Εργασία: Αλγόριθμοι Εύρεσης Μοτίβων σε Ακολουθίες DNA
Μάθημα: Ιατρική Πληροφορική
Ομάδα:
Τεριζή Χρύσα ΑΜ:2553
Ημερομηνία: 13 – 01 – 2017
Έτος: 2016 – 2017

Περιεχόμενα

1. Εισαγωγή

Θα εισάγουμε κάποιες βασικές έννοιες σχετικά με το τι είναι αλγόριθμος, DNA, βιοπληροφορική τα όποια χρειάζονται για να κάνουμε πιο εύκολα κατανοητές κάποιες έννοιες που είναι απαραίτητες.

1.1 Αλγόριθμος

Αλγόριθμος είναι μία ακολουθία εντολών που πρέπει κάποιος να εκτελέσει για να λύσει ένα καλά ορισμένο πρόβλημα.

Για να κατανοήσουμε την λειτουργία ενός αλγορίθμου θα χρησιμοποιήσουμε τον ψευδοκώδικα. Ο ψευδοκώδικας είναι μία γλώσσα που χρησιμοποιείται συχνά από τους επιστήμονες του κλάδου των υπολογιστών για την περιγραφή αλγορίθμων. Για την επίλυση βιολογικών προβλημάτων η φύση χρησιμοποιεί διαδικασίες τύπου - αλγορίθμου. [1]

1.2 DNA

Το δε(σ)οξυριβο(ζο)νουκλεΐ(νι)κό οξύ (DNA) είναι νουκλεικό οξύ που περιέχει τις γενετικές πληροφορίες που καθορίζουν τη βιολογική ανάπτυξη όλων των κυτταρικών ιών. Το DNA συνήθως έχει τη μορφή διπλής έλικας.

Δομή του DNA
Η διαμόρφωση των μεγάλων μορίων του DNA στο χώρο έχει τη μορφή δύο επιμήκων αλύσεων, οι οποίες συστρέφονται ελικοειδώς μεταξύ τους. Οι αζωτούχες βάσεις στο DNA είναι τέσσερις:

- κυτοσίνη C
- γουανίνη G
- θυμίνη T
- αδενίνη A

Οι αζωτούχες βάσεις, ανάλογα με τη σειρά αλληλουχίας τους σε τριάδες, κωδικοποιούν το μήνυμα για τη σύνθεση των αμινοξέων του κυττάρου στα ριβοσώματα. Εκεί τα αμινοξέα συνδυάζονται, με τη σειρά κατά την οποία μεταφέρθηκαν στο ριβόσωμα και συντίθενται έτσι οι διαφορετικές πρωτεΐνες.[2]

1.3 Βιοπληροφορική

Η **Βιοπληροφορική** (*bioinformatics*) είναι επιστημονικός κλάδος ο οποίος προέκυψε από τη συνεργασία των επιστημών της μοριακής βιολογίας και της πληροφορικής. Θεωρώντας τα βιολογικά δεδομένα (DNA, RNA, πρωτεΐνες) ως ψηφιακή πληροφορία, εφαρμόζει αλγορίθμους για την επεξεργασία τους και την παραγωγή χρήσιμων συμπερασμάτων με αποδοτικό τρόπο. Συνήθως χρησιμοποιούνται μέθοδοι κλάδων της τεχνητής νοημοσύνης, όπως η εξόρυξη δεδομένων και ο εξελικτικός υπολογισμός (π.χ. γενετικοί αλγόριθμοι). [3]

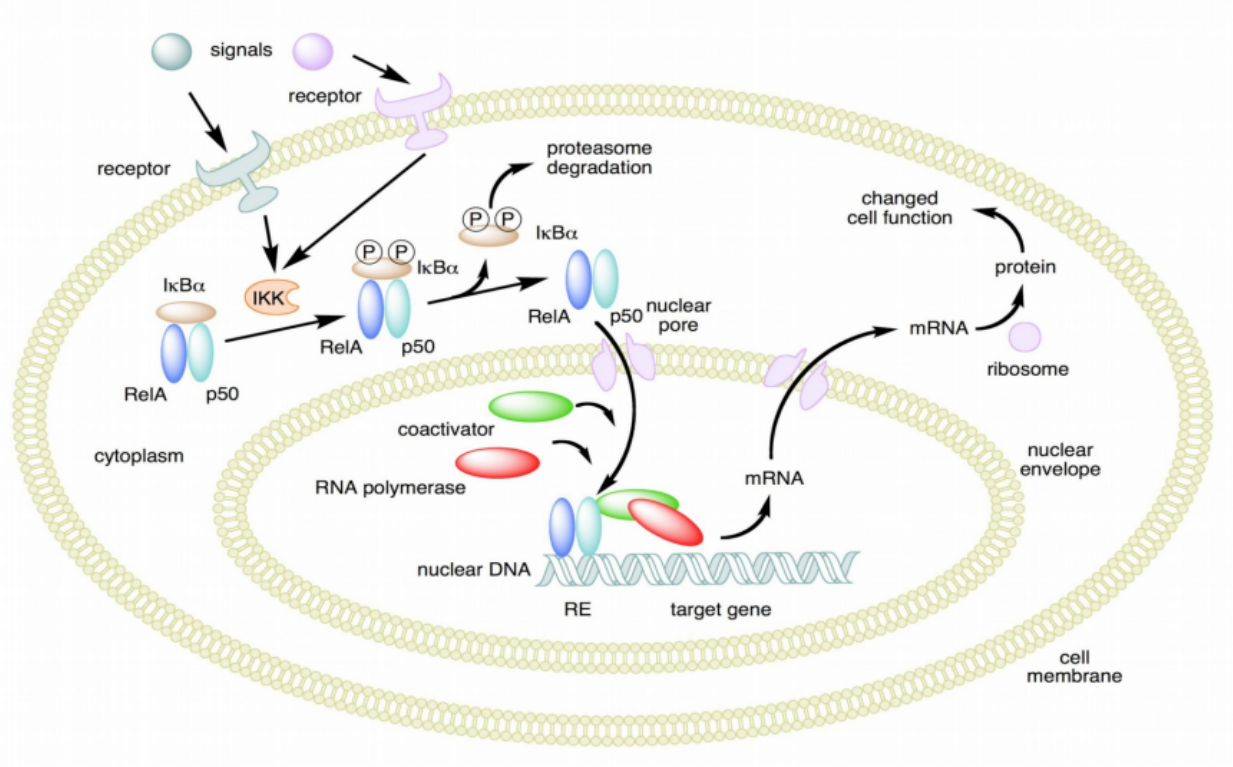
2. Μοτίβα σε βιολογικές αλληλουχίες

Υπάρχει μια ιεραρχία στην κωδικοποιήση των μηνυμάτων των βιολογικών αλληλουχιών που ξεπερνά τους απλούς περιορισμούς σύστασης και τις συνεμφανίσεις τους. Τα βιολογικά μηνύματα διαρθρώνονται σε συνδυασμούς “λέξεων” ανώτερης τάξης που οργανώνονται σε “φράσεις”, οι οποίες με τη σειρά τους συχνά επαναλαμβάνονται με σκοπό την απόδοση έμφασης και ενίοτε επανεμφανίζονται περιοδικά με την μορφή λογοτεχνικών μοτίβων. Σε αυτή την ενότητα θα δούμε πώς αναγνωρίζουμε, αξιολογούμε, αναπαριστούμε και μοντελοποιούμε τέτοια, ανώτερης τάξης, μηνύματα σε βιολογικές αλληλουχίες. Πιο συγκεκριμένα, το αντικείμενο αυτήν της ενότητας θα είναι τα μοριακά “σινιάλα” σε γονιδιωματικές αλληλουχίες, τα πρότυπα δηλαδή DNA ή RNA που “αναγνωρίζονται” από πρωτεΐνες ή άλλους παράγοντες που επιδρούν με διάφορους τρόπους πάνω στο γονιδίωμα επιφέροντας μεταβολές στη λειτουργία του. Συγκεκριμένα σε αυτή την ενότητα θα δούμε πως επαναλαμβανόμενα μοτίβα μέσα σε αλληλουχίες γονιδίων συνδέουν τα δομικά στοιχεία τους (τα εξόνια) και καθορίζουν τη δομή του τελικού μεταγράφου μέσω της συρραφής τους (ή αλλιώς του “ματίσματός” τους). Ή πώς η επανεμφάνιση μοτίβων πρόσδεσης πρωτεϊνών μεταγραφικών παραγόντων στους υποκινητές γονιδίων που συμμετέχουν σε μια κοινή λειτουργία, οργανώνει ένα ολόκληρο λειτουργικό πρόγραμμα σε κυτταρικό επίπεδο. Στο μαθηματικό και υπολογιστικό επίπεδο, θα δούμε πώς μπορούμε να εντοπίσουμε τέτοια μοτίβα σε αλληλουχίες με μια σχετική ακρίβεια, πώς θα

αξιολογήσουμε το λειτουργικό τους δυναμικό και πώς θα ανακαλύψουμε de novo άγνωστα μοτίβα. Πρίν από όλα αυτά όμως ας ξεκινήσουμε με ένα (όχι και τόσο απλό) βιολογικό πρόβλημα.

2.1 Ρυθμιστικά μοτίβα σε αλληλουχίες DNA

Το πείραμα με την "**μόλυνση των μυγών**" αποτέλεσε αφορμή για την εύρεση μοτίβων στο DNA. Το πρόβλημα είναι το εξής: οι βιολόγοι να μολύνουν τις μύγες με ένα βακτήριο, στη συνέχεια να τις κονιορτοιήσει και να μετρήσουν (ίσως με μία συστοιχία DNA) ποια γονίδια ενεργοποιούνται ως ανοσοαπάντηση. Από το συγκεκριμένο σύνολο γονιδίων, θα έβρισκαν τι πυροδοτεί την ενεργοποίηση τους. Έχει αποδειχθεί ότι πολλά γονίδια ανοσίας στο γονιδίωμα της μύγας έχουν συμβολοσειρές που θυμίζουν τη συμβολοσειρά TCGGGGATTTCC, η οποία βρίσκεται στα αριστερά της αρχής του γονιδίου. Οι μικρές αυτές συμβολοσειρές ονομάζονται θέσεις πρόσδεσης του NF-kB και αποτελούν σημαντικά παραδείγματα ρυθμιστικών μοτίβων που ενεργοποιούν γονίδια ανοσίας και άλλα γονίδια. Το πείραμα για τις μολύνσεις των μυγών θα θέλαμε να δώσει ένα σύνολο περιοχών DNA άνωθεν ευρισκομένων σε σχέση με τα γονίδια του γονιδιώματος, με κάθε περιοχή α περιέχει τουλάχιστον μία θέση πρόσδεσης NF-kB. Υποθέτουμε ότι το μοτίβο πρόσδεσης του NF-kB μας είναι άγνωστο, και δεν ξέρουμε ούτε που εντοπίζεται στις ανώθεν περιοχές του πειραματικού δείγματος. Για το πείραμα με την μόλυνση των μυγών χρειαζόμαστε έναν αλγόριθμο, ο οποίος θα παίρνει ως είσοδο ένα σύνολο αλληλουχιών από κάποιο γονιδίωμα και θα μπορεί να βρίσκει μικρές υποσυμβολοσειρές που εμφανίζονται με συχνότητα μεγαλύτερη της αναμενόμενης. Οι επιστήμονες της βιοπληροφορικής δανείστηκαν τη μέθοδο του κύριου Λεγκράντ, και μία δημοφιλής μέθοδος για την εύρεση μοτίβων βασίζεται στην υπόθεση ότι οι συχνές ή σπάνιες λέξεις αντιστοιχούν ενδεχομένως σε ρυθμιστικά μοτίβα στο DNA. Αυτή η "γλωσσολογική" προσέγγιση του DNA αποτελεί σημαντικό κομμάτι για τη μέθοδο εύρεσης σημάτων που οδηγείται από μοτίβα, η οποία βασίζεται στην απαρίθμηση όλων των πιθανών μοτίβων και την επιλογή εκείνου που εμφανίζεται πιο συχνά. Στην εικόνα που ακολουθεί φαίνεται όπου η αρχική αντίδραση ενεργοποίησης των υποδοχέων καταλήγει με την εισαγωγή μιας διμερούς πρωτεΐνης μέσα στον πυρήνα.



Η πρωτεΐνη αυτή έχει το περίεργο όνομα NF-κΒ όπως έχουμε μιλήσει για αυτή λίγο πιο πάνω. Οι μεταγραφικοί παράγοντες είναι πρωτεΐνες που προσδένονται στο γονιδιωματικό DNA και επιφέρουν αλλαγές πρώτα στη δομή του και έπειτα στη λειτουργία του. Η κύρια λειτουργία τους είναι η ενεργοποίηση της μεταγραφής γονιδίων, της εκκίνησης δηλαδή της διαδικασίας που θα παραγάγει ένα mRNA αντίγραφο ενός γονιδίου με σκοπό την παραγωγή μιας άλλης πρωτεΐνης 1 . Η πρόσδεση του NF-κΒ συμβαίνει σε πολλά, διαφορετικά σημεία του γονιδιώματος της μύγας με αποτέλεσμα να ενεργοποιείται η μεταγραφή όχι ενός αλλά μιας σειράς από γονίδια, που με τη σειρά τους θα μεταφραστούν σε πρωτεΐνες που θα συμμετέχουν στην οργάνωση της ανοσολογικης απόκρισης. Τα γονίδια-στόχοι, όπως τα λέμε, του NF-κΒ ξεπερνούν τα 500, κάνοντας τον NF-κΒ έναν από τους πιο σημαντικούς (και σίγουρα έναν από τους πιο καλά μελετημένους) μεταγραφικούς παράγοντες και περιλαμβάνουν εκτός από γονίδια του ανοσοποιητικού συστήματος, μεταβολικά ένζυμα, αυξητικούς παράγοντες ακόμα και άλλους μεταγραφικούς παράγοντες, που με τη σειρά τους θα ενεργοποιήσουν τη μεταγραφή άλλων γονιδίων διευρύνοντας έτσι τον “καταρράχτη” αντιδράσεων που συνιστούν την αντίδραση ενός οργανισμού στην επίθεση από έναν μολυσματικό παράγοντα. Μια λίστα από τα γονίδια τα οποία

ενεργοποιούνται από τον NF-κB έχει δημιουργηθεί από το εργαστήριο του Thomas Gilmore στο Πανεπιστήμιο της Βοστώνης (Gilmore 2006) και μπορεί να βρεθεί εδώ: http://www.bu.edu/nf-kb/gene-resources/target-genes/. Πώς όμως ακριβώς συμβαίνει η ενεργοποίηση της μεταγραφής γονιδίων από τον NF-κB (ή γενικότερα από μεταγραφικούς παράγοντες); Η διαδικασία περιλαμβάνει, όπως είδαμε παραπάνω, την πρόσδεση της πρωτεΐνης σε ένα συγκεκριμένο σημείο του γονιδιώματος. Το σημείο αυτό είναι συχνά πολύ κοντά στη θέση έναρξης της κωδικοποίησης ενός γονίδιού, σε μια περιοχή που ονομάζεται υποκινητής (promoter) και που ο ρόλος της είναι ακριβώς να παρέχει το χώρο και τις συνθήκες για την πρόσδεση πρωτεϊνών όπως οι μεταγραφικοί παράγοντες. Η ίδια η διαδικασία της πρόσδεσης δημιουργεί αρχικά μια τοπική παραμόρφωση στη δομή του DNA στον υποκινητή και συχνά ακολουθείται από την προσέλκυση άλλων πρωτεϊνών στη συγκεκριμένη θέση ή σε κοντινές περιοχές (Fiedler and Marc Timmers 2000). Η συσσώρευση πρωτεϊνών καταλήγει στην οργάνωση συμπλόκων που περιλαμβάνουν το ένζυμο της RNA πολυμεράσης ΙΙ που είναι υπεύθυνο για τη μεταγραφή των τμημάτων του DNA που κωδικοποιούν πρωτεϊνές σε mRNA. Λέμε τότε ότι η μεταγραφή έχει ενεργοποιηθεί ή ότι το γονίδιο “επάγεται” από την αρχική πρόσδεση του μεταγραφικού παράγοντα. Διαφορετικοί παράγοντες προσδένονται σε διαφορετικά σημεία του γονιδιώματος και ως αποτέλεσμα επάγουν διαφορετικά γονίδια. Κάποιοι από τους μεταγραφικούς παράγοντες δεν μπορούν να δράσουν παρά μόνο σε συνεργασία με άλλους (λέμε τότε πως είναι συν-ενεργοποιητές) και κάποιοι δε δρουν καν ενεργοποιητικά παρά κατασταλτικά, παρεμποδίζοντας ή και διακόπτοντας τη μεταγραφή γονιδίων-στόχων (λέμε τότε πως πρόκειται για καταστολείς). Η μελέτη της λειτουργίας των μεταγραφικών παραγόντων, ονομάζεται “γονιδιακή ρύθμιση” και αποτελεί έναν από τους πιο σημαντικούς κλάδους της μοριακής βιολογίας, καθώς βρίσκεται στη βάση ενός μεγάλου αριθμού κυτταρικών διεργασιών, από την επαγωγή ενός απλού γονιδίου σε ένα βακτήριο και την απόκριση στην έλλειψη κάποιου θρεπτικού συστατικού ως την ανοσολογική απόκριση ενός θηλαστικού σε μια μόλυνση. Τις περισσότερες φορές, όλα ξεκινούν από την πρόσδεση ενός μεταγραφικού παράγοντα στο DNA. [4]

2.2 Οι πρωτεΐνες δε διαβάζουν λατινικά

Οι μεταγραφικοί παράγοντες προσδένονται στο DNA σε συγκεκριμένες θέσεις που ονομάζονται εύλογα “σημεία πρόσδεσης μεταγραφικών παραγόντων” (transcription factor binding sites ή για συντομία TFBS). Οι θέσεις αυτές αποτελούν τμήματα DNA μήκους λίγων ζευγών βάσεων, συνήθως 4-12, τα οποία έχουν χαρακτηριστικές αλληλουχίες, ειδικές για κάθε παράγοντα, που

εμφανίζονται στα σημεία πρόσδεσης υπακούοντας σε συγκεκριμένους κανόνες συντήρησης, έχουν δηλαδή χαρακτηριστικά μοτίβου. Για τον NF-κΒ μια χαρακτηριστική αλληλουχία πρόσδεσης είναι η παρακάτω:

G G G A A T T C C C

Πρόκειται για ένα 10-νουκλεοτίδιο με χαρακτηριστικά συμμετρική αλληλουχία που είναι αλληλο-συμπληρωματική γύρω από έναν άξονα συμμετρίας. Αυτό συμβαίνει γιατί ο NF-κΒ προσδένεται στο DNA με τη μορφή διμερούς πρωτεΐνης η οποία “αναγνωρίζει” την ίδια υπο- αλληλουχία GGGAA/TTCCC (Ghosh et al. 1995). Τι εννοούμε όμως όταν λέμε ότι η πρωτεΐνη “αναγνωρίζει” μια αλληλουχία; Προφανώς και ένα πρωτεϊνικό μόριο δεν μπορεί να “διαβάσει” λατινικά. Αυτό που “αναγνωρίζεται” από τον κάθε μεταγραφικό παράγοντα είναι η στερεοχημική διαμόρφωση του DNA στο χώρο των τριών διαστάσεων και κυρίως την ύπαρξη συγκεκριμένων καταλοίπων στο DNA του σημείου πρόσδεσης, τα οποία εμφανίζουν μεγάλη χημική συγγένεια με συγκεκριμένα αμινοξέα του μεταγραφικού παράγοντα. Η συγγένεια αυτή, που μπορεί να είναι ηλεκτροστατική ή υδροφοβική αλληλεπίδραση ή να βασίζεται σε δεσμούς υδρογόνου ή δυνάμεις Van der Waals, οδηγεί τελικά στην πρόσδεση της πρωτεΐνης πάνω στο DNA στο σημείο που η συνολική ισχύς των αλληλεπιδράσεων είναι αρκετή για να ξεπεράσει την τάση όλων των μορίων για ελεύθερη διάχυση. Στην παρακάτω εικόνα φαίνεται σχηματικά η δομή του διμερούς πρωτεϊνικού παράγοντα (με τα στοιχεία της δομής του χρωματισμένα ως γαλάζιες β-πτυχωτές επιφάνειες και πορτοκαλί α-έλικες) καθώς βρίσκεται προσδεδεμένος σε μια περιοχή δίκλωνου DNA (με βιολετί) (Huang et al. 2001).

Η

“αναγνώριση” λοιπόν των σημείων πρόσδεσης βασίζεται σε χημική και δομική

συγγένεια μεταξύ της πρωτεΐνης και του DNA. Πόση όμως πρέπει να είναι η συγγένεια για να έχουμε επιτυχή πρόσδεση; Σε ποιο βαθμό θα πρέπει να υπάρχει “ταίριασμα” μεταξύ των αμινοξικών καταλοίπων του μεταγραφικού παράγοντα και του DNA; Σε ποιο βαθμό η DNA αλληλουχία του σημείου πρόσδεσης είναι περιορισμένη σε ό,τι αφορά τη διαδοχή των νουκλεοτιδίων; Γυρνώντας στο παράδειγμά μας, θα μπορούσε ο NF-κΒ να προσδεθεί στην παρακάτω αλληλουχία;

G G G A A T T T C C

Η συγκεκριμένη αλληλουχία διαφέρει από αυτήν που είδαμε παραπάνω στο 8ο κατάλοιπο (που τώρα είναι Τ αντί για C). Η απάντηση είναι πως ο NF-κΒ “αναγνωρίζει”, και συνεπώς προσδένεται, και σ' αυτήν την αλληλουχία με την ίδια ευκολία (για την ακρίβεια, ο NF-κΒ φαίνεται να προτιμά ελαφρώς αυτήν τη λιγότερο συμμετρική εκδοχή του σημείου πρόσδεσης). Οι δύο αυτές αλληλουχίες αποτελούν δύο χαρακτηριστικά παραδείγματα σημείων πρόσδεσης του NF-κΒ αλλά στην πραγματικότητα, ο μεταγραφικός παράγοντας που εξετάζουμε μπορεί να προσδεθεί σε οποιαδήποτε από τις παρακάτω αλληλουχίες:

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

οι οποίες διαφέρουν σε ένα ή μερικά νουκλεοτίδια μόνο χωρίς όμως να είναι πανομοιότυπες. Ο NF-κΒ μπορεί να προσδεθεί σε οποιαδήποτε από αυτές, κάτι που σημαίνει πως το γονιδίωμα μπορεί να χρησιμοποιήσει οποιαδήποτε από τις παραπάνω παραλλαγές για να κωδικοποιήσει την πρόσδεση του μεταγραφικού παράγοντα και τη συνεπακόλουθη ενεργοποίηση των γονιδίων- στόχων του. Οι παραπάνω αλληλουχίες συνθέτουν το “μοτίβο” πρόσδεσης (binding motif) του NF- κΒ και έχουν τα χαρακτηριστικά των μοτίβων που συζητήσαμε στην Εισαγωγή. Εμφανίζονται σε συγκεκριμένες θέσεις του γονιδιωματικού μηνύματος για να συνθέσουν ένα γενικότερο θέμα, το οποίο στο επίπεδο της μοριακής

βιολογίας είναι το “γονιδιακό πρόγραμμα έκφρασης” των γονιδίων- στόχων του NF-κB. Ερχόμαστε σε αυτό το σημείο, στο πρόβλημα που θα μας απασχολήσει και αυτό είναι:

- -

Πώς μπορούμε να εντοπίσουμε μοτίβα σε γονιδιωματικές αλληλουχίες;

- -

[5]

2.3 Τέσσερα προβλήματα

2.3.1 Ορισμός μοτίβου και αναζήτηση σημείων πρόσδεσης

Ανάλογα με το επίπεδο της πληροφορίας που διαθέτουμε για ένα συγκεκριμένο φαινόμενο το βασικό αυτό πρόβλημα μπορεί να διαιρεθεί σε επιμέρους προβλήματα. Αρχικά μπορούμε να υποθέσουμε ότι γνωρίζουμε κάποια από τα σημεία πρόσδεσης ενός μεταγραφικού παράγοντα. Κάτι τέτοιο μπορεί να έχει προκύψει από μια πειραματική διαδικασία που θα στοχεύει στο να εντοπίσει το αποτύπωμα μιας πρωτεΐνης στο DNA. Υπάρχουν πολλές διαφορετικές πειραματικές προσεγγίσεις για κάτι τέτοιο, από την πιο παραδοσιακή ανάλυση “αποτύπωσης DNA” (DNA footprinting), σε πιο σύγχρονες μεθοδολογίες που συνδυαζουν την ανοσοκατακρήμνιση χρωματίνης με αλληλούχιση DNA (π.χ. ChiPSeq) (Park 2009). Με δεδομένο έναν αριθμό αλληλουχιών σημείων πρόσδεσης τα βασικά ερωτήματα είναι:

- -

Δεδομένου ενός συνόλου σημείων πρόσδεσης ενός μεταγραφικού παράγοντα, πώς θα ορίσουμε το μοτίβο πρόσδεσης που μπορεί να τα παραγάγει;

Πώς θα χρησιμοποιήσουμε το μοτίβο αυτό για να εντοπίσουμε σημεία πρόσδεσης σε μια άγνωστη αλληλουχία;

- -

Το πρώτο από τα δύο ερωτήματα σχετίζεται με τη δυνατότητα αναγνώρισης από ένα μεταγραφικό παράγοντα περισσότερων του ενός συνδυασμούς νουκλεοτιδίων. Πράγματι υπάρχουν παράγοντες που είναι αυστηρά “επιλεκτικοί” και μπορούν να προσδεθούν μόνο σε πολύ καλά ορισμένα μοτίβα, που ουσιαστικά περιγράφονται με μια μοναδική αλληλουχία. Για παράδειγμα η πρωτεΐνη TBP (TATA-binding

protein) αναγνωρίζει πρακτικά μόνο το πεντανουκλεοτίδιο TATAA (από το οποίο παίρνει το όνομά της). Ο NF-κB εμφανίζει λιγότερους περιορισμούς, όπως είδαμε, κι έτσι μπορεί να προσδεθεί σε παραλλαγές ενός 10-νουκλεοτιδίου στο οποίο ένα ποσοστό μόνο από τις βάσεις χρειάζεται να παραμένουν πάντα οι ίδιες. Άλλες πρωτεΐνες εμφανίζουν ακόμα μεγαλύτερη “ανεκτικότητα”, με αποτέλεσμα οι αλληλουχίες όπου μπορούν να προσδεθούν να είναι ακόμα περισσότερες.

- -

***Ερώτηση**: Πώς εξηγούνται οι διαφορές μεταξύ των μεταγραφικών παραγόντων που έχουν ένα μοναδικό σημείο πρόσδεσης (όπως η TBP) και άλλων που έχουν περισσότερες παραλλαγές ενός σημείου πρόσδεσης (όπως ο NF-κB);*

- -

Το πρόβλημα που ανακύπτει είναι πώς θα δημιουργήσουμε ένα σύνολο από κανόνες που να τις συμπεριλαμβάνει όλες τις εναλλακτικές αλληλουχίες που μπορούν να λειτουργήσουν ως σημεία πρόσδεσης σε ένα μοτίβο. Από τη στιγμή που έχουμε ορίσει ένα μοτίβο, το δεύτερο από τα παραπάνω ερωτήματα προκύπτει σχεδόν αυτόματα. Πώς μπορούμε, γνωρίζοντας το σύνολο των κανόνων που ορίζουν τη σχέση συγγένειας μιας πρωτεΐνης με το DNA να εντοπίσουμε σημεία πρόσδεσης αυτής της πρωτεΐνης σε γονιδιωματικές αλληλουχίες; Το πρόβλημα αυτό έχει ιδιαίτερη σημασία, μιας και ο εντοπισμός θέσεων πρόσδεσης μεταγραφικών παραγόντων είναι συνήθως το πρώτο βήμα για τη διατύπωση πολύ χρήσιμων ερευνητικών υποθέσεων σχετικά με τη γονιδιακή ρύθμιση, τη σχέση μεταγραφικών παραγόντων και γονιδίων στόχων κλπ. Από τη στιγμή που ένα μοτίβο αποτελεί ουσιαστικά μια “γενίκευση” για μια συγκεκριμένη λειτουργία, η χρήση του για την αναζήτηση νέων, μέχρι πρότινος άγνωστων σημείων πρόσδεσης αποτελεί το επόμενο λογικό βήμα. [6]

2.3.2 Πληροφοριακό περιεχόμενο μοτίβου

Η αναζήτηση σημείων πρόσδεσης ενέχει έναν αριθμό από κινδύνους, ο κυριότερος από τους οποίους είναι ότι συχνά ένα μοτίβο είναι τόσο “ανεκτικό” από πλευράς περιορισμών στη σύσταση και τη διαδοχή των βάσεών του, που οι αλληλουχίες που το ικανοποιούν είναι εξαιρετικά μεγάλες σε αριθμό. Ένα “χαλαρό” μοτίβο που επιτρέπει πολλές αντικαταστάσεις σε πολλές από τις θέσεις, μπορεί πρακτικά να παραγάγει έναν τρομακτικό αριθμό αλληλουχιών και από αυτήν την άποψη δεν μας είναι ιδιαίτερα χρήσιμο. Αναζήτηση σημείων πρόσδεσης με ένα τέτοιο πρότυπο εντοπίζει πιθανούς στόχους τόσο συχνά και με τέτοια πυκνότητα στο γονιδίωμα που είναι πρακτικά αδύνατο να αξιολογηθούν. Λέμε

τότε ότι το μοτίβο αυτό έχει μικρό πληροφοριακό περιεχόμενο, καθώς δεν μας βοηθάει παρά ελάχιστα να διακρίνουμε μεταξύ σημείων πρόσδεσης και της γενικότερης αλληλουχίας. Από την άλλη πλευρά ένα μοτίβο με πολύ καλά ορισμένη σύσταση και σχέση διαδοχής βάσεων μπορεί να παραγάγει έναν περιορισμένο αριθμό αλληλουχιών οι οποίες αναμένεται να εντοπίζονται μέσα σε μια αλληλουχία με μια σχετική σπανιότητα. Η σπανιότητα αυτή αντανακλά τη δυνατότητα εξειδίκευσης του μοτίβου, το οποίο λέμε τώρα πως έχει υψηλό πληροφοριακό περιεχόμενο ή χαμηλή αβεβαιότητα. Είναι απαραίτητο να γνωρίζουμε το βαθμό της αβεβαιότητας που εμπεριέχει κάθε μοτίβο που δημιουργούμε από τα αρχικά μας δεδομένα. Το ερώτημα που καλούμαστε έτσι να απαντήσουμε είναι:

- -

Δεδομένου ενός μοτίβου που προκύπτει από ένα σύνολο σημείων πρόσδεσης, πόσο καλά ορισμένο είναι το μοτίβο αυτό;

- -

Στη συνέχεια θα δούμε με ποιον τρόπο μπορούμε να ποσοτικοποιήσουμε το βαθμό της αβεβαιότητας και τη σχέση της με το πληροφοριακό περιεχόμενο ενός μοτίβου.

2.3.3 Εύρεση άγνωστων μοτίβων

Μέχρι στιγμής έχουμε συζητήσει προβλήματα που προκύπτουν όταν γνωρίζουμε τις αλληλουχίες που αντιστοιχούν στα σημεία πρόσδεσης μιας πρωτεΐνης. Στις περισσότερες όμως περιπτώσεις δεν είμαστε τόσο τυχεροί. Ακόμα και αν διαθέτουμε πειραματικά δεδομένα footprinting ή ChIPSeq που είναι πολύ καλής ποιότητας, είναι εξαιρετικά δύσκολο να προσδιορίσουμε αλληλουχίες, όπως αυτές που είδαμε παραπάνω για τον NF-κΒ, κι αυτό γιατί ο βαθμός της διακριτικής ικανότητας ακόμα και των πιο εκλεπτυσμένων πειραματικών τεχνικών δεν είναι αρκετά μεγάλος. Πολύ συχνά διαθέτουμε ένα μεγάλο αριθμό αλληλουχιών, μήκους 500 βάσεων ή και περισσότερο μέσα στις οποίες καλούμαστε να εντοπίσουμε σημεία πρόσδεσης με μήκος μερικά νουκλεοτίδια. Η αναζήτηση αυτή συνιστά ένα πολύ δυσκολότερο πρόβλημα από αυτά που συζητήσαμε παραπάνω, το οποίο μπορεί να διατυπωθεί ως εξής:

- -

Δεδομένου ενός συνόλου αλληλουχιών που περιέχουν σημεία πρόσδεσης του ίδιου μεταγραφικού παράγοντα, πώς μπορούμε να προσδιορίσουμε το μοτίβο πρόσδεσής του;

- -

Το ερώτημα διαφέρει ουσιαστικά από αυτό που διατυπώθηκε νωρίτερα σε επίπεδο κλίμακας. Γνωρίζοντας τα σημεία πρόσδεσης έχουμε λύσει το μεγαλύτερο μέρος του προβλήματος και ο μετέπειτα προσδιορισμός του μοτίβου εξαρτάται από τη μεθοδολογία που θα επιστρατεύσουμε. Στην περίπτωση που συζητάμε όμως δεν γνωρίζουμε τίποτα για τα σημεία πρόσδεσης πέρα από το γεγονός ότι περιέχονται μέσα στις αλληλουχίες που έχουμε στα χέρια μας. Ανάλογα με το μέγεθος και το πλήθος αυτών των αλληλουχιών μπορούμε να πούμε πως ψάχνουμε για “βελόνες στα άχυρα”. [7]

2.4 Το πρώτο πρόβλημα

Ας ξεκινήσουμε να εξετάζουμε το πρώτο από τα προβλήματα που θέσαμε νωρίτερα:

- -

Δεδομένου ενός συνόλου σημείων πρόσδεσης ενός μεταγραφικού παράγοντα, πώς θα ορίσουμε το μοτίβο πρόσδεσης που μπορεί να τα παραγάγει;

- -

Ας εξετάσουμε το παράδειγμα των σημείων πρόσδεσης του NF-κB που είδαμε παραπάνω. Οι αλληλουχίες αυτές ήταν:

1 2 3 4 5 6 7 8 9 10
G G G A A T T C C C
G G G A A T T T C C
G G G G A T T C C C
G G G G A T T T C C
G G G A C T T C C C
G G G A C T T T C C
G G G G C T T C C C
G G G G C T T T C C

Ξέρουμε ότι το μοτίβο πρόσδεσης του NF-κΒ θα πρέπει να ικανοποιεί καθεμία από αυτές. Παρατηρούμε ότι για κάποιες θέσεις μέσα σε κάθε αλληλουχία, όπως οι 3 πρώτες, οι 2 τελευταίες, η 6η και η 7η, δεν αλλάζουν καθόλου. Για κάποιες άλλες αντίθετα, την 4η, την 5η και την 8η υπάρχουν παραλλαγές όπου περισσότερα από ένα νουκλεοτίδια μπορούν να συμπληρώσουν το σημείο πρόσδεσης. Αυτό που αναζητούμε, είναι το σύνολο των λογικών κανόνων που μπορούν να περιγράψουν όλες τις παραπάνω αλληλουχίες ταυτόχρονα. Ένας τρόπος για να το κάνουμε αυτό είναι να περιγράψουμε όλες τις αλληλουχίες εισάγοντας έναν ειδικό συμβολισμό για τις περιπτώσεις αμφισημίας. Μπορούμε π.χ. να ορίσουμε ότι:

- Στις θέσεις που δεν υπάρχει αβεβαιότητα και καλύπτονται από ένα μοναδικό κατάλοιπο θα αποδίδουμε αυτό το κατάλοιπο στο μοτίβο.
- Στις θέσεις όπου απαντώνται περισσότερα από ένα κατάλοιπα θα συμπεριλαμβάνονται όλες οι πιθανές παραλλαγές μέσα σε ένα ζεύγος αγκυλών [].

Ο πιο πάνω πίνακας των 8 αλληλουχιών μετατρέπεται έτσι στην παρακάτω έκφραση:

GGG[AG][AC]TT[TC]CC

Η απόδοση του μοτίβου με αυτόν τον τρόπο χρησιμοποιεί απλούς κανόνες με τη μορφή κανονικών εκφράσεων. Οι κανονικές εκφράσεις, είναι τυπικά πρότυπα που στηρίζονται σε μια σειρά από δεδομένους κανόνες και σκοπό έχουν να ομαδοποιήσουν σειρές χαρακτήρων με βάση τα κοινά δομικά τους στοιχεία. Εξαιτίας της κατασκευής τους οι κανονικές εκφράσεις μπορούν να παραγάγουν μια ομάδα από σειρές χαρακτήρων και για το λόγο αυτό είναι εξαιρετικά χρήσιμες σε λεξικογραφικές αναζητήσεις σειρών χαρακτήρων με αμφισημίες. Στο παράδειγμά μας, αν μεταφράσουμε νοερά τα περιεχόμενα των αγκυλών ως “χρησιμοποίησε έναν από τους χαρακτήρες που περιέχονται” τότε η παραπάνω έκφραση μπορεί να παραγάγει όλες τις αλληλουχίες του πίνακα των σημείων πρόσδεσης. Έχουμε καταφέρει έτσι σε κάποιο βαθμό να περιγράψουμε τις ιδιότητες αυτού του μοτίβου. [8]

2.5 Το πρόβλημα της εύρεσης μοτίβων

- -
Με δεδομένο ένα σύνολο αλληλουχιών DNA, βρείτε ένα σύνολο l - μερών, ένα από κάθε αλληλουχία, που μεγιστοποιεί τη συναινετική βαθμολογία.
Είσοδος: Η μήτρα DNA διαστάσεων t x n και το l, δηλαδή το μήκος του μοτίβου που πρέπει να βρεθεί.
Έξοδος: Ένας πίνακας με t αρχικές θέσεις $s = (s_1, s_2, \ldots, s_t)$ που μεγιστοποιεί τη βαθμολογία Score(s, DNA).
- -

Για να διατυπώσουμε το πρόβλημα της εύρεσης μοτίβου με σαφήνεια, πρέπει να ορίσουμε με ακρίβεια τι εννοούμε με τη λέξη "**μοτίβο**". Η αναπαράσταση ενός μοτίβου με μία μόνο συμβολοσειρά δεν είναι συχνά αντιπροσωπευτική των παραλλαγών του μοτίβου σε πραγματικές βιολογικές αλληλουχίες. Η αναπαράσταση του μοτίβου είναι πιο ευέλικτη όταν χρησιμοποιείται μήτρα προφίλ. Θεωρούμε ένα σύνολο t αλληλουχιών DNA με n νουκλεοτίδια σε κάθε αλληλουχία. Επιλέγουμε μία θέση σε κάθε μία από τις t αλληλουχίες , σχηματίζοντας έτσι έναν πίνακα $s = (s_1, s_2, \ldots, s_t)$ με $1 <= s_i <= n – l + 1$. Τα l – μέρη που αρχίζουν στις συγκεκριμένες θέσεις μπορούν να αναπαρασταθούν με μία μήτρα στοίχισης με διαστάσεις t x l, όπου το στοιχείο (i, j) είναι το νουκλεοτίδιο του $(s_i + j – 1)$ – οστού στοιχείου στην I – οστή αλληλουχία. Με βάση τη μήτρα στοίχισης, μπορούμε να υπολογίσουμε τη μήτρα προφίλ με διαστάσεις 4 x l, στην οποία το στοιχείο (i, j) είναι η συχνότητα εμφάνισης του νουκλεοτιδίου i στη στήλη j της μήτρας στοίχισης και το i παίρνει τιμές από 1 έως 4. Η μήτρα προφίλ δείχνει τη μεταβλητότητα της νουκλεοτιδικής σύνθεσης σε κάθε θέση για τη συγκεκριμένη επιλογή l – μερών. Για να συνοψίσουμε περισσότερο τη μήτρα προφίλ, μπορούμε να σχηματίσουμε μια ***συναινετική συμβολοσειρά***(βλέπε ενότητα 2.5.1) με το πιο συχνό στοιχείο από κάθε στήλη της μήτρας στοίχισης, το οποίο είναι το νουκλεοτίδιο με τη μεγαλύτερη συχνότητα στη μήτρα προφίλ. Αν μεταβάλλουμε τις αρχικές θέσεις στο s, μπορούμε να δημιουργήσουμε πολλές διαφορετικές μήτρες προφίλ από ένα δεδομένο δείγμα. Πρέπει να βρούμε κάποιον τρόπο να βαθμολογούμε κάθε μήτρα σε σχέση με τις υπόλοιπες. Κάποια προφίλ συμβολίζουν την υψηλή συντήρηση ενός μοτίβου, ενώ άλλα δείχνουν μη συντήρηση. Μια ανακριβής διατύπωση του προβλήματος της Εύρεσης Μοτίβου είναι ότι πρέπει να βρούμε τις αρχικές θέσεις s που αντιστοιχούν στο πιο καλά συντηρημένο προφίλ.

Αν η μήτρα προφίλ που αντιστοιχεί στις αρχικές θέσεις s συμβολίζεται με P(s), τότε η μεγαλύτερη μέτρηση στη στήλη j του P(s) θα συμβολίζεται με $M_{p(s)}(j)$.

- -

ΠΑΡΑΔΕΙΓΜΑ 1

Θα παρουσιάσουμε ένα παράδειγμα για να καταλάβουμε την σημασία των εννοιών(μήτρα στοίχισης, μήτρα προφίλ, συναινετική αλληλουχία) που έχουμε πει.

(α) Υπέρθεση των επτά υπογραμμισμένων 8 – μερών.

CGGGGCTATcCAgCTGGG…
...TAAggGCAACTCCAAAG…
GGATGgAtCTGATGCC…
AAGGAaGCAACcCCAGGAGC…
...CtTGgAACTTC
TGCATGCcAtTTTCAAC
...TGATGgcACTTGGATG…

(β) Η μήτρα στοίχισης, η μήτρα προφίλ και η συναινετική αλληλουχία που σχηματίζονται από τα 8 - μερή που αρχίζουν στις θέσεις s = (8, 19, 3, 5, 31, 27, 15). Τότε το Score(s) = 5 + 5 + 6 + 4 + 5 + 6 + 6 = 42.

Στοίχιση

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

Προφίλ

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

Συναινετική αλληλουχία

A T G C A A C T

- -

Για το προφίλ P(s) του παραδείγματος 1 έχουμε $M_{P(s)}(1) = 5$, $M_{P(s)}(2) = 5$, … $M_{P(s)}(8) = 6$. Με δεδομένες τις αρχικές θέσεις s, η συναινετική βαθμολογία ορίζεται ως $Score(s, DNA) = \Sigma^{l}_{j=1} M_{P(s)}(j)$. Η βαθμολογία Score(s, DNA) μπορεί να χρησιμοποιηθεί για τη μέτρηση της ισχύος ενός προφίλ που αντιστοιχεί στις αρχικές θέσεις s. Η συναινετική βαθμολογία l x t αντιστοιχεί στην καλύτερη δυνατή στοίχιση, όπου κάθε γραμμή μιας στήλης έχει το ίδιο γράμμα. Από την άλλη πλευρά η συναινετική βαθμολογία lt/4 αντιστοιχεί στη χειρότερη δυνατή στοίχιση, στην οποία όλα τα νουκλεοτίδια σε κάθε στήλη εμφανίζονται με την ίδια συχνότητα. Στην απλούστερη μορφή του, το πρόβλημα της Εύρεσης Μοτίβου μπορεί να διατυπωθεί ως η επιλογή των αρχικών θέσεων s από το δείγμα που μεγιστοποιεί τη βαθμολογία Score(s, DNA).

Μια διαφορετική οπτική είναι να αναδιατυπώσουμε το πρόβλημα της Εύρεσης Μοτίβου ως το πρόβλημα της εύρεσης μιας *μεσαίας συμβολοσειράς*. [9]

2.5.1 Συναινετικές αλληλουχίες

Οι αλληλουχίες που εξετάσαμε προκειμένου να περιγράψουμε το μοτίβο ως κανονική έκφραση ήταν μόλις 8 και οι μεταξύ τους αμφισημίες εύκολα παρατηρήσιμες. Έτσι μπορέσαμε εύκολα να τις καταγράψουμε και να τις κωδικοποιήσουμε σε μια κανονική έκφραση. Τι συμβαίνει όμως στην περίπτωση που είμαστε αρκετά τυχεροί ώστε να διαθέτουμε ένα μεγάλο αριθμό από σημεία πρόσδεσης; Στον πίνακα που βλέπουμε παρακάτω έχουμε συγκεντρώσει 104 σημεία πρόσδεσης (TFBS) του NF-κΒ όπως προέκυψαν από ένα πείραμα μεγάλης κλίμακας. Αν προσπαθήσουμε να περιγράψουμε με μια κανονική έκφραση το σύνολο αυτών των αλληλουχιών αυτή θα είναι:

GG[AG][AG][AG][AGCT][AGCT][AGCT][ACT][ACT][CT]

Βλέπετε ότι μεγαλύτερος αριθμός διαθέσιμων σημείων πρόσδεσης δημιουργεί προβλήματα σε αυτήν την προσέγγιση. Προσπαθώντας να συμπεριλάβουμε όλες τις πιθανές παραλλαγές καταλήγουμε σε μια πολύ γενική έκφραση με μικρό πληροφοριακό περιεχόμενο. Ωστόσο, μπορούμε να χρησιμοποιήσουμε αυτόν τον όγκο των δεδομένων με καλύτερο τρόπο χρησιμοποιώντας προσεγγίσεις που έχουμε συναντήσει στα προηγούμενα κεφάλαια και σχετίζονται με τις συχνότητες εμφάνισης νουκλεοτιδίων. Η έκτη θέση π.χ. μπορεί να περιέχει και τις τέσσερις

πιθανές βάσεις αλλά αυτό δε συμβαίνει με την ίδια συχνότητα εμφάνισης. Ας δούμε πώς μπορούμε να περιγράψουμε το μοτίβο με μεγαλύτερη ακρίβεια απ' ότι με μια απλή κανονική έκφραση χρησιμοποιώντας αυτές τις συχνότητες εμφάνισης.

```
GGGGCATTCC  GGGATATCCC  GGGAATTCCC  GGGAATGTCC  GGGATATTTC GGGGCCTCCC GGGAATTTCC GGGACTGCCC
GGGAAATTCC  GGGAAATCCC  GGGAATTCCC  GGGACTTACC  GGGGATTTCC GGGAATTTCC GGGACATTCC GGGAATTTCC
GGAAATTTCC  GGGAATTCCC  GGGGATTTCC  GGGGTTTCAC  GGGAAGGTCC GGGGCTTCCC GGGGCTTTCC GGGAAATTCC
GGGGCTTTCC  GGGACTTTCC  GGGACATTCC  GGGAATTTCC  GGGACATTCT GGGACAGCCC GGGGCTTTAC GGGACTTCCC
GGGAATTCAC  GGGAAATCCC  GGAGCTTTCC  GGGACTTTCC  GGGAAACCCC GGGGCTTCCC GGGAATTTCC GGGAAATTCC
GGGACTTCCC  GGGAATTTCT  GGGAATTCCC  GGGACTTCCC  GGGACTTTCC GGGGATTTCC GGGACATCCC GGGAAATCCC
GGGATGTTCC  GGGGTCTCCC  GGGACTGTCC  GGGAATTCCC  GGGACTTTAC GGGAATTTCC GGGACTTTCC GGGGCGTCCC
GGGGTTTCCC  GGGAATTTCC  GGGAATTTCC  GGGGATTTCC  GGGAATGCCC GGGGATTTCC GGGAATTTCC GGGATTTTCC
GGGGAATTCC  GGGACTTCCC  GGGATTTTCC  GGGAAGTCCC  GGGAAATTCC GGGAATTTCC GGGAATTTAC GGGAAATTCC
GGGGGTTTAC  GGGACTTTCC  GGGAATTTCC  GGGAATTTCC  GGGACATCCC GGGAATTCAC GGGACTTCCC GGGACTTTCC
GGGAATTTCC  GGGACTTTCC  GGGGACTTCC  GGGACTTTAC  GGGACTTTCC GGGATACTCC GGGGATGTAC GGGATATCCC
GGGAATTCCC  GGGACTTCCC  GGGACTTCAC  GGGGTTACCC  GGGAATCTCC GGGAATTTCC GGGACATCTC GGAAATTCCC
GGGAAACTCT  GGGGTTTCCC  GGGATTTTCC  GGGGCGTTCC  GGGAAACTCT GGGGTTTCCC GGGATTTTCC GGGGCGTTCC
```

Πιο συγκεκριμένα, μπορούμε να προσπαθήσουμε να απαντήσουμε στα εξής ερωτήματα:

- -

Ποιο νουκλεοτίδιο είναι το πιθανότερο σε κάθε θέση του μοτίβου;

Ποια είναι συνολικά η πιο πιθανή αλληλουχία πρόσδεσης;

Πόσο μεγάλη είναι η αμφισημία του μοτίβου;

- -

Ας ξεκινήσουμε με το πρώτο από αυτά. Η απάντηση μπορεί να δωθεί με μια απλή καταμέτρηση των νουκλεοτιδίων σε κάθε θέση ξεχωριστά, τη μετατροπή τους σε συχνότητα εμφάνισης και την εξαγωγή του νουκλεοτιδίου με τη μέγιστη τιμή συχνότητας εμφάνισης. Αν τώρα επιλέξουμε την αλληλουχία που θα ορίζεται από το πιθανότερο νουκλεοτίδιο για κάθε θέση του μοτίβου, θα έχουμε απαντήσει και στη δεύτερη από τις πιο πάνω ερωτήσεις. Συχνά αναφερόμαστε σε αυτήν την “πιο πιθανή” αλληλουχία ως συναινετική αλληλουχία (consensus sequence), καθώς είναι η αλληλουχία με την οποία συμφωνούν περισσότερο τα TFBS που έχουν χρησιμοποιηθεί για την εξαγωγή της. Ένας απλός αλγόριθμος για την εξαγωγή της συναινετικής αλληλουχίας από έναν πίνακα σημείων πρόσδεσης δίνεται παρακάτω:

- -

```
Αλγόριθμος :: Εξαγωγή Consensus
Δήλωση Πίνακα n Σημείων Πρόσδεσης μήκους l, TFBS[n,l];
Δήλωση Πίνακα Ν τεσσάρων νουκλεοτιδίων (A,G,C,T);
Δήλωση Πίνακα Ρ Πιθανοτήτων νουκλεοτιδίων (A,G,C,T);
Δήλωση Πίνακα Μ Μέγιστων Τιμών Πιθανοτήτων Μ[l];
Απαρίθμηση 1: Για θέση i = 1 έως i = l ανά 1;
        Δημιούργησε τη σειρά C=TFBS[1:n,i]; # όλα τα στοιχεία κάθε στήλης
        Απαρίθμηση 2: Για θέση j=1 έως j=n ανά 1;
                Διάβασε s=C[j];
                Αύξησε το πλήθος πίνακα νουκλεοτιδίων N[s]++;
        Τέλος: Απαρίθμηση 2
        #
        Συχνότητα: Για κάθε νουκλεοτίδιο s;
        Υπολογισε τη συχνότητα P[s]=N[s]/n; # διαίρεση με πλήθος σημείων
        Τέλος: Συχνότητα
        #
        Μέγιστο: max=0;
        Για κάθε νουκλεοτίδιο s;
                Έλεγξε: Αν P[s]>max
                Τότε:
                max=P[s];
                argmax=s; # argmax είναι το νουκλεοτίδιο με τη μέγιστη τιμή
        Τέλος Μέγιστο
        M[i]=argmax; # το νουκλεοτίδιο με τη μέγιστη τιμή για τη θέση i αποδίδεται
στο μ
        N=0; P=0; # αρχικοποίηση πινάκων πλήθους συχνοτήτων
Τέλος Απαρίθμηση 1
Απόδωσε αποτέλεσμα: Πίνακας Μ
Τερματισμός
```

- -

Η σειρά των πιο πάνω εντολών, αν και φαινομενικά περίπλοκη κάνει μια σειρά από απλούς υπολογισμούς που έχουμε ήδη δει. Διαβάζει έναν πίνακα σημείων πρόσδεσης ανα στήλη, καθώς ενδιαφερόμαστε να συγκρίνουμε τα κατάλοιπα για κάθε θέση μέσα στο μοτίβο ξεχωριστά και αφού υπολογίσει τη συχνότητα εμφάνισης του κάθε νουκλεοτιδίου εντός της στήλης, εξάγει αυτό με τη μεγαλύτερη συχνότητα και το αποδίδει σε έναν πίνακα Μ. Ο Μ με μήκος ίσο με αυτό των αλληλουχιών πρόσδεσης καταλήγει έτσι να περιέχει τη συναινετική

αλληλουχία, η οποία τα περιγράφει καλύτερα. Για το παράδειγμά μας των 104 TFBS που είδαμε παραπάνω αυτή είναι η:

G G G A A T T C C

Την αλληλουχία που αντιστοιχεί στη συναινετική αλληλουχία συναντήσαμε ήδη πιο πάνω και μάλιστα είχαμε ήδη από τότε αναφέρει πως είναι αυτή με την οποία ο NF-κB έχει τη μεγαλύτερη συγγένεια πρόσδεσης. Τι ακριβώς αναπαριστά η συναινετική αλληλουχία; Αρχικά είναι η αλληλουχία που αναμένεται να είναι και η πιο συχνά απαντώμενη ανάμεσα σε όλα τα σημεία πρόσδεσης. Πράγματι η GGGAATTTCC εμφανίζεται 14 φορές μέσα στο δείγμα μας (ένα ποσοστό 13%). Κατά κύριο λόγο όμως η συναινετική αλληλουχία είναι αυτή έναντι της οποίας το σύνολο των σημείων πρόσδεσης εμφανίζει τις λιγότερες διαφορές. Αν μπορούσαμε να ορίσουμε ένα μέτρο απόστασης μεταξύ αλληλουχιών και να το υπολογίσουμε για το σύνολο των αλληλουχιών θα μπορούσαμε να έχουμε ένα μέτρο της συμφωνίας με τη συναινετική αλληλουχία. Ένα τέτοιο μέτρο είναι η απόσταση Hamming. Ως απόσταση Hamming δύο σειρών χαρακτήρων ίδιου μήκους ορίζουμε το άθροισμα των χαρακτήρων στους οποίους διαφέρουν (Forney 1966). Έτσι οι:

```
G G G A A T T T C C
| |   | | | | | | |
G G C A A T T T C C
```

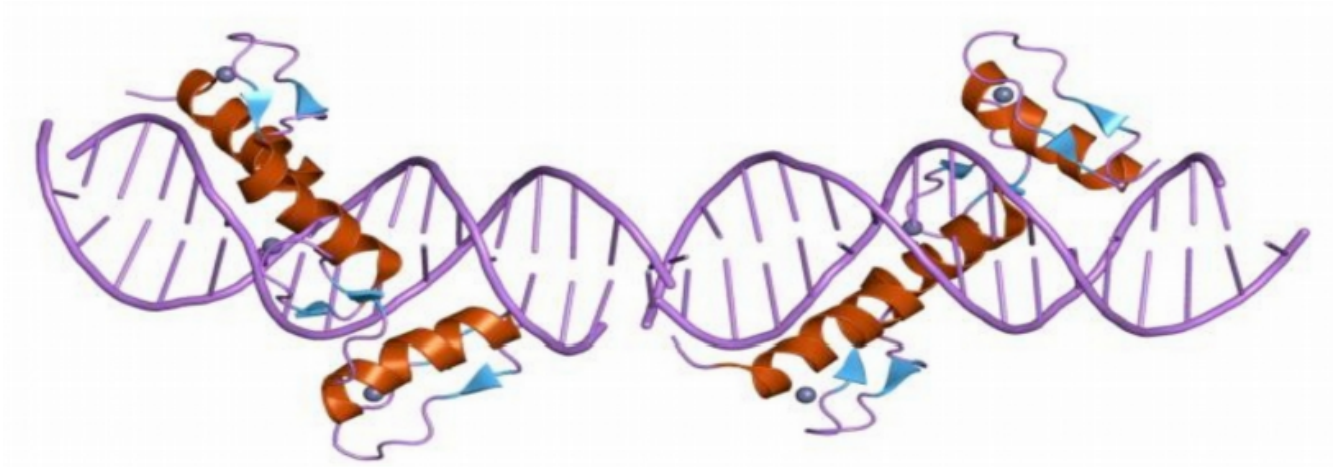
Έχουν απόσταση Hamming d=1 ενώ οι:

```
G G G A A T T T C C
| |   | | |     | |
G G C A A T A A C C
```

Έχουν απόσταση Hamming d=3.

Υπολογισμός των αποστάσεων Hamming για τα 104 σημεία πρόσδεσης του NF-κB του πίνακα που φαίνεται παραπάνω από τη συναινετική αλληλουχία τους, δίνει αποστάσεις Hamming μεταξυ 0 και 1, καμία δηλαδή αλληλουχία δεν διαφέρει από τη συναινετική περισσότερο από 1 κατάλοιπο. Το άθροισμα των αποστάσεων

Hamming για όλες τις αλληλουχίες είναι 86 και η μέση τιμή τους είναι 86/104=0.82. Αν αναλογιστούμε ότι η μέγιστη δυνατή απόσταση Hamming δύο σειρών μήκους N είναι ίση με N, τότε το 0.82 σε σχέση με το 10 είναι ένας αρκετά μικρός αριθμός. Μπορούμε από αυτό να εκτιμήσουμε ότι ο βαθμός αμφισημίας του μοτίβου είναι αρκετά μικρός, έχοντας έτσι απαντήσει και στο τρίτο ερώτημα.



Στο σημείο αυτό θα πρέπει να σκεφτούμε μια άλλη διάσταση του προβλήματος εξαγωγής της συναινετικής αλληλουχίας και υπολογισμού της αβεβαιότητας που σχετίζεται με τη θέση του μοτίβου. Όπως είδαμε και πιο πάνω, υπάρχουν θέσεις μέσα στο μοτίβο (π.χ. οι πρώτες 2) όπου η αβεβαιότητα είναι μηδενική, ενώ άλλες όπου η αβεβαιότητα εξαιρετικά μεγάλη. Τόσο η προσέγγιση της συναινετικής αλληλουχίας όσο και ο υπολογισμός αποστάσεων Hamming αντιμετωπίζει όλες τις θέσεις με την ίδια βαρύτητα. Κάτι τέτοιο όμως είναι προβληματικό, μεταξύ άλλων και από βιολογική άποψη. Φανταστείτε ένα μεταγραφικό παράγοντα όπως αυτός που φαίνεται στην εικόνα παραπάνω, ο οποίος αλληλεπιδρά με δύο διακριτές περιοχές του DNA μεταξύ των οποίων υπάρχει μια μικρή αλληλουχία με την οποία δεν υπάρχει καμία επαφή. Είναι λογικό να περιμένουμε πως τα νουκλεοτίδια που θα έχουν σημασία για το μοτίβο θα βρίσκονται στην αρχή και στο τέλος αλλά τα ενδιάμεσα θα μπορούν να μεταβάλλονται με λιγότερους περιορισμούς. Στην επόμενη ενότητα θα δούμε πώς μπορούμε να μελετήσουμε μια σειρά από σημεία πρόσδεσης με σκοπό να αναπαραστήσουμε ένα μοτίβο, στο οποίο οι διαφορές μεταξύ των θέσεων θα περιγράφονται πιο αναλυτικά.
[10]

2.5.2 Μεσαία συμβολοσειρά

- -

Με δεδομένο ένα σύνολο αλληλουχιών DNA, βρείτε τη μεσαία συμβολοσειρά.

Είσοδος: Η μήτρα DNA διαστάσεων t x n, και το l, δηλαδή το μήκος του μοτίβου που πρέπει να βρεθεί.

Έξοδος: Η συμβολοσειρά ν με l νουκλεοτίδια που ελαχιστοποιεί την απόσταση TotalDistance(ν, DNA) για όλες τις συμβολοσειρές με το συγκεκριμένο μήκος.

- -

Αν μας έχουν δοθεί δύο l – μερή ν και w, μπορούμε να υπολογίσουμε τη μεταξύ τους *απόσταση Hamming* $d_H(v, w)$ ως τον αριθμό των θέσεων που διαφέρουν στις δύο συμβολοσειρές. Έστω s = (s1, s2, …, st) είναι πίνακας αρχικών θέσεων και ότι το ν είναι κάποιο l – μερές. Θα συμβολίσουμε με $d_H(v, s)$ τη συνολική απόσταση Hamming ανάμεσα στο ν και στα l – μερή που αρχίζουν στις θέσεις **s**: $d_H(v, s) = \Sigma^t_{i=1} d_H(v, s_i)$ όπου $d_H(v, s_i)$ είναι η απόσταση Hamming ανάμεσα στο ν και το l – μερές που αρχίζει στη θέση s_i της i – οστής αλληλουχίας DNA. Θα χρησιμοποιήσουμε την σχέση $TotalDistance(v, DNA) = min_s(d_H(v, s))$ για να δηλώσουμε τη μικρότερη δυνατή απόσταση Hamming ανάμεσα σε μία δεδομένη συμβολοσειρά ν και οποιοδήποτε σύνολο αρχικών θέσεων στο DNA. Η εύρεση του TotalDistance αποτελεί απλό πρόβλημα: πρώτα πρέπει να βρει κανείς το καλύτερο ταίριασμα για τη συμβολοσειρά ν στην πρώτη αλληλουχία DNA, στη συνέχεια το καλύτερο ταίριασμα στη δεύτερη αλληλουχία DNA, κ.ο.κ. Δηλαδή, το ελάχιστο υπολογίζεται για όλες τις δυνατές αρχικές θέσεις **s**. Τέλος, η μεσαία συμβολοσειρά για το DNA ορίζεται ως η συμβολοσειρά ν που ελαχιστοποιεί την απόσταση TotalDistance(ν, DNA) η ελαχιστοποίηση υπολογίζεται για όλες τις 4^l συμβολοσειρές ν με μήκος l.

Βρίσκουμε μια συμβολοσειρά ν που ελαχιστοποιεί την απόσταση TotalDistance(ν, DNA) , η οποία είναι με τη σειρά της η μικρότερη απόσταση ανάμεσα σε όλες τις επιλογές των αρχικών θέσεων s στις αλληλουχίες DNA. Με άλλα λόγια υπολογίζουμε το

$$\min_{\text{όλες οι επιλογές των l – μερών ν}} \quad \min_{\text{όλες οι επιλογές των θέσεων έναρξης s}} \quad d_H(v, s)$$

Παρά το γεγονός ότι το πρόβλημα της Μεσαίας Συμβολοσειράς είναι πρόβλημα ελαχιστοποίησης και το πρόβλημα της Εύρεσης Μοτίβου είναι πρόβλημα

μεγιστοποίησης, τα δύο προβλήματα είναι υπολογιστικά ισοδύναμα. Έστω **s** ένα σύνολο αρχικών θέσεων με συναινετική βαθμολογία Score(s, DNA) και w η συναινετική συμβολοσειρά του αντίστοιχου προφίλ. Τότε ισχύει $d_H(w, s) = lt – Score(s, DNA)$. Εφόσον τα t και l είναι σταθερές, η μικρότερη τιμή της απόστασης d_H μπορεί να υπολογιστεί με τη μεγιστοποίηση της βαθμολογίας Score(s, DNA) για όλες τις επιλογές του **s**:

$$\min_{\text{όλες οι επιλογές των l – μερών v}} \min_{\text{όλες οι επιλογές των θέσεων έναρξης s}} d_H(v, s) =$$

$$lt - \max_{\text{όλες οι επιλογές του } \mathbf{s}} Score(s, DNA)$$

Με άλλα λόγια, η συναινετική συμβολοσειρά για τη λύση του προβλήματος της Εύρεσης Μοτίβου είναι η μεσαία συμβολοσειρά για το δείγμα DNA της εισόδου. Μπορούμε να χρησιμοποιήσουμε τη μεσαία συμβολοσειρά για το DNA για να δημιουργήσουμε ένα προφίλ που λύνει το πρόβλημα της Εύρεσης Μοτίβου, αναζητώντας σε κάθε μία από τις t αλληλουχίες την υποσυμβολοσειρά με τη μικρότερη απόσταση Hamming από τη μεσαία συμβολοσειρά. [11]

2.5.3 Πίνακες Βαρών ανά θέση

Ας προσπαθήσουμε να σκεφτούμε έναν τρόπο για να αναπαραστήσουμε καλύτερα την πληροφορία που χρησιμοποιήσαμε για την εξαγωγή της συναινετικής αλληλουχίας σε προηγούμενη ενότητα. Κατά την εφαρμογή του αλγορίθμου Εξαγωγή Consensus υπολογίζαμε τη συχνότητα εμφάνισης κάθε νουκλεοτιδίου ξεχωριστά για κάθε θέση του μοτίβου. Πώς θα μπορεί να ενσωματωθεί αυτή η πληροφορία σε μία δομή δεδομένων; Αν αναλογιστούμε ότι για κάθε θέση n εντός του μοτίβου υπάρχουν 4 ενδεχόμενα (τα τέσσερα νουκλεοτίδια) για τα οποία μπορούμε να υπολογίσουμε συχνότητες εμφάνισης, τότε μπορούμε να φανταστούμε έναν πίνακα n x 4, ο οποίος θα περιέχει τις τέσσερις συχνότητες εμφάνισης για κάθε θέση. Τέτοιοι πίνακες ονομάζονται, Πίνακες Βαρών ανά θέση (Positional Weight Matrices ή για συντομία PWM) καθώς περιέχουν πληροφορία ξεχωριστά για κάθε θέση του μοτίβου. Ο υπολογισμός ενός PWM μπορεί να γίνει με μια απλοποιημένη εκδοχή του αλγορίθμου των συναινετικών αλληλουχιών, ως εξής:

- -

Αλγόριθμος :: PWM

Δήλωση Πίνακα n Σημείων Πρόσδεσης μήκους l, TFBS[n,l];

Δήλωση Πίνακα N τεσσάρων νουκλεοτιδίων (A,G,C,T);

Δήλωση Πίνακα P Πιθανοτήτων νουκλεοτιδίων (A,G,C,T);

Δήλωση Πίνακα PWM[4, l];

Απαρίθμηση 1: Για θέση i = 1 έως i = l ανά 1;

Δημιούργησε τη σειρά C=TFBS[1:n,i]; # όλα τα στοιχεία κάθε στήλης

Απαρίθμηση 2: Για θέση j=1 έως j=n ανά 1;

Διάβασε s=C[j]

Αύξησε το πλήθος πίνακα νουκλεοτιδίων N[s]++;

Τέλος: Απαρίθμηση 2

#

Συχνότητα: Για κάθε νουκλεοτίδιο s;

Υπολογισε τη συχνότητα P[s]=N[s]/n; # διαίρεση με πλήθος σημείων

Απόδοση στον PWM[i,s]=P[s]

N=0; P=0; # αρχικοποίηση πινάκων πλήθους συχνοτήτων

Τέλος Απαρίθμηση 1

Απόδωσε αποτέλεσμα: Πίνακας PWM

Τερματισμός

- -

Η διαφορά του PWM με τον Consensus είναι ουσιαστικά ότι αντί να κρατάμε μόνο τη μέγιστη από τις τέσσερις συχνότητες εμφάνισης για κάθε θέση, τις ενσωματώνουμε όλες σε ένα πίνακα 2 διαστάσεων. Ο PWM που προκύπτει για τα 104 μοτίβα του πίνακα που αναφέραμε πιο πάνω είναι ο παρακάτω:

Νουκλεοτίδιο	1	2	3	4	5	6	7	8	9	10
A	0.00	0.00	0.03	0.76	0.49	0.23	0.01	0.01	0.10	0.00
C	0.00	0.00	0.00	0.00	0.37	0.03	0.04	0.38	0.88	0.97
G	1.00	1.00	0.97	0.24	0.01	0.05	0.07	0.00	0.00	0.00
T	0.00	0.00	0.00	0.00	0.13	0.69	0.88	0.61	0.02	0.03

Στον πίνακα που μόλις παρουσιάσαμε παραπάνω είναι σημειωμένα με κίτρινο τα βάρη που αντιστοιχούν στις μέγιστες τιμές συχνότητας ανά θέση. Ακολουθώντας τα για καθεμία από τις 10 θέσεις του μοτίβου μπορούμε να εξάγουμε εύκολα τη συναινετική αλληλουχία. Ωστόσο ο Πίνακας 3.2 πλεονεκτεί ως αναπαράσταση του μοτίβου καθώς περιέχει επιπλέον πληροφορία για όλα τα νουκλεοτίδια. Έτσι μπορούμε τώρα να εκτιμήσουμε ποσοτικά το βαθμό αβεβαιότητας της κάθε θέσης. Τόσο για τη θέση 4 όσο και για την 5 το πιθανότερο νουκλεοτίδιο είναι το A, ωστόσο για τη μεν 4 η πιθανότητα είναι 0.76 ενώ για την 5 μόλις 0.49. Ο PWM μας επιτρέπει έτσι να εκτιμήσουμε την αβεβαιότητα που εμπεριέχει το μοτίβο λαμβάνοντας υπόψη την κάθε θέση ξεχωριστά. Στη συνέχεια θα δούμε πώς μπορούμε να εκμεταλλευτούμε αυτό το γεγονός για να απαντήσουμε στο δεύτερο από τα αρχικά ερωτήματα, αυτό του εντοπισμού σημείων πρόσδεσης σε μια άγνωστη αλληλουχία. [12]

2.6 Εύρεση μοτίβων

Η προσέγγιση της ωμής βίας για να λυθεί το πρόβλημα της Εύρεσης Μοτίβου εξετάζει όλες τις πιθανές αρχικές θέσεις.

- -
BRUTEFORCEMOTIFSEARCH(DNA, t, n, l)
1. bestScore ← 0
2. **for each** (s_1, …, s_t) **from** (1, …, 1) **to** (n – l + 1, …, n – l + 1)
3. **if** Score(s, DNA) > bestScore
4. bestScore ← Score(s, DNA)
5. **bestMotif** ← (s_1, s_2, …, s_t)
6. **return bestMotif**
- -

Ο συνολικός αριθμός των θέσεων ισούται με $(n – l + 1)^t$ το οποίο είναι εκθετικό ως προς το t, δηλαδή τον αριθμό των αλληλουχιών. Για κάθε s, ο αλγόριθμος υπολογίζει τη βαθμολογία Score(s, DNA), η οποία απαιτεί O(l) πράξεις. Άρα η συνολική πολυπλοκότητα του αλγορίθμου εκτιμάται ότι είναι $O(ln^t)$.

Ένας άλλος αλγόριθμος ο οποίος καλυτερεύει την εντολή 2 είναι ο ακόλουθος,

```
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
BRUTEFORCEMOTIFSEARCHAGAIN(DNA, t, n, l)
1. s ← (1, 1, …, 1)
2. bestScore ← Score(s, DNA)
3. while για πάντα
4.     s ← NEXTLEAF(s, t, n – l + 1)
5.     if Score(s, DNA) > bestScore
6.         bestScore ← Score(s, DNA)
7.         bestMotif ← (s1, s2, …, st)
8.     if s = (1, 1, …, 1)
9.         return bestMotif
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
```

Ο αλγόριθμος NEXTLEAF λειτουργεί κατά τρόπο πολύ παρόμοιο με τη φυσική διαδικασία της μέτρησης αριθμών. Με δεδομένο L – μερές από αλφάβητο k γραμμάτων, η υπορουτίνα NEXTLEAF δείχνει πώς πραγματοποιείται ένα άλμα από κάποιο L – μερές **a** = $(a_1a_2…a_L)$ στο επόμενο L – μερές της προόδου.

```
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
NEXTLEAF(a, L, k)
1. for i ← L to 1
2.     if ai < k
3.         ai ← ai + 1
4.     return a
5.     ai ← 1
6.     return a
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
```

Για να προετοιμαστούμε για τη στρατηγική διακλάδωση και οριοθέτηση, χρειαζόμαστε την ισοδύναμη έκδοση SIMPLEMOTIFSEARCH που κάνει χρήση της υπορουτίνας NEXTVERTEX για να διερευνήσει κάθε φύλλο.

- -

```
SIMPLEMOTIFSEARCH(DNA, t, n, l)
1. s ← (1, 1, …, 1)
2. bestScore ← 0
3. i ← 1
4. while i > 0
5.     if i < t
6.         (s, i) ← NEXTVERTEX(s, i, t, n – l + 1)
7.     else
8.         if Score(s, DNA) > bestScore
9.             bestScore ← Score(s, DNA)
10.            bestMotif ← (s1, s2, …, st)
11.        (s, i) = NEXTVERTEX(s, i, t, n – l + 1)
12. return bestMotif
```

- -

Ο αλγόριθμος NEXTVERTEX χρησιμοποιείται για να διατρέξουμε το πλήρες δέντρο επαναληπτικά. Ο αλγόριθμος παίρνει ως είσοδο την κορυφή $a = (a_1, \ldots, a_L)$ στο επίπεδο i και επιστρέφει την επόμενη κορυφή του δέντρου. Στην πραγματικότητα, ο αλγόριθμος χρησιμοποιεί στο επίπεδο i μόνο τις τιμές $a_1, \ldots, a_i$ και αγνοεί τις τιμές $a_{i+1}, \ldots, a_L$. Οι είσοδοι του είναι παρόμοιοι μ τον αλγόριθμο NEXTLEAF με την εξαίρεση ότι το “τρέχον φύλλο” είναι πλέον “τρέχουσα κορυφή”, άρα χρησιμοποιεί την παράμετρο I για το επίπεδο το οποίο βρίσκεται η κορυφή. Με δεδομένα τα a, L, i και k, ο NEXTVERTEX επιστρέφει την επόμενη κορυφή του δέντρου ως το ζευγάρωμα ενός πίνακα και ενός επιπέδου. Ο αλγόριθμος θα επιστρέψει έναν αριθμό επιπέδου 0 όταν έχει διατρέξει ολόκληρο το δέντρο.

- -

```
NEXTVERTEX(a, i, L, k)
1. if i < L
2.     ai+1 ← 1
3.     return (a, i + 1)
4. else
5.     for j ← L to 1
6.         if aj < k
7.             aj ← aj + 1
8.             return (a, j)
9. return (a, 0)
```

- -

Όταν ισχύει η σχέση 1 ο αλγόριθμος κινείται στο επόμενο χαμηλότερο επίπεδο και εξετάζει το υποδέντρο του **a**. Αν i = L, ο αλγόριθμος κινείται κατά μήκος του χαμηλότερου επιπέδου με την προϋπόθεση ότι ισχύει $a_L < k$ ή μεταπηδάει σε υψηλότερο επίπεδο του δέντρου.

Με δεδομένο το σύνολο αρχικών θέσεων **s** = $(s_1, s_2, \ldots, s_t)$, ορίζουμε τη μερική συναινετική βαθμολογία Score(**s**, i, DNA) ως τη συναινετική βαθμολογία της μήτρας στοίχισης με διαστάσεις i x l, η οποία περιλαμβάνει μόνο τις πρώτες i γραμμές του DNA που αντιστοιχούν στις αρχικές θέσεις $(s_1, s_2, \ldots, s_i, -, -, \ldots, -)$. Στην περίπτωση αυτή, ο χαρακτήρας “-” υποδεικνύει ότι δεν έχουμε επιλέξει τιμή για τη συγκεκριμένη καταχώρηση στο **s**. Αν έχουμε τη μερική συναινετική βαθμολογία για τις θέσεις $s_1, s_2, \ldots, s_i$, οι εναπομένουσες t - i γραμμές μπορούν στην καλύτερη περίπτωση να βελτιώσουν τη συναινετική βαθμολογία μόνο κατά (t – i) x l. Αυτό συνεπάγεται ότι, αν η παράσταση Score(**s**, i, DNA) έχει μικρότερη τιμή από την τρέχουσα καλύτερη βαθμολογία bestScore, τότε δεν έχει νόημα να εξετάσουμε κάποια από τις εναπομένουσες t -i αλληλουχίες στο δείγμα – θα ήτανε χαμένος κόπος με τη συγκεκριμένη επιλογή του συνόλου $(s_1, s_2, \ldots, s_i)$. Κατά συνέπεια, το φράγμα Score(**s**, i, DNA) + (t – i) x l θα μπορούσε να μας βοηθήσει να αποφύγουμε την εξέταση $(n - l + 1)^{t-i}$ φύλλων. Παρότι η στρατηγική διακλάδωσης και οριοθέτησης βελτιώνει τον αλγόριθμο μας κάποια στιγμιότυπα του προβλήματος, δεν έχουμε βελτιώσει την αποδοτικότητα για τη χειρότερη περίπτωση: θα μπορούσαμε να σχεδιάσουμε ένα δείγμα με εμφυτευμένο μοτίβο που απαιτεί εκθετικό χρόνο για να βρεθεί.

- -

```
BRANCHANDBOUNDMOTIFSEARCH(DNA, t, n, l)
s ← (1, 1, …, 1)
bestScore ← 0
i ← 1
while i > 0
    if i < t
        optimisticScore ← Score(s, i, DNA) + (t – i) x l
        if optimisticScore < bestScore
            (s, i) ← BYPASS(s, i, t, n – l + 1)
        else
            (s, i) ← NEXTVERTEX(s, i, t, n – l + 1)
    else
```

```
12.         if Score(s, DNA) > bestScore
13.             bestScore ← Score(s)
14.             bestMotif ← (s1, s2, …, st)
15.         (s, i) ← NEXTVERTEX(s, i, t, n – l + 1)
16. return bestMotif
```

- -

Ο αλγόριθμος BYPASS είναι ο εξής,

- -

```
BYPASS(a, i, L, k)
1. for j ← i to 1
2.     if aj < k
3.         aj ← aj + 1
4.         return (a, j)
5. return (a, 0)
```

- -

Προηγουμένως υπαινιχθήκαμε ότι η χρήση της δενδρικής αναπαράστασης συμφέρει στη μείωση της εργασίας που εκτελούν οι αλγόριθμοι αναζήτησης ωμής βίας. Η γενική μέθοδος διακλάδωσης και οριοθέτησης θα μας δώσει τη δυνατότητα να αγνοήσουμε όλους τους απογόνους μιας κορυφής, αν μπορούμε να αποδείξουμε ότι δεν παρουσιάζουν ενδιαφέρον. Αν κανένας από τους απογόνους μιας κορυφής δεν μπορεί να έχει καλύτερη βαθμολογία από το καλύτερο φύλλο που έχει ήδη εξεταστεί, τότε δεν έχει νόημα να λάβουμε υπόψη μας τους απογόνους της συγκεκριμένης κορυφής. Σε κάθε κορυφή, υπολογίζουμε ένα φράγμα την πιο αισιόδοξη βαθμολογία optimisticScore για οποιοδήποτε φύλλο του υποδέντρου που έχει τη συγκεκριμένη κορυφή ως ρίζα και αποφασίζουμε αν θα εξετάσουμε τους απογόνους της ή όχι. Στην ουσία, η στρατηγική ονομάζεται διακλάδωση και οριοθέτηση για αυτόν ακριβώς το λόγο: υπολογίζουμε ένα φράγμα σε κάθε σημείο και αποφασίζουμε αν θα εξετάσουμε και την άλλη διακλάδωση ή όχι. Η τεχνική αυτή απαιτεί ότι μπορούμε να παραλείψουμε ένα ολόκληρο υποδέντρο που έχει ρίζα σε τυχαία κορυφή. Η υπορουτίνα NEXTVERTEX δεν μπορεί να το κάνει, αλλά ο αλγόριθμος BYPASS μας επιτρέπει να παραλείψουμε το υποδέντρο που έχει ρίζα στην κορυφή (**a**, i). Αν παραλείψουμε μια κορυφή στο επίπεδο i του δέντρου, μπορούμε να αυξήσουμε το a_i. Ο αλγόριθμος BYPASS έχει τους ίδιους τύπους εισόδου και εξόδου με τον αλγόριθμο NEXTLEAF. [13]

2.7 Το δεύτερο πρόβλημα

2.7.1 Εντοπισμός θέσεων πρόσδεσης σε άγνωστη αλληλουχία

Μέχρι στιγμής έχουμε δει πώς μπορούμε να ενσωματώσουμε την πληροφορία από πολλαπλά γνωστά σημεία πρόσδεσης ενός μεταγραφικού παράγοντα σε μοτίβα, που περιγράφουν συνολικά τα χαρακτηριστικά των θέσεων πρόσδεσης αλλά και που αποτυπώνουν τη σχετική αβεβαιότητα σε κάθε θέση. Πώς όμως θα μπορούσαμε να χρησιμοποιήσουμε το μοτίβο για να αναζητήσουμε νέα σημεία πρόσδεσης σε μια αλληλουχία;

Ας φανταστούμε μια αλληλουχία μεγάλου μήκους μέσα στην οποία αναζητούμε το μοτίβο. Ένας τρόπος να το κάνουμε είναι να τη διαβάζουμε σε υπο-αλληλουχίες μήκους ίσου με το μήκος του μοτίβου και να τις συγκρίνουμε με αυτό. Μπορούμε να φανταστούμε ένα “κυλιόμενο πλαίσιο” όπως αυτό που επιστρατεύσαμε στο Κεφάλαιο 2 και το οποίο θα έχει μήκος 10 νουκλεοτίδια. Στη συνέχεια μπορούμε να συγκρίνουμε κάθε υπο-αλληλουχία μήκους 10 με το μοτίβο. Πώς θα το κάνουμε αυτό; Μια προσέγγιση θα ήταν να υπολογίσουμε την απόσταση Hamming της κάθε υπο-αλληλουχίας από τη συναινετική αλληλουχία. Κάτι τέτοιο είναι προβληματικό για δύο λόγους. Πρώτο γιατί θα πρέπει να ορίσουμε κάτω από ποιο όριο απόστασης θα θεωρήσουμε ότι δεχόμαστε ότι η σύγκριση είναι επιτυχής. Θα είναι 1/10, 2/10 ή 3/10; Δεύτερο και κυριότερο, γιατί όπως είπαμε η απόσταση Hamming αντιμετωπίζει όλες τις θέσεις ως ισότιμες. Έτσι η υπο-αλληλουχία:

A G G A A T T T C C

Έχει απόσταση Hamming από τη συναινετική ίση με d=1, ωστόσο δεν θα μπορούσε να είναι σημείο πρόσδεσης καθώς από τον Πίνακα 3.2 προκύπτει ότι η πρώτη θέση θα πρέπει να περιέχει πάντοτε το νουκλεοτίδιο G (με πιθανότητα P=1.0). Μια αναζήτηση με βάση την απόσταση Hamming αναμένεται να εντοπίσει όλα τα σωστά σημεία πρόσδεσης αλλά θα εντοπίσει παράλληλα και πολλές θέσεις στις οποίες η αλληλουχία θα διαφέρει λίγο αλλά σε καίρια σημεία από το μοτίβο. Λέμε σε αυτήν την περίπτωση πως η μέθοδός μας έχει χαμηλή εξειδίκευση αφού εντοπίζει πολλά λανθασμένα σημεία πρόσδεσης. Πώς μπορούμε να ξεπεράσουμε το πρόβλημα των προτιμήσεων ανά θέση; Είδαμε ότι ο πίνακας προηγουμένως περιέχει τις συχνότητες εμφάνισης κάθε νουκλεοτιδίου ανά θέση στο μοτίβο. Θα μπορούσαμε να χρησιμοποιήσουμε τις τιμές του Πίνακα για να

βαθμολογήσουμε την ομοιότητα κάθε υπο-αλληλουχίας με το μοτίβο; Κάτι τέτοιο θα μπορούσε να γίνει με την παρακάτω διαδικασία:

```
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Αλγόριθμος :: PWM Αναζήτηση
Δήλωση Αλληλουχίας S μήκους n;
Δήλωση Πίνακα PWM[4,  l];
Δήλωση Πίνακα Score[n - l + 1]
Απαρίθμηση 1:
        Για θέση i = 1 έως i = n – l + 1 ανά 1;
                 Δημιούργησε την υποαλληλουχία s ← S[i : i + l - 1];
                #
                 μήκους = l
                Απαρίθμηση 2: Για κάθε θέση j = 1 έως j = l ανά 1;
                        Score[ i ]=Score[i ] +PWM[s [ j ], j ]
                Τέλος: Απαρίθμηση 2
Τέλος Απαρίθμηση 1
Απόδωσε αποτέλεσμα: Πίνακας Score Τερματισμός
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
```

Ο αλγόριθμος PWM Αναζήτηση (βλέπε 2.6) σαρώνει την αλληλουχία S σε επικαλυπτόμενα ανά 1 νουκλεοτίδιο, πλαίσια μήκους ίσου με το μήκος του μοτίβου l, δημιουργώντας σε κάθε επανάληψη μια υπο-αλληλουχία μήκους l. Στη συνέχεια υπολογίζει για κάθε μία από αυτές τις υποαλληλουχίες s, μια τιμή (score) ομοιότητας η οποία ισούται με το άθροισμα των συχνοτήτων εμφάνισης των νουκλεοτιδίων της s, όπως προκύπτει από τον Πίνακα 3.2. Έτσι π.χ. η υποαλληλουχία GAGTTACCCT θα έχει score

Score=P [G,1]+P [A, 2]+P [G,3]+P [T,4]+P [T,5]+P [A,6]+P [C,7]+P [C, 8] +P[C,9]+P [T, 10] Score=1+0+0.97+0+0.13+0. 23+0. 04+0 .38+0 .88+0 . 03=3. 66

Τι σημαίνει όμως η τιμή 3.66; Είναι υψηλή ή χαμηλή; Τι πληροφορία μπορούμε να πάρουμε από αυτήν; Θα μπορούσαμε να τη συγκρίνουμε με την τιμή που προκύπτει από τη συναινετική αλληλουχία (που είναι ίση με 8.25) αλλά και σε αυτήν την περίπτωση θα είχαμε το ίδιο πρόβλημα με τις αποστάσεις Hamming, θα μπορούσαμε δηλαδή να έχουμε ένα υψηλό score που θα προέκυπτε όμως από “απαγορευμένες” αλληλουχίες. Υπάρχει ωστόσο, ένα ακόμα πρόβλημα σε αυτήν την προσέγγιση. Λόγω των αυστηρών περιορισμών των τριών πρώτων και των τελευταίων δύο θέσεων, το μοτίβο μας συνολικά είναι εμπλουτισμένο σε βάσεις G

και C. Υπάρχει μια σαφής προτίμηση για βάσεις A και T στις μεσαίες θέσεις όμως αυτή δεν είναι τόσο ισχυρή όσο στα άκρα. Αυτό οδηγεί σε μια υπερεκπροσώπηση των βάσεων G+C μέσα στο μοτίβο και πράγματι, το GC περιεχόμενο των 104 σημείων πρόσδεσης του Πίνακα 3.1 είναι ίσο με GC=60.1%. Μια τέτοια τάση σημαίνει ότι αλληλουχίες που είναι πλούσιες σε βάσεις G και C θα τείνουν πιο εύκολα να δίνουν υψηλά PWM-score απλώς και μόνο τυχαία, χωρίς να σημαίνει πως περιέχουν πραγματικά σημεία πρόσδεσης. Αντίθετα, πραγματικά σημεία πρόσδεσης από περιοχές με χαμηλό GC περιεχόμενο θα έχουν μειωμένη πιθανότητα να σημειώσουν υψηλό score απλώς και μόνο λόγω της σύστασής τους.

Πώς μπορούμε να ξεπεράσουμε αυτό το πρόβλημα; Στη συνέχεια θα δούμε πως μπορούμε να βελτιώσουμε τη στρατηγική των Πινάκων PWM λαμβάνοντα υπόψη γενικότερα χαρακτηριστικά νουκλεοτιδικής σύστασης υποβάθρου. [14]

2.7.2 Πίνακες Βαθμονόμησης ανά θέση

Αν φανταστούμε πως αναζητούμε ένα ολιγονουκλεοτίδιο (για την ακρίβεια μια οικογένεια ολιγονουκλεοτιδίων που αντιστοιχεί στο μοτίβο) είναι λογικό να περιμένουμε πως αυτό θα συμβεί πιο εύκολα σε περιοχές του γονιδιώματος όπου η νουκλεοτιδική σύσταση είναι παρόμοια με τη σύσταση του μοτίβου, απ' ό,τι σε άλλες.

Μοτίβο NF-κB ***(P)***

Νουκλεοτίδιο	1	2	3	4	5	6	7	8	9	10
A	0.00	0.00	0.03	0.76	0.49	0.23	0.01	0.01	0.10	0.00
C	0.00	0.00	0.00	0.00	0.37	0.03	0.04	0.38	0.88	0.97
G	1.00	1.00	0.97	0.24	0.01	0.05	0.07	0.00	0.00	0.00
T	0.00	0.00	0.00	0.00	0.13	0.69	0.88	0.61	0.02	0.03

Πίνακας Υποβάθρου ***(Q)***

Νουκλεοτίδιο	1	2	3	4	5	6	7	8	9	10
A	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16
C	0.27	0.27	0.27	0.27	0.27	0.27	0.27	0.27	0.27	0.27
G	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33
T	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.23

$$R = \log_2\left(P_{i,j} / Q_{i,j}\right)$$

Νουκλεοτίδιο	1	2	3	4	5	6	7	8	9	10
A	-7.3	-7.3	-2.4	2.2	1.6	0.5	-3.9	-3.9	-0.7	-7.3
C	-8.1	-8.1	-8.1	-8.1	0.5	-3.1	-2.7	0.5	1.7	1.8
G	1.6	1.6	1.6	-0.5	-4.9	-2.7	-2.2	-8.4	-8.4	-8.4
T	-7.8	-7.8	-7.8	-7.8	-0.8	1.6	1.9	1.4	-3.5	-2.9

Position-Specific Scoring Matrix, PSSM

Στην περίπτωση ενός μοτίβου δεν έχουμε μόνο προτιμήσεις στις συχνότητες εμφάνισης δι-, τρι- ή ν- νουκλεοτιδίων αλλά έναν πίνακα πιθανοτήτων ανά θέση

που περιγράφει πλήρως τις πιθανότητες διαδοχής βάσεων μέσα σε ένα 10-νουκλεοτίδιο (στο παράδειγμα του NF-κB). Αυτό δε σημαίνει πως πρέπει να παραβλέψουμε χαρακτηριστικά που σχετίζονται με τη σύσταση της άγνωστης αλληλουχίας.

Γνωρίζουμε ότι το μοτίβο πρόσδεσης του NF-κB εμπεριέχεται στον πίνακα πιο πάνω. Θα ονομάσουμε αυτόν τον πίνακα, Πίνακα Μοτίβου P. Αυτό που πρέπει να κάνουμε είναι να δημιουργήσουμε έναν αντίστοιχο πίνακα PWM που θα αντιστοιχεί σε τυχαίες αλληλουχίες. Για την ακριβεία είναι καλύτερο να δημιουργήσουμε έναν PWM που θα αντιστοιχεί σε τυχαίες αλληλουχίες που όμως θα έχουν συνολικά την ίδια νουκλεοτιδική σύσταση με το μοτίβο του NF-κB. Ο PWM αυτός, που θα τον ονομάσουμε Πίνακα Υποβάθρου Q, είναι ένας Πίνακας ίδιων διαστάσεων με αυτόν του μοτίβου, στον οποίον οι συχνότητες εμφάνισης σε κάθε θέση είναι ίσες με τις συνολικές συχνότητες εμφάνισης νουκλεοτιδίων στο μοτίβο. Ο πίνακας υποβάθρου διατηρεί έτσι μόνο την πληροφορία της συνολικής σύστασης χωρίς να κρατά κανένα από τα χαρακτηριστικά προτίμησης θέσεων του μοτίβου. Τα “βάρη” για κάθε θέση είναι τα ίδια2 . Στη συνέχεια θα συγκρίνουμε τους δύο πίνακες P, Q με τον τρόπο που φαίνεται στην εικόνα με τους τρεις πίνακες, διαιρώντας κάθε στοιχείο του P με το αντίστοιχο του Q και παίρνοντας το δυαδικό λογάριθμο του λόγου Pi,j/Qi,j. Η διαδικασία αυτή δημιουργεί έναν νέο πίνακα που ονομάζουμε Πίνακα Βαθμονόμησης ανα Θέση (Position-specific scoring Matrix, ή για συντομία PSSM). Ο πίνακας PSSM έχει κάποια ιδιαίτερα χαρακτηριστικά. Το βασικότερο από αυτά είναι ότι ενσωματώνει την πληροφορία για τη σύσταση του μοτίβου με έναν τρόπο που είναι ειδικός ανά θέση (από την ιδιότητά του αυτή παίρνει και το όνομά του). Στον PSSM οι διαφορές μεταξύ των συχνοτήτων εμφάνισης του PWM μεγεθύνονται ή συρρικνώνονται ανάλογα με τη συνολική σύσταση της αλληλουχίας του υποβάθρου. Έτσι ένας PSSM πίνακας μπορεί να διακρίνει καλύτερα τα σημεία πρόσδεσης ανεξάρτητα από τη σύσταση της υποκείμενης αλληλουχίας. [15]

3. Συνδυαστικό ταίριασμα μοτίβου

Στην προηγούμενη ενότητα μελετήσαμε το πρόβλημα της Εύρεσης Μοτίβου, το οποίο αφορά την εύρεση κάποιου μοτίβου που εμφανίζεται πολύ συχνά σε ένα δείγμα DNA. Δεν αναζητάμε κάποιο συγκεκριμένο μοτίβο, αντιθέτως πρέπει να συμπεράνουμε με ποιο είναι το μοτίβο από το δείγμα. Το συνδυαστικό ταίριασμα μοτίβου, από την άλλη πλευρά ψάχνει για ακριβείς ή προσεγγιστικές εμφανίσεις ενός δεδομένου μοτίβου σε μακροσκελές κείμενο. Παρόλο που το ταίριασμα μοτίβου είναι απλούστερο από την εύρεση μοτίβου αφού γνωρίζουμε τι ψάχνουμε,

το μεγάλο μέγεθος των γονιδιωμάτων δυσκολεύει το πρόβλημα στην πράξη. Σε αυτήν την ενότητα θα αναπτύξουμε αρκετές μεθόδους που καθιστούν πρακτικό το ταίριασμα μοτίβου σε μεγάλες συμβολοδειρές. Ένα αντικείμενο που θα επαναλαμβάνουμε σε αυτήν την ενότητα είναι ο τρόπος οργάνωσης των δεδομένων σε αποδοτικές δομές δεδομένων. Οι ταξινομημένες λίστες αποτελούν έναν από τους πολλούς τύπους των δομών δεδομένων.

3.1 Εύρεση επαναλήψεων

Πολλές γενετικές ασθένειες συνδέονται με αφαιρέσεις, επαναλήψεις, και αναδιατάξεις μεγάλων χρωμοσωμικών περιοχών. Για παράδειγμα, το σύνδρομο DiGeorge που προκαλεί συχνά εξασθένηση του ανοσοποιητικού συστήματος και καρδιακές ανωμαλίες, σχετίζεται με μια μεγάλη αφαίρεση 3 μεταβάσεων στο ανθρώπινο χρωμόσωμα 22. Μια αφαίρεση τέτοιου μεγέθους είναι πιθανό ότι θα διαγράψει σημαντικά γονίδια και θα προκαλέσει κάποια ασθένεια. Σε τέτοιες δραματικές αλλαγές της γονιδιωματικής αρχιτεκτονικής ένα ζεύγος αλληλουχιών με μεγάλες ομοιότητες εμφανίζεται συχνά εκατέρωθεν του τμήματος που έχει αφαιρεθεί. Οι αλληλουχίες αυτές σχηματίζουν μια επανάληψη στο DNA και η εύρεση όλων των επαναλήψεων στα γονιδιώματα έχει μεγάλη σημασία. Οι επαναλήψεις στο DNA κρύβουν πολλά εξελικτικά μυστικά. Ο μεγάλος αριθμός επαναλήψεων σε πολλά γονιδιώματα αποτελεί ένα εντυπωσιακό και ανεξήγητο φαινόμενο: οι επαναλήψεις καλύπτουν το 50% του ανθρώπινου γονιδιώματος. Η πιο απλή μέθοδος για να εντοπίσει κανείς *ακριβείς* επαναλήψεις είναι να κατασκευάσει έναν πίνακα που αποθηκεύει όλες τις θέσεις κάθε l – μέρους στη γονιδιωματική αλληλουχία του DNA. Ένας τέτοιος πίνακας θα είχε 4^l κλιμάκια , με κάθε κλιμάκιο να περιέχει έναν αριθμό θέσεων μεταξύ 0 και M, όπου M είναι η συχνότητα εμφάνισης του πιο συχνού l – μερούς στο γονιδίωμα. Ο μέσος αριθμός στοιχείων σε κάθε κλιμάκιο είναι ίσος με $n/4^l$ όπου n είναι το μήκος του γονιδιώματος. Σε πολλές εφαρμογές η παράμετρος l παίρνει τιμές από 10 μέχρι 13, άρα το μέγεθος του πίνακα δεν είναι τέτοιο που να τον κάνει δύσχρηστο. Παρόλο που αυτή η προσέγγιση της πινακοποίησης μας επιτρέπει να βρούμε γρήγορα όλες τις επαναλήψεις μήκους l, τέτοιες επαναλήψεις μικρού μήκους δεν παρουσιάζουν μεγάλο ενδιαφέρον για μέγιστες επαναλήψεις με μεγάλο μήκος, δηλαδή επαναλήψεις που δεν μπορούν να επεκταθούν προς τα αριστερά ή προς τα δεξιά. Για να εντοπίσουμε μέγιστες επαναλήψεις με μήκος μεγαλύτερο από κάποια προκαθορισμένη παράμετρο L, πρέπει να επεκτείνουμε όλες τις ακριβείς επαναλήψεις μήκους l προς τα αριστερά ή προς τα δεξιά για να διαπιστώσει αν είναι ενσωματωμένες σε επανάληψη με μήκος μεγαλύτερο από L. Αφού ισχύει συνήθως $l << L$, ο αριθμός των επαναλήψεων που πρέπει να εντοπιστούν.

Επομένως, σε αυτή τη μέθοδο εύρεσης επαναλήψεων, ο περισσότερος χρόνος ξοδεύεται σε άσκοπες επεκτάσεις μικρών επαναλήψεων. Ο δημοφιλής αλγόριθμος REPuter αποφεύγει το συγκεκριμένο εμπόδιο με χρήση του επιθεματικού δένδρου, μίας δομής δεδομένων. [16]

3.2 REPuter αλγόριθμος

REPuter είναι ένα εργαλείο λογισμικού το οποίο μπορεί να υπολογίζει τις ακριβείς επαναλήψεις και τα παλίνδρομα σε ολόκληρο το γονιδίωμα πολύ αποτελεσματικά. Το εργαλείο αυτό μπορεί κάποιος να το βρει μέσω του Bielefeld Bioinformatics Server. Οι επιστήμονες της Πληροφορικής έχουν αναπτύξει μεθόδους για να εντοπίζουν επαναλαμβανόμενα κομμάτια/μέρη διαφορετικών ειδών. Έχουν αναπτυχθεί αρκετά εργαλεία λογισμικού τα οποία εντοπίζουν επαναλαμβανόμενα μέρη όπως είναι το Devereux, το Agarwal and States και το Rivals. Ωστόσο, με τις γνώσεις, όλα τα διαθέσιμα εργαλεία λογισμικού έχουν αυστηρά όρια για το μέγιστο μήκος της ακολουθίας εισόδου που επιτρέπουν να επεξεργαστούν. Για παράδειγμα, η επαναληψιμότητα της GCG επιτρέπει μόνο είσοδο μήκους μέχρι 350.000 βάσεις. Έχει αναπτυχθεί ένα εργαλείο το REPuter το οποίο επιτρέπει τον προσδιορισμό όλων των ακριβών επαναλαμβανόμενων συμβολοσειρών που περιέχονται στο com- plete γονιδίωμα. Ακριβής επαναλήψεις είναι μόνο ένα μικρό μέρος όλων των επαναλήψεων του βιολογικού ενδιαφέροντος. Το REPuter μπορεί επίσης να χρησιμοποιηθεί ως μια γρήγορη υπορουτίνα για τα προγράμματα που ανιχνεύουν τις ακριβής επαναλήψεις που συνήθως αποτελούν τις βασικές μονάδες της κατά προσέγγισης επαναλήψεων όπως για παράδειγμα το Leung. Ο αλγόριθμος είναι γραμμικός και ως προς την είσοδο αλλά και ως προς την έξοδο. Η μέθοδος που χρησιμοποιείται είναι ασυμπτωτικά βέλτιστη. Το κύριο πλεονέκτημα σε σχέση με τα προηγούμενα προγράμματα είναι η μείωση του χρόνου και του χώρου πολυπλοκότητας των εργαλείων. Έτσι κάποιος μπορεί να χειριστεί πολύ περισσότερες ακολουθίες εισόδου. Η τρέχουσα έκδοση είναι σε θέση να επεξεργάζεται ακολουθίες εισόδου που αποτελούνται από έως και 67 εκατομμύρια βάσεις. Για παράδειγμα, για να υπολογίσουμε όλες τις μέγιστες επαναλήψεις μήκους τουλάχιστον 20 που περιέχεται στο cerevisiae γονιδίωμα S. (12 147 818 βάσεις) διαρκεί περίπου 46 seconds σε ένα Pentium 350 MHz υπολογιστή, χρησιμοποιώντας 160 Mbyte χώρου. Ο REPuter αποτελείται από δύο προγράμματα, μια μηχανή αναζήτησης και ένα οπτικό αποτέλεσμα. Η μηχανή αναζήτησης επεξεργάζεται μία ακοουθία DNA δίνοντας ο χρήστης την ακολουθία σε Fasta-format και επιστρέφει μια εκπροσώπηση όλων των μέγιστων

επαναλήψεων σε ένα απλό ASCII αρχείο. Το άλλο κομμάτι, το οπτικό επεξεργάζεται την έξοδο της μηχανής αναζήτησης και παράγει μια επισκόπηση του αριθμού, του μήκους, και τη θέση των επαναλαμβανόμενων συμβολοσειρών. Για την διευκόλυνση του χρήστη έχουνε υπολογίσει από πριν τις μέγιστες επαναλήψεις για ορισμένα γονιδιώματα. Η μηχανή αναζήτησης μπορεί να μεταφορτωθεί ως εκτελέσιμο δυαδικό για διάφορες πλατφόρμες. [17]

3.3 Ακριβές ταίριασμα μοτίβου

Ένα συνήθες πρόβλημα στη βιοπληροφορική είναι η αναζήτηση μιας γνωστής αλληλουχίας σε μια βάση δεδομένων με αλληλουχίες. Με δεδομένες τη συμβολοσειρά μοτίβου **p** = $p_1 \ldots p_n$ και μια μεγαλύτερη συμβολοσειρά κειμένου **t** = $t_1 \ldots t_m$ το πρόβλημα του Ταιριάσματος Μοτίβου είναι η εύρεση όλων των εμφανίσεων του μοτίβου **p** στο κείμενο **t**.

- -
Με δεδομένα ένα μοτίβο και ένα κείμενο, βρείτε όλες τις εμφανίσεις του μοτίβου στο κείμενο.
Είσοδος: Το μοτίβο **p** = $p_1 \ldots p_n$ και το κείμενο **t** = $t_1 \ldots t_m$
Έξοδος: Όλες οι θέσεις $1 <= i <= m - n + 1$ έτσι ώστε η υποσυμβολοσειρά n γραμμάτων του **t** που αρχίζει στη θέση i να συμπίπτει με το **p**.
- -

Για παράδειγμα, αν **t** = ATGGTCGGT και **p** = GGT, τότε το αποτέλεσμα ενός αλγορίθμου που επιλύει το πρόβλημα του Ταιριάσματος Μοτίβου θα είναι οι θέσεις 3 και 7. Θα χρησιμοποιήσουμε τη σημειογραφία $t_i = t_i \ldots t_{i+n-1}$ για να συμβολίσουμε μια υποσυμβολοσειρά μήκους n από το **t** που αρχίζει στη θέση i. Αν $\mathbf{t_i = p}$, τότε έχουμε βρει μια εμφάνιση του μοτίβου στο κείμενο. Αν ελέγξουμε όλες τις δυνατές τιμές του i με αύξουσα σειρά, σαρώνουμε ουσιαστικά το **t** με κυλιόμενο παράθυρο μήκους n από τα αριστερά προς τα δεξιά, και σημειώνουμε τη θέση του παραθύρου κατά την εμφάνιση του μοτίβου **p**. Ένας αλγόριθμος ωμής βίας που λύνει το το πρόβλημα του Ταιριάσματος Μοτίβου εκτελεί αυτήν ακριβώς τη λειτουργία.

- -
PATTERNMATCHING(**p**, **t**)
1. n ← μήκος μοτίβου **p**
2. m ← μήκος κειμένου **t**
3. **for** i ← 0 **to** m – n + 1

4. **if t_i = p**
5. **return i**

- -

Σε κάθε θέση, η υπορουτίνα PATTERNMATCHING χρειάζεται μέχρι n πράξεις για να επαληθεύσει αν το μοτίβο **p** βρίσκεται στο παράθυρο ελέγχοντας αν ισχύουν οι σχέσεις $t_i = p_1$, $t_{i+1} = p_2$ κ.ο.κ. Για τυπικά στιγμιότυπα, ο αλγόριθμος καταναλώνει το μεγαλύτερο μέρος του χρόνου του για να ανακαλύψει ότι το μοτίβο δεν εμφανίζεται στη θέση i του κειμένου. Αυτός ο έλεγχος μπορεί να χρειαστεί μία μόνο πράξη, που δείχνει οριστικά ότι το **p** δεν εμφανίζεται στη θέση i του t ωστόσο, μπορεί να χρειαστούν μέχρι και * πράξεις για το συγκεκριμένο έλεγχο. Συνεπώς, ο χρόνος εκτέλεσης της χειρότερης περίπτωσης για τον αλγόριθμο PATTERNMATCHING εκτιμάται ότι είναι O(nm). Το σενάριο της χειρότερης περίπτωσης συμβαίνει όταν αναζητάμε το μοτίβο **p** = AAAAT στο κείμενο **t** = AA...AAA. Αν το m είναι ίσο με 1 εκατομμύριο, όχι μόνο απαιτεί η αναζήτηση 5 εκατομμύρια πράξεις, αλλά ολοκληρώνεται χωρίς έξοδο, το οποίο είναι λίγο απογοητευτικό. Για να υπολογίσουμε τον χρόνο του παραπάνω αλγορίθμου πρώτων, υπάρχει μεγάλη πιθανότητα ότι ο πρώτος έλεγχος θα είναι ασυμφωνία, και έτσι δεν χρειάζεται να ελέγξουμε
τα υπόλοιπα n - 1 γράμματα του **p**. Γενικά, η πιθανότητα να ταιριάζει το πρώτο γράμμα του μοτίβου είναι ίση με 1/A. Ομοίως η πιθανότητα να ταιριάζουν τα δύο πρώτα γράμματα με το κείμενο είναι ίση με 1/A^2, ενώ η πιθανότητα να ταιριάζει το πρώτο γράμμα και να μην ταιριάζει το δεύτερο γράμμα είναι ίση με A – 1/A^2. Η πιθανότητα ότι ο αλγόριθμος ταιριάζει ακριβώς j από n χαρακτήρες του **p** που αρχίζουν στο t_i είναι ίση με A – 1/A^i για τιμές του j από 1 μέχρι n – 1. Εφόσον ο αλγόριθμος αυτός ελαττώνεται γρήγορα καθώς το j αυξάνεται, μπορούμε να διαπιστώσουμε ότι μειώνεται πολύ η πιθανότητα να πραγματοποιήσει ο αλγόριθμος μεγάλες δοκιμές. Οι πράξεις που απαιτούνται για τον έλεγχο της παρουσίας του **p** είναι τάξης O(m) παρά τον απαισιόδοξο χρόνο εκτέλεσης O(nm) της χειρότερης περίπτωσης. Το 1973 ο * επινόησε μια μεγαλοφυή δομή δεδομένων που ονομάζεται *επιθεματικό δέντρο* και λύνει το πρόβλημα του Ταιριάσματος Μοτίβου σε γραμμικό χρόνο O(m) για οποιοδήποτε κείμενο και μοτίβο. Φαίνεται ότι το μέγεθος του μοτίβου δεν παίζει κανένα ρόλο όσον αφορά την πολυπλοκότητα του προβλήματος του Ταιριάσματος Μοτίβου. [18]

3.4 Δέντρα – Επιθεματικά δέντρα

- -

Το πρόβλημα του Ταιριάσματος Πολλών Μοτίβων:
Με δεδομένα ένα σύνολο μοτίβων και ένα κείμενο, βρίσουμε όλες τις εμφανίσεσι οποιουδήποτε από τα μοτίβα στο κείμενο.
Είσοδος: Το σύνολο k μοτίβων $\mathbf{p^1}$, $\mathbf{p^2}$, …, $\mathbf{p^k}$ και το κείμενο **t** = t1… tm.
Έξοδος: Όλες οι θέσεις 1 <= i <= m έτσι ώστε η υποσυμβολοσειρά του **t** που αρχίζει στη θέση i να συμπίπτει με το μοτίβο $\mathbf{p^j}$ για 1 <= j <= k.

- -

Βέβαια, το πρόβλημα του Ταιριάσματος Πολλών Μοτίβων για k μοτίβα μπορεί να αναχθεί σε k μεμονωμένα προβλήματα Ταιριάσματος Μοτίβου και να επιλυθεί σε χρόνο O(knm), όπου n είναι το μήκος του μεγαλύτερου από τα k μοτίβα, μετά από k εφαρμογές του αλγορίθμου PATTERNMATCHING (κανονικά ο χρόνος θα ήτανε O(Nm), όπου N είναι το συνολικό μήκος των μοτίβων $\mathbf{p^1}$, $\mathbf{p^2}$, …, $\mathbf{p^k}$). Αν αντικαταστήσουμε τον PATTERNMATCHING με αλγόριθμο γραμμικού χρόνου για το ταίριασμα μοτίβου, το πρόβλημα του Ταιριάσματος Πολλών Μοτίβων μπορεί να επιλυθεί σε χρόνο O(km). Όμως, υπάρχει ένας ταχύτερος τρόπος επίλυσης του προβλήματος σε χρόνο O(N + m), όπου N είναι το συνολικό μήκος των μοτίβων $\mathbf{p^1}$, $\mathbf{p^2}$, …, $\mathbf{p^k}$. Προτάθηκε η δομή δεδομένων του δέντρου με λέξεις - κλειδιά για την επίλυση του προβλήματος. Ένας περισσότερο τυπικός ορισμός είναι ο εξής: το δέντρο με λέξεις - κλειδιά για ένα σύνολο μοτίβων $\mathbf{p^1}$, $\mathbf{p^2}$, …, $\mathbf{p^k}$ είναι ένα δέντρο με ρίζα και ετικέτες που ικανοποιεί τις ακόλουθες συνθήκες, αν υποθέσουμε για λόγους απλότητας ότι κανένα μοτίβο του συνόλου δεν αποτελεί πρόθεμα κάποιου άλλου μοτίβου:

- κάθε ακμή του δέντρου έχει ως ετικέτα ένα γράμμα του αλφαβήτου,
- οποιεσδήποτε δύο εξερχόμενες ακμές της ίδιας κορυφής έχουν διαφορετικές ετικέτες,
- κάθε μοτίβο $\mathbf{p^i}$(1 <= i <= k) από το σύνολο των μοτίβων αποτελείται από γράμματα που βρίσκονται σε κάποια διαδρομή μεταξύ της ρίζας και ενός φύλλου.

Μπορούμε να κατασκευάσουμε το δέντρο με λέξεις – κλειδιά σε χρόνο O(N) επεκτείνοντας σταδιακά το δέντρο για τα πρώτα j μοτίβα στο δέντρο για τα j + 1 μοτίβα. Μπορούμε να χρησιμοποιήσουμε το δέντρο με λέξεις – κλειδιά για να βρούμε αν υπάρχει κάποιο μοτίβο στο σύνολο $\mathbf{p^1}$, $\mathbf{p^2}$, …, $\mathbf{p^k}$, το οποίο ταιριάζει με το κείμενο που αρχίζει σε μια σταθερή θέση i του κειμένου. Για να γίνει αυτό, πρέπει απλώς να διατρέξουμε το δέντρο με λέξεις – κλειδιά χρησιμοποιώντας τα γράμματα $t_i t_{i+1} t_{i+2}$… του κειμένου για να αποφασίσουμε πού θα μετακινηθούμε

σε κάθε βήμα. Η διαδικασία αυτή είναι ολοκληρωμένη σε ένα φύλλο, οπότε υπάρχει ταίριασμα με το μοτίβο που αναπαριστάται από το φύλλο, είτε διακόπτεται πριν φτάσει σε φύλλο, οπότε δεν υπάρχει ταίριασμα που αρχίζει στη θέση i. Αν το μήκος του μεγαλύτερου μοτίβου είναι ίσο με n, τότε το πρόβλημα του Ταιριάσματος Πολλών Μοτίβων μπορεί να επιλυθεί σε χρόνο O(N + nm) που απαιτείται για να κατασκευάσουμε το δέντρο με λέξεις – κλειδιά και να το χρησιμοποιήσουμε στη συνέχεια για να ψάξουμε το κείμενο. Ο αλγόριθμος Aho-Corasick μειώνει περισσότερο το χρόνο εκτέλεσης του προβλήματος του Ταιριάσματος Πολλών Μοτίβων σε O(N + m).

- -

Τα επιθεματικά δέντρα επιτρέπουν την προεπεξεργασία ενός κειμένου με τέτοιον τρόπο ώστε να μπορούμε να απαντήσουμε στο ερώτημα αν οποιοδήποτε μοτίβο μήκους n εμφανίζεται στο κείμενο ή όχι, σε χρόνο O(n) και ανεξάρτητα από το μέγεθος του κειμένου. Ένα επιθεματικό δέντρο για κείμενο μήκους m μπορεί να κατασκευαστεί σε χρόνο O(m). Γεγονός που οδηγεί σε ένα γραμμικό αλγόριθμο O(n + m) για το πρόβλημα του Ταιριάσματος Μοτίβου. Το επιθεματικό δέντρο για ένα κείμενο **t** = $t_1 \ldots t_m$ είναι δέντρο με ρίζα και ετικέτες, το οποίο έχει m φύλλα και ικανοποιεί τις ακόλουθες συνθήκες:

- κάθε ακμή έχει μια υποσυμβολοσειρά του κειμένου ως ετικέτα,
- κάθε εσωτερική κορυφή έχει τουλάχιστον δύο απογόνους,
- οποιεσδήποτε δύο εξερχόμενες ακμές της ίδιας κορυφής αρχίζουν με διαφορετικό γράμμα,
- όλα τα επιθέματα του κειμένου **t** αποτελούνται από γράμματα που βρίσκονται σε μια διαδρομή μεταξύ ρίζας και κάποιου φύλλου.

Τα επιθεματικά δέντρα οδηγούν αμέσως σε ένα γρήγορο αλγόριθμο για το ταίριασμα μοτίβου. Ο αλγόριθμος SUFFIXTREEPATTERNMATCHING αναζητάει ακριβείς εμφανίσεις του **p** στο **t**.

- -

SUFFIXTREEPATTERNMATCHING(**p**, **t**)
1. Κατασκευή του επιθεματικού δέντρου για το κείμενο **t**
2. Νημάτωση του μοτίβου **p** μέσω του επιθεματικού δέντρου
3. **if** η νημάτωση είναι πλήρης
4. **output** θέσεις κάθε φύλλου **p –** ταιριάσματος στο δέντρο
5. **else**
6. **output ‘**το μοτίβο δεν εμφανίζεται στο κείμενο**’**

- -

Νημάτωση: Νημάτωση ενός μοτίβου **p** μέσω του επιθεματικού δέντρου Τ ως το ταίριασμα χαρακτήρων από το **p** κατά μήκος της μοναδικής διαδρομής στο Τ μέχρι να ταιριάξουν όλοι οι χαρακτήρες του **p**, ή να μην υπάρχουν άλλα δυνατά ταιριάσματα. [19]

3.5 Ευρετικοί αλγόριθμοι αναζήτησης ομοιοτήτων

Ο παραπάνω αλγόριθμος επιθεματικού δέντρου είναι γρήγορος, αλλά μπορεί να βρει μόνο ακριβείς και όχι προσεγγιστικές εμφανίσεις ενός γονιδίου σε μια βάση δεδομένων. Πολλές ευρετικές μέθοδοι για τη γρήγορη αναζήτηση σε βάση δεδομένων στη μοριακή βιολογία χρησιμοποιούν την ιδέα *φιλτραρίσματος*. Το φιλτράρισμα βασίζεται στην παρατήρηση ότι μια καλή στοίχιση περιλαμβάνει συνήθως μικρά τμήματα με απόλυτη ή μεγάλη ομοιότητα. Άρα,, μπορεί κάποιος να αναζητήσει μικρά ακριβή ταιριάσματα, χρησιμοποιώντας έναν πίνακα κατακερματισμού ή ένα επιθεματικό δέντρο, και να χρησιμοποιήσει τα μικρά ταιριάσματα ως φύτρα για επιπλέον ανάλυση. Ένας αλγόριθμος που προτάθηκε για το πρόβλημα του ταιριάσματος μοτίβου ήτανε βασισμένος στην παρατήρηση ότι οι συμβολοσειρές που διαφέρουν κατά μία ασυμφωνία πρέπει να ταιριάζουν απόλυτα στο πρώτο ή το δεύτερο μισό. Οι βιολόγοι απεικονίζουν συχνά τις ομοιότητες μεταξύ δύο αλληλουχιών με τη μορφή των *μητρών κουκκίδων*. [20]

3.6 Προσεγγιστικό ταίριασμα μοτίβου

- -

Βρίσκουμε όλες τις προσεγγιστικές εμφανίσεις ενός μοτίβου σε κάποιο κείμενο.
Είσοδος: Το μοτίβο **p** = $p_1p_2\ldots p_n$, το κείμενο **t** = $t_1t_2\ldots t_m$ και η παράμετρος k, δηλαδή ο μέγιστος αριθμός των ασυμφωνιών.
Έξοδος: Όλες οι θέσεις $1 <= i <= m – n + 1$ έτσι ώστε τα $t_it_{i+1}\ldots t_{i+n-1}$ και $p_1p_2\ldots p_n$ να έχουν το πολύ k ασυμφωνίες.
- -

Η εύρεση των προσεγγιστικών ταιριασμάτων ενός μοτίβου σε κάποιο κείμενο αποτελεί σημαντικό πρόβλημα στην υπολογιστική μοριακή βιολογία. Ο απλοϊκός αλγόριθμος ωμής βίας για το προσεγγιστικό ταίριασμα μοτίβου εκτελείται σε χρόνο O(nm). Ο παρακάτω αλγόριθμος θα παραγάγει ως έξοδο όλες τις θέσεις εμφάνισης του **p** στο **t** με όχι περισσότερες από k ασυμφωνίες. Ανακαλύφθηκε ένας αλγόριθμος για το προσεγγιστικό ταίριασμα συμβολοσειράς σε χρόνο

εκτέλεσης O(km) για τη χειρότερη περίπτωση. Παρόλο που ο αλγόριθμος αυτός παράγει την καλύτερη γνωστή απόδοση για τη χειρότερη περίπτωση, δεν είναι απαραίτητα ο καλύτερος στην πράξη. Αρκετές μέθοδοι με βάση το φιλτράρισμα έχουν καλύτερους χρόνους εκτέλεσης στην πράξη. Το πρόβλημα του Ταιριάσματος Ερωτήματος γενικεύει περισσότερο το πρόβλημα του Προσεγγιστικού Ταιριάσματος Μοτίβου. Η διαφορά ανάμεσα στο πρόβλημα του Προσεγγιστικού Ταιριάσματος Μοτίβου και το πρόβλημα του Ταιριάσματος Ερωτήματος είναι ότι το πρώτο συγκρίνει ολόκληρο το μοτίβο μήκους n με το κείμενο, ενώ το δεύτερο συγκρίνει όλες τις υποσυμβολοσειρές μήκους n της αρχικής αλληλουχίας με το κείμενο. Η χρήση αλγορίθμων φιλτραρίσματος για το προσεγγιστικό ταίριασμα ερωτήματος περιλαμβάνει μια διαδικασία δύο σταδίων. Το πρώτο στάδιο προεπιλέγει ένα σύνολο θέσεων στο κείμενο που είναι ενδεχομένως όμοιες με το ερώτημα. Το δεύτερο στάδιο επαληθεύει κάθε πιθανή θέση, απορρίπτοντας πιθανές θέσεις με περισσότερες από k ασυμφωνίες. Αν ο αριθμός των πιθανώς ταιριασμάτων είναι μικρός και η επαλήθευση των πιθανών ταιριασμάτων δεν είναι πολύ αργή, η μέθοδος παράγει ένα γρήγορο αλγόριθμο ταιριάσματος ερωτήματος για τις περισσότερες εισόδους που προκύπτουν στην πράξη. [21]

4. Τέσσερα προβλήματα (1)

4.1.1 Πληροφοριακό περιεχόμενο μοτίβου

Χρειαζόμαστε όμως έναν αντικειμενικό τρόπο για να αξιολογήσουμε τα αποτελέσματα της αναζήτησής μας. Στη συγκεκριμένη ενότητα θα δούμε πώς μπορούμε να υπολογίσουμε μια ποσότητα που ονομάζεται πληροφοριακό περιεχόμενο (information content) ενός μοτίβου και στη συνέχεια θα εφαρμόσουμε μια μέθοδο υπολογισμού αυτής της ποσότητας στα σημεία πρόσδεσης που προκύπτουν τόσο από την αναζήτηση με τον PWM όσο και με τον PSSM πίνακα. Πριν περιγράψουμε τη μέθοδο αυτή, όμως θα πρέπει να ορίσουμε μαθηματικά την έννοια της πληροφορίας γενικά και την πληροφορία μιας αλληλουχίας ειδικότερα.

Μαθηματικό Ιντερμέδιο Ι. Εντροπία Shannon και Πληροφοριακό Περιεχόμενο

Η ερευνήτρια Ν είναι στη μέση ενός επίπονου πειράματος. Το ίδιο και ο συναδελφός της Β. Οι δυο τους ξέρουν πως θα βρίσκονται στο θάλαμο μικροσκοπίας για τις επόμενες ώρες παρατηρώντας τομές από ιστούς ποντικιών. Αποφασίζουν να πολεμήσουν τη βαρεμάρα τους με ένα απλό παιχνίδι. Ο Β σκέφτεται έναν ακέραιο αριθμό x από το 1 έως το 1000 και η Ν προσπαθεί να το μαντέψει κάνοντας τις λιγότερες δυνατές ερωτήσεις στον Β. Μετά από μερικά παιχνίδια η Ν καταλήγει σε μια έξυπνη στρατηγική που βασίζεται στην απλή διαδικασία: Κάθε φορά ρωτάει τον Β αν ο αριθμός είναι μεγαλύτερος από το 500. Αν ο Β της απαντήσει “ναι” τότε υπογίζει το μέσο της απόστασης μεταξύ 500 και 1000 και ρωτάει τον Β αν ο αριθμός είναι μεγαλύτερος από το 750, ενώ αν της απαντήσει “όχι” μοιράζει στη μέση την απόσταση μεταξύ 1 και 500 και ρωτάει τον Β αν ο αριθμός x είναι μεγαλύτερος από το 250. Συνεχίζει με αυτόν τον τρόπο διαιρώντας διαρκώς δια δύο το διάστημα μέσα στο οποίο βρίσκεται ο αριθμός. Η στρατηγική της μπορεί να περιγραφεί με τον παρακάτω αναδρομικό αλγόριθμο:

- -

Αλγόριθμος :: Μάντεψε Χ
Δήλωση Πλήθους αριθμών Ν;
Μυστικός αριθμός Χ;
Υπολογισμός g=1/2 * N;
Επανάληψη Α:
    Έλεγξε: Αν Χ > g;
        Τότε g = g + ½ * g; Επιστροφή στο Α
    Έλεγξε: Αν Χ < g;
        Τότε g = g - 1/2 * g; Επιστροφή στο Α
    Έλεγξε: Αν Χ = g;
        Τότε “Το βρήκες”; Πήγαινε στον Τερματισμό
Τερματισμός

- -

Μπορείτε να σκεφτείτε γιατί η λογική που ακολουθεί η Ν είναι πολύ καλύτερη από το να προσπαθεί να μαντέψει συνεχώς στην τύχη; Ας αναλογιστούμε αρχικά τι πιθανότητα έχει η Ν να βρει τον αριθμό κάνοντας τυχαίες προβλέψεις, του τύπου “Είναι ο αριθμός σου ο 671;” Δεδομένου ότι ο αριθμός είναι μεταξύ 1 και 1000, η πιθανότητα να το μαντέψει με την πρώτη είναι 1/1000. Αν δεν το βρει με την πρώτη, τότε στη δεύτερη επιλογή της έχει να διαλέξει μεταξύ 999 αριθμών και συνεπώς η πιθανότητα είναι τώρα 1/999. Με τον ίδιο ρυθμό, αν προσπαθήσει να συνεχίσει μαντεύοντας θα πρέπει να κάνει 900 ερωτήσεις για να φτάσει να έχει 1/100 πιθανότητες για να βρει τον αριθμό. Πραγματικά μοιάζει σαν να ψάχνει

βελόνα στα άχυρα! Αυτό που συμβαίνει είναι ότι ουσιαστικά με κάθε ερώτηση που κάνει δεν αποκλείει παρά μόνο έναν αριθμό από το σύνολο των 1000. Η πληροφορία δηλαδή που αποκομίζει από κάθε απάντηση του Β. σε αυτές τις ερωτήσεις είναι ελάχιστη. Αντίθετα, με τη στρατηγική της διαίρεσης του διαστήματος μεταξύ 1 και 1000 σε ολοένα μικρότερα, η πληροφορία που έχει κάθε απάντηση του Β είναι πολύ μεγαλύτερη. Μετά την πρώτη κιόλας ερώτηση έχει πιθανότητα 1/500 να βρει τον αριθμό, μετά τη δεύτερη έχει 1/250 κ.ο.κ.

Πώς συνδέεται η πληροφορία που παίρνει η Ν. σε κάθε ερώτηση με την πιθανότητα να μαντέψει σωστά; Με κάθε ερώτηση η Ν μπορεί πρακτικά να χωρίσει το σύνολο των αριθμών από το 1 έως το 1000 σε δύο υποσύνολα. Αυτό που περιέχει τον x και αυτό που δεν τον περιέχει. Γιατί όμως είναι προτιμότερο για την Ν. να κάνει ερωτήσεις που χωρίζουν το διάστημα σε δύο ίσα μέρη; Αν υποθέσουμε ότι η Ν ρωτήσει τον Β αν ο αριθμός είναι μεγαλύτερος από το 250, χωρίζοντας έτσι το διάστημα σε ¼ και ¾, τότε υπάρχουν τα εξής ενδεχόμενα. Αν ο Β της πει πως δεν είναι, έχει αυξήσει την αρχική πιθανότητα απο 1/1000 σε 1/250, αν όμως της πει πως είναι τότε η πιθανότητα έχει αυξηθεί μόνο από 1/1000 στο 1/750. Την ίδια στιγμή, η πιθανότητα της δεύτερης περίπτωσης (να της απαντήσει δηλαδή ο Β “ναι”) είναι 3 φορές μεγαλύτερη από το να της απαντήσει “όχι”. Συνεχίζοντας με αυτόν τον τρόπο (διαιρώντας το διάστημα σε λόγο 3/1) θα χρειαστεί στατιστικά μεγαλύτερο αριθμό ερωτήσεων για να βρει τον αριθμό x. Δεδομένου ότι η αρχική επιλογή του x είναι εντελώς τυχαία, ο καλύτερος τρόπος είναι η διαρκής διαίρεση του διαστήματος σε δύο υποσύνολα ίσου μεγέθους κι αυτό γιατί η μέγιστη πληροφορία αποκομίζεται όταν τα δύο υποσύνολα που θα δημιουργήσει η κάθε ερώτηση έχουν την ίδια πιθανότητα, καθώς η συνολική αβεβαιότητα μειώνεται περισσότερο.

Το 1948, ο μαθηματικός και μηχανικός Claude Shannon, αναλογιζόμενος ένα ανάλογο πρόβλημα όρισε μαθηματικά αυτήν την αβεβαιότητα με μια ποσότητα που ονομάστηκε προς τιμή του Εντροπία Shannon (H) (C. E. Shannon 1948). Ορίζουμε Εντροπία Shannon ενός συστήματος με Ν ενδεχόμενα καθένα από τα οποία έχει πιθανότητα p την εξής ποσότητα:

$$H = -\sum_{i=1}^{N} p_i \log\left(p_i\right)$$

Γιατί “Εντροπία”; Γιατί με τον ίδιο τρόπο που η θερμοδυναμική εντροπία αποτελεί ένα μέτρο της αταξίας ενός φυσικού συστήματος, η Εντροπία Shannon αποτελεί ένα μέτρο της αβεβαιότητας που υπάρχει σε ένα σύστημα μετάδοσης μηνύματος δεδομένων των πιθανοτήτων παρατήρησης του συνόλου των ενδεχομένων του. Ας δούμε ποια είναι η Εντροπία Shannon για το πρόβλημα της Ν. Αν η Ν δοκιμάσει να χωρίζει το διάστημα σε ίσα τμήματα, τότε με κάθε ερώτηση η πιθανότητα του x να βρίσκεται σε κάθε ένα από αυτά είναι p=½. Ισχύει ότι:

$$H=-\sum_{i=1}^{2} p_i \log\left(p_i\right)=-\left(0.5\log\left(0.5\right)+0.5\log\left(0.5\right)\right)=0.6932$$

αν αντίθετα χωρίσει το διάστημα σε μήκη με σχέση ¼ και ¾ τότε η 3.1 δίνει:

$$H=-\sum_{i=1}^{2} p_i \log\left(p_i\right)=-\left(0.25\log\left(0.25\right)+0.75\log\left(0.75\right)\right)=0.5623$$

Βλέπουμε δηλαδή ότι η εντροπία μιας επιλογής με ίσες πιθανότητες είναι μεγαλύτερη από αυτήν που προκύπτει όταν για ίδιο αριθμό ενδεχομένων υπάρχουν διαφορετικές πιθανότητες για το καθένα.

Ισχύει γενικότερα ότι:
Ένα σύστημα με N ενδεχόμενα επιτυγχάνει τη μέγιστη Εντροπία όταν η πιθανότητα κάθε ενδεχομένου είναι ίση με 1/N και αυτή είναι ίση με -Nlog(N-1).

Τι σημαίνει όμως για ένα οποιοδήποτε σύστημα ή μέγιστη Εντροπία και ποια είναι η σχέση της με το πληροφοριακό περιεχόμενο; Ισχύει ότι στην κατάσταση μέγιστης εντροπίας το σύστημα έχει τη μέγιστη δυνατότητα μετάδοσης ενός μηνύματος. Όσο μεγαλύτερη είναι δηλαδή η μεταβολή της Εντροπίας, τόσο μεγαλύτερη η πληροφορία που αποκομίζουμε. [22]

4.1.2 Εντροπία Μοτίβων

Καθώς η ώρα στο θάλαμο μικροσκοπίας περνάει, η Ν και ο Β αποφασίζουν να κάνουν το παιχνίδι τους λίγο πιο πολύπλοκο. Έτσι ο Β προσκαλεί τώρα την Ν να

μαντέψει όχι έναν αριθμό από το 1 έως το 1000 αλλά ένα δεκανουκλεοτίδιο (τα μεγάλα σε διάρκεια πειράματα έχουν συχνά παρενέργειες!) Η Ν. είναι αρκετά συνετή ώστε να μην μπει καν στον κόπο να προσπαθήσει να μαντέψει ένα από τα 410 πιθανά ενδεχόμενα, ωστόσο αναλογίζεται ποια θα ήταν η ανάλογη στρατηγική για ένα τόσο δύσκολο πρόβλημα. Σκέφτεται πως θα ήταν καλύτερο να προσπαθήσει να μαντέψει το κάθε νουκλεοτίδιο χωριστά. Υποθέτοντας πως ο Β έχει επιλέξει τελείως τυχαία, κάθε νουκλεοτίδιο έχει πιθανότητα 0.25 να καταλάβει καθεμία από τις 10 θέσεις. Η αβεβαιότητα για κάθε θέση δίνεται από την Εντροπία Shannon της εξίσωσης 3.1 και είναι ίση με:

$$H=-\sum_{i=1}^{4} 0.25 \log_2(0.25)=2$$

Στην προκειμένη περίπτωση χρησιμοποιήσαμε το δυαδικό λογάριθμο αντί για το φυσικό, για λόγους συμμετρίας του προβλήματος. Μπορούμε να φανταστούμε την Ν. να κάνει ερωτήσεις στο β. για να μαντέψει καθένα από τα νουκλεοτίδια. Επειδή αρκούν δύο ερωτήσεις για να βρει το σωστό διαλέγουμε τη βάση του λογαρίθμου έτσι ώστε η Εντροπία να ταυτίζεται με τον αριθμό των ερωτήσεων που απαιτούνται (κάτι που είναι πολύ κοντά στο φυσικό νόημά της). Η συνολική αβεβαιότητα είναι ίση με 10 x 2=20 κάτι που σημαίνει πως στην Ν. αρκούν 20 ερωτήσεις για να βρει το σωστό δεκανουκλεοτίδιο!

- -

***Ερώτηση**: Ποιες είναι οι δύο ερωτήσεις που επαρκούν στην Ν. για να μαντέψει ένα από τα τέσσερα νουκλεοτίδια;*

- -

Ας αναλογιστούμε τι σημαίνει αυτό για το παράδειγμά μας. Είδαμε πως η αρχική αβεβαιότητα για κάθε θέση σε ένα ολιγονουκλεοτίδιο είναι 2. Με αντίστοιχο τρόπο η αρχική τιμή αβεβαιότητας για ένα σύστημα όπου υπάρχουν Ν ενδεχόμενα είναι η μέγιστη τιμή της Εντροπίας Shannon που δίνεται από την ακόλουθη εξίσωση:

$$H_{max}=-\sum_{i=1}^{N} \frac{1}{N} \log_2\left(\frac{1}{N}\right)$$

Η τιμή Hmax αντιστοιχεί στη περίπτωση που όλα τα ενδεχόμενα είναι ισοπίθανα. Ωστόσο στην περίπτωση ενός μοτίβου αλληλουχίας, αυτό δε συμβαίνει. Κάποια ενδεχόμενα (κατάλοιπα) είναι πολύ πιο πιθανά από καποια άλλα σε συγκεκριμένες θέσεις. Μπορούμε λοιπόν να πούμε ότι σε κάθε θέση υπάρχει ένα συγκεκριμένο πληροφοριακό περιεχόμενο που μπορεί να υπολογιστεί ως:

$$I_{position} = H_{max} - \sum_{i=1}^{N} P_i \log_2(P_i)$$

Για την περίπτωση τώρα ενός μοτίβου νουκλεοτιδίων η προηγούμενη σχέση γίνεται:

$$I_{position} = 2 - \sum_{i=1}^{4} P_i \log_2(P_i)$$

Που σημαίνει πώς μπορούμε να αξιολογήσουμε το πληροφοριακό περιεχόμενο ενός μοτίβου δεδομένου ενός πίνακα PWM. Έτσι στον PWM πίνακα η πρώτη θέση έχει πληροφοριακό περιεχόμενο ίσο με:

$$I_1 = 2 - \left(P[A] \log_2(P[A]) + P[G] \log_2(P[G]) + P[C] \log_2(P[C]) + P[T] \log_2(P[T]) \right)$$

$$I_1 = 2 - \left(0 \log_2(0) + 1 \log_2(1) + 0 \log_2(0) + 0 \log_2(0) \right) = 2$$

Στον υπολογισμό αυτό θεωρούμε ότι η τιμή 0log2(0) είναι ίση με 0. Βλέπουμε ότι για τη συγκεκριμένη θέση το πληροφοριακό περιεχόμενο είναι ίσο με 2, που είναι και η μέγιστη τιμή που μπορεί να πάρει δεδομένων τεσσάρων πιθανών ενδεχομένων. Πράγματι, για τη θέση 1, η πιθανότητα εμφάνισης του G είναι 1.0 που σημαίνει ότι η αβεβαιότητα αυτής της θέσης είναι μηδενική. Αντίστοιχα, το πληροφοριακό περιεχόμενο της θέσης 5 θα είναι ίσο με:

$$I_5 = 2 - \left(0.49 \log_2(0.49) + 0.37 \log_2(0.37) + 0.01 \log_2(0.01) + 0.13 \log_2(0.13) \right) = 0.52$$

τιμή που είναι σημαντικά μικρότερη από αυτήν της θέσης 1 και που αντανακλά το μεγαλύτερο βαθμό αβεβαιότητάς της.

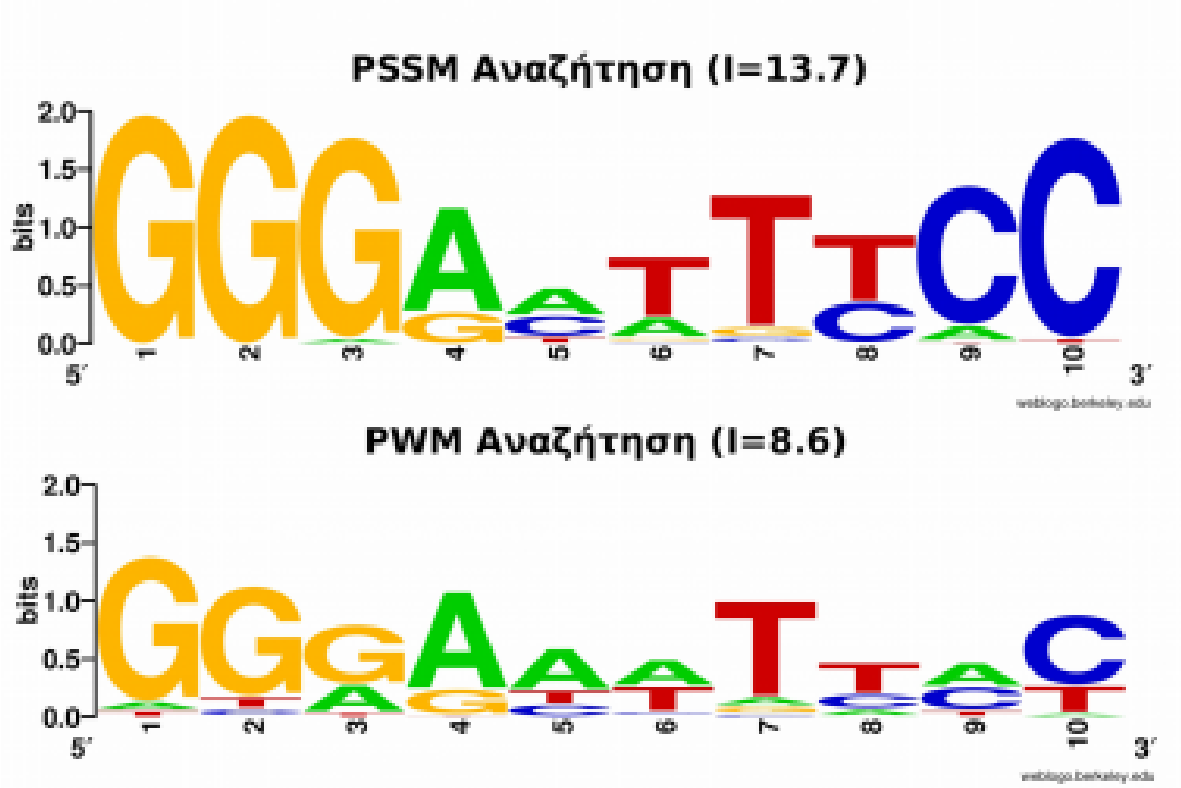
Πώς μπορούμε να χρησιμοποιήσουμε αυτήν την προσέγγιση για να αξιολογήσουμε το πληροφοριακό περιεχόμενο του μοτίβου όπως αυτό ορίζεται από έναν πίνακα PWM; Κάτι τέτοιο μπορεί να γίνει αναλυτικά με την εξής διαδικασία:

```
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Αλγόριθμος :: Πληροφοριακό Περιεχόμενο PWM
Δήλωση PWM διαστάσεων NxM; N=έκταση PWM, M=αριθμός ενδεχομένων
Υπολογισμός Hmax=-M * (1/M*log2(1/M));
Επανάληψη 1 : Για i = 1 έως i = N ανά 1:
        Επανάληψη 2: για j = 1 έως j = M ανά 1:
                H[i, j]=-PWM[i, j] * log2(PWM[i, j]);
                I[i]=Hmax-H[i, j];
        Τέλος Επανάληψης 2
Τέλος Επανάληψης 1
Επανάληψη 3: Για k = 1 έως k = N ανά 1:
        Επανάληψη 4: για j = 1 έως j = M ανά 1:
                P[i, j]=PWM[i, j] * I[i];
        Τέλος Επανάληψης 4
Τέλος Επανάληψης 3
Εκτύπωσε P[i, j]
Τερματισμός
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
```

Ο παραπάνω αλγόριθμος υπολογίζει την τιμή του πληροφοριακού περιεχομένου ανά θέση (I[i]), κρατώντας σε έναν πίνακα H[i,j] τη συνεισφορά εντροπίας σε κάθε ενδεχομένου j στην κάθε θέση i του μοτίβου. Το πληροφοριακό περιεχόμενο μπορεί έτσι να αξιολογηθεί ανά θέση, όπως είδαμε και πιο πάνω στα παραδείγματα για τις θέσεις 1 και 5. Ο αλγόριθμος πληροφοριακό Περιεχόμενο PWM κάνει και κάτι ακόμα κι αυτό είναι ότι για κάθε νουκλεοτίδιο j σε κάθε θέση i υπολογίζει μια ποσότητα P που αντιστοιχεί στη σταθμισμένη Εντροπία του i,j με βάση τη συχνότητα εμφάνισης PWM[i, j]. Ο πίνακας P[i, j] που επιστρέφει ο αλγόριθμος ως αποτέλεσμα είναι ο παρακάτω:

Θέση	1	2	3	4	5	6	7	8	9	10
A	0.00	0.00	0.05	0.92	0.25	0.18	0.01	0.01	0.14	0.00
C	0.00	0.00	0.00	0.00	0.19	0.02	0.05	0.37	1.23	1.75
G	2.00	2.00	1.75	0.29	0.01	0.04	0.09	0.00	0.00	0.00
T	0.00	0.00	0.00	0.00	0.07	0.53	1.16	0.59	0.03	0.05
I(θέσης)	2.00	2.00	1.81	1.20	0.52	0.78	1.32	0.97	1.39	1.81

Ένας τέτοιος πίνακας μπορεί να χρησιμοποιηθεί σε μια πολύ διαδεδομένη μορφή αναπαράστασης μοτίβων που ονομάζεται logo αλληλουχιών (sequence logo) (Schneider and Stephens 1990). Τα logo αλληλουχιών αναπαριστούν το σύμβολο κάθε καταλοίπου σε κάθε θέση με ύψος ανάλογο της σταθμισμένης του Εντροπίας, ενώ ταυτόχρονα αποδίδουν το πληροφοριακό περιεχόμενο σε κάθε θέση στο συνολικό ύψος όλων των συμβόλων. Τα logo μπορούν έτσι παράλληλα να οπτικοποιήσουν ένα μοτίβο αλλά και να αξιολογήσουν την πληροφορία που αυτό έχει. Στην εικόνα που ακολουθεί βλέπουμε τα logo αλληλουχιών που προκύπτουν από τα σημεία πρόσδεσης με τα 100 καλύτερα score της αναζήτησης μέσω PSSM και PWM αντίστοιχα όπως αυτή παρουσιάστηκε παραπάνω.



Είναι προφανές πως τα σημεία πρόσδεσης που προκύπτουν από την αναζήτηση μέσω PSSM αποδίδουν ένα μοτίβο με μεγαλύτερο πληροφοριακό περιεχόμενο (I=13.7 έναντι μόλις 8.6 για την περίπτωση των PWM). Η ίδια η εικόνα του logo που προκύπτει μέσω PSSM- αναζήτησης δείχνει πιο έντονες “κορυφές” για τις περισσότερες θέσεις του μοτίβου, κάτι που είναι ενδεικτικό μεγαλύτερου πληροφοριακού περιεχομένου. Μια σειρά από άλλα χαρακτηριστικά των μοτίβων αποδίδονται παραστατικά από τα logo όπως είναι σε ποιες θέσεις υπάρχει αυξημένη αβεβαιότητα, ποια είναι τα πιθανότερα κατάλοιπα σε κάθε θέση και ποια είναι η συνεισφορά τους στο πληροφοριακό της περιεχόμενο κλπ. Έτσι, το ύψος του κάθε καταλοίπου στην κάθε θέση του logo είναι συνάρτηση του πόσο συχνά εμφανίζεται το συγκεκριμένο κατάλοιπο σε αυτή τη θέση, ενώ το συνολικό υψος των καταλοίπων σε κάθε θέση ισούται με το πληροφοριακό περιεχόμενο σε δυαδικές μονάδες εντροπίας (bits). Συνοψίζοντας, μπορούμε να πούμε πως η χρήση της μαθηματικής έννοιας της Εντροπίας Shannon αποτελεί έναν πολύ καλό τρόπο αξιολόγησης του πληροφοριακού περιεχομένου και κατ' επέκταση της σημασίας ενός μοτίβου. [23]

5. Τυχαιοκρατικοί αλγόριθμοι - Το τέταρτο πρόβλημα

Θα περάσουμε τώρα στο τέταρτο και τελευταίο ερώτημα που θέσαμε στην αρχή αυτού του κεφαλαίου το οποίο ήταν:

- -

Δεδομένου ενός συνόλου αλληλουχιών που περιέχουν σημεία πρόσδεσης του ίδιου μεταγραφικού παράγοντα, πώς μπορούμε να προσδιορίσουμε το μοτίβο πρόσδεσής του;

- -

Το συγκεκριμένο πρόβλημα είναι αρκετά δυσκολότερο από τα τρία προηγούμενα για το βασικό λόγο ότι δεν προϋποθέτει καμία γνώση των περιορισμών των σημείων πρόσδεσης. Σε αντίθεση δηλαδή με τα ερωτήματα που αντιμετωπίσαμε ως τώρα δεν έχουμε καμία προηγούμενη γνώση για το μοτίβο, πέρα από το ότι το σύνολο των αλληλουχιών που διαθέτουμε περιέχει τουλάχιστο μια θέση πρόσδεσης ανά αλληλουχία. Η κατάσταση αυτή είναι αρκετά ρεαλιστική για μια μεγάλη κατηγορία πειραμάτων στα οποία αναζητούμε τα σημεία πρόσδεσης μεταγραφικών παραγόντων που δεν έχουν μελετηθεί αρκετά ώστε να γνωρίζουμε το μοτίβο πρόσδεσής τους. Πώς θα προσεγγίσουμε ένα τόσο πολύπλοκο πρόβλημα; Αρχικά μπορούμε να το διατυπώσουμε πιο αναλυτικά ως εξής:

- -

Δεδομένου ενός συνόλου Τ αλληλουχιών που περιέχουν ένα τουλάχιστο σημείο πρόσδεσης ενός μεταγραφικού παράγοντα, να βρεθεί μια συλλογή ολιγονουκλεοτιδίων μήκους n που να αποτελεί το μοτίβο m για το οποίο το πληροφοριακό περιεχόμενο είναι το μέγιστο.

- -

Στην πιο πάνω διατύπωση έχουμε περιγράψει το πρόβλημα αρκετά πιο αναλυτικά ώστε να αρχίσουμε να σκεφτόμαστε τρόπους για την επίλυσή του. Προσέξτε ότι έχουμε κάνει μια παραδοχή για το μήκος του μοτίβου (n) το οποίο είναι μέρος του προβλήματος καθώς, όταν δε γνωρίζουμε τίποτα για τα σημεία πρόσδεσης, δεν μπορούμε να γνωρίζουμε a priori ούτε την έκτασή του. Για λόγους απλότητας ωστόσο θα θεωρήσουμε στη συνέχεια ότι αναζητούμε μοτίβα με συγκεκριμένο μήκος. Η διαφορά στην περίπτωση που τα μήκη δεν είναι συγκεκριμένα είναι ότι κανείς απλώς θα πρέπει να δοκιμάσει την ίδια προσέγγιση για όλα τα διαφορετικά μήκη μέσα σε ένα εύρος. Ας δούμε τώρα πώς θα μπορούσαμε να εντοπίσουμε αυτό το μοτίβο διαβάζοντας τις αλληλουχίες. Αν ορίσουμε b[T] το σύνολο των πραγματικών σημείων πρόσδεσης και PWMb[t] τον πίνακα βαρών που προκύπτει

από αυτά τότε ισχύει ότι για οποιαδήποτε συλλογή r[T] ολιγονουκλεοτιδίων μήκους n από το σύνολο των αλληλουχιών θα ισχύει ότι:

$$I\left(PWM_{b[T]}\right) > I\left(PWM_{r[T]}\right) \text{για κάθε } r$$

Το πρόβλημα ουσιαστικά συνίσταται στο να εντοπίσουμε ένα σύνολο ολιγονουκλεοτιδίων b για το οποίο να ισχύει η ανισότητα 3.6. Ας υποθέσουμε ότι προσπαθούμε να αντιμετωπίσουμε αυτό το πρόβλημα με εξαντλητικό τρόπο, μέσω μιας προσέγγισης brute force. Αυτό σημαίνει ότι θα πρέπει να υπολογίσουμε όλους τους πιθανους PWMb[t] για το σύνολο των T αλληλουχιών. Τι σημαίνει αυτό από πλευράς υπολογισμών; Για μήκος ολιγονουκλεοτιδίου n και μήκος της κάθε αλληλουχίας l τα πιθανά ολιγονουκλεοτίδια είναι l – n + 1 για κάθε αλληλουχια. Οι συνδυασμοί τους για T αλληλουχίες είναι $(l - n + 1)^T$, για καθέναν από τους οποίους θα χρειαστεί να κάνουμε nT πράξεις για να εξάγουμε τον αντίστοιχο PWM. Συνολικά θα απαιτηθούν $nT(l – n + 1)^T$ υπολογισμοί, ένας αριθμός που ακόμα και για μικρά n, l, T είναι τόσο μεγάλος που καθιστά την προσέγγιση αυτή απαγορευτική. Πώς μπορούμε να επιταχύνουμε τη διαδικασία; Χρησιμοποιούμε δειγματοληψία Gibbs.

Οι τυχαικρατικοί αλγόριθμοι λαμβάνουν τυχαίες αποφάσεις σε όλη τη διάρκεια της λειτουργίας τους. Το γεγονός ότι ένας τυχαιοκρατικός αλγόριθμος εκτελεί μια μη αιτιοκρατική ακολουθία ενεργειών σημαίνει συχνά ότι σε αντίθεση με τους αιτιοκρατικούς αλγορίθμους, δεν υπάρχει είσοδος η οποία να επιστρέφει πάντοτε αποτελέσματα χειρότερης περίπτωσης. Στην ενότητα αυτή θα δούμε πως οι τυχαικρατικοί αλγόριθμοι θα λύσουν το πρόβλημα της Εύρεσης Μοτίβων.

5.1 Δειγματοληψία Gibbs

Το 1993, ο Chip Lawrence και οι συνεργάτες του πρότειναν τη χρήση της δειγματοληψίας *Gibbs* για την εύρεση μοτίβων σε αλληλουχίες DNA. Με δεδομένο ένα σύνολο με t αλληλουχίες που έχουν μήκος n νουκλεοτίδια και έναν ακέραιο l, ο δειγματολήπτης Gibbs προσπαθεί να λύσει το πρόβλημα της Εύρεσης Μοτίβου, δηλαδή της εύρεση ενός l - μερούς σε κάθε μία από τις t αλληλουχίες έτσι ώστε να μεγιστοποιηθεί η ομοιότητα μεταξύ αυτών των l - μερών. Έστω **s** = $(s_1, \ldots, s_t)$ οι αρχικές θέσεις των επιλεγμένων l - μερών σε t αλληλουχίες. Οι συγκεκριμένες υποσυμβολοσειρές σχηματίζουν μια μήτρα στοίχισης t x l και το

αντίστοιχο προφίλ **P(s)** = (p_{ij}) με διαστάσεις 4 x l. Θεωρούμε το προφίλ ως τη συχνότητα των συμβόλων σε μια στοίχιση και όχι το πλήθος των συμβόλων. Άρα, κάθε στήλη της μήτρας προφίλ σχηματίζει μια κατανομή πιθανοτήτων. Με δεδομένα το προφίλ **P** και ένα τυχαίο l - μερές **a** = $a_1a_2\ldots a_l$, θεωρούμε την ποσότητα P(**a**|**P**) = $\Pi^{l}_{i=1}$pi, δηλαδή την πιθανότητα ότι το **a** έχει παραχθεί από το **P**. Τα l - μέρη που παρουσιάζουν ομοιότητα με τη συνεναιτική συμβολοσειρά του προφίλ θα έχουν μεγαλύτερες πιθανότητες ενώ τα l - μερή που δεν παρουσιάζουν ομοιότητα θα έχουν μικρότερες πιθανότητες. Με δεδομένο το προφίλ **P,** μπορούμε επομένως να υπολογίσουμε την πιθανότητα κάθε l - μερούς στην αλληλουχία i, και να βρούμε το l - μερές που έχει την μεγαλύτερη πιθανότητα να έχει παραχθεί από το **P** -- το συγκεκριμένο l - μερές θα ονομάζεται το πιο πιθανό κατά **P** l - μερές της αλληλουχίας. Αυτό αποτελεί κίνητρο για τον παρακάτω αλγόριθμο GREEDYPROFILEMOTIFSEARCH, ο οποίος λύνει το πρόβλημα της Εύρεσης Μοτίβου:

- -

GREEDYPROFILEMOTIFSEARCH(DNA, t, n, l)
1. Τυχαία επιλογή των αρχικών θέσεων (s_1, …, s_t) στο DNA
2. Σχηματισμός του προφίλ **P** από το **s**
3. bestScore ← 0
4. **while** Score(**s**, DNA) > bestScore
5. bestScore ← Score(**s**, DNA)
6. **for** i ← 1 **to** t
7. Εύρεση του πιο πιθανού κατά **P** l – μερούς **a** από την i – οστή αλληλουχία
8. s_i ← Αρχική θέση του **a**
9. **return** bestScore

- -

Ο αλγόριθμος GREEDYPROFILEMOTIFSEARCH ξεκινάει από τυχαίο φύτρο επιλέγοντας τις αρχικές θέσεις **s** τυχαία με βάση την ομοιόμορφη κατανομή, και προσπαθεί να βελτιώσει την επιλογή του με χρήση άπληστης στρατηγικής. Αφού ο υπολογιστικός χώρος των αρχικών θέσεων είναι τεράστιος, τα τυχαία επιλεγμένα φύτρα θα αντιστοιχούν σπάνια σε βέλτιστο μοτίβο, και υπάρχει μικρή πιθανότητα να μας οδηγήσουν στη βέλτιστη λύση μέσω της άπληστης στρατηγικής. Έτσι, ο GREEDYPROFILEMOTIFSEARCH εκτελείται πολλές φορές με την ελπίδα ότι μία από τις χιλιάδες εκτελέσεις θα παράγει ένα φύτρο που προσεγγίζει κατά τύχη το βέλτιστο. Είναι απίθανο ο αλγόριθμος GREEDYPROFILEMOTIFSEARCH θα βρει το βέλτιστο μοτίβο. Ο συγκεκριμένος αλγόριθμος αλλάζει τις αρχικές θέσεις (s_1,

s_2, …, s_t) σε κάθε επανάληψη, και μπορεί να αλλάξει μέχρι και t θέσεις σε μία επανάληψη. Η δειγματοληψία Gibbs είναι μία επαναληπτική διαδικασία που απορρίπτει σε κάθε επανάληψη ένα l - μερές από τη στοίχιση και το αντικαθιστά με νέο. Με άλλα λόγια, αλλάζει τουλάχιστον μια θέση στο **s** σε κάθε επανάληψη, και έτσι κινείται με περισσότερη προσοχή στο χώρο όλων των αρχικών θέσεων. Όπως και ο αλγόριθμος GREEDYPROFILEMOTIFSEARCH, η δειγματοληψία Gibbs αρχίζει με τυχαία επιλογή l - μερών σε κάθε μία από τις t αλληλουχίες DNA, αλλά η επιλογή που κάνει σε κάθε επανάληψη είναι τυχαία και όχι άπληστη.

1. Τυχαία επιλογή των αρχικών θέσεων **s** = (s_1, …, s_t) στο DNA και σχηματισμός του συνόλου των l - μερών που αρχίζουν σε αυτές τις θέσεις.
2. Τυχαία επιλογή μίας αλληλουχίας από t αλληλουχίες DNA.
3. Δημιουργία ενός προφίλ **P** από τα l - μερή στις υπόλοιπες t - 1 αλληλουχίες.
4. Για κάθε θέση i στην επιλεγμένη αλληλουχία DNA, υπολογισμός της πιθανότητας p_i να παράγεται το l - μερές που αρχίζει στη συγκεκριμένη θέση από το προφίλ **P**($1 <= i <= n – l + 1$).
5. Τυχαία επιλογή της νέας αρχικής θέσης στην επιλεγμένη αλληλουχία DNA, σύμφωνα με την κατανομή που είναι ανάλογη με (p_1, p_2, …, p_{n-l+1}).
6. Επανάληψη μέχρι να υπάρξει σύγκλιση.

Παρότι η δειγματοληψία Gibbs είναι αποτελεσματική σε πολλές περιπτώσεις, ενδέχεται να συγκλίνει σε τοπικό αντί για ολικό μέγιστο, ειδικά για δύσκολα προβλήματα αναζήτησης με δυσδιάκριτα μοτίβα. Η εύρεση μοτίβων γίνεται ιδιαίτερα δύσκολη αν η κατανομή των νουκλεοτιδίων στο δείγμα είναι μη ομοιόμορφη, δηλαδή αν κάποια νουκλεοτίδια στο δείγμα εμφανίζονται πιο συχνά από τα υπόλοιπα. Σε αυτή τη περίπτωση, η αναζήτηση για το σήμα με το μέγιστο αριθμό εμφανίσεων μπορεί να οδηγήσει σε μοτίβα που αποτελούνται από τα πιο συχνά νουκλεοτίδια και τα οποία έχουν μικρή βιολογική σημαντικότητα. Για παράδειγμα, αν το A έχει συχνότητα 70%, και τα T, G και C έχουν συχνότητα 10%, τότε το πολύ (A) μπορεί να είναι το πιο συχνό μοτίβο, αποκρύπτοντας έτσι το βιολογικά σχετικό μοτίβο.
Όσον αφορά τον εντοπισμό μοτίβων σε μεροληπτικά δείγματα, ορισμένοι αλγόριθμοι χρησιμοποιούν τη σχετική εντροπία για να επιλέξουν το μοτίβο ανάμεσα σε εκείνα που αποτελούνται από συχνά νουκλεοτίδια. Με δεδομένο ένα προφίλ μήκους l, η σχετική εντροπία ορίζεται ως

$$\Sigma^{l}_{j=1} \Sigma \; p_{rj} \log_2 (p_{rj} / b_r)$$

όπου p_{rj} είναι η συχνότητα του νουκλεοτιδίου r στη θέση j της στοίχισης και b_r είναι η συχνότητα υποβάθρου του r. Η δειγματοληψία Gibbs μπορεί να προσαρμοστεί για να λειτουργεί με σχετικές εντροπίες. [24]

4.2 Τυχαίες προβολές

Ο αλγόριθμος RANDOMPROJECTIONS είναι μια ακόμη τυχαικρατική μέθοδος για την εύρεση μοτίβων. Αν ένα μοτίβο μήκους l "εμφυτευτεί" σε αλληλουχίες DNA χωρίς μεταλλάξεις, τότε η εύρεση του μοτίβου ανάγεται απλώς στην καταμέτρηση των εμφανίσεων των l - μερών από το δείγμα. Όμως, η εύρεση του μοτίβου γίνεται πολύ δύσκολη όταν το μοτίβο εμφυτεύεται με μεταλλάξεις. Σε οποιοδήποτε στιγμιότυπο ενός μοτίβου με μήκος για παράδειγμα 8 και δύο μεταλλαγμένες θέσεις έξι θέσεις δεν επηρεάζονται από τις μεταλλάξεις και μπαίνουμε μ στον πειρασμό να χρησιμοποιήσουμε τις συγκεκριμένες θέσεις ως βάση για την εύρεση του μοτίβου. Αυτή η μέθοδος έχει δύο επιπλοκές. Πρώτον, οι έξι σταθερές θέσεις δεν σχηματίζουν απαραίτητων μια συνεχόμενη συμβολοσειρά: τα μεταλλαγμένα νουκλεοτίδια μπορεί να βρίσκονται στις θέσεις 3 και 7, με αποτέλεσμα οι υπόλοιπες έξι συντηρημένες θέσεις να σχηματίζουν ένα μοτίβο με κενά. Δεύτερον, τα διαφορετικά στιγμιότυπα του μοτίβου μπορούν να μεταλλαχθούν σε διαφορετικές θέσεις. Για παράδειγμα, τρία διαφορετικά στιγμιότυπα του μεταλλαγμένου μοτίβου μπορεί να διαφέρουν στις θέσεις 3 και 7, 3 και 6, και 2 και 6 αντίστοιχα. Η βασική παρατήρηση είναι ότι, παρόλο που και τα τρία μεταλλαγμένα μοτίβα **X***X*, **X**X**, *X***X** είναι διαφορετικά μεταξύ τους, το "συναινετικό" τους μοτίβο με κενά *XX**XX* δεν επηρεάζεται από τις μεταλλάξεις. Αν γνωρίζαμε ποιο μοτίβο με κενά δεν επηρεάστηκε από τις μεταλλάξεις, θα μπορούσαμε να το χρησιμοποιήσουμε για την αναζήτηση του μοτίβου σαν να ήταν εμφυτευμένο μοτίβο με κενά χωρίς μεταλλάξεις. Στον πραγματικό κόσμο, όλες οι l - θέσεις ενδέχεται να επηρεαστούν από μια μετάλλαξη σε κάποια στιγμιότυπα του μοτίβου. Ωστόσο, είναι πολύ πιθανό ότι υπάρχει ένα σχετικά μεγάλο σύνολο στιγμιοτύπων που μοιράζονται τις ίδιες μη μεταλλαγμένες θέσεις, και ο αλγόριθμος RANDOMPROJECTIONS τα χρησιμοποιεί για να εντοπίσει μοτίβα. Όμως το πρόβλημα είναι ότι οι τέσσερις θέσεις που δεν επηρεάζονται στο *XX**XX* είναι άγνωστες. Ο αλγόριθμος RANDOMPROJECTIONS παρακάμπτει το πρόβλημα, δοκιμάζοντας διαφορετικά τυχαία επιλεγμένα σύνολα με k θέσεις για να αποκαλύψει το αρχικό εμφυτευμένο μοτίβο. Αυτά τα σύνολα θέσεων ονομάζονται *προβολές*. Θα ορίσουμε το (k, l) - πρότυπο ως οποιοδήποτε σύνολο k διαφορετικών ακεραίων $1 <= t_1 <= \ldots < t_k <= l$.

Για ένα (k, l) - πρότυπο **t** = (t_1, …, t_k) και ένα l - μερές **a** = a_1… a_l ορίζουμε ως Projection(**a, t**) = a_{t1}, a_{t2}, …, a_{tk} τη συνένωση των νουκλεοτιδίων από το **a**, όπως αυτή ορίζεται από το πρότυπο **t**. Για παράδειγμα, αν **a** = ATGCATT και **t** = (2, 5, 7), τότε προκύπτει ότι Projection(**a, t**) = TAT. Ο αλγόριθμος RANDOMPROJECTIONS παρακάτω επιλέγει ένα τυχαίο (k, l) - πρότυπο και προβάλλει κάθε l - μερές του δείγματος πάνω σε αυτό, τα k - μερή που προκύπτουν καταγράφονται μέσω ενός πίνακα κατακερματισμού. Αναμένουμε ότι τα k - μέρη που αντιστοιχούν σε προβολές του εμφυτευμένου μοτίβου θα εμφανίζονται περισσότερο συχνά από τα υπόλοιπα k – μερή. Κατά συνέπεια, τα k-μερή που εμφανίζονται πολλές φορές ως προβολές των l - μερών από το δείγμα αναπαριστούν πιθανές προβολές του εμφιτευμένου μοτίβου. Φυσικά, τα παραπάνω εξαρτώνται από το θόρυβο, και ένα μόνο (k, l) - πρότυπο δεν θα αποκαλύψει απαραιτήτως το εμφυτευμένο μοτίβο. Ο αλγόριθμος RANDOMPROJECTIONS επιλέγει επανειλημμένα ένα δεδομένο αριθμό m του τυχαίου (k, l) - προτύπου και συγκεντρώνει τα δεδομένα που λαμβάνονται για όλες τις m επαναλήψεις. Καθώς ο αλγόριθμος επιλέγει διαφορετικά τυχαία πρότυπα, οι τοποθεσίες του εμφυτευμένου μοτίβου γίνονται πιο προφανείς.

Σε σύγκριση με άλλους αλγορίθμους εύρεσης μοτίβων, ο αλγόριθμος RANDOMPROJECTIONS απαιτεί τις εξής πρόσθετες παραμέτρους: k ο αριθμός των θέσεων στο πρότυπο, θ το κατώφλι που καθορίζει ποια κλίμακα στον πίνακα κατακερματισμού πρέπει να ληφθούν υπόψη μετά από την προβολή όλων των l - μερών, και m ο αριθμός των επιλεγμένων τυχαίων προτύπων. Ο αλγόριθμος δημιουργεί τον πίνακα **Bins** με μέγεθος 4^k έτσι ώστε κάθε δυνατή προβολή k - μερές αν αντιστοιχεί σε μοναδική διεύθυνση του πίνακα. Για δεδομένο (k, l) - πρότυπο **r**, το **Bins(x)** περιέχει το πλήθος των l - μερών **a** στο DNA έτσι ώστε να ισχύει η σχέση Projection(**a, r**) = **x**.

- -

```
RANDOMPROJECTIONS(DNA, t, n, l, k, θ, m)
1. δημιουργία του πίνακα motifs διαστάσεων t x n και συμπλήρωση σου με μηδενικά
2. for m επαναλήψεις
3.     δημιουργία του πίνακα Bins μεγέθους 4^k και η συμπλήρωση του με μηδενικά
4.     r ← τυχαίο (k, l) - πρότυπο
5.     for i ← 1 to t
6.         for j ← 1 to n – l + 1
7.             a ← j - οστό l – μερές στην i – οστή αλληλουχία DNA
8.             Bins(Projection(a, r)) = Bins(Projection(a, r)) + 1
```

9. **for** i ← 1 **to** t
10. **for** j ← 1 **to** n – l + 1
11. **a** ← j - οστό l – μερές στην i – οστή αλληλουχία DNA
12. **if Bins**(Projection(**a, r**)) > θ
13. **motifs** $_{i, j}$ ← **motifs** $_{i, j}$ + 1
14. **for** Ii ← 1 **to** t
15. s_i ← δείκτης του μεγαλύτερου στοιχείου στη γραμμή i του **motifs**
16. **return s**

- -

Ο αλγόριθμος δεν παρέχει καμία εγγύηση για την επιστροφή του σωστού μοτίβου, αλλά μπορούμε να αποδείξουμε ότι επιστρέφει το σωστό μοτίβο με μεγάλη πιθανότητα, αν υποθέσουμε ότι οι παράμετροι επιλέγονται με λογικό τρόπο. Η κύρια διαφορά ανάμεσα σε αυτόν τον απλό αλγόριθμο και τον πρακτικό αλγόριθμο **Projection** είναι ο τρόπο με τον οποίο ο αλγόριθμος αξιολογεί τα αποτελέσματα από τον κατακερματισμό όλων των προβολών των l - μερών. Η μέθοδος που παρουσιάσαμε εδώ για την επιλογή των θέσεων (s_1, s_2, …, s_t) είναι στοιχειώδης, ενώ ο αλγόριθμος **Projection** κάνει χρήση μιας ευρετικής μεθόδου που είναι πιο δύσκολο να ξεγελαστεί από τυχαία μεγάλες τιμές του πλήθους στον πίνακα **motifs**. Συγκεκριμένα χρησιμοποιούν έναν αλγόριθμο Μεγιστοποίησης Αναμενόμενης Τιμής, ο οποίος αποτελεί μια τεχνική τοπικής αναζήτησης που χρησιμοποιείται σε πολλούς αλγορίθμους βιοπληροφορικής. [25]

6. Συμπεράσματα

Ένα μεγάλο μέρος από τις προσεγγίσεις που εξετάσαμε σε αυτές τις ενότητες σχετίζονται με συγκρίσεις μεταξύ δύο αλληλουχιών στο επίπεδο της πρωτοταγούς δομής. Τόσο η απόσταση Hamming όσο και η σύγκριση μέσω PWM και PSSM αποτελούν τρόπους αξιολόγησης της ομοιότητας μεταξύ δύο αλληλουχιών, μιας ιδιότητας που έχει εξαιρετική σημασία για πολλούς λόγους. Λόγω της σχέσης που έχει η πρωτοταγής αλληλουχία με τη διαμόρφωση στον τριδιάστατο χώρο (και που θα συζητήσουμε σε επόμενο κεφάλαιο) δύο αλληλουχίες με παρόμοια διαδοχή καταλοίπων είναι πολύ πιθανό να έχουν την ίδια λειτουργία. Σε αυτό το κεφάλαιο χρησιμοποιήσαμε αυτήν την ιδιότητα για να αναζητήσουμε σημεία πρόσδεσης μεταγραφικών παραγόντων, ενώ επιπλέον είδαμε πως η ταυτόχρονη μελέτη της ομοιότητας αλληλουχιών μεταξύ ειδών μπορεί να βελτιώσει ακόμα περισσότερο τη δυνατότητα εντοπισμού τους. Μπορούμε εύλογα να

αναλογιστούμε ότι η στρατηγική της σύγκρισης της πρωτοταγούς δομής θα μπορούσε να επεκταθεί και σε αλληλουχίες μεγαλύτερου μήκους με σκοπό την εξαγωγή συμπερασμάτων τόσο για τη λειτουργία τους, όσο και για την εξελικτική τους προέλευση. Ωστόσο, το πρόβλημα της σύγκρισης δύο αλληλουχιών μήκους εκατοντάδων νουκλεοτιδίων διαφέρει από αυτό της σύγκρισης μοτίβων όχι μόνο ποσοτικά αλλά και ποιοτικά.

7. MatLab

Έχω υλοποιήσει σε MatLab τον αλγόριθμο BRUTEFORCEMOTIFSEARCH:

```
function [] = motifSearch(A, l, k, p)

    bestScore = 0;
    athroismaMax = 0;

    for i = 1:l
      letterA(1, i) = 0;
      letterT(1, i) = 0;
      letterG(1, i) = 0;
      letterC(1, i) = 0;
    end

    epanalipseis = k - l + 1;
    for i = 1:(epanalipseis^p)
        for j = 1:p
           example(i, j) = cellstr('Waaaaaa');
        end
    end

    step = epanalipseis^p;

    for i = 1:p
      step = step/epanalipseis;
      for j = 1:epanalipseis
        x = A(i,[j:j + l - 1]);
        for n = 1:((epanalipseis^(p-1)/step))
```

```
        start = (n-1)*step*epanalipseis + j;
        example(start:start+step-1, i) = cellstr(x);
      end
    break
   end
end
example

for i = 1:(epanalipseis^p)
    for ii = 1:l
      letterA(1, ii) = 0;
      letterT(1, ii) = 0;
      letterG(1, ii) = 0;
      letterC(1, ii) = 0;
    end
    for j = 1:p
      triada = char(cellstr(example(i,j)))
      for m = 1:l
        gramma = triada(m);
        if gramma == 'A'
          letterA(1, m) = letterA(1, m) + 1;
        end
        if gramma == 'T'
          letterT(1, m) = letterT(1, m) + 1;
        end
        if gramma == 'G'
          letterG(1, m) = letterG(1, m) + 1;
        end
        if gramma == 'C'
          letterC(1, m) = letterC(1, m) + 1;
        end
      end
      letterA
      letterT
      letterG
      letterC
```

```
        disp('--------------------')
      end
      correctLetter = '';
      for gg = 1:l
        if letterA(1,gg) > letterT(1,gg)
          maxGramma = 'A';
          maxValue = letterA(1,gg);
        else
          maxGramma = 'T';
          maxValue = letterT(1,gg);
        end
        if letterG(1,gg) > maxValue
          maxGramma = 'G';
          maxValue = letterG(1,gg);
        end
        if letterC(1,gg) > maxValue
          maxGramma = 'C';
          maxValue = letterC(1,gg);
        end
        correctLetter = strcat(correctLetter,maxGramma);
      end
      max1 = max(max(letterA(1,1), letterT(1,1)), max(letterG(1,1),
letterC(1,1)));
      max2 = max(max(letterA(1,2), letterT(1,2)), max(letterG(1,2),
letterC(1,2)));
      max3 =  max(max(letterA(1,3), letterT(1,3)), max(letterG(1,3),
letterC(1,3)));
      athroisma = max1 + max2 + max3;
      if athroisma > bestScore
        bestScore = athroisma;
        correctLetter
        athroisma = 0;
      end
    end

end
```

8.Βιβλιογραφία

[αριθμος]	Πηγή
[1], [2], [3]	Βικιπαίδεια
[4], [5], [6], [7], [8], [10], [12], [14], [15], [22], [23]	Αποθετήριο kallipos
[9], [11], [13], [16], [18], [19], [20], [21], [24], [25]	Εισαγωγή στους αλγορίθμους βιοπληροφορικής Jones & Pevzner
[17]	Bionformatics applications note vol.15, no. 5 1999, pages. 426 - 427