

Περιεχόμενα

Σημείωμα του επιμελητή	xi
Πρόλογος στην ελληνική έκδοση	xv
Πρόλογος	xix
1 Εισαγωγή	1
1.1 Ομοιότητα, ομολογία και στοίχιση αλληλουχιών	2
1.2 Επισκόπηση του βιβλίου	4
1.3 Πιθανότητες και πιθανοκρατικά μοντέλα	7
1.4 Προτάσεις για περαιτέρω ανάγνωση	17
2 Στοίχιση ανά ζεύγη	18
2.1 Εισαγωγή	18
2.2 Το μοντέλο βαθμολόγησης	21
2.3 Αλγόριθμοι στοίχισης	29
2.4 Δυναμικός προγραμματισμός με πιο περίπλοκα μοντέλα	51
2.5 Αλγόριθμοι ευριστικής στοίχισης	57
2.6 Στοιχίσεις γραμμικού χώρου	62
2.7 Η σημαντικότητα των βαθμολογιών	65
2.8 Εξαγωγή των παραμέτρων της βαθμολόγησης από τα δεδομένα της στοίχισης	74
2.9 Προτάσεις για περαιτέρω ανάγνωση	82
3 Ακολουθίες Markov και κρυμμένα μοντέλα Markov	83
3.1 Αλυσίδες Markov	86
3.2 Κρυμμένα μοντέλα Markov	92
3.3 Εκτίμηση παραμέτρων για HMM	107
3.4 Δομή μοντέλων HMM	117
3.5 Πιο πολύπλοκες ακολουθίες Markov	123
3.6 Σταθερότητα αλγόριθμων HMM	131
3.7 Προτάσεις για περαιτέρω ανάγνωση	134

4	Ζευγαρωτή στοίχιση με τη χρήση HMM	136
4.1	Ζευγαρωτά HMM	137
4.2	Η πλήρης πιθανότητα x, y αθροίζοντας για όλες τις διαδρομές καταστάσεων	146
4.3	Υποβέλτιστη στοίχιση	149
4.4	Η εκ των υστέρων πιθανότητα του x_i να στοιχιστεί με το y_j	154
4.5	Ζευγαρωτά HMM έναντι FSA για αναζητήσεις	158
4.6	Προτάσεις για περαιτέρω ανάγνωση	162
5	Προφίλ HMM για οικογένειες αλληλουχιών	164
5.1	Πίνακες βαθμολόγησης χωρίς κενά	167
5.2	Προσθέτοντας καταστάσεις εισδοχής και απαλοιφής για τη δημιουργία προφίλ HMM	168
5.3	Κατασκευή προφίλ HMM από πολλαπλές στοίχισεις	172
5.4	Αναζήτηση με προφίλ HMM	177
5.5	Παραλλαγές προφίλ HMM για τοπικές στοίχισεις	184
5.6	Περισσότερα για την εκτίμηση πιθανοτήτων	187
5.7	Κατασκευή βέλτιστων μοντέλων	197
5.8	Αποδίδοντας βάρη σε αλληλουχίες εκμάθησης	201
5.9	Προτάσεις για περαιτέρω ανάγνωση	213
6	Μέθοδοι στοίχισης πολλαπλών αλληλουχιών	215
6.1	Τι σημαίνει πολλαπλή στοίχιση;	217
6.2	Βαθμολογώντας τις πολλαπλές στοίχισεις	221
6.3	Πολυδιαστατικός δυναμικός προγραμματισμός	227
6.4	Μέθοδοι προοδευτικής στοίχισης	231
6.5	Πολλαπλή στοίχιση με εκπαίδευση προφίλ κρυμμένων μοντέλων Markov	242
6.6	Προτάσεις για περαιτέρω ανάγνωση	257
7	Κατασκευάζοντας φυλογενετικά δέντρα	258
7.1	Το δέντρο της ζωής	259
7.2	Εισαγωγή στα δέντρα	261
7.3	Κατασκευάζοντας ένα δέντρο από αποστάσεις ανά ζεύγη	265
7.4	Φειδωλότητα	276
7.5	Αξιολογώντας τα δέντρα: το bootstrap	284
7.6	Ταυτόχρονη στοίχιση και φυλογένεση	285
7.7	Προτάσεις για περαιτέρω ανάγνωση	297
7.8	Παράρτημα: Απόδειξη του θεωρήματος neighbour-joining	299

8	Πιθανοκρατικές προσεγγίσεις στη φυλογένεση	303
8.1	Εισαγωγή	303
8.2	Πιθανοκρατικά μοντέλα εξέλιξης	305
8.3	Υπολογίζοντας την πιθανοφάνεια για στοιχίσεις χωρίς κενά	310
8.4	Χρησιμοποιώντας την πιθανοφάνεια για συναγωγή	320
8.5	Προς πιο ρεαλιστικά εξελικτικά μοντέλα	334
8.6	Σύγκριση πιθανοκρατικών και μη πιθανοκρατικών μεθόδων	347
8.7	Προτάσεις για περαιτέρω ανάγνωση	358
9	Μετασχηματιστικές γραμματικές	360
9.1	Μετασχηματιστικές γραμματικές	362
9.2	Κανονικές γραμματικές	367
9.3	Γραμματικές ανεξάρτητες από τα συμφραζόμενα	375
9.4	Γραμματικές εξαρτώμενες από τα συμφραζόμενα	381
9.5	Στοχαστικές γραμματικές	384
9.6	Στοχαστικές ανεξάρτητες από τα συμφραζόμενα γραμματικές για μοντελοποίηση αλληλουχιών	388
9.7	Προτάσεις για περαιτέρω ανάγνωση	397
10	Ανάλυση της δομής του RNA	399
10.1	Το RNA	400
10.2	Πρόβλεψη της δευτεροταγούς δομής του RNA	413
10.3	Μοντέλα συνδιακύμανσης: προφίλ RNA βασισμένα σε SCFG	425
10.4	Προτάσεις για περαιτέρω ανάγνωση	456
11	Το υπόβαθρο των πιθανοτήτων	458
11.1	Κατανομές πιθανοτήτων	458
11.2	Εντροπία	468
11.3	Συναγωγή	476
11.4	Δειγματοληψία	480
11.5	Προσδιορισμός των πιθανοτήτων από καταμετρήσεις	488
11.6	Ο αλγόριθμος μεγιστοποίησης της προσδοκίας	494
	Βιβλιογραφία	499
	Επίμετρο στην ελληνική έκδοση	521
	Γλωσσάρι	557
	Λεξικό συνωνύμων	563
	Ευρετήριο όρων	565

Σημείωμα του επιμελητή

Η παρούσα έκδοση αποτελεί το δεύτερο βιβλίο της σειράς «Μαθηματικά και Πληροφορική» των εκδόσεων Πεδίο και δείχνει με τον καλύτερο τρόπο το εύρος που φιλοδοξεί να καλύψει η νέα αυτή σειρά, την οποία έχω τη χαρά και την τιμή να επιμελούμαι. Πέρα από μεταφράσεις θεμελιωδών συγγραμμάτων, όπως η *Ανάλυση Βιολογικών Αλληλουχιών*, η σειρά στοχεύει εξίσου στην έκδοση πρωτότυπων συγγραμμάτων που θα εμπλουτίσουν σημαντικά την ελληνική βιβλιογραφία.

Το βιβλίο που κρατάτε στα χέρια σας ξεφεύγει από τα αυστηρά όρια των Μαθηματικών και της Πληροφορικής και απευθύνεται στον φοιτητή και τον ερευνητή της Βιολογίας που ενδιαφέρεται για την αλγοριθμική και τη μαθηματική ανάλυση βιολογικών αλληλουχιών, περιλαμβανομένων μεθόδων για τη στοίχιση αλληλουχιών, την πρόβλεψη φυλογενετικών σχέσεων ή την πρόβλεψη της δομής του RNA. Με τη βοήθεια των ηλεκτρονικών υπολογιστών, η προσέγγιση αυτή έχει επεκταθεί σήμερα σε όλους τους τομείς της βιολογικής έρευνας, έχοντας οδηγήσει σε εντυπωσιακά αποτελέσματα και έχοντας δημιουργήσει την επιστημονική περιοχή της Βιοπληροφορικής και της Υπολογιστικής Βιολογίας. Πρόκειται για ένα συναρπαστικό πεδίο γνώσης, με κρίσιμες εφαρμογές, το οποίο έχει πλέον ωριμάσει σε αυτόνομη περιοχή. Δυστυχώς, όμως, διαθέτει ελάχιστα συγγράμματα στην ελληνική βιβλιογραφία, κάτι που μας παρακίνησε να ασχοληθούμε με το βιβλίο αυτό.

Από την άλλη μεριά, το βιβλίο θα αποτελέσει μια ευχάριστη έκπληξη για τον φοιτητή και τον ερευνητή της Πληροφορικής και των Μαθηματικών, καθώς

θα διαπιστώσει τον πρωταρχικό ρόλο που οι επιστήμες αυτές διαδραματίζουν στη Βιοπληροφορική, με τρόπο μάλιστα που συχνά δεν φανταζόμαστε, ξεκινώντας από τον δυναμικό προγραμματισμό και τα πιθανοτικά μοντέλα μέχρι τις γραμματικές ανεξάρτητες συμφοραζομένων και τις ανελίξεις Markov. Για να βοηθήσουμε όσους διαθέτουν βιολογικό υπόβαθρο να κατανοήσουν τις αλγοριθμικές και μαθηματικές έννοιες, αλλά και αντίστροφα, αυτούς που έρχονται από την Πληροφορική και τα Μαθηματικά να καταλάβουν την απαραίτητη βιολογία, έχουμε προσθέσει επεξηγήσεις των σημαντικότερων όρων και, στο τέλος του βιβλίου, ένα Επίμετρο και ένα Γλωσσάρι για μια σύντομη εισαγωγή στη σχετική επιστημονική περιοχή.

Το βιβλίο αυτό αποτελεί τη βίβλο στην ανάλυση γονιδιωματικών και πρωτεϊνικών αλληλουχιών και, παρά την ηλικία του (εκδόθηκε πρώτη φορά το 1998), παραμένει σήμερα ένα κορυφαίο εργαλείο για να εισαγάγει τον προπτυχιακό ή μεταπτυχιακό φοιτητή της Βιοπληροφορικής σε νέες έννοιες, αλλά και κορυφαίο εγχειρίδιο αναφοράς για τον ερευνητή. Οι συγγραφείς υπήρξαν πρωτεργάτες της έρευνας που οδήγησε στα αποτελέσματα που περιγράφονται στα κεφάλαια του βιβλίου, με προεξάρχον παράδειγμα την εφαρμογή των κρυμμένων μοντέλων Markov (Hidden Markov Models) στη βιολογία. Δεν είναι τελικά τυχαίο πως η πρώτη έκδοσή του συμπίπτει περίπου με τις απαρχές της Βιοπληροφορικής ως αυτοδύναμης επιστήμης.

Η αντιμετώπιση βιολογικών προβλημάτων με αλγοριθμικές, μαθηματικές και στατιστικές μεθόδους, ο συνδυασμός ορολογίας από όλες αυτές τις επιστήμες και η πυκνότητα του πρωτότυπου κειμένου προσέθεσαν στη δυσκολία της μετάφρασης. Γι' αυτό τον λόγο αποφασίσαμε να απευθυνθούμε σε εξειδικευμένους συναδέλφους, οι οποίοι ανέλαβαν τις επιμέρους ενότητες κεφαλαίων. Επιθυμώ εδώ να συγχαρώ όλους τους μεταφραστές για το έργο τους και να ευχαριστήσω, για την προσεκτική ανάγνωση ορισμένων κεφαλαίων, τον Θάνο Παπαθανασίου και τον Αναξαγόρα Φωτόπουλο, αποφοίτους του Διδρυματικού Προγράμματος Μεταπτυχιακών Σπουδών «Τεχνολογίες Πληροφορικής στην Ιατρική και τη Βιολογία». Καθώς η ελληνική βιβλιογραφία έχει αρκετά κενά ή αντιφάσεις στη σχετική ορολογία, πραγματοποιήσαμε τις επιλογές μας με γνώμονα την πίστη στο πρωτότυπο και την ορολογία που παραδοσιακά έχει επικρατήσει στην ελληνική βιβλιογραφία. Ζητάμε την κατανόηση του αναγνώστη για τυχόν αβλεψίες.

Τέλος, συγχαρητήρια αξίζουν και στις εκδόσεις Πεδίο για τη στήριξη, με αυτό τον τρόπο, της επιστημονικής δραστηριότητας, η οποία αποτελούσε ανέκαθεν πολυτέλεια στη χώρα μας, πολύ περισσότερο στους χαλεπούς καιρούς που βιώνουμε.

Ιωάννης Ζ. Εμίρης
Καθηγητής, Τμήμα Πληροφορικής και Τηλεπικοινωνιών
Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
Σεπτέμβριος 2015

Πρόλογος στην ελληνική έκδοση

Η σύγχρονη αντίληψη της οργάνωσης της ζωής δεν αρκείται στη μέχρι σήμερα περιγραφή της ως ενός μοναδικού φαινομένου βιοχημείας, κληρονομικότητας και εξέλιξης της μέσω της φυσικής επιλογής, αλλά αναγνωρίζει πλέον την ανάγκη να προσεγγίζει τη λογική της ζωής ως ενός *πολύπλοκου συστήματος διαχείρισης πληροφοριών*. Με την ανάπτυξη των «ομικών» τεχνολογιών, που επιτάχυναν δραματικά τη δυνατότητα αλληλούχισης του DNA, και τη μεγάλης κλίμακας ανάλυση πληθυσμών RNA και πρωτεϊνών σε συγκεκριμένα κύτταρα και ιστούς, ήρθε και η ανάγκη της ολοκληρωμένης καταγραφής και κατανόησης της ζωής σε επίπεδο συστήματος. Είναι πλέον σαφές ότι τα μακρομοριακά συστατικά της ζωής δεν δρουν μεμονωμένα και γραμμικά, αλλά αλληλεπιδρούν συνθετικά και διαχειρίζονται τη ροή πληροφοριών με τη μορφή δυναμικών, πολύπλοκων και προσαρμοζόμενων δικτύων. Τέτοια δίκτυα πληροφοριών δημιουργούνται δυναμικά μέσα από μοριακούς κώδικες, μεταφέρουν μηνύματα από και προς το περιβάλλον των κυττάρων και των οργανισμών και διαμορφώνουν τη φυσιολογία και ομοιόσταση του συστήματος.

Η πολυπλοκότητα αυτού του βιολογικού «αλληλεπιδρώματος» (interactome) είναι τεράστια και τα διαθέσιμα δεδομένα ήδη τόσο πολλά, που ακόμα και η αποθήκευσή τους αποτελεί φλέγον υπολογιστικό πρόβλημα. Ο Sydney Brenner, στη διάλεξή του για το βραβείο Νομπέλ το 2002, είπε πολύ εύστοχα ότι *«κολυμπάμε σε μια θάλασσα δεδομένων, αλλά διψάμε για νέα γνώση»*. Η σύγχρονη επιστήμη αδυνατεί προς το παρόν να διαχειριστεί τον τεράστιο αυτό όγκο πληροφορίας με αποτελεσματικό και συστηματικό τρόπο, και

αυτό ακριβώς το πρόβλημα έρχεται να αντιμετωπίσει η Βιοπληροφορική. Η Βιολογία σήμερα δεν νοείται πλέον χωρίς τη συμβολή της Πληροφορικής. Οι δικές μου πρώτες επαφές με τη μεθοδολογία και τα εργαλεία της Πληροφορικής έγιναν στα πρώτα χρόνια της ερευνητικής μου δραστηριότητας ως βιολόγου (αρχές δεκαετίας του 1980), όταν χρησιμοποίησα τις «διάτρητες κάρτες» των πρώιμων ηλεκτρονικών υπολογιστών που ήταν διαθέσιμοι τότε, επιχειρώντας να προβλέψω τη δευτεροταγή δομή μιας πρωτεΐνης από την πρωτοταγή αλληλουχία των αμινοξέων της. Οι δυνατότητες που προσέφερε η Πληροφορική για την προσέγγιση θεμελιωδών βιολογικών ερωτημάτων ήταν ήδη από τότε προφανείς. Το βιβλίο που κρατάτε στα χέρια σας αποτελεί μία από τις πρώτες προσπάθειες καταγραφής και παρουσίασης μεθοδολογιών για την υπολογιστική ανάλυση αλληλουχιών και είναι από μια άποψη ιστορικό, αν αναλογιστούμε ότι τον καιρό που οι DEKM συνέτασσαν τα κεφάλαιά του (μόλις τη δεκαετία του 1990), ολόκληρη η GENBANK, η μεγαλύτερη και πληρέστερη βάση δεδομένων γονιδιωματικών αλληλουχιών, κυκλοφορούσε σε μια σειρά από δισκέτες χωρητικότητας αρχικά 720 kb και κατόπιν 1,5 Mb! Σήμερα, η ίδια η GENBANK περιέχει 200 δισεκατομμύρια νουκλεοτιδικές βάσεις. Μαζί με τον απαραίτητο σχολιασμό των αλληλουχιών, καταλαμβάνει το ισοδύναμο 10 terabytes. Την ίδια στιγμή, μόνο από αλληλουχημένα πλήρη γονιδιώματα διαθέτουμε περισσότερο από 1 τρισεκατομμύριο βάσεις, αριθμός που διπλασιάζεται περίπου κάθε δεκαοχτώ μήνες! Παρά το γεγονός ότι από τη συγγραφή του βιβλίου έχει μεσολαβήσει (σχεδόν κυριολεκτικά) μια κοσμογονία δεδομένων, τόσο τα εννοιολογικά προβλήματα όσο και οι τεχνικές μεθοδολογίες που αυτό παρουσιάζει παραμένουν επίκαιρα.

Η Βιοπληροφορική σήμερα όμως είναι κάτι περισσότερο από ένα σύνολο τεχνικών εργαλείων για την ανάλυση βιολογικών δεδομένων. Μαζί με την Υπολογιστική Βιολογία αποτελεί μάλλον ένα εννοιολογικό (conceptual) όργανο, έναν τρόπο σκέψης που ορίζει σε μεγάλο βαθμό τη γενικότερη μεθοδολογική προσέγγιση (modus operandi) στην εποχή της βιολογίας των μεγάλων δεδομένων. Στις δεκαετίες που έρχονται θα οδηγηθούμε στην ανάγκη διατύπωσης ενός νέου Επιστημολογικού Παραδείγματος για την κατανόηση της αρχιτεκτονικής της ζωής, μιας νέας θεωρίας της Φυσιολογίας, ενοποιημένης με την υπολογιστική μηχανική των δικτύων και των δυναμικά μεταβαλλόμενων συστημάτων, στα πλαίσια της οποίας οι νέοι επιστήμονες θα κληθούν να διατυπώσουν νέες υποθέσεις και να σχεδιάσουν και ερμηνεύσουν

πειράματα σε ένα νέο πλαίσιο ιδεών και σκέψεων. Έτσι λοιπόν η μετάφραση του βιβλίου των DEKM υποστηρίζει και έναν πιο φιλόδοξο στόχο: την εξοικείωση των νέων βιοεπιστημόνων με τις υπολογιστικές μεθοδολογίες, ώστε να γίνουν αυτοί που θα επινοήσουν την επόμενη γενιά αλγορίθμων και τις προσεγγίσεις εκείνες που στην εποχή της βιολογίας των μεγάλων δεδομένων θα δώσουν την αποφασιστική ώθηση για τον μετασχηματισμό της ανεπεξέργαστης πληροφορίας σε ουσιαστική γνώση και δι' αυτής ίσως στη βαθύτερη κατανόηση της αρχιτεκτονικής της ζωής.

Γιώργος Κόλλιας, Ακαδημαϊκός
Καθηγητής Φυσιολογίας, Ιατρική Σχολή,
Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
Συνεργαζόμενος Ερευνητής, Ερευνητικό Κέντρο Βιοϊατρικών
Επιστημών «Αλέξανδρος Φλέμιγκ»
Οκτώβριος 2015

Πρόλογος

Το 1992, σε ένα συνέδριο στο Snowbird για τα νευρωνικά δίκτυα, ο David Haussler και οι συνεργάτες του στο Πανεπιστήμιο της California στη Santa Cruz (συμπεριλαμβανομένου του A. Krogh) περιέγραψαν τα πρώιμα αποτελέσματά τους στη μοντελοποίηση στοιχίσεων πολλών ταυτόχρονα αλληλουχιών από πρωτεΐνες με πιθανοκρατικά μοντέλα τα οποία ονομάζονται κρυμμένα μοντέλα Markov (hidden Markov models – HMM). Το σχετικό τεχνικό δελτίο που συνέταξαν κυκλοφόρησε ευρύτατα. Κάποια αντίγραφα κατέληξαν στο Εργαστήριο Μοριακής Βιολογίας του Συμβουλίου Ιατρικών Ερευνών (MRC) στο Cambridge, όπου ο R. Durbin και ο G. J. Mitchison μόλις άλλαζαν ερευνητικά ενδιαφέροντα από τη μοντελοποίηση νευρωνικών δικτύων στην ανάλυση γονιδιωματικών αλληλουχιών με H/Y. Παράλληλα, τον ίδιο χώρο είχε επιλέξει πρόσφατα ο S. R. Eddy ως νέος μεταδιδακτορικός ερευνητής με υπόβαθρο στην πειραματική μοριακή γενετική, αλλά με ενδιαφέροντα σχετικά με την ανάλυση σε H/Y. Αργότερα, και ο A. Krogh επισκέφτηκε το Cambridge και παρέμεινε εκεί έναν χρόνο.

Σύντομα, όλοι μας υιοθετήσαμε τις ιδέες της πιθανοκρατικής μοντελοποίησης. Είχαμε πειστεί ότι τα κρυμμένα Μαρκοβιανά μοντέλα και τα στοχαστικά τους ανάλογα στη γραμματική είναι όμορφα μαθηματικά αντικείμενα, ταιριαστά και ικανά να συλλάβουν την κρυμμένη πληροφορία των βιολογικών αλληλουχιών. Έτσι, η ομάδα της Santa Cruz και η ομάδα του Cambridge ανέπτυξαν, ανεξάρτητα η μία από την άλλη, δύο πακέτα λογισμικού για την ανάλυση αλληλουχιών με HMM, τα οποία διέθεσαν δωρεάν. Ξεχωριστά επί-

σης επέκτειναν τις μεθόδους HMM για να τις εφαρμόσουν στη στοχαστική και ανεξάρτητη από τα συμφοραζόμενα γραμματική ανάλυση των δευτεροταγών δομών του RNA. Μια άλλη ομάδα, στο Εργαστήριο Αεριοπροώθησης (JPL) του Caltech, με επικεφαλής τον Pierre Baldi, περίπου την ίδια εποχή εμπνεύστηκε, και αυτή, από την εργασία που παρουσιάστηκε στο συνέδριο του Snowbird και εργάστηκε πάνω σε προσεγγίσεις HMM.

Προς το τέλος του 1995 θεωρήσαμε ότι είχαμε συγκεντρώσει αρκετή εμπειρία στις τεχνικές της πιθανοκρατικής μοντελοποίησης. Από την άλλη, νιώθαμε ότι μικρό μόνο κομμάτι από την εργασία μας σχετικά με το θέμα αυτό είχε παρουσιαστεί στην επιστημονική κοινότητα. Τα HMM προκαλούσαν παντού το ενδιαφέρον, όμως οι περισσότεροι εξακολουθούσαν να τα θεωρούν μαύρα κουτιά των μαθηματικών αντί για φυσικά μοντέλα κατάλληλα για τα προβλήματα της στοίχισης αλληλουχιών. Πολλά από τα καλύτερα άρθρα που περιέγραφαν με λεπτομέρειες τις ιδέες και τις μεθόδους των HMM ήταν από φορμαλιστική άποψη μη προσβάσιμα στους περισσότερους βιολόγους. Παράλληλα, γίνονταν φανερό και σε εμάς και σε διάφορες άλλες ερευνητικές ομάδες ότι οι ίδιες ιδέες μπορούσαν να εφαρμοστούν σε ένα πολύ πιο ευρύ πεδίο βιολογικών θεμάτων, που περιλάμβανε τη μοντελοποίηση της δομής των πρωτεϊνών, την εύρεση/ταυτοποίηση γονιδίων και τη φυλογενετική ανάλυση. Έτσι, στις χριστουγεννιάτικες διακοπές του 1995-1996, παρακινήμενοι λίγο από ματαιοδοξία, λίγο από αφέλεια και λίγο από την ηρεμία των ημερών, αποφασίσαμε να γράψουμε ένα βιβλίο για την ανάλυση των βιολογικών αλληλουχιών με έμφαση στην πιθανοκρατική μοντελοποίηση. Τα δύο τελευταία χρόνια, οι αρχικές μεγάλες βλέψεις μας έδωσαν λοιπόν ως απόσταγμα αυτό που ελπίζουμε ότι είναι ένα πρακτικό βιβλίο.

Πρόκειται για υποκειμενικό βιβλίο γραμμένο από ισχυρογνώμονες συγγραφείς. Δεν αποτελεί ένα βήμα προς βήμα βοήθημα για την ανάλυση αλληλουχιών. Ο κύριος σκοπός μας είναι να προσφέρουμε μια κατανοητή εισαγωγή στις αρχές της ανάλυσης αλληλουχιών και να δείξουμε γιατί θεωρούμε χρήσιμη την πιθανοκρατική προσέγγιση. Προσπαθήσαμε να αποφύγουμε την ενασχόληση με συγκεκριμένα προγράμματα H/Y και αντ' αυτού να επικεντρωθούμε στους αλγόριθμους και στις αρχές πίσω από αυτούς.

Κάναμε προσεκτική επιλογή στις αναφορές μας στις εργασίες άλλων ερευνητών, οι οποίες επηρέασαν τον τρόπο σκέψης μας. Βέβαια, είμαστε σίγουροι ότι παραλείψαμε άλλους, τις εργασίες των οποίων θα έπρεπε να

διαβάσουμε, και λυπόμαστε γι' αυτό. Επίσης, για ένα βιβλίο που αναγκαστικά άπτεται πεδίων με φάσμα από την εξελικτική βιολογία και τη θεωρία πιθανοτήτων έως τη βιοφυσική, είχαμε να αντιμετωπίσουμε τους περιορισμούς του χρόνου, των δυνάμεων αλλά και της ατελούς γνώσης μας για ένα πλήθος θεμάτων, τα οποία τελικά θίξαμε επιδερμικά.

Η υπολογιστική βιολογία είναι διεπιστημονικό πεδίο. Όσοι την ασκούν –και εμείς μαζί– στηρίζονται σε ποικίλα υπόβαθρα μοριακής βιολογίας, μαθηματικών, επιστήμης υπολογιστών και φυσικής. Το κοινό στο οποίο απευθυνόμαστε είναι οποιοσδήποτε πτυχιούχος ή τελειόφοιτος αυτών των επιστημών. Επιδιώξαμε λοιπόν μια συνοπτική και λίγο διαισθητική παρουσίαση, η οποία να μην είναι ούτε απαγορευτικά μαθηματική ούτε υπερβολικά βιολογική-τεχνική.

Θεωρούμε προϋπόθεση ότι οι αναγνώστες είναι ήδη εξοικειωμένοι με τις βασικές αρχές της μοριακής γενετικής, όπως το Κεντρικό Δόγμα της Βιολογίας σύμφωνα με το οποίο από το DNA «φτιάχνεται» RNA και από το RNA πρωτεΐνες και ότι τα νουκλεϊκά οξέα είναι αλληλουχίες (ακολουθίες) από τέσσερα διαφορετικά νουκλεοτίδια, ενώ οι πρωτεΐνες είναι ακολουθίες από 20 διαφορετικά αμινοξέα. Όπου χρειάζεται δίνονται λεπτομερέστερες πληροφορίες γύρω από τη μοριακή γενετική. Προσπαθήσαμε να τις συγκεντρώσουμε στο τέλος κάθε κεφαλαίου και με γενικότερη μορφή στο τέλος του βιβλίου. Το τελευταίο Κεφάλαιο 11 καλύπτει κάποια θέματα της θεωρίας πιθανοτήτων που σχετίζονται με την προηγούμενη ύλη του υπόλοιπου βιβλίου.

Είμαστε ευγνώμονες σε πολλούς ανθρώπους που υπήρξαν πρόθυμοι να ελέγξουν τμήματα του χειρόγραφού μας και μάλιστα σε σύντομο χρονικό διάστημα. Ευχαριστούμε ιδιαίτερα τους Ewan Birney, Bill Bruno, David MacKay, Cathy Eddy, Jotun Hein και Søren Riis. Οι Bret Larget και Robert Mau μας πρόσφεραν πολύ χρήσιμες πληροφορίες σχετικά με τις μεθόδους δειγματοληψίας που χρησιμοποιούσαν για τη φυλογενετική. Ο David Haussler δεν δίστασε να χρησιμοποιήσει ένα πρώιμο χειρόγραφο μας (για το οποίο αισθανόμαστε άβολα) σε ένα σεμινάριο στη Santa Cruz το φθινόπωρο του 1996. Ευχαριστούμε αυτόν και όλη την τάξη του για τις πολύ χρήσιμες παρατηρήσεις που μας έκαναν. Είμαστε επίσης ευγνώμονες στον David που μας ενέπνευσε να ασχοληθούμε με αυτό το πεδίο. Ήταν επίσης χαρά μας να συνεργαστούμε με τους David Tranah και Maria Murphy στον εκδοτικό οίκο του Cambridge και με τη Sue Glover της εκδοτικής SG για την παραγωγή του βιβλίου. Επέ-

δειξαν όλοι εξαιρετική ικανότητα στη διόρθωση και τη στοιχειοθεσία με το σύστημα LATEX, σε ένα βιβλίο φορτωμένο με εξισώσεις, αλγόριθμους και ψευδοκώδικα, καθώς επίσης μεγάλη ανοχή στα άκρως αισιόδοξα αλλά παντελώς ανακριβή χρονοδιαγράμματά μας. Είμαστε σίγουροι ότι μας έχουν ξεφύγει ορισμένα λάθη, θα ήταν όμως πολλά περισσότερα χωρίς τη βοήθεια όλων αυτών των ανθρώπων.

Θέλουμε επίσης να ευχαριστήσουμε όσους χρηματοδότησαν την έρευνα και την εργασία μας γι' αυτό το βιβλίο: τη Wellcome Trust, το Ινστιτούτο Έρευνας του Ανθρώπινου Γονιδιώματος του NIH¹, το NATO, την Eli Lilly & Co, τον οργανισμό Human Frontiers Science Program και το Εθνικό Ίδρυμα Ερευνών της Δανίας. Ευχαριστούμε επίσης τους δικούς μας οργανισμούς: το Κέντρο Sanger (R. Durbin), τη Σχολή Ιατρικής του Πανεπιστημίου της Washington (S. Eddy), το Κέντρο Ανάλυσης Βιολογικών Αλληλουχιών (A. Krogh) και το Εργαστήριο Μοριακής Βιολογίας του Συμβουλίου Ιατρικών Ερευνών (G. Mitchison). Ευχαριστούμε ακόμη τον Sim και την Anne Durbin, που μας παραχώρησαν με προθυμία το σπίτι τους στο Λονδίνο τον Φεβρουάριο του 1997, όπου σε έναν οίστρο γραψίματος και αναθεωρήσεων καταλήξαμε στη σχεδόν τελική μορφή του βιβλίου. Ευχαριστούμε τους φίλους, τις οικογένειες και τις ερευνητικές μας ομάδες για την ανοχή τους όσο εμείς γράφαμε, και τους S. Eddy και A. Mitchison για τα ταξίδια τους στο Λονδίνο. Υποσχόμαστε να μην εμπλακούμε σε καινούρια μεγαλεπίβολα έργα, τουλάχιστον όχι σύντομα.

1. Σ.τ.μ.: National Institutes of Health: Εθνικό Ινστιτούτο Υγείας των ΗΠΑ.