

### Πρόβλημα 6.12

Δύο παίκτες παίζουν το παρακάτω παιχνίδι με δύο «χρωμοσώματα» που έχουν μήκος  $n$  και  $m$  νουκλεοτίδια αντίστοιχα. Σε κάθε γύρο του παιχνιδιού, ένας παίκτης μπορεί να καταστρέψει ένα από τα χρωμοσώματα και να διαχωρίσει το άλλο σε δύο μη κενά τμήματα. Για παράδειγμα, ο πρώτος παίκτης μπορεί να καταστρέψει ένα χρωμόσωμα μήκους  $n$  και να διαχωρίσει ένα άλλο χρωμόσωμα σε δύο χρωμοσώματα με μήκη  $\frac{n}{3}$  και  $\frac{m}{3}$ . Ο παίκτης που διαγράφει το τελευταίο νουκλεοτίδιο κερδίζει. Ποιος θα κερδίσει; Περιγράψτε τη νικηφόρα στρατηγική για όλες τις τιμές των  $n$  και  $m$ .

### Πρόβλημα 6.13

Δύο παίκτες παίζουν το ακόλουθο παιχνίδι με δύο αλληλουχίες που έχουν μήκος  $n$  και  $m$  νουκλεοτίδια αντίστοιχα. Σε κάθε γύρο του παιχνιδιού, ένας παίκτης μπορεί να αφαιρέσει έναν τυχαίο αριθμό νουκλεοτιδίων από τη μία αλληλουχία ή τον ίδιο (αλλά και πάλι τυχαίο) αριθμό νουκλεοτιδίων και από τις δύο αλληλουχίες. Ο παίκτης που αφαιρεί το τελευταίο νουκλεοτίδιο κερδίζει. Ποιος θα κερδίσει; Περιγράψτε τη νικηφόρα στρατηγική για όλες τις τιμές των  $n$  και  $m$ .

### Πρόβλημα 6.14

Δύο παίκτες παίζουν το εξής παιχνίδι με δύο αλληλουχίες που έχουν μήκος  $n$  και  $m$  νουκλεοτίδια αντίστοιχα. Σε κάθε γύρο του παιχνιδιού, ένας παίκτης μπορεί να αφαιρέσει δύο νουκλεοτίδια από τη μία αλληλουχία (είτε την πρώτη είτε τη δεύτερη) και ένα νουκλεοτίδιο από την άλλη. Ο παίκτης που δεν μπορεί να κάνει κίνηση κερδίζει. Ποιος θα κερδίσει; Περιγράψτε τη νικηφόρα στρατηγική για όλες τις τιμές των  $n$  και  $m$ .

### Πρόβλημα 6.15

Δύο παίκτες παίζουν το παρακάτω παιχνίδι με μια νουκλεοτιδική αλληλουχία που έχει μήκος  $n$ . Σε κάθε γύρο του παιχνιδιού, ένας παίκτης μπορεί να αφαιρέσει είτε ένα είτε δύο νουκλεοτίδια από την αλληλουχία. Ο παίκτης που αφαιρεί

το τελευταίο γράμμα κερδίζει. Ποιος θα κερδίσει; Περιγράψτε τη νικηφόρα στρατηγική για όλες τις τιμές του  $n$ .

### Πρόβλημα 6.16

Δύο παίκτες παίζουν το ακόλουθο παιχνίδι με μια νουκλεοτιδική αλληλουχία που έχει μήκος  $n = n_A + n_T + n_C + n_G$ , όπου  $n_A, n_T, n_C$ , και  $n_G$  είναι το πλήθος των A, T, C, και G στην αλληλουχία. Σε κάθε γύρο του παιχνιδιού, ένας παίκτης μπορεί να αφαιρέσει ένα ή δύο νουκλεοτίδια από την αλληλουχία. Ο παίκτης που μένει με μια μονο-νουκλεοτιδική αλληλουχία τυχαίου μήκους (δηλαδή, την αλληλουχία που περιέχει μόνο το ένα από τα 4 δυνατά νουκλεοτίδια) χάνει. Ποιος θα κερδίσει; Περιγράψτε τη νικηφόρα στρατηγική για όλες τις τιμές των  $n_A, n_T, n_C$ , και  $n_G$ .

### Πρόβλημα 6.17

Ποια είναι η βέλτιστη καθολική στοίχιση για τις συμβολοσειρές APPLE και HAPPE; Δείξτε όλες τις βέλτιστες στοιχίσεις και τις αντίστοιχες διαδρομές για μπόνους ταιριάσματος +1, ποινή ασυμφωνίας -1, και ποινή προσθαφαίρεσης -1.

### Πρόβλημα 6.18

Ποια είναι η βέλτιστη καθολική στοίχιση για τις συμβολοσειρές MOAT και BOAT; Δείξτε όλες τις βέλτιστες στοιχίσεις και τις αντίστοιχες διαδρομές για την παρακάτω μήτρα βαθμολόγησης και ποινή προσθαφαίρεσης -1.

	A	B	M	O	S	T
A	1	-1	-1	-2	-2	-3
B		1	-1	-1	-2	-2
M			2	-1	-1	-2
O				1	-1	-1
S					1	-1
T						2

### Πρόβλημα 6.19

Συμπληρώστε τη μήτρα δυναμικού προγραμματισμού της καθολικής στοίχισης για τις συμβολοσειρές AT και AAGT με συγγενική συνάρτηση βαθμολόγησης που ορίζεται από τα εξής: μπόνους ταιριάσματος 0, ποινή ασυμφωνίας -1, ποινή ανοίγματος κενού -1, και ποινή επέκτασης κενού -1. Βρείτε όλες τις βέλτιστες καθολικές στοιχίσεις.

### Πρόβλημα 6.20

Θεωρήστε τις αλληλουχίες  $\mathbf{v} = \text{TACGGGTAT}$  και  $\mathbf{w} = \text{GGACGTACG}$ . Υποθέστε ότι το μπόνους ταιριάσματος είναι +1 και ότι οι ποινές ασυμφωνίας και προσθαφαίρεσης είναι -1.

- Συμπληρώστε τον πίνακα δυναμικού προγραμματισμού για μια καθολική στοίχιση μεταξύ των  $\mathbf{v}$  και  $\mathbf{w}$ . Σχεδιάστε βέλη στα κελιά του πίνακα για να αποθηκεύσετε τις πληροφορίες οπισθοδρόμησης. Ποια είναι η βαθμολογία της βέλτιστης καθολικής στοίχισης και σε ποια στοίχιση αντιστοιχεί η συγκεκριμένη βαθμολογία;
- Συμπληρώστε τον πίνακα δυναμικού προγραμματισμού για μια τοπική στοίχιση μεταξύ των  $\mathbf{v}$  και  $\mathbf{w}$ . Σχεδιάστε βέλη στα κελιά του πίνακα για να αποθηκεύσετε τις πληροφορίες οπισθοδρόμησης. Ποια είναι η βαθμολογία της βέλτιστης τοπικής στοίχισης σε αυτή την περίπτωση και ποια στοίχιση επιτυγχάνει τη συγκεκριμένη βαθμολογία;
- Έστω ότι χρησιμοποιούμε συγγενική ποινή κενού όπου τα κόστη για το άνοιγμα και την επέκταση ενός κενού είναι ίσα με -20 και -1, αντίστοιχα. Οι βαθμολογίες για τα ταιριάσματα και τις ασυμφωνίες είναι αμετάβλητες. Ποια είναι η βέλτιστη καθολική στοίχιση σε αυτή την περίπτωση και τι βαθμολογία επιτυγχάνει;

### Πρόβλημα 6.21

Για ζεύγος συμβολοσειρών  $\mathbf{v} = v_1 \dots v_n$  και  $\mathbf{w} = w_1 \dots w_m$ , ορίζουμε ως  $M(\mathbf{v}, \mathbf{w})$  τη μήτρα για την οποία η καταχώριση  $(i, j)$  αντιστοιχεί στη βαθμολογία της βέλτιστης καθολικής στοίχισης που στοιχίζει το χαρακτήρα  $v_i$  με το χαρακτήρα  $w_j$ . Διατυπώστε έναν αλγόριθμο  $O(nm)$  που υπολογίζει τη μήτρα  $M(\mathbf{v}, \mathbf{w})$ .

Ορίζουμε ότι η στοίχιση επικάλυψης μεταξύ δύο αλληλουχιών  $\mathbf{v} = v_1 \dots v_n$  και  $\mathbf{w} = w_1 \dots w_m$  είναι η στοίχιση ανάμεσα σε ένα πρόθεμα της  $\mathbf{v}$  και ένα επίθεμα της  $\mathbf{w}$ . Για παράδειγμα, αν  $\mathbf{v} = \text{TATATA}$  και  $\mathbf{w} = \text{AAATTT}$ , τότε μια (όχι απαραιτήτως βέλτιστη) στοίχιση επικάλυψης μεταξύ των  $\mathbf{v}$  και  $\mathbf{w}$  είναι η

ATA  
AAA

Η βέλτιστη στοίχιση επικάλυψης είναι η στοίχιση που μεγιστοποιεί τη βαθμολογία της καθολικής στοίχισης μεταξύ των  $v_1, \dots, v_n$  και  $w_1, \dots, w_m$ , όπου το μέγιστο υπολογίζεται για όλα τα πρόθεμα  $v_1, \dots, v_n$  της  $\mathbf{v}$  και όλα τα επίθεμα  $w_1, \dots, w_m$  της  $\mathbf{w}$ .

### Πρόβλημα 6.22

Διατυπώστε έναν αλγόριθμο που υπολογίζει τη βέλτιστη στοίχιση επικάλυψης και εκτελείται σε χρόνο  $O(nm)$ .

Έστω ότι έχουμε τις αλληλουχίες  $\mathbf{v} = v_1 \dots v_n$  και  $\mathbf{w} = w_1 \dots w_m$ , όπου η  $\mathbf{v}$  έχει μεγαλύτερο μήκος από την  $\mathbf{w}$ . Θέλουμε να βρούμε μια υποσυμβολοσειρά της  $\mathbf{v}$  που ταιριάζει καλύτερα με ολόκληρη την  $\mathbf{w}$ . Η καθολική στοίχιση δεν θα έχει αποτέλεσμα επειδή θα προσπαθήσει να στοιχίσει ολόκληρη τη  $\mathbf{v}$ . Η τοπική στοίχιση δεν θα έχει αποτέλεσμα επειδή ενδέχεται να μην στοιχίσει ολόκληρη τη  $\mathbf{w}$ . Επομένως, αυτό είναι ένα διαφορετικό πρόβλημα που αποκαλείται πρόβλημα της *Προσαρμογής*. Η *προσαρμογή* μιας αλληλουχίας  $\mathbf{w}$  σε μια αλληλουχία  $\mathbf{v}$  είναι το πρόβλημα της εύρεσης μιας υποσυμβολοσειράς  $v'$  της  $\mathbf{v}$  αλληλουχίας  $\mathbf{v}$  που έχει τη μέγιστη βαθμολογία στοίχισης  $s(v', \mathbf{w})$  ανάμεσα σε όλες τις υποσυμβολοσειρές της  $\mathbf{v}$ . Για παράδειγμα, αν  $\mathbf{v} = \text{GTAGGCTTAAGGTTA}$  και  $\mathbf{w} = \text{TGATA}$ , οι καλύτερες στοιχίσεις ίσως είναι

	καθολική	τοπική	προσαρμογή
$\mathbf{v}$	GTAGGCTTAAGGTTA	TAG	TAGGCTTA
$\mathbf{w}$	-TAG- - - - A - - - T-A	TAG	TAGA- -TA
βαθμολογία	-3	3	2

Οι βαθμολογίες υπολογίζονται ως 1 για το ταίριασμα και -1 για την ασυμφωνία ή την προσθαφαίρεση. Προσέξτε ότι η βέλτιστη τοπική στοίχιση δεν είναι έγκυρη στοίχιση προσαρμογής. Από την άλλη πλευρά, η βέλτιστη καθολική στοίχιση περιέχει μια έγκυρη στοίχιση προσαρμογής, αλλά επιτυγχάνει μη βέλτιστη βαθμολογία σε σχέση με όλες τις στοιχίσεις προσαρμογής.

### Πρόβλημα 6.23

Διατυπώστε έναν αλγόριθμο που υπολογίζει τη βέλτιστη στοίχιση προσαρμογής. Εξηγήστε πώς συμπληρώνεται η πρώτη γραμμή και η πρώτη στήλη του πίνακα δυναμικού προγραμματισμού και γράψτε μια σχέση επανάληψης για τη συμπλήρωση του υπόλοιπου πίνακα. Παρουσιάστε μια μέθοδο που βρίσκει την καλύτερη στοίχιση αφού συμπληρωθεί ο πίνακας. Ο αλγόριθμος θα πρέπει να εκτελείται σε χρόνο  $O(nm)$ .

Έχουμε μελετήσει δύο μεθόδους για τη στοίχιση αλληλουχιών: την καθολική και την τοπική στοίχιση. Υπάρχει και η μέση κατάσταση, μια μέθοδος που είναι γνωστή ως ημικαθολική στοίχιση. Στην ημικαθολική στοίχιση, ολόκληρες οι αλληλουχίες στοιχίζονται (όπως στην καθολική στοίχιση). Το στοιχείο που την καθιστά ημικαθολική είναι ότι συμπεριλαμβάνονται τα «εσωτερικά κενά» της στοίχισης, αλλά δεν ισχύει το ίδιο για τα «κενά στα άκρα». Για παράδειγμα, θεωρήστε τις ακόλουθες δύο εναλλακτικές στοιχίσεις:

Αλληλουχία 1: CAGCA-CTTGGATTCTCGG

Αλληλουχία 2: ---CAGCGTG---

Αλληλουχία 1: CAGCACTTGGATTCTCGG

Αλληλουχία 2: CAGC----G-T---GG

Η πρώτη στοίχιση έχει 6 ταιριάσματα, 1 ασυμφωνία, και 12 κενά. Η δεύτερη στοίχιση έχει 8 ταιριάσματα, 0 ασυμφωνίες, και 10 κενά. Αν χρησιμοποιήσουμε την πιο απλή μέθοδο βαθμολόγησης (+1 για το ταίριασμα, -1 για την ασυμφωνία, -1 για το κενό), η βαθμολογία της πρώτης στοίχισης είναι ίση με -7 και η βαθμολογία της δεύτερης στοίχισης είναι ίση με -2, άρα θα προτιμούσαμε τη δεύτερη. Παρόλα αυτά, η πρώτη στοίχιση είναι πιο ρεαλιστική από βιολογικής άποψης. Δεν υπολογίζουμε τα κενά «στα άκρα» για να προκύψει ένας αλγόριθμος που επιλέγει την πρώτη από τη δεύτερη στοίχιση.

Με βάση την καινούρια («ημικαθολική») μέθοδο, η πρώτη στοίχιση θα είχε 6 ταιριάσματα, 1 ασυμφωνία, και 1 κενό, ενώ η δεύτερη στοίχιση θα εξακολουθούσε να έχει 8 ταιριάσματα, 0 ασυμφωνίες, και 10 κενά. Η πρώτη στοίχιση θα είχε πλέον βαθμολογία ίση με 4 και η δεύτερη βαθμολογία ίση με -2, άρα η πρώτη θα είχε την καλύτερη βαθμολογία.

Προσέξτε ότι οι ομοιότητες και οι διαφορές ανάμεσα στο πρόβλημα της Προσαρμογής και το πρόβλημα της Ημικαθολικής Στοίχισης φαίνονται από την ημικαθολική — αλλά χωρίς προσαρμογή — στοίχιση της αλληλουχίας ACGTCAT με την αλληλουχία TCATGCA:

$$\begin{array}{ll} \text{Αλληλουχία 1:} & \text{ACGTCAT---} \\ \text{Αλληλουχία 2:} & \text{---TCATGCA} \end{array}$$

### Πρόβλημα 6.24

Επινοήστε έναν αποδοτικό αλγόριθμο για το πρόβλημα της Ημικαθολικής Στοίχισης και δείξτε πώς λειτουργεί για τις αλληλουχίες ACAGATA και AGT. Όσον αφορά τη βαθμολόγηση, χρησιμοποιήστε μπόνους ταιριάσματος +1, ποινή ασυμφωνίας -1, και ποινή προσθαφαίρεσης -1.

Ορίζουμε την καθολική στοίχιση χωρίς αφαιρέσεις ως τη στοίχιση μεταξύ δύο αλληλουχιών  $\mathbf{v} = v_1 v_2 \dots v_n$  και  $\mathbf{w} = w_1 w_2 \dots w_m$ , όπου επιτρέπονται μόνο ταιριάσματα, ασυμφωνίες και προσθήκες. Δηλαδή, δεν μπορούν να υπάρχουν αφαιρέσεις από τη  $\mathbf{v}$  στην  $\mathbf{w}$  (όλα τα γράμματα της  $\mathbf{w}$  εμφανίζονται στη στοίχιση χωρίς κενά διαστήματα). Είναι προφανές ότι πρέπει να ισχύει  $m \geq n$  και έστω ότι  $k = m - n$ .

### Πρόβλημα 6.25

Διατυπώστε έναν αλγόριθμο  $O(nk)$  για την εύρεση της βέλτιστης καθολικής στοίχισης χωρίς αφαιρέσεις (παρατηρήστε τη βελτίωση σε σύγκριση με τον αλγόριθμο  $O(nm)$  όταν το  $k$  έχει μικρές τιμές).

### Πρόβλημα 6.26

Οι υποσυμβολοσειρές  $v_{i,...}, v_{i+k}$  και  $v_{i',...}, v_{i'+k}$  της συμβολοσειράς  $v_1, \dots, v_n$  δημιουργούν ένα ζεύγος υποσυμβολοσειρών αν ισχύει ότι  $i' - i + k > MinGap$ , όπου  $MinGap$  είναι μια παράμετρος. Ορίστε τη βαθμολογία του ζεύγους υποσυμβολοσειρών ως τη βαθμολογία της (καθολικής) στοίχισης των  $v_{i,...}, v_{i+k}$  και  $v_{i',...}, v_{i'+k}$ . Σχεδιάστε

έναν αλγόριθμο που βρίσκει ένα ζεύγος υποσυμβολοσειρών με μέγιστη βαθμολογία.

### Πρόβλημα 6.27

Για μια παράμετρο  $k$ , υπολογίστε την καθολική στοίχιση δύο συμβολοσειρών, με τον περιορισμό ότι η στοίχιση περιέχει το πολύ  $k$  κενά (μπλοκ με συνεχόμενες προσθαφαρέσεις).

Οι νουκλεοτιδικές αλληλουχίες γράφονται μερικές φορές με αλφάριθμο πέντε χαρακτήρων: A, T, G, C, και N, όπου το N συμβολίζει ένα απροσδιόριστο νουκλεοτίδιο (στην ουσία ένα σύμβολο μπαλαντέρ). Οι βιολόγοι μπορούν να χρησιμοποιήσουν το N όταν ο προσδιορισμός της αλληλουχίας δεν τους επιτρέπει να συμπεράνουν ξεκάθαρα την ταυτότητα ενός νουκλεοτιδίου σε μια συγκεκριμένη θέση. Μια αλληλουχία με ένα N αναφέρεται ως εκφυλισμένη συμβολοσειρά· για παράδειγμα, η συμβολοσειρά ATTNG αντιστοιχεί σε τέσσερις διαφορετικές ερμηνείες: ATTAG, ATTTG, ATTGG, και ATTCG. Γενικά, μια αλληλουχία με  $k$  απροσδιόριστα νουκλεοτίδια N θα έχει  $4^k$  διαφορετικές ερμηνείες.

### Πρόβλημα 6.28

Με δεδομένες μια μη εκφυλισμένη συμβολοσειρά  $v$  και μια εκφυλισμένη συμβολοσειρά  $w$  που περιέχει το σύμβολο N  $k$  φορές, επινοήστε μια μέθοδο για την εύρεση της καλύτερης ερμηνείας της  $w$  σύμφωνα με τη  $v$ . Δηλαδή, από όλες τις  $4^k$  πιθανές ερμηνείες της  $w$ , βρείτε την  $w'$  με τη μέγιστη βαθμολογία στοίχισης  $s(w', v)$ .

### Πρόβλημα 6.29

Με δεδομένες μια μη εκφυλισμένη συμβολοσειρά  $v$  και μια εκφυλισμένη συμβολοσειρά  $w$  που περιέχει το σύμβολο N  $k$  φορές, επινοήστε μια μέθοδο για την εύρεση της χειρότερης ερμηνείας της  $w$  σύμφωνα με τη  $v$ . Δηλαδή, από όλες τις  $4^k$  πιθανές ερμηνείες της  $w$ , βρείτε την  $w'$  με την ελάχιστη βαθμολογία στοίχισης  $s(w', v)$ .

### Πρόβλημα 6.30

Με δεδομένες δύο συμβολοσειρές  $v_1$  και  $v_2$ , εξηγήστε πώς μπορεί να δημιουργηθεί μια συμβολοσειρά  $w$  που ελαχιστοποιεί την παράσταση

$$|d(v_1, w) - d(v_2, w)|$$

έτσι ώστε να ισχύει

$$d(v_1, w) + d(v_2, w) = d(v_1, v_2)$$

όπου  $d(\cdot, \cdot)$  είναι η απόσταση μετασχηματισμού μεταξύ δύο συμβολοσειρών.

βλημα, αφού η στοίχιση που προκύπτει ενδέχεται να αντιστοιχεί σε επικαλυπτόμενες υπο-  
συμβολοσειρές.

### Πρόβλημα 6.36

Σχεδιάστε έναν αλγόριθμο για το πρόβλημα της Βέλτιστης Μη Ακριβούς Επανά-  
ληψης.

Στο πρόβλημα της Χιμαιρικής Στοίχισης, δίνονται μια συμβολοσειρά  $\mathbf{v}$  και ένα σύνολο συμβο-  
λοσειρών  $\{\mathbf{w}_1, \dots, \mathbf{w}_N\}$ , και πρέπει να βρεθεί το  $\max_{1 \leq i, j \leq N} s(\mathbf{v}, \mathbf{w}_i \circ \mathbf{w}_j)$ , όπου  $\mathbf{w}_i \circ \mathbf{w}_j$  είναι η συ-  
νένωση των  $\mathbf{w}_i$  και  $\mathbf{w}_j$  και το  $s(\cdot, \cdot)$  συμβολίζει τη βαθμολογία της βέλτιστης καθολικής στοίχι-  
σης.

### Πρόβλημα 6.37

Επινοήστε έναν αποδοτικό αλγόριθμο για το πρόβλημα της Χιμαιρικής Στοίχισης.

Ένας ιός μολύνει ένα βακτήριο και τροποποιεί μια διεργασία αναδιπλασιασμού στο βακτήριο  
προσθέτοντας

- σε κάθε A, ένα πολυA με μήκος από 1 έως 5·
- σε κάθε C, ένα πολυC με μήκος από 1 έως 10·
- σε κάθε G, ένα πολυG με τυχαίο μήκος  $\geq 1$ ·
- σε κάθε T, ένα πολυT με τυχαίο μήκος  $\geq 1$ .

Δεν επιτρέπονται κενά ή άλλες προσθήκες στο DNA που έχει τροποποιηθεί από τον ιό. Για  
παράδειγμα, η αλληλουχία AAATAAAAGGGGCCCTTTTCC αποτελεί μολυσμένη έκδοση της  
ATAGCTC.

### Πρόβλημα 6.38

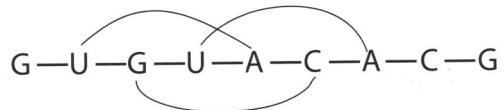
Με δεδομένες τις αλληλουχίες  $\mathbf{v}$  και  $\mathbf{w}$ , περιγράψτε έναν αποδοτικό αλγόριθμο  
που θα προσδιορίσει αν η  $\mathbf{v}$  αποτελεί μολυσμένη έκδοση της  $\mathbf{w}$ .

### Πρόβλημα 6.39

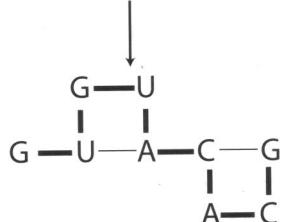
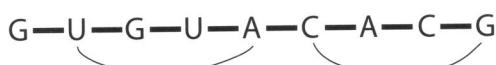
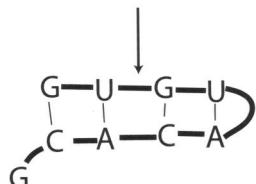
Υποθέστε τώρα ότι για κάθε νουκλεοτίδιο (A, C, G, T), ο ιός είτε θα αφαιρέσει ένα  
γράμμα είτε θα προσθέσει μια ακολουθία επανάληψης του γράμματος με τυχαίο  
μήκος. Διατυπώστε έναν αποδοτικό αλγόριθμο που ανιχνεύει αν η  $\mathbf{v}$  θα μπορού-  
σε να αποτελεί μολυσμένη έκδοση της  $\mathbf{w}$  κάτω από αυτές τις συνθήκες.

### Πρόβλημα 6.40

Ορίστε την ομοαφαίρεση ως την ενέργεια που αφαιρεί (διαγράφει) μια ακολου-  
θία επανάληψης του ίδιου νουκλεοτιδίου και την ομοπροσθήκη ως την ενέργεια  
που προσθέτει (εισάγει) μια ακολουθία επανάληψης του ίδιου νουκλεοτιδίου.  
Για παράδειγμα, η αλληλουχία ACAAAAAAGCTTTA προκύπτει από την  
ACGCTTTA με ομοπροσθήκη μιας ακολουθίας με έξι A, ενώ η αλληλουχία



(α) Εναλλασσόμενοι δεσμοί



(β) Μη εναλλασσόμενοι δεσμοί

**Σχήμα 6.30** Εναλλασσόμενοι και μη εναλλασσόμενοι δεσμοί στην αναδίπλωση RNA.

### Πρόβλημα 6.51

Αναπτύξτε έναν αλγόριθμο δυναμικού προγραμματισμού για την εύρεση του μεγαλύτερου συνόλου μη εναλλασσόμενων δεσμών, όταν δίνεται η αλληλουχία RNA.

Το ανθρώπινο γονιδίωμα μπορεί να θεωρηθεί ως μια συμβολοσειρά με  $n$  ( $\approx 3$  δισεκατομμύρια) νουκλεοτίδια, η οποία έχει διαμεριστεί σε υποσυμβολοσειρές που αναπαριστούν τα χρωμοσώματα. Ωστόσο, οι βιολόγοι χρησιμοποιούσαν για πολλές δεκαετίες μια διαφορετική αναπαράσταση του γονιδιώματος με ζώνες, που προκύπτει με την παραδοσιακή οπτική μικροσκοπία. Στο Σχήμα 6.31 παρουσιάζονται 48 ζώνες (όπως φαίνονται στο χρωμόσωμα 4) από τις 862 παρατηρήσιμες ζώνες για ολόκληρο το ανθρώπινο γονιδίωμα. Παρόλο που αρκετοί



**Σχήμα 6.31** Μοτίβα ζωνών στο ανθρώπινο χρωμόσωμα 4.

παράγοντες (π.χ., η τοπική συχνότητα G/C) έχουν θεωρηθεί ότι μπορεί να διέπουν το σχηματισμό αυτών των ζωνών, ο μηχανισμός σχηματισμού τους έχει κατανοθεί ελάχιστα. Μια αντιστοίχιση ανάμεσα στη γονιδιωματική αλληλουχία του ανθρώπου (η οποία ήταν διαθέσιμη μόνο μετά το 2001) και την αναπαράσταση με πρότυπα ζωνών θα χρησίμευε στην εκμετάλλευση γονιδιακών πληροφοριών στο επίπεδο της αλληλουχίας για την αντιμετώπιση ασθενειών που έχουν συσχετιστεί με ορισμένες θέσεις ζωνών. Παρόλα αυτά, δεν έχει βρεθεί μέχρι πρόσφατα κάποια αντιστοίχιση ανάμεσα σε αυτές τις δύο αναπαραστάσεις του γονιδιώματος.

Το πρόβλημα του Εντοπισμού Ζωνών είναι η εύρεση των θέσεων του αρχικού και του τελικού νουκλεοτιδίου για κάθε ζώνη στο γονιδίωμα (για λόγους απλότητας υποθέτουμε ότι όλα τα χρωμοσώματα συνενώνονται για να σχηματίσουν ένα σύστημα συντεταγμένων). Με άλλα λόγια, το παραπάνω πρόβλημα αφορά την εύρεση ενός αύξοντος πίνακα  $start(b)$ , ο οποίος περιέχει τη θέση του αρχικού νουκλεοτιδίου για κάθε ζώνη  $b$  στο γονιδίωμα. Κάθε ζώνη  $b$  αρχίζει στο νουκλεοτίδιο που δίνεται από τον  $start(b)$  και τελειώνει στο  $start(b + 1) - 1$ .<sup>35</sup>

Μια απλοϊκή μέθοδος για την επίλυση του προβλήματος θα ήταν να χρησιμοποιήσουμε τα παρατηρούμενα δεδομένα του εύρους της ζώνης για να υπολογίσουμε τις θέσεις των νουκλεοτιδίων. Όμως, η λύση δεν είναι ακριβής επειδή υποθέτει ότι το εύρος της ζώνης είναι τέλεια συσχετισμένο με το μήκος της σε νουκλεοτίδια. Στην πραγματικότητα, η συσχέτιση είναι συχνά αρκετά μικρή και πρέπει να βρεθεί μια διαφορετική μέθοδος.

Κατά την τελευταία δεκαετία, ο βιολόγοι έχουν διεξαγάγει μεγάλο αριθμό πειραμάτων *FISH* (*fluorescent in situ hybridization, in situ υβριδοποίηση φθορισμού*) που μπορούν να συμβάλλουν στην επίλυση του προβλήματος του Εντοπισμού Ζωνών. Τα δεδομένα FISH αποτελούνται από ζεύγη  $(x, b)$ , όπου  $x$  είναι μια θέση στο γονιδίωμα και  $b$  είναι ο δείκτης της ζώνης που περιέχει τη θέση  $x$ . Τα δεδομένα FISH παρουσιάζουν συχνά πειραματικά σφάλματα, άρα κάποια σημεία FISH ενδέχεται να αλληλοαναιρούνται.

Με δεδομένη τη λύση  $start(b)$  για το πρόβλημα του Εντοπισμού Ζωνών, όπου  $1 \leq b \leq 862$ , ορίζουμε την ποιότητα *FISH* της λύσης ως τον αριθμό των πειραμάτων FISH που υποστηρίζει, δηλαδή τον αριθμό των πειραμάτων FISH  $(x, b)$  έτσι ώστε να ισχύει  $start(b) \leq x \leq start(b + 1)$ .

<sup>35</sup> Για λόγους απλότητας, υποθέτουμε ότι  $start(863) = n + 1$ , γεγονός που συνεπάγεται ότι η τελευταία 862η ζώνη αρχίζει στο νουκλεοτίδιο  $start(862)$  και τελειώνει στο  $n$ .