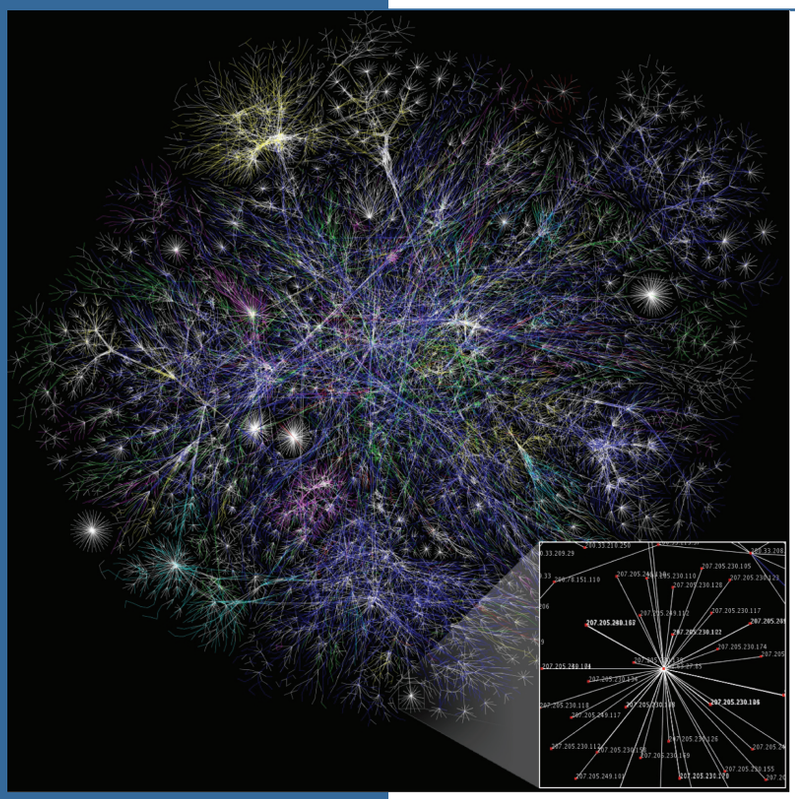
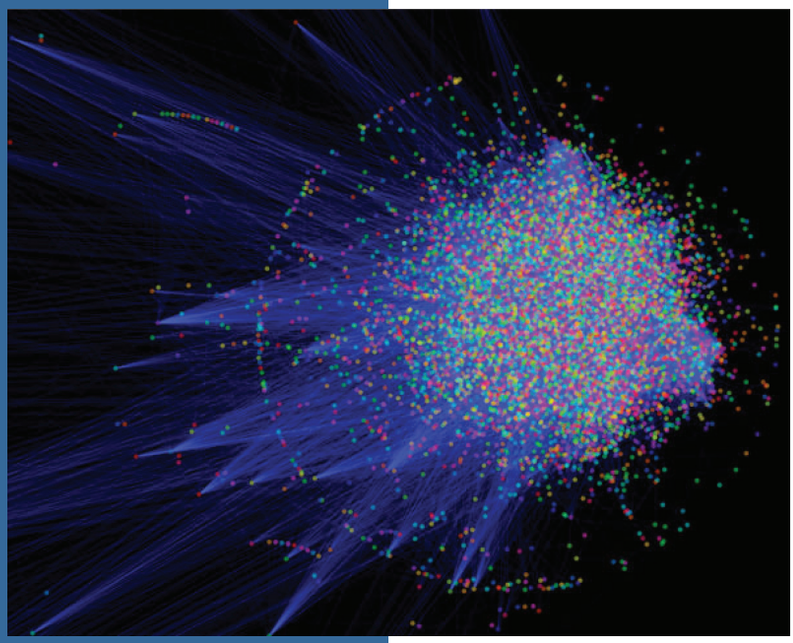
ΜΕΡΟΣ

I

Ανάλυση αλληλουχιών DNA, RNA και πρωτεϊνών





Το βιβλίο αυτό καλύπτει βασικά θέματα της βιοπληροφορικής. Στο Κεφάλαιο 1 θα δούμε μια επισκόπηση των προσεγγίσεων που ακολουθούνται, συμπεριλαμβανομένης της χρήσης λογισμικού το οποίο βασίζεται στο διαδίκτυο και *λογισμικού γραμμής εντολών* (command-line software). Στο Κεφάλαιο 2 θα μάθουμε πώς αποκτούμε πρόσβαση στις αλληλουχίες πρωτεϊνών και νουκλεϊκών οξέων. Στο Κεφάλαιο 3 θα περιγράψουμε πώς στοιχίζουμε τις αλληλουχίες κατά ζεύγη, ενώ στο Κεφάλαιο 4 θα δούμε πώς, χρησιμοποιώντας το BLAST, συγκρίνουμε μια αλληλουχία με τις αλληλουχίες που βρίσκονται καταχωρισμένες στις βάσεις δεδομένων. Κατόπιν, θα δούμε πώς πραγματοποιούμε εξειδικευμένες αναζητήσεις σε βάσεις δεδομένων πρωτεϊνών ή DNA (Κεφάλαιο 5). Στη συνέχεια, θα περιγράψουμε πώς εκτελούμε πολλαπλές στοιχίσεις αλληλουχιών (Κεφάλαιο 6) και πώς απεικονίζουμε αυτές τις στοιχίσεις ως φυλογενετικά δέντρα υιοθετώντας μια εξελικτική αντίληψη (Κεφάλαιο 7).

Στην πάνω εικόνα φαίνονται οι συνδέσεις υπολογιστών στο διαδίκτυο (από την καταχώριση της Wikipedia για το διαδίκτυο), ενώ στην κάτω εικόνα παρουσιάζεται ένας χάρτης των αλληλεπιδράσεων των ανθρώπινων πρωτεϊνών (από την καταχώριση της Wikipedia για τις αλληλεπιδράσεις μεταξύ πρωτεϊνών). Χρησιμοποιώντας τα εργαλεία της βιοπληροφορικής, επιδιώκουμε να κατανοήσουμε τις αρχές της βιολογίας στην κλίμακα ολόκληρου του γονιδιώματος.

Πηγές: (*Πάνω*) Dcrjsr (2002), Creative Commons Attribution 3.0 Unported license. (*Κάτω*) The Opte Project (2006), Creative Commons Attribution 2.5 Generic lisence.

Κεφάλαιο 1

Εισαγωγή

Διεισδύοντας σε τόσο πολλά μυστικά, παύουμε να πιστεύουμε στο άγνωστο. Ωστόσο, αυτό κάθεται ήρεμο και παραμονεύει.

H. L. Mencken

ΜΑΘΗΣΙΑΚΟΙ ΣΤΟΧΟΙ

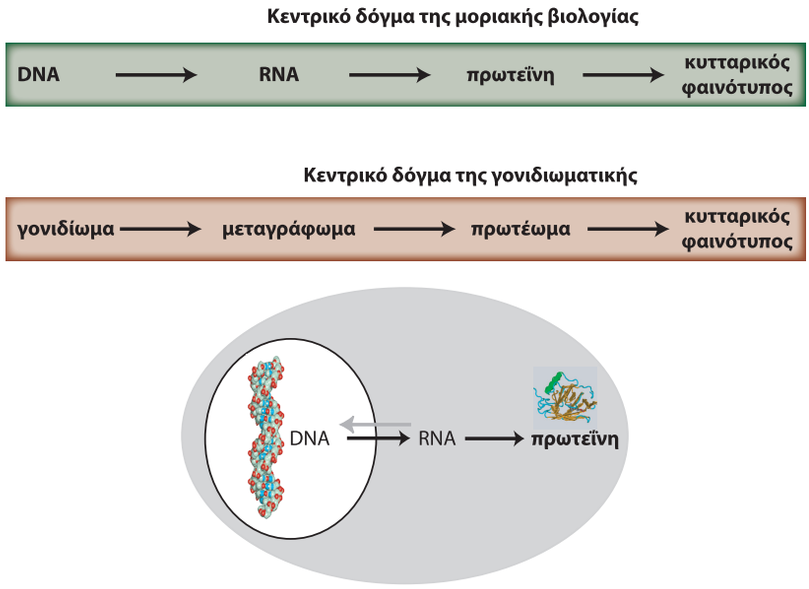
Μετά τη μελέτη αυτού του κεφαλαίου θα πρέπει:

- να κατανοείτε τι είναι η βιοπληροφορική,
- να είστε σε θέση να περιγράψετε το αντικείμενο της βιοπληροφορικής,
- να μπορείτε να εξηγήσετε γιατί η μελέτη των σφαιρινών προσφέρει ένα καλό παράδειγμα που βοηθά να γίνει κατανοητό το αντικείμενο της βιοπληροφορικής,
- να είστε σε θέση να αντιπαραβάλετε τις προσεγγίσεις που βασίζονται σε διαδικτυακό λογισμικό με αυτές που βασίζονται σε λογισμικό γραμμής εντολών.

Η βιοπληροφορική αντιπροσωπεύει ένα νέο πεδίο στη διεπαφή της μοριακής βιολογίας και των υπολογιστών. Ορίζω τη βιοπληροφορική ως τη χρήση αλγορίθμων και βάσεων δεδομένων υπολογιστών για την ανάλυση πρωτεϊνών, γονιδίων, αλλά και για την ανάλυση του συνόλου του DNA ενός οργανισμού (δηλαδή ολόκληρων γονιδιωμάτων). Μία σημαντική πρόκληση στη βιολογία είναι να κατανοήσουμε τις πληροφορίες που κρύβονται στις τεράστιες ποσότητες αφενός μεν βιολογικών αλληλουχιών και αφετέρου δεδομένων που αφορούν τη δομή των πρωτεϊνών. Οι πληροφορίες αυτές παράγονται σήμερα μαζικά από τα προγράμματα αλληλούχισης γονιδιωμάτων, από μελέτες πρωτεωμικής και από άλλες ευρείας κλίμακας αναλύσεις της μοριακής βιολογίας. Στα εργαλεία της βιοπληροφορικής περιλαμβάνονται προγράμματα ηλεκτρονικών υπολογιστών τα οποία βοηθούν στην αποκάλυψη των θεμελιωδών μηχανισμών που βρίσκονται στη βάση ερωτημάτων που θέτει η βιολογία. Μερικά από τα βιολογικά ερωτήματα τα οποία μελετά η βιοπληροφορική σχετίζονται με τη δομή και τη λειτουργία των μακρομορίων, με τις βιοχημικές οδούς, με τις ασθένειες και με την εξέλιξη.

Σύμφωνα με τον ορισμό του αμερικανικού οργανισμού NIH (National Institutes of Health), η βιοπληροφορική είναι «η έρευνα, η ανάπτυξη ή η εφαρμογή υπολογιστικών προσεγγίσεων (συμπεριλαμβανομένων εργαλείων που αφορούν την απόκτηση, την αποθήκευση, την οργάνωση, την ανάλυση και την απεικόνιση πληροφοριών) για τη διεύρυνση της χρήσης βιολογικών, ιατρικών ή συμπεριφορικών δεδομένων». Το συναφές πεδίο της υπολογιστικής βιολογίας είναι «η ανάπτυξη και εφαρμογή μεθόδων ανάλυσης πληροφοριών, θεωρητικών προσεγγίσεων, μαθηματικών μοντέλων και υπολογιστικών τεχνικών προσομοίωσης στη μελέτη βιολογικών, συμπεριφορικών και κοινωνικών συστημάτων». Σύμφωνα με έναν άλλο ορισμό που προέρχεται από το NHGRI (National Human Genome Research Institute), «η βιοπληροφορική είναι ο κλάδος της βιολογίας που ασχολείται με την απόκτηση, αποθήκευση, απεικόνιση και ανάλυση των πληροφοριών που περιέχουν οι αλληλουχίες νουκλεϊκών οξέων και πρωτεϊνών».

Στον ιστότοπο του βιβλίου (http://bioinfbook.org) μπορείτε να βρείτε τους ορισμούς της βιοπληροφορικής σύμφωνα με το NIH και το NHGRI (WebLink 1.1 και WebLink 1.2).



Εικόνα 1.1 Μια πρώτη οπτική γωνία της βιοπληροφορικής αφορά τη διερεύνηση βιολογικών προβλημάτων στο επίπεδο του κυττάρου. Η βιοπληροφορική έχει εξελιχθεί ως επιστημονικός κλάδος, καθώς η συσσώρευση δεδομένων που αφορούν μοριακές αλληλουχίες οδηγεί στον μετασχηματισμό της βιολογίας. Βάσεις δεδομένων όπως αυτές του EMBL (European Molecular Biology Laboratory), η GenBank, η SRA (Sequence Read Archive) και η DDBJ (DNA Database of Japan) χρησιμεύουν ως αποθετήρια για αλληλουχίες συνολικού μεγέθους της τάξης των τετράκις εκατομμυρίων (10^{15}) νουκλεοτιδίων DNA (βλ. Κεφάλαιο 2). Υπάρχουν αντίστοιχες βάσεις δεδομένων για μετάγραφα RNA και για πρωτεΐνες. Μια κύρια κατεύθυνση του πεδίου της βιοπληροφορικής αφορά την άντληση πληροφοριών από τις αλληλουχίες αυτές, με σκοπό την απόκτηση γνώσεων που θα συμβάλουν στην επίλυση ενός ευρέος φάσματος βιολογικών προβλημάτων.

Ο Russ Altman (1998) και οι Altman και Dugan (2003) περιγράφουν δύο οπτικές γωνίες της βιοπληροφορικής. Η πρώτη αντιμετωπίζει τα βιολογικά ερωτήματα στο επίπεδο του κυττάρου και μέσα από το κεντρικό δόγμα της μοριακής βιολογίας, το οποίο περιγράφει τη ροή της πληροφορίας στα βιολογικά συστήματα (**Εικόνα 1.1**). Η δεύτερη θέτει ερωτήματα των οποίων η αντιμετώπιση, πέρα από την ερμηνεία πειραμάτων, απαιτεί επίσης την ανάπτυξη και τη διάδοση ειδικού λογισμικού, την αποθήκευση και την ανταλλαγή δεδομένων, καθώς και την εκτέλεση τυποποιημένων σειρών διαδοχικών ερευνητικών βημάτων.

Η βιοπληροφορική επικεντρώνεται στην ανάλυση αλληλουχιών πρωτεϊνών και νουκλεϊκών οξέων. Συναφή με αυτή επιστημονικά πεδία είναι η γονιδιωματική και η λειτουργική γονιδιωματική. Στόχος της γονιδιωματικής είναι ο προσδιορισμός και η ανάλυση της πλήρους αλληλουχίας DNA (δηλαδή του γονιδιώματος) των οργανισμών. Τα γονίδια, που αποτελούν τμήματα του DNA, μεταγράφονται σε ριβονουκλεϊκό οξύ (RNA), το οποίο, στη συνέχεια, συνήθως μεταφράζεται σε πρωτεΐνη. Η λειτουργική γονιδιωματική μελετά τη γονιδιακή και την πρωτεϊνική λειτουργία στο επίπεδο ολόκληρου του γονιδιώματος. Για τους ανθρώπους και άλλα είδη είναι τώρα δυνατό να χαρακτηριστεί το γονιδίωμα, το σύνολο του RNA (μεταγράφωμα) ή το σύνολο των πρωτεϊνών (πρωτέωμα) ενός ατόμου. Είναι ακόμη δυνατό να προσδιοριστούν οι μεταβολίτες, οι επιγενετικές αλλαγές και ο κατάλογος των συμβιωτικών μικροοργανισμών (το μικροβίωμα) ενός ατόμου (Topol, 2014).

Σκοπός αυτού του βιβλίου είναι να εξηγήσει τόσο τη θεωρία όσο και την εφαρμογή στην πράξη της βιοπληροφορικής και της γονιδιωματικής. Ο σχεδιασμός του αποσκοπεί στο να βοηθήσει τον φοιτητή βιολογίας να χρησιμοποιεί προγράμματα υπολογιστών και βάσεις δεδομένων για την επίλυση βιολογικών προβλημάτων που σχετίζονται με πρωτεΐνες, γονίδια και γονιδιώματα. Η βιοπληροφορική είναι μια συνθετική επιστήμη. Μέσω της μελέτης μεμονωμένων πρωτεϊνών και γονιδίων στοχεύει στην κατανόηση ευρύτερων θεμάτων της βιολογίας, όπως είναι η σχέση της δομής με τη λειτουργία, η ανάπτυξη και οι ασθένειες.

Σε ό,τι αφορά τους επιστήμονες υπολογιστών, αυτό το βιβλίο αποσκοπεί στο να τους εξηγήσει τη σημασία δημιουργίας αλγορίθμων και βάσεων δεδομένων στη βιολογική έρευνα.

1.1 ΟΡΓΑΝΩΣΗ ΤΟΥ ΒΙΒΛΙΟΥ

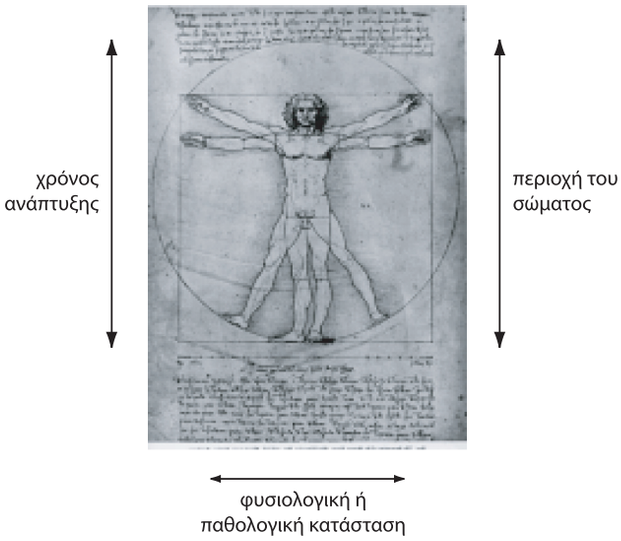
Το βιβλίο διαιρείται σε δύο κύρια μέρη. Στο Μέρος I (Κεφάλαια 2-7) εξηγείται πώς μπορούμε να αποκτήσουμε πρόσβαση σε αλληλουχίες DNA και πρωτεϊνών (Κεφάλαιο 2). Αφού μάθουμε πώς λαμβάνονται οι αλληλουχίες, θα δούμε πώς συγκρίνουμε δύο αλληλουχίες μεταξύ τους (στοίχιση κατά ζεύγη, Κεφάλαιο 3) και πώς συγκρίνουμε μια αλληλουχία με τις αλληλουχίες μιας βάσης δεδομένων (Κεφάλαια 4 και 5) χρησιμοποιώντας κυρίως το Βασικό Εργαλείο Αναζήτησης Τοπικής Στοίχισης ή BLAST (Basic Local Alignment Search Tool). Στη συνέχεια, θα δούμε πώς στοιχίζουμε πολλαπλές αλληλουχίες (πολλαπλή στοίχιση, Κεφάλαιο 6) και πώς μπορούμε να παρουσιάσουμε αυτές τις στοιχισμένες αλληλουχίες DNA και πρωτεϊνών σε φυλογενετικά δέντρα (Κεφάλαιο 7). Έτσι, στο Κεφάλαιο 7 εισάγεται το θέμα της μοριακής εξέλιξης.

Στο Μέρος II περιγράφονται προσεγγίσεις λειτουργικής γονιδιωματικής στη μελέτη αλληλουχιών DNA, RNA και πρωτεϊνών οι οποίες αποσκοπούν στον προσδιορισμό της γονιδιακής λειτουργίας (Κεφάλαια 8-14). Σύμφωνα με το κεντρικό δόγμα της βιολογίας, το DNA μεταγράφεται σε RNA και κατόπιν μεταφράζεται σε πρωτεΐνη. Το Κεφάλαιο 8 θα μας εισαγάγει στα ευκαρυωτικά χρωμοσώματα, ενώ στο Κεφάλαιο 9 παρουσιάζεται η τεχνολογία αλληλούχισης επόμενης γενιάς (με έμφαση στην ανάλυση των πειραματικών αποτελεσμάτων στην πράξη). Στο Κεφάλαιο 10 θα εξετάσουμε τις βιοπληροφορικές προσεγγίσεις στη μελέτη του RNA (τόσο του κωδικού όσο και του μη κωδικού). Θα περιγραφούν η ποσοτικοποίηση του mRNA και η δημιουργία προφίλ γονιδιακής έκφρασης τόσο με μικροσυστοιχίες όσο και με αλληλούχιση του mRNA (RNA-seq). Στο Κεφάλαιο 11 θα παρουσιάσουμε την ανάλυση στην πράξη των πειραματικών δεδομένων που συλλέγονται από μικροσυστοιχίες και RNA-seq. Στη συνέχεια, θα εξετάσουμε τις πρωτεΐνες από τις οπτικές γωνίες της μελέτης πρωτεϊνικών οικογενειών και της ανάλυσης μεμονωμένων πρωτεϊνών (Κεφάλαιο 12). Στο Κεφάλαιο 13 θα μελετήσουμε τη δομή των πρωτεϊνών. Το Μέρος II του βιβλίου ολοκληρώνεται με μια επισκόπηση του ταχέως αναπτυσσόμενου τομέα της λειτουργικής γονιδιωματικής (Κεφάλαιο 14), που ενσωματώνει σύγχρονες προσεγγίσεις για τον χαρακτηρισμό του γονιδιώματος, του μεταγραφώματος και του πρωτεώματος.

1.2 ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ: Η ΜΕΓΑΛΗ ΕΙΚΟΝΑ

Τα πεδία της βιοπληροφορικής και της γονιδιωματικής αντιμετωπίζουν τα βιολογικά ερωτήματα από τρεις οπτικές γωνίες. Η πρώτη είναι από τη σκοπιά του κυττάρου (**Εικόνα 1.1**). Εδώ ακολουθούμε το κεντρικό δόγμα της μοριακής βιολογίας. Η βιοπληροφορική εξετάζει στο σύνολό τους τις αλληλουχίες του DNA (το γονιδίωμα), του RNA (το μεταγράφωμα) και των πρωτεϊνών (το πρωτέωμα). Τα τετράκις εκατομμύρια μοριακών αλληλουχιών προσφέρουν μεγάλες ευκαιρίες, αλλά και μεγάλες προκλήσεις. Μια τέτοια βιοπληροφορική προσέγγιση περιλαμβάνει την αξιοποίηση των βάσεων δεδομένων και την ανάλυση των αλληλουχιών τους με αλγορίθμους υπολογιστών, στο πλαίσιο της διερεύνησης ερωτημάτων που θέτουν η μοριακή και η κυτταρική βιολογία. Η προσέγγιση αυτή αναφέρεται μερικές φορές ως λειτουργική γονιδιωματική και αποτελεί τυπικό παράδειγμα βιοπληροφορικής: τα βιολογικά ερωτήματα προσεγγίζονται σε ποικίλα επίπεδα, που ξεκινούν από τη μελέτη μεμονωμένων γονιδίων/πρωτεϊνών, διευρύνονται στη μελέτη μονοπατιών και δικτύων γονιδίων/πρωτεϊνών και φτάνουν μέχρι την ανάλυση στο επίπεδο ολόκληρου του γονιδιώματος. Στόχος μας είναι να κατανοήσουμε πώς μπορούμε να μελετήσουμε μεμονωμένα γονίδια/πρωτεΐνες, αλλά και ομάδες χιλιάδων γονιδίων/πρωτεϊνών.

Η δεύτερη οπτική γωνία της βιοπληροφορικής αντιμετωπίζει τα βιολογικά ερωτήματα από τη σκοπιά μεμονωμένων οργανισμών (**Εικόνα 1.2**). Κάθε οργανισμός κατά τα διάφορα στάδια της ανάπτυξής του υφίσταται μεταβολές, ενώ οι πολυκύτταροι οργανισμοί παρουσιάζουν διαφοροποιήσεις και μεταξύ των διαφορετικών περιοχών του σώματός τους. Ενώ μερικές φορές αντιμετωπίζουμε τα γονίδια ως στατικές οντότητες που καθορίζουν χαρακτηριστικά όπως το χρώμα των ματιών ή το ύψος, στην πραγματικότητα



Εικόνα 1.2 Μια δεύτερη οπτική γωνία της βιοπληροφορικής αφορά τη διερεύνηση βιολογικών προβλημάτων στο επίπεδο του οργανισμού. Διευρύνοντας τη θεώρησή μας από το επίπεδο του κυττάρου στο επίπεδο του οργανισμού, μπορούμε να εξετάσουμε τη δυναμική του γονιδιώματος (το οποίο περιλαμβάνει το σύνολο των γονιδίων του οργανισμού που μεταγράφονται σε RNA) και των πρωτεϊνών ενός οργανισμού. Για έναν μεμονωμένο οργανισμό, μπορούμε λοιπόν, χρησιμοποιώντας εργαλεία της βιοπληροφορικής, να περιγράψουμε τις αλλαγές που συμβαίνουν κατά την ανάπτυξή του, τις διαφοροποιήσεις που χαρακτηρίζουν τις διάφορες περιοχές του σώματός του και τέλος τις μεταβολές στη φυσιολογία του λόγω μιας ποικιλίας εγγενών και περιβαλλοντικών ερεθισμάτων ή λόγω παθολογικών καταστάσεων.

τα γονίδια ρυθμίζονται δυναμικά και η έκφρασή τους μεταβάλλεται τόσο στον χρόνο (κατά τα διάφορα στάδια της ανάπτυξης του οργανισμού) όσο και στον χώρο (από ιστό σε ιστό). Επίσης, η έκφραση των γονιδίων επηρεάζεται από μια ποικιλία σημάτων (εγγενών και περιβαλλοντικών) και μπορεί να μεταβάλλεται λόγω κάποιας ασθένειας. Πολλά από τα διαθέσιμα εργαλεία της βιοπληροφορικής εστιάζονται στη διερεύνηση ερωτημάτων σε επίπεδο μεμονωμένου οργανισμού: υπάρχουν πολλές υπολογιστικές βάσεις με δεδομένα που αφορούν τη διαφορική έκφραση των γονιδίων σε διαφορετικές φάσεις της ανάπτυξης του οργανισμού ή σε διαφορετικούς ιστούς και συνθήκες. Ένα από τα ισχυρότερα εργαλεία της λειτουργικής γονιδιωματικής είναι η χρήση μικροσυστοιχιών DNA ή RNA-seq για τον ταυτόχρονο προσδιορισμό της έκφρασης χιλιάδων γονιδίων σε βιολογικά δείγματα.

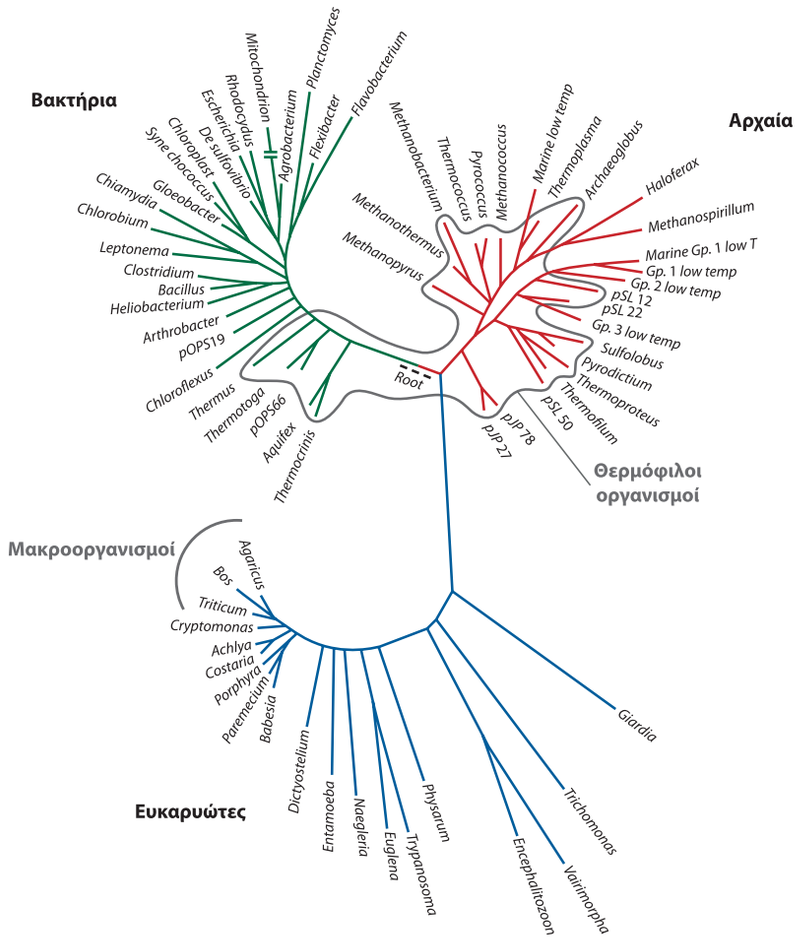
Η τρίτη οπτική γωνία της βιοπληροφορικής αντιμετωπίζει τα βιολογικά ερωτήματα σε μεγαλύτερη κλίμακα, στο επίπεδο του δέντρου της ζωής (**Εικόνα 1.3**). Τα πολλά εκατομμύρια είδη οργανισμών που υπάρχουν κατατάσσονται στους τρεις κύριους κλάδους του δέντρου της ζωής, αυτούς των βακτηρίων, των αρχαίων και των ευκαρυωτών. Οι υπολογιστικές βάσεις μοριακών αλληλουχιών περιέχουν σήμερα αλληλουχίες DNA από ~300.000 διαφορετικά είδη. Υπάρχουν διαθέσιμες οι πλήρεις αλληλουχίες του γονιδιώματος χιλιάδων οργανισμών. Ένα από τα κύρια διδάγματα που αντλούμε από αυτές είναι η θεμελιώδης ενότητα της ζωής σε μοριακό επίπεδο. Μελετώντας αυτές τις αλληλουχίες, διαπιστώνουμε επίσης την ισχύ της συγκριτικής γονιδιωματικής (της σύγκρισης γονιδιωμάτων μεταξύ τους). Μέσω της ανάλυσης των αλληλουχιών DNA μαθαίνουμε πώς τα χρωμοσώματα εξελίσσονται και πώς διαμορφώνονται μέσω διαφόρων διαδικασιών, όπως είναι οι διπλασιασμοί, οι απαλοιφές και οι αναδιατάξεις χρωμοσωμικών τμημάτων, καθώς και οι διπλασιασμοί ολόκληρου του γονιδιώματος (Κεφάλαιο 8).

Στην **Εικόνα 1.4** παρουσιάζεται το περιεχόμενο αυτού του βιβλίου υπό τα τρία παραπάνω πρίσματα της βιοπληροφορικής.

Ένα χαρακτηριστικό παράδειγμα: οι σφαιρίνες

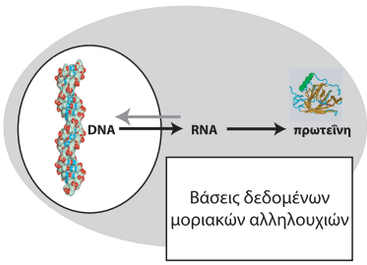
Σε διάφορα σημεία αυτού του βιβλίου θα επικεντρωθούμε στην οικογένεια των γονιδίων των σφαιρινών, που αποτελεί ένα καλό παράδειγμα για να γίνουν αντιληπτές οι έννοιες της βιοπληροφορικής και της γονιδιωματικής. Η οικογένεια των σφαιρινών είναι μία από τις καλύτερα χαρακτηρισμένες στη βιολογία.

- Ιστορικά, η αιμοσφαιρίνη είναι μία από τις πρώτες πρωτεΐνες που μελετήθηκαν, καθώς περιγράφηκε στις δεκαετίες του 1830 και του 1840 από τους Gerardus Johannes Mulder, Justus Liebig και άλλους.



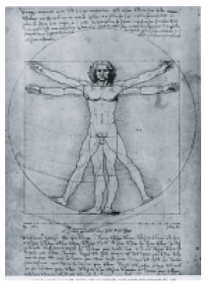
Εικόνα 1.3 Μια τρίτη οπτική γωνία της βιοπληροφορικής αφορά τη διερεύνηση βιολογικών προβλημάτων στο επίπεδο του δέντρου της ζωής. Το πεδίο της βιοπληροφορικής μελετά όλες τις μορφές ζωής στη Γη, δηλαδή και τους τρεις κύριους κλάδους του δέντρου της ζωής: τα βακτήρια, τα αρχαία και τους ευκαρυώτες. Οι ιοί, που βρίσκονται στο όριο του ορισμού της ζωής, δεν απεικονίζονται εδώ. Για όλα τα είδη, η συλλογή και η ανάλυση των μοριακών αλληλουχιών τους μας επιτρέπει να περιγράψουμε το γονιδίωμά τους (το σύνολο του DNA τους). Μπορούμε να μελετήσουμε περαιτέρω τις παραλλαγές που παρατηρούνται μεταξύ των ειδών ή μεταξύ των ατόμων ενός είδους και έτσι να συναγάγουμε την εξελικτική ιστορία της ζωής στη Γη. Προσαρμοσμένη από τις δημοσιεύσεις Barns et al. (1996), Hugenholtz and Pace (1996) και Pace (1997).

- Η μυοσφαιρίνη, μια σφαιρίνη που δεσμεύει το οξυγόνο στον μυϊκό ιστό, ήταν η πρώτη πρωτεΐνη της οποίας προσδιορίστηκε η δομή. Ο προσδιορισμός της δομής της πραγματοποιήθηκε μέσω ανάλυσης με κρυσταλλογραφία ακτίνων Χ (Κεφάλαιο 13).
- Η αιμοσφαιρίνη, ένα τετραμερές τεσσάρων υπομονάδων σφαιρίνης (κυρίως $\alpha_2\beta_2$ σε ενήλικα άτομα), είναι ο κύριος φορέας οξυγόνου στο αίμα των σπονδυλωτών. Η δομή της ήταν επίσης μία από τις πρώτες που χαρακτηρίστηκαν. Η σύγκριση των αλληλουχιών της μυοσφαιρίνης, της α-σφαιρίνης και της β-σφαιρίνης αντιπροσωπεύει ένα από τα πρώτα παραδείγματα πολλαπλής στοίχισης αλληλουχιών (Κεφάλαιο 6) και οδήγησε στην ανάπτυξη των πινάκων αντικατάστασης αμινοξέων που χρησιμοποιούνται για την απόδοση ενός αριθμού (δηλαδή για τον προσδιορισμό ενός σκορ) στην ομοιότητα μεταξύ πρωτεϊνών (Κεφάλαιο 3).



Μέρος I: Βιοπληροφορική: αναλύοντας το DNA, το RNA και τις πρωτεΐνες

Κεφάλαιο 1: Εισαγωγή
Κεφάλαιο 2: Πώς να αποκτήσετε πρόσβαση σε αλληλουχίες
Κεφάλαιο 3: Πώς να συγκρίνετε δύο αλληλουχίες
Κεφάλαια 4 και 5: Πώς να συγκρίνετε μία αλληλουχία με τις αλληλουχίες των βάσεων δεδομένων
Κεφάλαιο 6: Πώς να πραγματοποιήσετε πολλαπλές στοιχίσεις αλληλουχιών
Κεφάλαιο 7: Πώς οι πολλαπλές στοιχίσεις αλληλουχιών παρουσιάζονται ως φυλογενετικά δέντρα



Μέρος II: Λειτουργική γονιδιωματική: από το DNA στο RNA και στην πρωτεΐνη

Κεφάλαιο 8: DNA: Το ευκαρυωτικό χρωμόσωμα
Κεφάλαιο 9: Ανάλυση DNA: αλληλούχιση επόμενης γενιάς
Κεφάλαιο 10: Βιοπληροφορικές προσεγγίσεις στο RNA
Κεφάλαιο 11: Ανάλυση μικροσυστοιχιών και RNA-seq
Κεφάλαιο 12: Ανάλυση πρωτεϊνών και πρωτεϊνικών οικογενειών
Κεφάλαιο 13: Πρωτεϊνική δομή
Κεφάλαιο 14: Λειτουργική γονιδιωματική

Εικόνα 1.4 Επισκόπηση των κεφαλαίων αυτού του βιβλίου.

- Καθώς η τεχνολογία αλληλούχισης του DNA αναπτύχθηκε στη δεκαετία του 1980, οι γενετικοί τόποι της σφαιρίνης στα ανθρώπινα χρωμοσώματα 16 (για την α-σφαιρίνη) και 11 (για τη β-σφαιρίνη) ήταν από τους πρώτους που αλληλουχήθηκαν και αναλύθηκαν. Τα γονίδια της σφαιρίνης ρυθμίζονται περίπλοκα στον χρόνο (κατά τα διάφορα στάδια της ανάπτυξης του οργανισμού, οπότε πραγματοποιείται και η μετάβαση από τις εμβρυϊκές σφαιρίνες στις σφαιρίνες των ενηλίκων), ενώ η έκφρασή τους εμφανίζει επίσης ιστοειδικότητα. Θα επανέλθουμε σε αυτούς τους γενετικούς τόπους όταν θα μιλήσουμε για τον έλεγχο της γονιδιακής έκφρασης (Κεφάλαια 10 και 14).
- Ενώ η αιμοσφαιρίνη και η μυοσφαιρίνη αποτελούν τις πιο καλά χαρακτηρισμένες σφαιρίνες, η οικογένεια των σφαιρινών περιλαμβάνει επίσης διάφορες κατηγορίες φυτικών σφαιρινών, αιμοσφαιρίνες των ασπονδύλων (μερικές από τις οποίες φέρουν πολλαπλές επικράτειες σφαιρίνης εντός ενός πρωτεϊνικού μορίου), βακτηριακές ομοδιμερείς αιμοσφαιρίνες (αποτελούμενες από δύο υπομονάδες σφαιρίνης) και φλαβοαιμοσφαιρίνες που εμφανίζονται σε βακτήρια, αρχαία και μύκητες (βλ. Κεφάλαιο 5, **Εικόνα 5.5**).

1.3 ΟΡΓΑΝΩΣΗ ΤΩΝ ΚΕΦΑΛΑΙΩΝ

Στόχος αυτού του βιβλίου είναι να αποτελέσει τόσο έναν θεωρητικό οδηγό για τη βιοπληροφορική όσο και έναν πρακτικό οδηγό για τη χρήση των βάσεων δεδομένων και των αλγορίθμων υπολογιστών. Σε κάθε κεφάλαιο αναφέρονται οι σχετικές πηγές του διαδικτύου. Στο τέλος των κεφαλαίων υπάρχουν μικρές ενότητες που ονομάζονται «Εποπτική εικόνα», «Παγίδες» και «Συμβουλές προς τους φοιτητές». Η ενότητα «Εποπτική εικόνα» αναφέρεται στον ρυθμό με τον οποίο μεγαλώνει το αντικείμενο στο οποίο εστιάζεται κάθε κεφάλαιο. Για παράδειγμα, στο Κεφάλαιο 2, το οποίο αναφέρεται στην πρόσβαση σε αλληλουχίες DNA και πρωτεϊνών, επισημαίνεται ότι η ποσότητα της πληροφορίας στις υπολογιστικές βάσεις αλληλουχιών DNA αυξάνεται με εκρηκτικό ρυθμό. Από την άλλη πλευρά, το αντικείμενο της στοίχισης αλληλουχιών κατά ζεύγη, που είναι βασικό για ολόκληρο το πεδίο της βιοπληροφορικής (Κεφάλαιο 3), αναπτύχθηκε τις δεκαετίες του 1970 και του 1980. Ωστόσο, ακόμη και για θεμελιώδεις διαδικασίες όπως αυτή της πολλαπλής στοίχισης αλληλουχιών (Κεφάλαιο 6) και αυτή της μοριακής φυλογένεσης (Κεφάλαιο 7), εισάγονται διαρκώς δεκάδες καινοτόμες προσεγγίσεις που συνεχώς προκαλούν βελτιώσεις. Για παράδειγμα, τα *κρυπτομαρκοβιανά μοντέλα* (hidden Markov models) και οι *μπεϋζιανές προσεγγίσεις* (Bayesian approaches) εφαρμόζονται σε ένα ευρύ φάσμα προβλημάτων βιοπληροφορικής.

Η ενότητα «Παγίδες» αφορά ορισμένες συνήθεις δυσκολίες που αντιμετωπίζουν οι βιολόγοι όταν χρησιμοποιούν εργαλεία της βιοπληροφορικής. Μερικά λάθη μπορεί να φαίνονται τετριμμένα, όπως η σύγκριση μιας πρωτεϊνικής αλληλουχίας με αυτές μιας υπολογιστικής βάσης αλληλουχιών DNA. Άλλα είναι λιγότερο προφανή, όπως ορισμένα σφάλματα που εισάγονται από προγράμματα πολλαπλής στοίχισης αλληλουχιών ανάλογα με τον τρόπο με τον οποίο ρυθμίζονται διάφορες από τις παραμέτρους τους. Πράγματι, ενώ το πεδίο της βιοπληροφορικής εξαρτάται σε μεγάλο βαθμό από την ανάλυση αλληλουχιών, είναι σημαντικό να αντιληφθούμε ότι υπάρχουν πολλοί τύποι σφαλμάτων που συνδέονται με την παραγωγή, συλλογή, αποθήκευση και ανάλυση δεδομένων. Σε μια ποικιλία αναζητήσεων και αναλύσεων αντιμετωπίζουμε τα προβλήματα των ψευδώς θετικών και των ψευδώς αρνητικών αποτελεσμάτων.

Σε κάθε κεφάλαιο περιλαμβάνονται ερωτήσεις πολλαπλής επιλογής στην ενότητα «Κουίζ αυτοαξιολόγησης» για να ελέγξετε τις γνώσεις σας. Υπάρχουν επίσης προβλήματα που απαιτούν από εσάς να χρησιμοποιήσετε τις έννοιες που παρουσιάζονται σε κάθε κεφάλαιο. Αυτά τα προβλήματα μπορούν να αποτελέσουν τη βάση ενός εργαστηρίου ηλεκτρονικών υπολογιστών στο πλαίσιο ενός μαθήματος βιοπληροφορικής.

Οι βιβλιογραφικές αναφορές στο τέλος κάθε κεφαλαίου προηγούνται μιας σύντομης παρουσίασης ορισμένων προτεινόμενων άρθρων. Αυτή η ενότητα, υπό τον τίτλο «Προτάσεις για περαιτέρω μελέτη», περιλαμβάνει κλασικά άρθρα που δείχνουν πώς ανακαλύφθηκαν οι επιστημονικές αρχές που περιγράφονται σε κάθε κεφάλαιο. Τα πιο χρήσιμα από τα άρθρα ανασκόπησης και τα ερευνητικά άρθρα επισημαίνονται.

1.4 ΠΡΟΤΑΣΕΙΣ ΓΙΑ ΦΟΙΤΗΤΕΣ ΚΑΙ ΚΑΘΗΓΗΤΕΣ: ΑΣΚΗΣΕΙΣ, ΒΡΕΣ ΕΝΑ ΓΟΝΙΔΙΟ ΚΑΙ ΧΑΡΑΚΤΗΡΙΣΕ ΕΝΑ ΓΟΝΙΔΙΟ

Το βιβλίο αυτό αποτελεί μια εισαγωγή στη βιοπληροφορική (Μέρη Ι και ΙΙ, Κεφάλαια 1-14). Κατά μία έννοια, η βιοπληροφορική υπηρετεί τη βιολογία, διευκολύνοντας τη διατύπωση ερωτήσεων και στη συνέχεια τη διερεύνηση απαντήσεων σε θέματα σχετικά με τις πρωτεΐνες, τα γονίδια και τα γονιδιώματα.

Οι φοιτητές συχνά δείχνουν ιδιαίτερο ενδιαφέρον για κάποιο θέμα, π.χ. ένα γονίδιο, μια φυσιολογική διαδικασία, μια ασθένεια ή ένα γονιδίωμα. Ελπίζω ότι, μέσω της μελέτης των σφαιρινών και άλλων συγκεκριμένων πρωτεϊνών και γονιδίων σε διάφορα σημεία του βιβλίου αυτού, οι φοιτητές θα διδαχθούν πώς να εφαρμόζουν τις αρχές της βιοπληροφορικής για να βρουν απαντήσεις στα δικά τους ερευνητικά ερωτήματα.

Οι ιστοσελίδες που περιγράφονται σε αυτό το βιβλίο περιλαμβάνονται στον ιστότοπό του (http://www.bioinfbook.org) ως WebLinks. Αυτός ο ιστότοπος περιέχει πάνω από 900 διευθύνσεις ιστοσελίδων, οργανωμένες ανά κεφάλαιο. Σε κάθε κεφάλαιο υπάρχουν επίσης αναφορές σε διαδικτυακά αρχεία που είναι διαθέσιμα στον ιστότοπο του βιβλίου. Έτσι, όταν βλέπετε, για παράδειγμα, μια εικόνα ενός φυλογενετικού δέντρου ή μιας στοίχισης αλληλουχιών, μπορείτε να ανακτήσετε εύκολα τα πρωτογενή δεδομένα και να την αναπαραγάγετε μόνοι σας.

Ένα άλλο χαρακτηριστικό του μαθήματος της βιοπληροφορικής στο Πανεπιστήμιο Johns Hopkins είναι ότι μέχρι την τελευταία ημέρα του μαθήματος κάθε φοιτητής υποχρεούται να έχει ανακαλύψει ένα νέο γονίδιο. Ο φοιτητής πρέπει να ξεκινήσει με οποιαδήποτε πρωτεϊνική αλληλουχία τον ενδιαφέρει και να διεξαγάγει έρευνες σε βάσεις δεδομένων για τον εντοπισμό του γονιδιωματικού DNA που κωδικοποιεί μια πρωτεΐνη την οποία κανείς δεν έχει περιγράψει προηγουμένως. Η σχετική μεθοδολογία περιγράφεται λεπτομερώς στο Κεφάλαιο 4 (και συνοψίζεται στο διαδικτυακό αρχείο 4.5 στην ιστοσελίδα http://www.bioinfbook.org/chapter4). Ο φοιτητής επιλέγει λοιπόν το όνομα του γονιδίου και της αντίστοιχης πρωτεΐνης και συλλέγει πληροφορίες για τον οργανισμό από τον οποίο προέρχονται, καθώς και στοιχεία που αποδεικνύουν ότι το συγκεκριμένο γονίδιο δεν έχει περιγραφεί προηγουμένως. Στη συνέχεια, πραγματοποιεί μια πολλαπλή στοίχιση αλληλουχιών με τη νέα πρωτεΐνη (ή το νέο γονίδιο) και δημιουργεί ένα φυλογενετικό δέντρο που δείχνει τη σχέση της με άλλες γνωστές αλληλουχίες.

Κάθε χρόνο, η πραγματοποίηση αυτής της άσκησης προκαλεί μια μικρή ανησυχία σε μερικούς φοιτητές. Στο τέλος όμως όλοι επιτυγχάνουν να τη διεκπεραιώσουν. Ένα πλεονέκτημα αυτής της άσκησης είναι ότι απαιτεί από τον φοιτητή να χρησιμοποιεί στην πράξη τις μεθόδους της βιοπληροφορικής. Πολλοί φοιτητές επιλέγουν ένα γονίδιο (ή μια πρωτεΐνη) σχετικό με τα δικά τους ερευνητικά ενδιαφέροντα.

Η διδασκαλία της βιοπληροφορικής και της γονιδιωματικής αποτελεί μια πρόκληση, δεδομένης της αξιοσημείωτης ποικιλότητας των φοιτητών που διδάσκονται αυτούς τους νέους κλάδους. Κάθε κεφάλαιο παρέχει ένα βασικό υπόβαθρο σχετικά με το θέμα του. Για τους πιο προχωρημένους φοιτητές, στο τέλος κάθε κεφαλαίου παρατίθενται ορισμένες βασικές ερευνητικές δημοσιεύσεις. Σε αυτές τις δημοσιεύσεις χρησιμοποιούνται ποικίλες τεχνικές και η μελέτη τους, σε συνδυασμό με το αντίστοιχο κεφάλαιο του βιβλίου, συμβάλλει στη βαθύτερη κατανόηση του αντικειμένου.

1.5 ΛΟΓΙΣΜΙΚΟ ΒΙΟΠΛΗΡΟΦΟΡΙΚΗΣ: ΔΥΟ ΠΡΟΣΕΓΓΙΣΕΙΣ

Υπάρχουν δύο ουσιωδώς διαφορετικές προσεγγίσεις στη βιοπληροφορική: η μία βασίζεται στη χρήση διαδικτυακού λογισμικού, ενώ η άλλη βασίζεται στη χρήση λογισμικού γραμμής εντολών (**Εικόνα 1.5**). Τα εργαλεία διαδικτυακού λογισμικού, που μερικές φορές ονομάζονται point-and-click, δεν απαιτούν γνώσεις προγραμματισμού και είναι άμεσα προσβάσιμα. Τα εργαλεία λογισμικού γραμμής εντολών ενδέχεται να έχουν μια πιο απότομη καμπύλη μάθησης, αλλά σχεδόν πάντα προσφέρουν περισσότερες επιλογές. Είναι πιο κατάλληλα για την ανάλυση δεδομένων μεγάλης κλίμακας που πολύ συχνά συναντάμε στη βιοπληροφορική. Ακόμη και για τα μικρότερης κλίμακας δεδομένα, η διαχείρισή τους μέσω της γραμμής εντολών μπορεί να σας προσφέρει μεγαλύτερη ευελιξία και ακρίβεια κατά την πραγματοποίηση της εργασίας σας, αλλά και να καταστήσει την έρευνά σας πιο αξιόπιστη, καθώς μπορείτε να τεκμηριώσετε κάθε βήμα που ακολουθήσατε κατά την ανάλυση των δεδομένων σας.

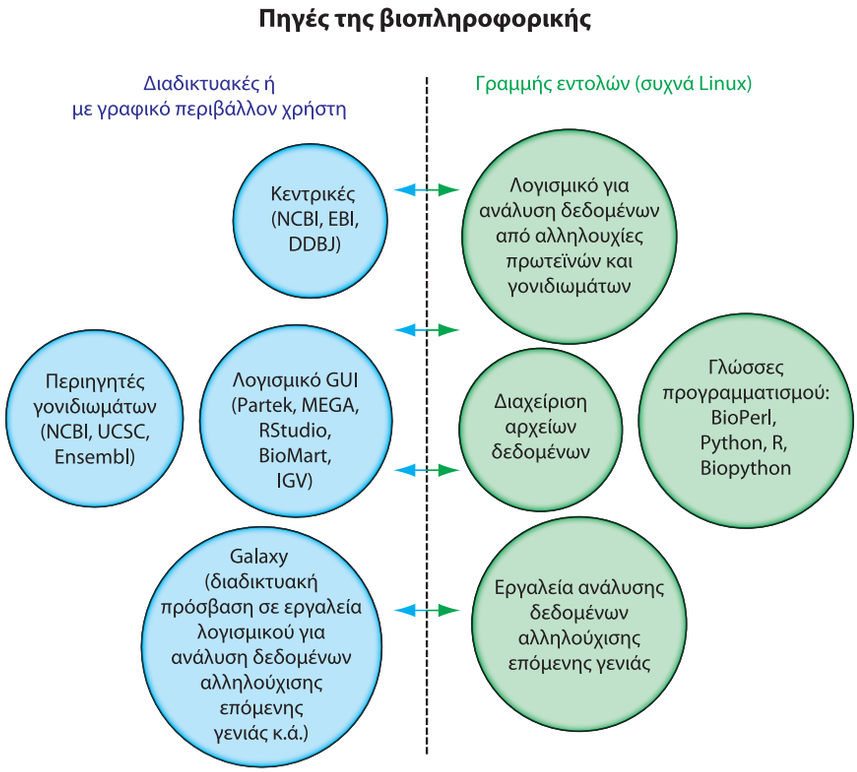
Διαδικτυακό λογισμικό

Το πεδίο της βιοπληροφορικής εξαρτάται σε μεγάλο βαθμό από το διαδίκτυο, μέσω του οποίου εξασφαλίζεται η πρόσβαση σε αλληλουχίες και σε λογισμικό που είναι χρήσιμο για την ανάλυση μοριακών δεδομένων. Μία άλλη χρήσιμη υπηρεσία που προσφέρουν διάφοροι ιστότοποι αφορά την ομαδοποίηση μιας ποικιλίας βιολογικών δεδομένων και πληροφοριών από διαφορετικές πηγές και την παρουσίασή τους με ενοποιημένο τρόπο. Στο βιβλίο αυτό θα περιγράψουμε μια ποικιλία ιστότοπων. Αρχικά, θα επικεντρωθούμε στις κύριες δημόσιες βάσεις δεδομένων, που χρησιμεύουν ως αποθετήρια δεδομένων για το DNA και για τις πρωτεΐνες. Σε αυτές περιλαμβάνονται:

(1) το NCBI (National Center for Biotechnology Information), το οποίο φιλοξενεί την GenBank και άλλες πηγές πληροφορίας,
(2) το EBI (European Bioinformatics Institute),
(3) η Ensembl, στην οποία συμπεριλαμβάνονται *περιηγητές γονιδιωμάτων* (genome browsers) και πηγές πληροφορίας για τη μελέτη δεκάδων γονιδιωμάτων,
(4) ο ιστότοπος βιοπληροφορικής του Πανεπιστημίου της Καλιφόρνιας στη Santa Cruz (UCSC, University of California at Santa Cruz), συμπεριλαμβανομένων ενός περιηγητή γονιδιωμάτων και ενός *περιηγητή πινάκων* (table browser) που επιτρέπει τη διαχείριση δεδομένων για μια ποικιλία διαφορετικών ειδών, τα οποία είναι οργανωμένα σε πίνακες.

Στο βιβλίο αυτό παρουσιάζονται συνολικά σχεδόν 1.000 επιπλέον ιστοσελίδες που σχετίζονται με τη βιοπληροφορική. Τα κύρια πλεονεκτήματα που προσφέρουν οι ιστοσελίδες είναι η

Οι URL αυτών των ιστοσελίδων είναι: NCBI, http://ncbi.nlm.nih.gov (WebLink 1.3), EBI, http://www.ebi.ac.uk/ (WebLink 1.4), Ensembl, http://www.ensembl.org/ (WebLink 1.5) και UCSC, http://genome.ucsc.edu/ (WebLink 1.6). Σχετικά με τον τεράστιο αριθμό των διαθέσιμων βάσεων δεδομένων, μπορείτε να βρείτε πληροφορίες στο ετήσιο τεύχος του Ιανουαρίου του επιστημονικού περιοδικού *Nucleic Acids Research*, http://nar.oxfordjournals.org/ (WebLink 1.7).



Εικόνα 1.5 Πηγές της βιοπληροφορικής. Αριστερά εμφανίζονται πηγές πληροφορίας που βασίζονται στο διαδίκτυο (συχνά αναφέρονται ως point-and-click). Σε αυτές περιλαμβάνονται οι βασικότερες διαδικτυακές πύλες εισόδου σε δεδομένα βιοπληροφορικής [το NCBI (National Center for Biotechnology Information) και το EBI (European Bioinformatics Institute)], οι σημαντικότεροι περιηγητές γονιδιωμάτων [αυτός της Ensembl και αυτός του UCSC (University of California at Santa Cruz)], βάσεις δεδομένων και εξειδικευμένες ιστοσελίδες. Δεξιά εμφανίζονται πηγές πληροφορίας που βασίζονται στη γραμμή εντολών. Σε αυτές περιλαμβάνονται οι γλώσσες προγραμματισμού (όπως η Biopython, η BioPerl και η R) και το λογισμικό γραμμής εντολών (η πρόσβαση στο οποίο κατά κανόνα γίνεται μέσω του λειτουργικού συστήματος Linux). (GUI: Graphical User Interface, Γραφικό Περιβάλλον Χρήστη.)

εύκολη πρόσβαση, η γρήγορη ενημέρωση και η ευκολία στη χρήση (καθώς σε γενικές γραμμές η χρήση τους δεν προϋποθέτει δεξιότητες προγραμματισμού, γνώση της γραμμής εντολών και ικανότητα χρήσης λειτουργικών συστημάτων που βασίζονται στο Linux). Εξαιτίας των χαρακτηριστικών που διαθέτουν, η χρήση τους στην κοινότητα των βιοπληροφορικάριων είναι ιδιαίτερα διαδεδομένη.

Λογισμικό γραμμής εντολών

Τα εργαλεία γραμμής εντολών είναι σαφές πως προσφέρουν πολλά πλεονεκτήματα. Οι σύγχρονες προσεγγίσεις υψηλής απόδοσης στη βιολογία οδηγούν στην ταχεία συσσώρευση δεδομένων τόσο ευρείας όσο και μικρότερης κλίμακας. Η διαχείριση των δεδομένων αυτών απαιτεί περίπλοκες αναλύσεις. Μπορούμε να δούμε το λογισμικό γραμμής εντολών με διάφορους τρόπους.

(1) Το λειτουργικό σύστημα που συνήθως χρησιμοποιείται είναι το Linux, του οποίου το περιβάλλον μοιάζει με του Unix. Το macOS είναι συμβατό με το Linux (και με το POSIX). Αντί-

Η λέξη POSIX είναι το ακρωνύμιο της φράσης Portable Operating System Interface for Unix. Στο POSIX προβλέπονται πρότυπα για τη διατήρηση της συμβατότητας μεταξύ λειτουργικών συστημάτων.

Στην ηλεκτρονική διεύθυνση http://bioinfbook.org/chapter1 μπορείτε να βρείτε πληροφορίες σχετικά με συνδέσμους εκμάθησης του Unix.

θετα, τα Windows, ενώ είναι δημοφιλή, δεν είναι κατάλληλα για την πλειοψηφία των προγραμμάτων γραμμής εντολών. Σε αυτό το βιβλίο θεωρώ ότι ο αναγνώστης δε γνωρίζει πώς να χρησιμοποιεί το Unix. Ξεκινώντας από το Κεφάλαιο 2, παρουσιάζονται ορισμένες βασικές οδηγίες για να εξοικειωθείτε με το Linux. Για τον σκοπό αυτό παρουσιάζονται παραδείγματα εντολών για ένα ευρύ φάσμα λογισμικού.

(2) Στη βιοπληροφορική, για τη διαχείριση των δεδομένων χρησιμοποιούνται συχνά γλώσσες προγραμματισμού, όπως η Perl (ή το παράγωγο της BioPerl – Stajich, 2007), η Python (ή το παράγωγο της Biopython) και η R. Η εκμάθηση τέτοιων γλωσσών είναι σημαντική, καθώς είναι εξαιρετικά χρήσιμο να γράφετε *σενάρια* (scripts) και να τα χρησιμοποιείτε για τη διεκπεραίωση ποικίλων εργασιών. Για εκατοντάδες εφαρμογές βιοπληροφορικής υπάρχουν διαθέσιμα *προγραμματιστικά πακέτα* (modules). Για παράδειγμα, η ιστοσελίδα BioConductor περιλαμβάνει σήμερα >1.000 προγραμματιστικά πακέτα που είναι χρήσιμα για την επίλυση πολλών τύπων προβλημάτων. Καθώς η R έχει απότομη καμπύλη μάθησης, παρέχω προτάσεις για βιβλία, άρθρα και ιστότοπους που μπορείτε να χρησιμοποιήσετε προκειμένου να εξοικειωθείτε μαζί της. Σημειώστε επίσης πως είναι δυνατή η χρήση ενός πακέτου `R` από ανθρώπους που δε γνωρίζουν σε βάθος αυτή τη γλώσσα προγραμματισμού. Για παράδειγμα, στο Κεφάλαιο 8 χρησιμοποιούμε το πακέτο `R Biostrings` για να εξαγάγουμε πληροφορίες σχετικά με τα χαρακτηριστικά των χρωμοσωμάτων, ενώ στο Κεφάλαιο 11 χρησιμοποιούμε πακέτα `R` για να αναλύσουμε δεδομένα γονιδιακής έκφρασης που έχουν ληφθεί με μικροσυστοιχίες ή με αλληλούχιση επόμενης γενιάς. Αφού αρχικά μάθετε να χρησιμοποιείτε μερικά πακέτα, θα είστε στη συνέχεια σε θέση να μάθετε με γοργούς ρυθμούς να χρησιμοποιείτε πολύ περισσότερα.

(3) Η γραμμή εντολών του Unix προσφέρει το Bash, ένα προεπιλεγμένο κέλυφος για λειτουργικά συστήματα Linux και macOS. Σε αυτό το βιβλίο παρουσιάζουμε μια ποικιλία σεναρίων Bash. Το Bash περιλαμβάνει μια σειρά βοηθητικών προγραμμάτων που μπορούν να εκτελέσουν διάφορες εργασίες, όπως την ταξινόμηση των δεδομένων ενός πίνακα, την αντιμετάθεσή του, την καταμέτρηση του αριθμού των γραμμών και των στηλών του, τη συγχώνευση δεδομένων ή την εργασία με κανονικές εκφράσεις. Θα δούμε εντολές Bash, για παράδειγμα, στο **Πλαίσιο 2.3** και στο Κεφάλαιο 9, όπου χρησιμοποιούνται για την ανάλυση δεδομένων αλληλούχισης επόμενης γενιάς.

Όμως ποιο λειτουργικό σύστημα πρέπει να χρησιμοποιείτε; Το Linux είναι απαραίτητο για πολλούς ειδικούς στη βιοπληροφορική, συχνά επειδή χρησιμοποιείται για πρόσβαση σε πολύ μεγάλα σύνολα δεδομένων (π.χ. terabytes δεδομένων) με μεγάλες ποσότητες μνήμης RAM. Για παράδειγμα, σας συνιστώ να εγκαταστήσετε σε έναν φορητό υπολογιστή το Bio-Linux, είτε απευθείας είτε ως *εικονική μηχανή* (virtual machine). Για πολλούς φοιτητές που ξεκινούν να μάθουν βιοπληροφορική, το macOS αποτελεί μια καλή επιλογή επειδή προσφέρει ένα τερματικό που μοιάζει με Unix. Για τους χρήστες των Windows, το λογισμικό Cygwin παρέχει ένα περιβάλλον παρόμοιο με το Unix. Αν έχετε πρόσβαση σε *διακομιστή Linux* (Linux server), μπορείτε να συνδεθείτε μαζί του από περιβάλλον Windows ή macOS χρησιμοποιώντας κατάλληλο λογισμικό, όπως το PuTTY.

Μπορείτε να βρείτε το λογισμικό Bio-Linux 8 (έκδοση του Ιουλίου 2014) στην ηλεκτρονική διεύθυνση http://environmentalomics.org/biolinux/ (WebLink 1.8), το λογισμικό Cygwin στην ηλεκτρονική διεύθυνση http://www.cygwin.com (WebLink 1.9) και το λογισμικό PuTTY στην ηλεκτρονική διεύθυνση http://www.putty.org (WebLink 1.10).

Μπορούμε να κάνουμε μια περαιτέρω διάκριση, αυτή μεταξύ της χρήσης λογισμικού γραμμής εντολών και της χρήσης μιας γλώσσας προγραμματισμού. Η εκμάθηση της Perl, της Python ή άλλων γλωσσών προσφέρει τεράστια οφέλη (Dudley and Butte, 2009). Ωστόσο, ακόμη και αν δε μάθετε να προγραμματίζετε, πρέπει τουλάχιστον να αποκτήσετε βασικές δεξιότητες σχετικά με τον τρόπο απόκτησης, αποθήκευσης, χειρισμού και εξερεύνησης μεγάλων αρχείων. Πολλά αρχεία που χρησιμοποιούνται στη βιοπληροφορική και στη γονιδιωματική είναι πάρα πολύ μεγάλα, με συνέπεια η διαχείρισή τους από διαδικτυακό λογισμικό ή από λογισμικό που διαθέτει *γρα-*

φικό περιβάλλον χρήστη (GUI, Graphical User Interface) να είναι αναποτελεσματική (αν όχι αδύνατη). Πολλά αρχεία που παράγονται από ορισμένα εργαλεία λογισμικού απαιτούν κάποιου τύπου αναδιάρθρωση πριν μελετηθούν περαιτέρω (π.χ. πρέπει να αναλυθούν με πρόσθετο λογισμικό). Για πολλούς φοιτητές καταλήγει να είναι απαραίτητο να μάθουν να χειρίζονται αρχεία μέσω της γραμμής εντολών.

Γεφυρώνοντας τις δύο προσεγγίσεις

Υπάρχουν διαθέσιμα πολλά εργαλεία βιοπληροφορικής, τα οποία στοχεύουν να γεφυρώσουν τις δύο προσεγγίσεις που παρουσιάζονται σε αυτό το βιβλίο (**Πίνακας 1.1**): αυτή του διαδικτυακού λογισμικού και αυτή του λογισμικού γραμμής εντολών. Για παράδειγμα, η NCBI διαθέτει τη διαδικτυακή βάση δεδομένων Entrez, που σας επιτρέπει να εισάγετε έναν όρο αναζήτησης και να αντλείτε σχετικές πληροφορίες. Η NCBI διαθέτει επίσης το EDirect, μια συλλογή λογισμικών γραμμής εντολών για την πρόσβαση στις βάσεις δεδομένων (βλ. Κεφάλαιο 2). Ομοίως, η Ensembl παρέχει τη δυνατότητα πρόσβασης στα δεδομένα της μέσω *διεπαφών προγραμματισμού εφαρμογών* (API, Application Programming Interfaces) της Perl. Ένα άλλο ανάλογο παράδειγμα αφορά το Galaxy, που παρέχει ένα ευρύ φάσμα εργαλείων διαδικτυακού λογισμικού τα οποία, εναλλακτικά, είναι διαθέσιμα και σε μορφή λογισμικού γραμμής εντολών που εκτελείται σε περιβάλλον Linux.

Ποια είναι η καλύτερη προσέγγιση; Ανάλογα με το πρόβλημα που θέλετε να λύσετε, πρέπει να επιλέγετε τα κατάλληλα εργαλεία. Αν εργάζεστε με δεδομένα αλληλούχισης επόμενης γενιάς, είναι απαραίτητο να μάθετε να χρησιμοποιείτε εργαλεία λογισμικού στο λειτουργικό σύστημα Linux. Αν δε γνωρίζετε να χρησιμοποιείτε Linux, θα μπορούσατε να χρησιμοποιήσετε τα πιο προσιτά εργαλεία του Galaxy, ώστε να αρχίσετε να εξοικειώνεστε με τους τύπους δεδομένων και αλγορίθμων που θα συναντήσετε κατά τη μετάβασή σας σε εργαλεία που βασίζονται στο Linux. Αν πραγματοποιείτε φυλογενετικές μελέτες, μπορείτε να ξεκινήσετε με το λογισμικό MEGA για να γνωρίσετε μια ποικιλία προσεγγίσεων, πριν προχωρήσετε στη χρήση λογισμικού γραμμής εντολών για την εκτέλεση μπεϋζιανών αναλύσεων (βλ. Κεφάλαιο 7).

Σε αυτό το βιβλίο χρησιμοποιούνται διάφορα παραδείγματα προκειμένου να βοηθηθεί ο αναγνώστης να γεφυρώσει τις δύο προσεγγίσεις. Στο Κεφάλαιο 8 παρουσιάζονται τόσο η BioMart (μια διαδικτυακή πηγή της Ensembl που διασυνδέει εκατοντάδες βάσεις δεδομένων) όσο και το `biomaRt` (ένα πακέτο R που διεκπεραιώνει αναζητήσεις της BioMart).

Πίνακας 1.1 Συνοπτική παρουσίαση μερικών παραδειγμάτων διαδικτυακού λογισμικού (ή λογισμικού που διαθέτει GUI) και λογισμικού γραμμής εντολών που χρησιμοποιούνται σε διάφορα κεφάλαια αυτού του βιβλίου.

Τμήμα: Κεφάλαιο	Αντικείμενο	Διαδικτυακό λογισμικό ή λογισμικό που διαθέτει GUI	Λογισμικό γραμμής εντολών
I: 2	Πρόσβαση σε πληροφορία	BioMart Genome Workbench	EDirect
I: 3	Στοίχιση κατά ζεύγη	BLAST	BLAST+ Biopython needle (EMBOSS) water (EMBOSS)
I: 4	BLAST	BLAST	BLAST+
I: 5	Αναζήτηση σε βάση δεδομένων	DELTA-BLAST Megablast	HMMER
I: 6	Πολλαπλή στοίχιση αλληλουχιών	Pfam, MUSCLE	MAFFT
I: 7	Φυλογένεση	MEGA	MrBayes
II: 8	Χρωμοσώματα	Galaxy	geecee (EMBOSS) isochore (EMBOSS)
II: 9	Αλληλούχιση επόμενης γενιάς	Galaxy, SIFT, PolyPhen2	SAMTools, tabix, VCFtools
II: 10	RNA	RNAfam, tRNAscan	
II: 11	RNA-seq	Galaxy	affy (R package), RSEM
II: 12	Πρωτεωμική	ExPASy	pepstats (EMBOSS)
II: 13	Δομή πρωτεϊνών	Cn3D, Pymol	psiphi (EMBOSS)
II: 14	Λειτουργική γονιδιωματική	FLink, Cytoscape	

Πίνακας 1.2 Διαγωνισμοί επίλυσης επιστημονικών προβλημάτων του πεδίου της βιοπληροφορικής.

Όνομα/ Ακρωνύμιο	Διαγωνισμός	Κεφάλαιο
Alignathon	Σύγκριση μεθόδων στοίχισης γονιδιωμάτων (Compare whole-genome alignment methods)	6
EGASP	Αξιολόγηση μεθόδων γονιδιωματικού υπομνηματισμού στο ENCODE (ENCODE Genome Annotation Assessment Project)	8
Assemblathon	Σύγκριση της επίδοσης των μεθόδων συναρμολόγησης γονιδιωμάτων (Compare the performance of genome assemblers)	9
GAGE	Αξιολόγηση εδραιωμένων μεθόδων συναρμολόγησης γονιδιωμάτων (Genome Assembly Gold-standard Evaluations)	9
ABRF	Ένωση Μονάδων Βιομοριακών Πηγών (Association of Biomolecular Resource Facilities) – Αξιολόγηση της φωσφορυλίωσης (Assessment of phosphorylation)	12
CASP	Κριτική αξιολόγηση δομικών προβλέψεων (Critical Assessment of Structure Prediction)	13
CAFA	Κριτική αξιολόγηση της πρωτεϊνικής λειτουργίας (Critical Assessment of Protein Function)	14
CAGI	Κριτική αξιολόγηση μεθόδων ερμηνείας της ποικιλότητας του γονιδιώματος (Critical Assessment of Genome Interpretation)	14

Θα δούμε επίσης ότι η κοινότητα της βιοπληροφορικής βελτιώνει συνεχώς το υπάρχον λογισμικό και αναπτύσσει νέες μεθόδους. Υπάρχουν συχνά διαγωνισμοί στους οποίους οι διοργανωτές βραβεύουν την επίλυση κάποιου επιστημονικού προβλήματος, για παράδειγμα τον προσδιορισμό μιας πρωτεϊνικής δομής ή τη συναρμολόγηση ενός γονιδιώματος. Τα μέλη της κοινότητας καλούνται να ανταγωνιστούν μεταξύ τους για να λύσουν το πρόβλημα μέσα σε κάποιο χρονικό διάστημα. Με τη σύγκριση των διαφόρων αποτελεσμάτων είναι δυνατόν να εκτιμηθεί η επίδοση κάθε λογισμικού (με άλλα λόγια ποια από τα ευρήματά του είναι αληθώς ή ψευδώς θετικά ή αρνητικά). Έτσι καθορίζεται η *ευαισθησία* (sensitivity) και η *ειδικότητα* (specificity) κάθε λογισμικού και εκτιμάται ποιο είναι το καταλληλότερο για την επίλυση του επιστημονικού προβλήματος. Σχετικά παραδείγματα παρουσιάζονται στον **Πίνακα 1.2**.

Νέα παραδείγματα σχετικά με την εκμάθηση του προγραμματισμού στη βιοπληροφορική

Είναι εξαιρετική ιδέα να μάθετε μια γλώσσα προγραμματισμού, καθώς θα διευκολύνει τη δουλειά σας στη βιοπληροφορική. Ίσως θελήσετε να εκτελέσετε προγράμματα που είναι γραμμένα σε R ή σε Python (όπως συμβαίνει σε αυτό το βιβλίο) ή μπορεί να θελήσετε να γράψετε τον δικό σας κώδικα προκειμένου να επεξεργαστείτε ορισμένα δεδομένα και να απαντήσετε σε κάποιο ερώτημα. Εκτός από τα διαθέσιμα βιβλία και τα παραδοσιακά μαθήματα, πολλοί ιστότοποι προσφέρουν εξ αποστάσεως μαθήματα ή σεμινάρια. Ο David Searls (2012a, 2014) έχει παρουσιάσει πολλές τέτοιες πηγές του διαδικτύου, στις οποίες περιλαμβάνονται διάφορα Μαζικά Ανοιχτά Διαδικτυακά Μαθήματα (MOOC, Massive Open Online Courses) όπου μπορούν να εγγραφούν δεκάδες χιλιάδες φοιτητές. Στη δημοσίευση Searls (2012b) προτείνονται δέκα κανόνες για την εκμάθηση μέσω διαδικτύου, τους οποίους αναφέρουμε συνοπτικά: να κάνετε ένα σχέδιο, να είστε επιλεκτικοί, να οργανώνετε το μαθησιακό περιβάλλον σας, να μελετάτε το διαθέσιμο υλικό, να κάνετε τις ασκήσεις, να κάνετε τα τεστ αξιολόγησης, να εκμεταλλευτείτε τα πλεονεκτήματα που σας προσφέρονται (π.χ. την ευκολία της όλης διαδικασίας), να αλληλεπιδράτε με τους άλλους, να προσδιορίζετε τα επιτεύγματά σας και να θέτετε ρεαλιστικούς μαθησιακούς στόχους. Αυτοί οι κανόνες ισχύουν επίσης και για την εκμάθηση από ένα σύγγραμμα όπως αυτό.

Εξαιρετικές ιστοσελίδες για την εκμάθηση μιας γλώσσας είναι η Code School (https://www.codeschool.com, WebLink 1.11), η Code Academy (http://www.codecademy.com, WebLink 1.12), Data Camp (https://www.datacamp.com, WebLink 1.13) η Software Carpentry (http://software-carpentry.org, WebLink 1.14), καθώς και η Rosalind (http://rosalind.info/problems/locations/, WebLink 1.15), στην οποία η εκπαίδευση πραγματοποιείται μέσω της επίλυσης προβλημάτων.

Η έρευνα στη βιοπληροφορική πρέπει να είναι αναπαραγώγιμη

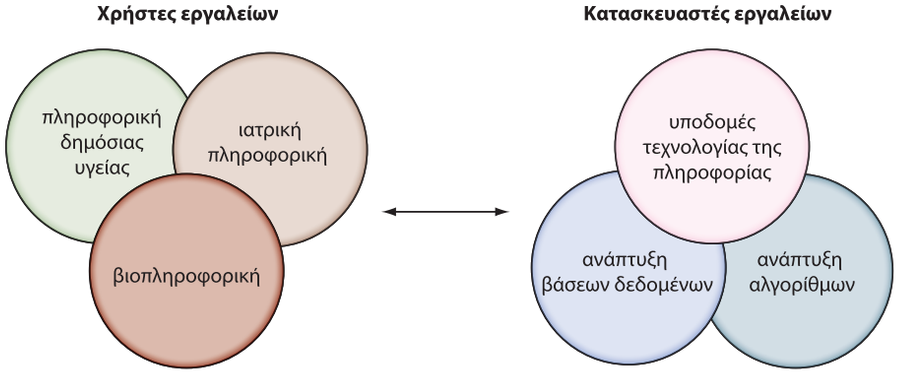
Η επιστήμη από τη φύση της είναι σωρευτική και προοδευτική. Είτε χρησιμοποιείτε διαδικτυακά εργαλεία είτε βασίζεστε στη γραμμή εντολών, πρέπει να διεξάγετε την έρευνά σας κατά τρόπο που να μπορεί να αναπαραχθεί από εσάς αλλά και από άλλους επιστήμονες. Έτσι διευκολύνεται η σωρευτική και προοδευτική φύση της δουλειάς σας. Στον τομέα της βιοπληροφορικής, αυτό σημαίνει τα ακόλουθα:

- Μια πορεία εργασίας πρέπει να είναι καλά αρχειοθετημένη. Αυτό μπορεί να περιλαμβάνει τη διατήρηση αρχείων κειμένου στον υπολογιστή σας, στα οποία έχετε αντιγράψει και επικολλήσει σύνθετες εντολές και διευθύνσεις URL ή έχετε καταγράψει άλλου τύπου δεδομένα προκείμενου να μπορείτε να επαναλάβετε με ακρίβεια τις αναλύσεις που έχετε πραγματοποιήσει. Πολλοί άνθρωποι επιλέγουν να διατηρούν ένα παραδοσιακό εργαστηριακό τετράδιο στο οποίο γράφουν με το χέρι, όμως καθίσταται όλο και πιο έντονη η ανάγκη αυτό να συνοδεύεται από κάποια μορφή ηλεκτρονικής αρχειοθέτησης.
- Για να διευκολύνετε την εργασία σας, οι πληροφορίες που αποθηκεύονται σε έναν υπολογιστή πρέπει να είναι καλά οργανωμένες. Στο **Πλαίσιο 2.3** παρουσιάζουμε μια δημοσίευση (Noble, 2009) στην οποία αναφέρονται οδηγίες για τον τρόπο οργάνωσης των αρχείων σας.
- Τα δεδομένα πρέπει να είναι διαθέσιμα στους άλλους. Ειδικά για την αποθήκευση δεδομένων από ευρείας κλίμακας αναλύσεις, υπάρχουν διαθέσιμα κατάλληλα αποθετήρια (repositories). Σχετικά παραδείγματα αποτελούν στο NCBI το GEO (Gene Expression Omnibus) και το SRA (Sequence Read Archive) και στο EBI το ArrayExpress και το ENA (European Nucleotide Archive).
- Τα μεταδεδομένα (metadata) μπορεί να είναι εξίσου σημαντικά με τα δεδομένα. Τα μεταδεδομένα είναι πληροφορίες σχετικές με τα σετ δεδομένων. Για ένα βακτηριακό γονιδίωμα το οποίο έχει αλληλουχηθεί, τα μεταδεδομένα μπορεί να περιλαμβάνουν την περιοχή από την οποία απομονώθηκε το βακτήριο, τις συνθήκες καλλιέργειάς του και το αν είναι παθογόνο. Για μια μελέτη της γονιδιακής έκφρασης στον ανθρώπινο εγκέφαλο, τα μεταδεδομένα μπορεί να περιλαμβάνουν το μεταθανάτιο διάστημα, το φύλο, κλινικές πληροφορίες και τη μέθοδο απομόνωσης του RNA. Τα μεταδεδομένα παρέχουν κρίσιμες πληροφορίες που επιτρέπουν στον ερευνητή να αναλύσει με στατιστικές μεθόδους την επίδραση διαφόρων παραμέτρων στις μετρήσεις που έχει πραγματοποιήσει.
- Οι βάσεις δεδομένων που χρησιμοποιούνται θα πρέπει να καταγράφονται. Δεδομένου ότι τα περιεχόμενα των βάσεων δεδομένων αλλάζουν με την πάροδο του χρόνου, είναι σημαντικό να καταγράφεται ο αριθμός έκδοσής τους και η ημερομηνία ή οι ημερομηνίες κατά τις οποίες πραγματοποιήθηκε η πρόσβαση σε αυτές.
- Το λογισμικό πρέπει επίσης να καταγράφεται. Για τα συνήθη πακέτα λογισμικού πρέπει να καταγράφεται ο αριθμός έκδοσής τους. Η περαιτέρω αρχειοθέτηση των βημάτων που ακολουθήθηκαν είναι απαραίτητη για να μπορούν άλλοι ερευνητές να επαναλάβουν ανεξάρτητα τις αναλύσεις. Σε μια προσπάθεια να μοιραστούν λογισμικό, πολλοί ερευνητές χρησιμοποιούν αποθετήρια όπως το GitHub.

Το GitHub (https://github.com, WebLink 1.16) είναι το πιο δημοφιλές αποθετήριο για τη διαμοίραση λογισμικού. Προσφέρει στους επιστήμονες πρόσβαση σε συγκεκριμένες εκδόσεις λογισμικού. Φιλοξενεί ελεύθερα προσβάσιμο, καθώς και ιδιωτικό λογισμικό. Στις αρχές του 2015 διέθετε σχεδόν 20 εκατομμύρια λογισμικά και 8 εκατομμύρια χρήστες.

1.6 ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ ΚΑΙ ΑΛΛΑ ΠΕΔΙΑ ΤΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ

Τα τελευταία χρόνια έχει πραγματοποιηθεί μεγάλη ανάπτυξη διαφόρων άλλων τομέων της πληροφορικής, όπως η ιατρική πληροφορική, η πληροφορική της υγειονομικής περίθαλψης, η νοσηλευτική πληροφορική και η πληροφορική βιβλιοθηκών (**Εικόνα 1.6**). Η βιοπληροφορική παρουσιάζει επικάλυψη με αυτούς τους κλάδους, αλλά διακρίνεται από την έμφαση που δίνει στο DNA και σε άλλα βιομόρια. Μπορούμε επίσης να διακρίνουμε τους χρήστες εργαλείων (π.χ. βιολόγους που χρησιμοποιούν λογισμικό βιοπληροφορικής για τη μελέτη της γονιδιακής λειτουργίας ή επιστήμονες του χώρου της υγείας που χρησιμοποιούν ηλεκτρονικά αρ-



Εικόνα 1.6 Χρήστες εργαλείων και κατασκευαστές εργαλείων. Η πληροφορική βρίσκει εφαρμογές σε όλο και περισσότερα επιστημονικά πεδία τα τελευταία χρόνια, με αποτέλεσμα τη δημιουργία νέων κλάδων της, όπως η βιοπληροφορική, η πληροφορική δημόσιας υγείας, η ιατρική πληροφορική και η πληροφορική βιβλιοθηκών. Καθένας από αυτούς τους τομείς ασχολείται με τη συστηματοποίηση και την ανάλυση όλο και μεγαλύτερων σετ δεδομένων. Η βιοπληροφορική και η γονιδιωματική επικεντρώνονται κυρίως στις πρωτεΐνες, στα γονίδια και στα γονιδιώματα.

χεία) από τους κατασκευαστές εργαλείων (π.χ. εκείνους που δημιουργούν υποδομές πληροφορικής ή κατασκευάζουν βάσεις δεδομένων ή γράφουν λογισμικό). Στη βιοπληροφορική, περισσότερο από ό,τι σε άλλους κλάδους της πληροφορικής, οι χρήστες εργαλείων εξελίσσονται σταδιακά σε κατασκευαστές εργαλείων.

1.7 ΣΥΜΒΟΥΛΕΣ ΠΡΟΣ ΤΟΥΣ ΦΟΙΤΗΤΕΣ

Τα πεδία της βιοπληροφορικής και της γονιδιωματικής είναι εξαιρετικά ευρέα. Θα πρέπει να αποφασίσετε ποιου τύπου προβλήματα θέλετε να μελετήσετε και ποιες τεχνικές είναι οι πλέον κατάλληλες για την αντιμετώπισή τους. Στην **Εικόνα 1.5** παρουσιάζεται ένα ευρύ φάσμα διαθέσιμων εργαλείων και προσεγγίσεων. Καθώς προχωράτε στην ανάγνωση αυτού του βιβλίου, είναι πιθανό να καθίσταται όλο και πιο σαφές ποια από τα εργαλεία αυτά είναι τα καταλληλότερα για εσάς. Σας ενθαρρύνω να προσεγγίσετε αυτό το εγχειρίδιο όσο το δυνατόν πιο ενεργά. Όταν συζητάμε για έναν ιστότοπο ή για ένα πακέτο λογισμικού, αδράξτε την ευκαιρία να το εξερευνήσετε σε βάθος.

Το Biostars ξεκίνησε το 2009 από τον Istvan Albert του Penn State University. Επισκεφθείτε το στην ηλεκτρονική διεύθυνση http://www.biostars.org (WebLink 1.17)

Αν χρειαστείτε βοήθεια, έχετε διάφορες επιλογές. Δοκιμάστε να χρησιμοποιήσετε το Biostars, ένα διαδικτυακό φόρουμ στο οποίο μπορείτε μεταξύ άλλων να δημοσιεύσετε ερωτήσεις, να λάβετε απαντήσεις από την κοινότητα, να εξερευνήσετε πηγές διδακτικού υλικού (Parnell et al., 2011). Μέχρι το 2015 περισσότεροι από 16.000 εγγεγραμμένοι χρήστες δημιούργησαν >125.000 αναρτήσεις. Προσπαθήστε να συμμετάσχετε στο Biostars ή σε άλλα φόρουμ βιοπληροφορικής, για να βρείτε άλλους που έχουν παρόμοιες αναζητήσεις με τις δικές σας.

ΠΡΟΤΑΣΕΙΣ ΓΙΑ ΠΕΡΑΙΤΕΡΩ ΜΕΛΕΤΗ

Η δημοσίευση Dudley and Butte (2009) παρέχει έναν εξαιρετικό οδηγό για την ανάπτυξη δεξιοτήτων προγραμματισμού στη βιοπληροφορική (συμπεριλαμβανομένης της χρήσης λογισμικού ανοιχτού κώδικα και του Unix). Υπήρξαν σχετικά λίγες σφαιρικές ανασκοπήσεις του πεδίου της βιοπληροφορικής τα τελευταία πέντε χρόνια, πιθανώς λόγω της διεύρυνσής του. Υπάρχουν ωστόσο χιλιάδες άρθρα ανασκόπησης που καλύπτουν εξειδικευμένα θέματα. Σε καθένα από τα επόμενα κεφάλαια παραπέμπω σε μια σειρά από αυτά.

Το 2011 ο Eric Green, ο Mark Guyer και οι συνεργάτες τους στο NHGRI (National Human Genome Research Institute) δημοσίευσαν το εξαιρετικό άρθρο «Charting a course for genomic medicine from base pairs to

bedside» (Green et al., 2011), στο οποίο περιγράφονται τα επιτεύγματα στη γονιδιωματική και οι προοπτικές της για την επόμενη δεκαετία.

Κάθε Ιανουάριο το περιοδικό *Nucleic Acids Research* δημοσιεύει ένα τεύχος αφιερωμένο στις βάσεις, όπου περιγράφονται διάφορες πηγές πληροφορίας που είναι ιδιαίτερα σημαντικές για τη βιοπληροφορική (Fernández-Suárez et al., 2014). Το περιοδικό αυτό, μέσω της ιστοσελίδας του, παρέχει πρόσβαση σε έναν τεράστιο αριθμό δημοσιεύσεων.

Επισκεφθείτε το τεύχος του NAR με τις βάσεις δεδομένων στην ηλεκτρονική διεύθυνση http://nar.oxfordjournals.org/ (WebLink 1.18).

ΒΙΒΛΙΟΓΡΑΦΙΑ

Altman, R.B. 1998. Bioinformatics in support of molecular medicine. *Proceedings of AMIA Symposium* **1998**, 53-61. PMID: 9929182.

Altman, R.B., Dugan, J.M. 2003. Defining bioinformatics and structural bioinformatics. *Methods of Biochemical Analysis* **44**, 3-14. PMID: 12647379.

Barns, S.M., Delwiche, C.F., Palmer, J.D., Pace, N.R. 1996. Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. *Proceedings of the National Academy of Sciences, USA* **93**(17), 9188-9193. PMID: 8799176.

Dudley, J.T., Butte, A.J. 2009. A quick guide for developing effective bioinformatics programming skills. *PLoS Computational Biology* **5**(12), e1000589. PMID: 20041221.

Fernández-Suárez, X.M., Rigden, D.J., Galperin, M.Y. 2014. The 2014 Nucleic Acids Research Database Issue and an updated NAR online Molecular Biology Database Collection. *Nucleic Acids Research* **42**(1), D1-6. PMID: 24316579.

Green, E.D., Guyer, M.S. 2011. National Human Genome Research Institute. Charting a course for genomic medicine from base pairs to bedside. *Nature* **470**(7333), 204-213. PMID: 21307933.

Hugenholtz, P., Pace, N.R. 1996. Identifying microbial diversity in the natural environment: a molecular phylogenetic approach. *Trends in Biotechnology* **14**, 190-197. PMID: 8663938.

Noble, W.S. 2009. A quick guide to organizing computational biology projects. *PLoS Computational Biology* 5(7), e1000424. PMID: 19649301.

Pace, N.R. 1997. A molecular view of microbial diversity and the biosphere. *Science* **276**, 734-740. PMID: 9115194.

Parnell, L.D., Lindenbaum, P., Shameer, K. *et al*. 2011. BioStar: an online question & answer resource for the bioinformatics community. *PLoS Computational Biology* **7**(10), e1002216. PMID: 22046109.

Searls, D.B. 2012a. An online bioinformatics curriculum. *PLoS Computational Biology* **8**(9), e1002632. PMID: 23028269.

Searls, D.B. 2012b. Ten simple rules for online learning. *PLoS Computational Biology* **8**(9), e1002631. PMID: 23028268.

Searls, D.B. 2014. A new online computational biology curriculum. *PLoS Computational Biology* **10**(6), e1003662. PMID: 24921255.

Stajich, J.E. 2007. An Introduction to BioPerl. *Methods in Molecular Biology* **406**, 535-548. PMID: 18287711.

Topol, E.J. 2014. Individualized medicine from prewomb to tomb. *Cell* **157**(1), 241-253. PMID: 24679539.