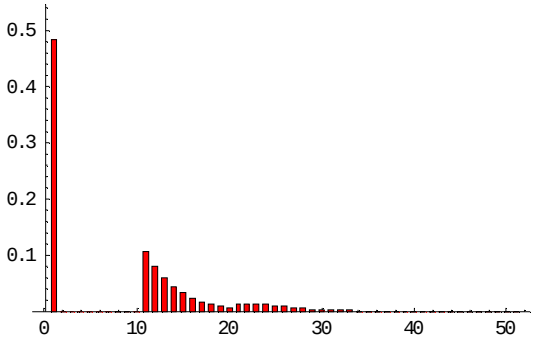
$n = 50, \quad k = 5, \quad p = 0.75$



$n = 50, \quad k = 2, \quad p = 0.75$



$n = 50, \quad k = 10, \quad p = 0.75$



Από τα παραπάνω σχεδιαγράμματα παρατηρείται ότι διατηρώντας σταθερές τιμές των n, k και μεταβάλλοντας την τιμή του p, όπως για τα ζεύγη τιμών (25,2) και (50,2), η κατανομή της $S_{n,k}$ προσεγγίζει την Κανονική Κατανομή καθώς η τιμή της πιθανότητας p πλησιάζει στη μονάδα. Αυτό είναι περισσότερο εμφανές στην περίπτωση που ισχύει $n = 50$ σε σύγκριση με την περίπτωση όπου $n = 25$, στην οποία η κατανομή της $S_{n,k}$ είναι λοξή προς τα αριστερά. Επίσης, η κατανομή της $S_{n,k}$ προσεγγίζει την Κανονική Κατανομή για μικρές τιμές της ποσότητας k/n. Στο Σχήμα 3.2 για λόγους σύγκρισης, δείχνεται πως διαφοροποιείται η κατανομή για μια συγκεκριμένη τιμή του $p = 0.8$, καθώς μεταβάλλονται τα n και k.

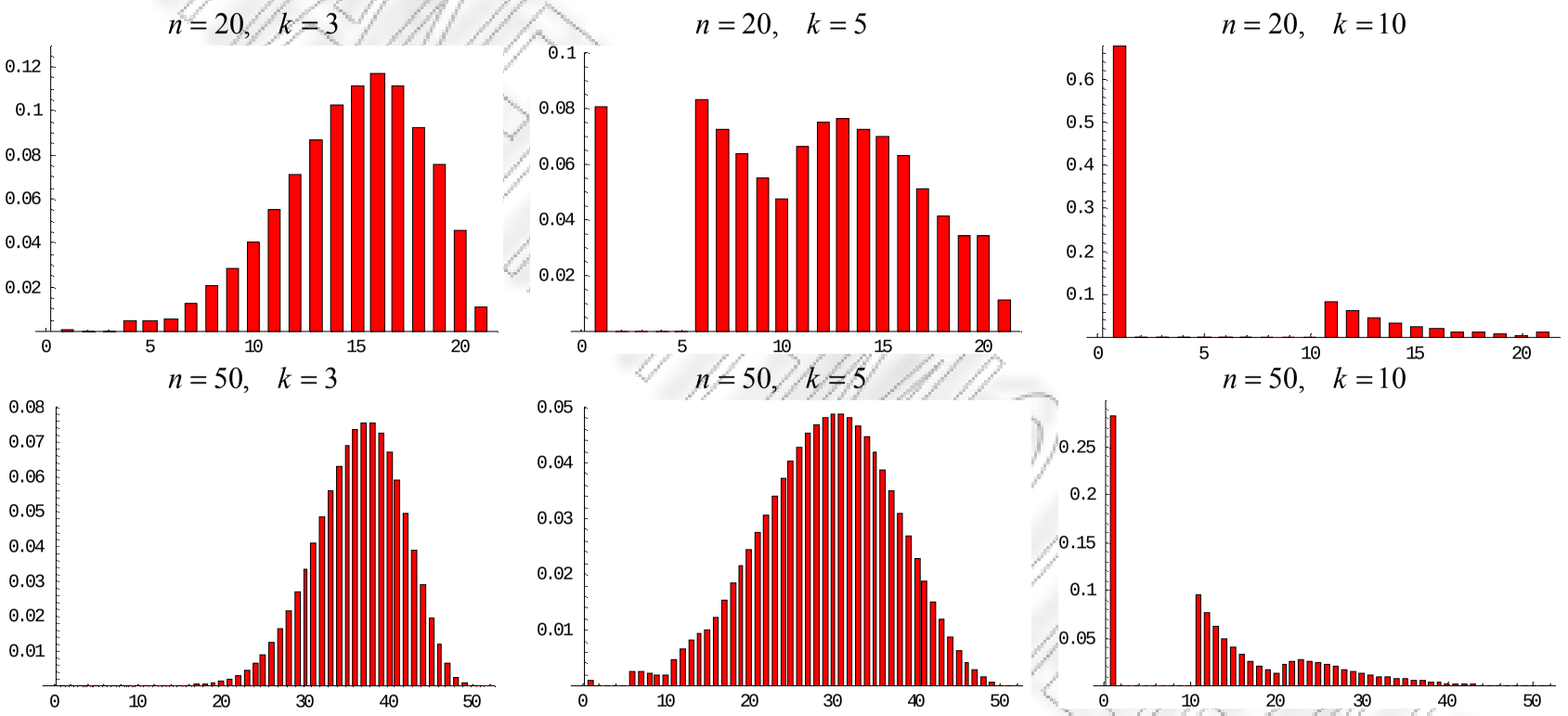
Η εύρεση της συνάρτησης πιθανότητας της $S_{n,k}$ υλοποιείται στο *Mathematica* ως εξής:

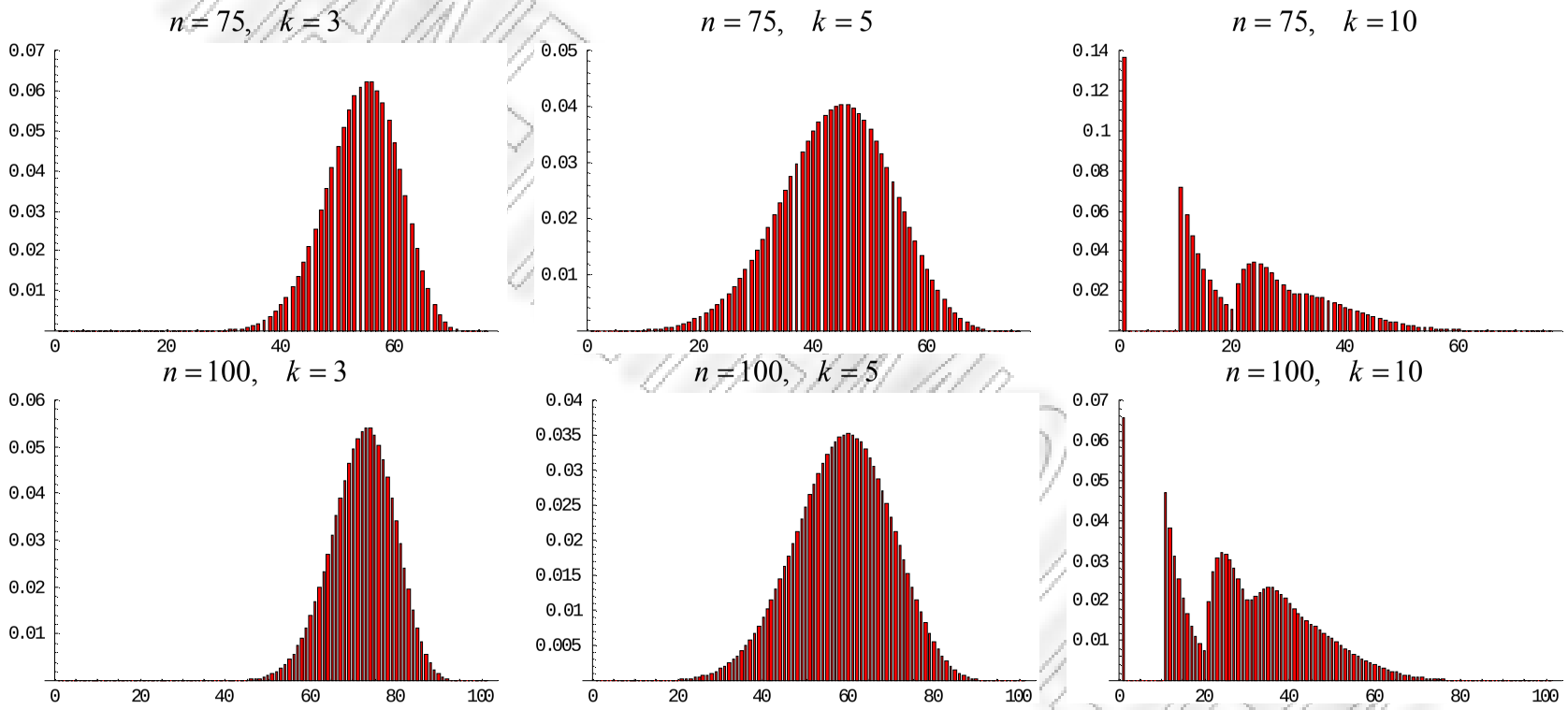
```
n = 20;
k = 3;
p = 0.8;
q = 1 - p;
list1 = {};
Array[g, {n, n}, {-n, -n}];

(* First Condition *)

Do[
  g[t, x] = 0;
  , {x, -n, -1}, {t, -n, n}];
Do[
  Do[
   g[t, x] = 0;
   {x, t + 1, n}], {t, -n, n - 1}];

(* Second Condition *)

Do[
  g[t, 0] = 1, {t, 0, k - 1}];
Do[
  Do[
   g[t, x] = 0, {x, 1, n}], {t, 0, k - 1}];
```

```
(* Third Condition *)

Do[
  pr1[t, 0] = 1 - (p ^k);
  Do[
   pr1[t, x] = 0;
   , {x, 1, k - 1}];
  pr1[t, k] = (p^k);
  Do[
   pr1[t, x] = 0;
   , {x, k + 1, n}];
  , {t, k, k}];

(* Fourth Condition *)

Do[
  pr1[t, 0] = 1 - p ^k (1 + q);
  Do[
   pr1[t, x] = 0;
   , {x, 1, k - 1}];

  pr1[t, k] = 2 q p ^k;
  pr1[t, k + 1] = p ^ (k + 1);

  Do[
   pr1[t, x] = 0;
   , {x, k + 2, n}];
  , {t, k + 1, k + 1}];

(* General Reccurence *)

Do[
  Do[
   g[t, x] = g[t - 1, x] + (p) g[t - 1, x - 1] - (p) g[t - 2, x - 1]
      - (q p ^k) (g[t - k - 1, x] - g[t - k - 1, x - k])
      - (q p ^ (k + 1)) (g[t - k - 2, x - k] - g[t - k - 2, x - 1]);
   , {x, 0, n}], {t, k + 2, n}];

Do[
  AppendTo[list1, N[g[n, x], 3]], {x, 0, n}];

<< Graphics`Graphics`

BarChart[list1, BarSpacing → -.2, BarGroupSpacing → .6,
 PlotRange → {0, 0.125}, Ticks → Automatic]
```

Σχήμα 3.2

Συνάρτηση πιθανότητας της $S_{n,k}$ για $p = 0.8$ και διάφορες τιμές των παραμέτρων n, k




$n = 75, \quad k = 3$
$n = 75, \quad k = 5$
$n = 75, \quad k = 10$
$n = 100, \quad k = 3$
$n = 100, \quad k = 5$
$n = 100, \quad k = 10$

Επίλογος

Στο 1ο Κεφάλαιο της διπλωματικής αυτής εργασίας παρουσιάστηκε η αλυσίδα DNA ως προς τη δομή της και έγινε μια εισαγωγή της σημαντικότητας των επαναλαμβανόμενων σχηματισμών (*tandem repeats*) στην παρουσία κάποιων συγκεκριμένων ασθενειών. Επίσης, περιγράφηκαν οι τρεις μεγάλες Βάσεις Δεδομένων για αλυσίδες DNA οι οποίες είναι σήμερα διαθέσιμες στην ερευνητική κοινότητα .

Στο 2ο Κεφάλαιο παρουσιάστηκαν οι Αριθμητικοί Αλγόριθμοι Αντιστοίχισης δύο ακολουθιών DNA. Αρχικά, περιγράφηκαν η Μέθοδος της Απόστασης και η Μέθοδος της Ομοιότητας, που αποτελούν δύο μεθόδους εύρεσης της βέλτιστης αντιστοίχισης μεταξύ δύο ακολουθιών. Οι δύο μέθοδοι εφαρμόστηκαν στη Συνολική Αντιστοίχιση 2 ακολουθιών, που προϋποθέτει τη συμμετοχή όλων των βάσεων που απαρτίζουν την κάθε ακολουθία καθώς και στην προσαρμογή μιας ακολουθίας μικρού μήκους σε μια ακολουθία μεγάλου μήκους. Η μέθοδος της Ομοιότητας χρησιμοποιήθηκε στην περίπτωση της Τοπικής αντιστοίχισης. Μια τροποποιημένη μορφή προγραμματισμού που καλείται *Wraparound Dynamic Programming* περιγράφηκε και εφαρμόστηκε για την αντιστοίχιση ακολουθίας DNA με επιλεγμένο επαναλαμβανόμενο σχηματισμό. Ακόμα, παρουσιάστηκαν κάποιες εναλλακτικές μορφές προγραμματισμού που απαιτούν λιγότερο υπολογιστικό χρόνο αλλά και χώρο στη μνήμη ενός Η/Υ. Επίσης, περιγράφηκε η μεθοδολογία παραγωγής αντιστοιχίσεων με τη χρήση των *Tracebacks*. Στην τελευταία παράγραφο του κεφαλαίου αναφερθήκαμε στην έννοια των αναστροφών και περιγράφηκαν δύο αλγόριθμοι για την βέλτιστη τοπική αντιστοίχιση επιτρέποντας την εμφάνιση των αναστροφών κατά τη διαδικασία της αντιστοίχισης. Όλοι οι παραπάνω αλγόριθμοι υλοποιήθηκαν με τη βοήθεια του *Mathematica*.

Στο 3ο Κεφάλαιο παρουσιάστηκε η στατιστική συνάρτηση $S_{n,k}$ που εκφράζει το συνολικό αριθμό επιτυχιών στις ροές επιτυχίας μεγέθους τουλάχιστον k και μελετήθηκε ως προς την ακριβή της κατανομή για την περίπτωση ανεξάρτητων και ισόνομων δίτιμων αποτελεσμάτων. Συγκεκριμένα, περιγράφηκε η τεχνική της εμφύτευσης σε πεπερασμένη αλυσίδα *Markov* για την προσαρμογή της τυχαίας μεταβλητής $S_{n,k}$ σε Μαρκοβιανή αλυσίδα και δόθηκε ο τύπος

υπολογισμού της συνάρτησης πιθανότητάς της. Επίσης, μελετήθηκε η κατανομή της τ.μ. με τη μέθοδο της εμφύτευσης μεταβλητών πολυωνυμικού τύπου σε Μαρκοβιανή αλυσίδα. Δόθηκαν οι τύποι υπολογισμού της μονής και της διπλής γεννήτριας συνάρτησης, της συνάρτησης πιθανότητας, της μέσης τιμής καθώς και της γεννήτριας συνάρτησης των μέσων τιμών. Η συνάρτηση πιθανότητας της τ.μ. $S_{n,k}$ αποδόθηκε γραφικά μέσω *Mathematica* για διάφορες τιμές των παραμέτρων n, k και p. Από τα γραφήματα που προέκυψαν, διαπιστώθηκε ότι η κατανομή της τ.μ. μπορεί να προσεγγιστεί από την Κανονική Κατανομή για ακολουθίες μεγάλου μήκους, πιθανότητα αντιστοίχισης ομοίων βάσεων κοντά στη μονάδα και αρκετά μικρή τιμή της ποσότητας k/n.

Βιβλιογραφία

Antzoulakos, D.L., Bersimis, S. and Koutras, M.V. (2003). On the distribution of the total number of run lengths, *Annals of the Institute of Statistical Mathematics*, **55**, 865-884.

Arratia, R., Gordon, L. and Waterman, M.S. (1986). An extreme value theory for sequence matching, *The Annals of Statistics*, **14**, 971-993.

Arratia, R., Gordon, L. and Waterman, M.S. (1990b). The Erdös- Rényi law in distribution, for coin tossing and sequence matching, *The Annals of Statistics*, **18**, 539-570.

Benson, G. (1997). Sequence alignment with tandem duplication, *Journal of Computational Biology*, **4**, 351-367.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences, *Nucleic Acids Research*, **27**, 573-580.

Benson, G. and Su, X. (1998). On the distribution of *k*-tuple matches for sequence homology: a constant time exact calculation of the variance, *Journal of Computational Biology*, **5**, 87-100.

Benson, G. and Waterman, M.S. (1994). A method for fast database search for all k-nucleotide repeats, *Nucleic Acids Research*, **22**, 4828-4836.

Campuzano, V., Montermini, L., Molto, M.D., Pianese, L. and Cossee, M. (1996). Friedreich's Ataxia: Autosomal Recessive Disease Caused by an Intronic GAA Triplet Repeat Expansion, *Science*, **271**, 1423-1427.

Clote, P. and Backofen, R. (2000). *Computational Molecular Biology – An Introduction*, John Wiley & Sons.

Fischetti, V., Landau, G., Schmidt, J. and Sellers, P. (1992). Identifying periodic occurences of a template with applications to a protein structure. *Proceedings of the 3rd Annual Symposium on Combinatorial Pattern Matching, Lecture Notes in Computer Science*, **644**, 111-120.

Fitch, W.M. and Smith, T.F. (1983). Optimal sequence alignments, *Proceedings of the National Academy of Sciences of USA*, **80**, 1382-1386.

Fu, J.C. (1986). Reliability of consecutive-k-out-of-n:F system with $k-1$ step Markov dependence, *IEEE Transactions on Reliability*, **35**, 602-606.

Fu, J.C. and Koutras, M.V. (1994). Distribution theory of runs: A Markov chain approach, *Journal of the American Statistical Association*, **89**, 1050-1058.

Fu, J.C., Lou, W.Y.W., Bai, Z.-D. and Li, G. (2002). The exact and limiting distributions for the number of successes in success runs within a sequence of Markov-dependent two-state trials, *Annals of the Institute of Statistical Mathematics*, **54**, 719-730.

Fu, Y.-H., Pizzuti, A., Fenwick Jr., R.G., King, J., Rajnarayan, S., Dunne, P.W., Dubel, J., Nasser, G.A., Ashizawa, T., DeJong, P., Wieringa, B., Korneluk, R., Perryman, M.B., Epstein, H.F. and Caskey, C.T. (1992). An unstable triplet repeat in a gene related to myotonic muscular dystrophy, *Science*, **255**, 1256-1258.

Galil, Z. and Giancarlo, R. (1989). Speeding up dynamic programming with applications to molecular biology, *Theoretical Computer Science*, **64**, 107-118.

Goldstein L. and Waterman, M.S. (1992). Poisson, Compound Poisson and Process approximations for testing statistical significance in sequence comparisons, *Bulletin of Mathematical Biology*, **54**, 785-812.

Gordon, L., Schilling, M.F. and Waterman, M.S. (1986). An extreme value theory for long head runs, *Probability Theory and Related Fields*, **72**, 279-287.

Gotoh, O. (1982). An improved algorithm for matching biological sequences, *Journal of Molecular Biology*, **162**, 705-708.

Gribskov, M., McLachlan, A.D. and Eisenberg, D. (1987). Profile Analysis: Detection of distantly related proteins, *Proceedings of the National Academy of Sciences of USA*, **84**, 4355-4358.

Griggs, J.R., Hanlon, P.J. and Waterman, M.S. (1986). Sequence alignments with matched sections, *SIAM Journal of Algebraic Discrete Methods,* **7**, 604-608.

Hirschberg, D.S. (1975). A linear space algorithm for computing maximal common subsequences, *Communications of the ACM*, **18**, 341-343.

Howe, C.J., Barker, R.F., Bowman, C.M. and Dyer, T.A. (1988). Common features of three inversions in wheat chloroplast DNA, *Current Genetics*, **13**, 343-349.

Huang, X. and Miller, W. (1991). A time-efficient, linear-space local similarity algorithm, *Advances in Applied Mathematics*, **12**, 337-357.

Huntington's Disease Collaborative Research Group (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes, *Cell*, **72**, 971-983.

Koutras, M.V. and Alexandrou, V.A. (1995). Runs, scans and urn model distributions: A unified Markov chain approach, *Annals of the Institute of Statistical Mathematics*, **47**, 743-766.

Koutras, M.V. and Alexandrou, V.A. (1995). Runs, scans and urn model distributions: A unified Markov chain approach, *Annals of the Institute of Statistical Mathematics*, **47**, 743-766.

Koutras, M.V. and Alexandrou, V.A. (1997). Nonparametric statistical randomness tests based on success runs of fixed length, *Statistics and Probability Letters*, **32**, 393-404.

Kruskal, J.B. (1983). An Overview of sequence comparison, In *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, Eds D. Sankoff and J.B. Kruskal, 1-40, Addison-Wesley, London.

Kruskal, J.B. and Sankoff, D. (1983). *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, 265-310, Addison-Wesley, London.

Landau, G.M. and Schmidt, J.P. (1993). An algorithm for approximate tandem repeats, *Lecture Notes in Computer Science*, *Combinatorial Pattern Matching*, **684**, 120-133.

Laquer, H.T. (1981). Asymptotic limits for a two-dimensional recursion, *Studies in Applied Mathematics*, **64**, 271-277.

La Spada, A.R., Wilson, E.M., Lubahn, D.B., Harding, A.E. and Fischbeck, K.H.(1991). Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy, *Nature*, **352**, 77-79.

Lange, K. (2002). *Mathematical and Statistical Methods for Genetic Analysis (2nd edition)*, Statistics for Biology and Health, Springer – Verlag.

Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii Nauk USSR*, **163**, 845–848 (Russian), (English Translation: (1966), *Cybernetics and Control Theory*, **10**, 707–710).

Lou, W.Y.W. (2003). The exact distribution of the *k*-tuple statistic for sequence homology, *Statistics and Probability Letters*, **61**, 51-59.

Martin, D.E.K. (2005). Distribution of the number of successes in success runs of length at least *k* in hight-order Markovian sequences, *Methodology and Computing in Applied Probability*, **7**, 543-554.

Myers, E.W. and Miller, W. (1988a). Sequence comparison with concave weighting functions, *Bulletin of Mathematical Biology*, **50**, 97-120.

Myers, E.W. and Miller, W. (1988b). Optimal alignments in linear space, *Computer Applications in the Biosciences (CABIOS)*, **4**, 11-17.

Myers, E.W. and Miller, W. (1989). Approximate matching of regular expression, *Bulletin of Mathematical Biology*, **51**, 5-37.

Needleman, S.B. and Wunsch, C.D. (1970).A general method applicable to the search for similarities in the amino acid sequences of two proteins, *Journal of Molecular Biology*, **48**, 443-453.

Schöniger, M. and Waterman, M.S. (1992). A local algorithm for DNA sequence alignment with inversions, *Bulletin of Mathematical Biology*, **54**, 521-536.

Sellers, P.H. (1974). On the theory and computation of evolutionary distances, *SIAM Journal on Applied Mathematics*, **26**, 787-793.

Sellers, P.H. (1979). Pattern recognition in genetic sequences, *Proceedings of the National Academy of Sciences of USA*, **76**, 3041.

Sellers, P.H. (1980). The theory and computation of evolutionary distances: Pattern recognition, *Journal of Algorithms*, **1**, 359-373.

Smith, T.F. and Waterman, M.S. (1981a). Identification of common molecular subsequences, *Journal of Molecular Biology*, **147**, 195-197.

Smith, T.F. and Waterman, M.S. (1981b). Comparison of biosequences, *Advances in Applied Mathematics*, **2**, 482-489.

Stanton, R.G. and Cowan, D.D. (1970). Short Notes: Note on a "Square" functional equation, *SIAM Review*, **12**, 277-279.

Ukkonen, E. (1983). On approximate string matching, *Proceedings of the International Conference on Foundations of Computer Theory, Lecture Notes in Computer Science*, **158**, 487-495.

Ukkonen, E. (1984). Algorithms for approximate string matching, *Information and Control*, **64**, 100-118.

Ulam, S.M. (1972). Some ideas and prospects in biomathematics, *Annual Review of Biophysics and Bioengineering*, **1**, 277-292.

Verkerk, A., Pieretti, M., Sutcliffe, J., Fu,Y., Kuhl, D., Pizzuti, A., Reiner, O., Richards, S., Victoria, M., Zhang, F., Eussen, B., van Ommen, G., Blonden, A., Riggins, G., Chastain, J.,Kunst, C., Galjaard, H., Caskey, C., Nelson, D., Oostra, B. and Warren, S. (1991). Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome, *Cell*, **65**, 905-914.

Vingron, M. and Argos, P. (1990). Determination of reliable regions in protein sequence alignments, *Protein Engineering*, **3**, 565-569.

Wagner, R.A. (1975). On the complexity of the extended string-to-string correction problem,

Proceedings of the 7th Annual ACM Symposium on Theory of Computing, 218-223.

Wagner, R.A. and Fischer, M.J. (1974). The string-to-string correction problem, *Journal of the Association for Computing Machinery*, **21**, 168-173.

Waterman, M.S. (1983). Sequence alignments in the neighborhood of the optimum with general application to dynamic programming, *Proceedings of the National Academy of Sciences of USA*, **80**, 3123-3124.

Waterman, M.S. (1984a). Efficient Sequence Alignment Algorithms, *Journal of Theoretical Biology*, **108**, 333-337.

Waterman, M.S. (1984b). General Methods of sequence comparison, *Bulletin of Mathematical Biology*, **46**, 473-500.

Waterman, M.S. (1995). *Introduction to Computational Biology: Maps, Sequences and Genomes*, Chapman & Hall.

Waterman, M.S. and Eggert, M. (1987). A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons, *Journal of Molecular Biology*, **197**, 723-728.

Waterman, M.S., Smith, T.F. and Beyer, W.A. (1976). Some biological sequence metrics, *Advances in Mathematics*, **20**, 367-387.

Watson, J.D. and Crick, F.H. (1953). Molecular Structure of Nucleic Acids, *Nature*, **171**, 737-738.

Zhou, D.X., Massenet, O., Quigley, F., Marion, M.J., Monéger, F., Huber, P. and Mache, R. (1988). Characterization of a large inversion in the spinach chloroplast genome relative to Marchantia: A possible transposon-mediated origin, *Current Genetics*, **13**, 433-439.