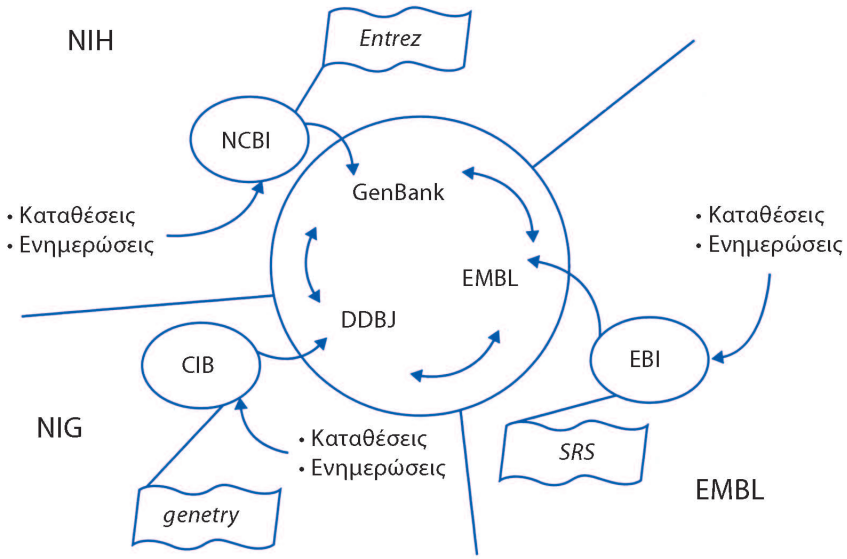
ΕΙΣΑΓΩΓΗ

Εδώ και τρεις δεκαετίες περίπου, καταβάλλονται πυρετώδεις προσπάθειες να κατανοηθεί, στο πιο θεμελιώδες επίπεδο, το υλικό που αποτελεί το βασικό «βιβλίο της ζωής». Οι βιολόγοι (και οι επιστήμονες εν γένει) προσπαθούν να κατανοήσουν τους τρόπους με τους οποίους τα εκατομμύρια ή δισεκατομμύρια των βάσεων στο γονιδίωμα ενός οργανισμού περιέχουν όλες τις πληροφορίες που χρειάζεται το κύτταρο για να εκτελέσει το πλήθος των μεταβολικών διεργασιών που απαιτούνται για να επιβιώσει ο οργανισμός, πληροφορίες που μεταβιβάζονται από γενιά σε γενιά. Για μια βασική κατανόηση του τρόπου με τον οποίο το σύνολο των επιμέρους νουκλεοτιδικών βάσεων κινεί τη ζωική μηχανή, πρέπει να συλλεχθούν, να αποθηκευτούν και να αναλυθούν μεγάλες ποσότητες δεδομένων, με τρόπο ώστε να διευκολύνεται η αναζήτηση και η ανάλυσή τους. Έτσι, έχει καταβληθεί μεγάλη προσπάθεια για τη σχεδίαση και τη συντήρηση βάσεων δεδομένων βιομοριακών ακολουθιών. Αυτές οι βάσεις δεδομένων έχουν επιδράσει σημαντικά στην προώθηση της κατανόησης της βιολογίας, όχι μόνο από υπολογιστικής άποψης αλλά και μέσω της συνδυασμένης χρήσης τους με πειραματικές μελέτες που διενεργούνται στον πάγκο του εργαστηρίου.

Η ιστορία των βάσεων δεδομένων ακολουθιών ξεκίνησε στις αρχές της δεκαετίας του 1960, όταν η Margaret Dayhoff και οι συνεργάτες της, στην Πηγή Πληροφοριών Πρωτεϊνών (Protein Information Resource , PIR), συνέλεξαν όλες τις πρωτεϊνικές ακολουθίες που ήταν γνωστές εκείνη την εποχή και τις δημοσίευσαν σε έντυπη μορφή με το όνομα *Άτλαντας Πρωτεϊνικών Ακολουθιών και Δομών* (*Atlas of Protein Sequence and Structure*, Dayhoff et al., 1978). Όταν συγκεντρώθηκε ένας σημαντικός αριθμός νουκλεοτιδικών ακολουθιών, αυτές προστέθηκαν στον *Άτλαντα*. (Είναι σημαντικό να σημειωθεί ότι, σε αυτήν τη φάση της ιστορίας της, η βιολογία εστίαζε περισσότερο στην αλληλούχηση *πρωτεϊνών* μέσω παραδοσιακών τεχνικών όπως η αποικοδόμηση Edman, και όχι στην αλληλούχηση του DNA.) Καθώς εξελισσόταν ο *Άτλαντας*, συμπεριλήφθηκαν περιγραφικά κείμενα για τις πρωτεϊνικές ακολουθίες, καθώς και πληροφορίες σχετικά με την εξέλιξη πολλών πρωτεϊνικών οικογενειών. Η εργασία αυτή ήταν ουσιαστικά η πρώτη σχολιασμένη (annotated) βάση δεδομένων ακολουθιών, παρ' όλο που ήταν σε έντυπη μορφή. Μέχρι το 1972 τα δεδομένα που περιέχονταν στον *Άτλαντα* είχαν αυξηθεί υπερβολικά και ήταν πλέον εμφανής η ανάγκη δημιουργίας μιας ηλεκτρονικής μορφοποίησης του έργου. Το περιεχόμενο του *Άτλαντα* διανέμονταν ηλεκτρονικά από το PIR σε μαγνητική ταινία (μαγνητοταινία), στην οποία περιλαμβάνονταν κάποια στοιχειώδη προγράμματα για την αναζήτηση και την αξιολόγηση μακρινών εξελικτικών σχέσεων.

Η ανάπτυξη των βάσεων δεδομένων DNA το 1982, που ξεκίνησε στο Ευρωπαϊκό Εργαστήριο Μοριακής Βιολογίας (European Molecular Biology Laboratory, EMBL) και στην οποία συνεισέφερε λίγο αργότερα η GenBank, οδήγησε στην επόμενη φάση της ιστορίας των βάσεων δεδομένων ακολουθιών: στην εκρηκτική αύξηση του αριθμού των βάσεων δεδομένων νουκλεοτιδικών ακολουθιών που ήταν διαθέσιμες στους ερευνητές. Τόσο το EMBL (με έδρα τη Χαϊδελβέργη) όσο και το Εθνικό Κέντρο Πληροφοριών για τη Βιοτεχνολογία (National Center for Biotechnology Information, NCBI), τμήμα της Εθνικής Βιβλιοθήκης για την Ιατρική (National Library of Medicine) των Εθνικών Ινστιτούτων Υγείας (National Institutes of Health) των Η.Π.Α., συνεισέφεραν στην εισαγωγή δεδομένων, δηλαδή στη μεταγραφή και στη μετατροπή των δεδομένων που δημοσιεύονταν σε έντυπα περιοδικά, σε μια ηλεκτρονική μορφοποίηση καταλληλότερη για χρήση από υπολογιστές. Η ιαπωνική βάση δεδομένων DNA (DNA Databank of Japan, DDBJ) συνεργάστηκε στη συλλογή δεδομένων λίγα χρόνια αργότερα. Το 1998, μετά από συνάντηση των τριών αυτών ομάδων [(υπό την ονομασία International Nucleotide Sequence Database Collaboration (Διεθνής Συνεργασία Βάσεων Δεδομένων Νουκλεοτιδικών Ακολουθιών)] συμφωνήθηκε ότι θα χρησιμοποιείται μια κοινή μορφοποίηση για τα στοιχεία δεδομένων που περιέχονται σε μια μοναδιαία εγγραφή (unit record) και ότι κάθε βάση δεδομένων θα ενημερώνει μόνο τις εγγραφές εκείνες που κατατίθενται απευθείας σε αυτή. Σήμερα και τα τρία κέντρα το National Institute of Genetics στη Mishima της Ιαπωνίας, το European Bioinformatics Institute (EBI) στο Hinxton της Βρετανίας και το NCBI στην Bethesda του Maryland] συλλέγουν τα δεδομένα που κατατίθενται άμεσα (βλ. Πλαίσιο 1.3) και τα διανέμουν έτσι, ώστε κάθε κέντρο να έχει αντίγραφα όλων των ακολουθιών, δηλαδή λειτουργούν ως κύριο κέντρο διανομής για τις ακολουθίες αυτές. Ωστόσο, η διαχείριση των εγγραφών γίνεται από τη βάση δεδομένων στην οποία κατατέθηκαν και η ενημέρωση γίνεται μόνο από εκείνη τη βάση δεδομένων. Οι εγγραφές DDBJ/EMBL/GenBank ενημερώνονται αυτομάτως κάθε 24 ώρες και στα τρία κέντρα και, συνεπώς, κάθε ακολουθία που υπάρχει στην DDBJ θα υπάρχει και στο EMBL και στην GenBank και αντιστρόφως (βλ. Εικόνα 1.1).

Κατά παράλληλο τρόπο, τα θεμέλια των βάσεων δεδομένων πρωτεϊνικών ακολουθιών τέθηκαν στις αρχές της δεκαετίας του 1980, όταν ο Amos Bairoch, στο Πανεπιστήμιο της Γενεύης, μετέτρεψε τον *Άτλαντα* του PIR σε μορφοποίηση παρόμοια με αυτήν που χρησιμοποιεί το EMBL για την νουκλεοτιδική βάση δεδομένων. Σε αυτή την πρώτη έκδοση, που ονομάστηκε PIR+, προστέθηκαν συμπληρωματικές πληροφορίες για κάθε πρωτεΐνη, γεγονός που βελτίωσε την αξία της ως μία επιμελημένη και αναλυτικά σχολιασμένη πηγή πληροφοριών για τις πρωτεΐνες. Το καλοκαίρι του 1986, ο Bairoch άρχισε να διανέμει το PIR+ στο US BIONET (πρόδρομο του Διαδικτύου), με το όνομα Swiss-Prot. Εκείνη την εποχή, η βάση δεδομένων περιείχε μόνο 3900 πρωτεϊνικές ακολουθίες, αριθμός που θεωρούν τότε τεράστιο, αλλά απέχει πολύ από τα σημερινά δεδομένα. Καθώς η Swiss-Prot και το EMBL χρησιμοποιούσαν παρόμοια πρότυπα, αναπτύχθηκε μια φυσική συνεργασία μεταξύ αυτών των δύο ευρωπαϊκών ομάδων, η οποία ενισχύθηκε όταν η έδρα και των δύο μεταφέρθηκε στο EBI του EMBL, στο Hinxton της Μ. Βρετανίας. Ένα από τα πρώτα προγράμματα που ανέλαβε η συνεργασία αυτή ήταν η δημιουργία μιας επέκτασης στη Swiss-Prot. Η διαφύλαξη της υψηλής ποιότητας των καταχωρίσεων της Swiss-Prot είναι χρονο-



ΕΙΚΟΝΑ 1.1 Ροή δεδομένων από νέες καταθέσεις και ενημερώσεις μεταξύ των τριών βάσεων δεδομένων. Βλέπε το κείμενο για λεπτομέρειες.

βόρα, καθώς απαιτεί εκτεταμένη ανάλυση ακολουθιών και λεπτομερή επιμέλεια από ειδικούς σχολιαστές (Apweiler, 2001). Για να επιταχυνθεί η δημοσίευση πρωτεϊνικών ακολουθιών που ακόμη δεν είχαν σχολιαστεί σύμφωνα με τα αυστηρά πρότυπα της Swiss-Prot, δημιουργήθηκε μια νέα βάση δεδομένων που ονομάστηκε TrEMBL [«Translation of EMBL nucleotide sequences», Μετάφραση Νουκλεοτιδικών Ακολουθιών EMBL). Αυτή η επέκταση στη Swiss-Prot αποτελούνταν αρχικά από καταχωρίσεις υπολογιστικά σχολιασμένες, προερχόμενες από τη μετάφραση των κωδικών ακολουθιών (coding sequences, CDS) της βάσης δεδομένων DDBJ/EMBL/GenBank, στις οποίες υπήρχαν μόνο δεδομένα που δεν περιλαμβάνονταν ήδη στη Swiss-Prot.

ΠΡΩΤΟΓΕΝΕΙΣ ΚΑΙ ΔΕΥΤΕΡΟΓΕΝΕΙΣ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ

Προτού προχωρήσουμε στην αναλυτική περιγραφή των κυριότερων βάσεων δεδομένων ακολουθιών, πρέπει να γίνει διάκριση μεταξύ των πρωτογενών βάσεων δεδομένων (βάσεις αρχειοθέτησης) και των δευτερογενών βάσεων δεδομένων (επιμελημένες βάσεις). Η σημαντικότερη συμβολή των βάσεων δεδομένων ακολουθιών στην κοινότητα των βιολόγων είναι η δυνατότητα πρόσβασης στις ίδιες τις ακολουθίες. Οι πρωτογενείς βάσεις δεδομένων περιλαμβάνουν κυρίως πειραματικά αποτελέσματα (με κάποια ερμηνεία), αλλά δεν υφίστανται επιμέλεια και αναθεώρηση. Επιμέλεια και αναθεώρηση (curated review) πραγματοποιείται στις δευτερογενείς βάσεις δεδομένων. Οι νουκλεοτιδικές ακολουθίες στην DDBJ/EMBL/GenBank προέρχονται από την αλληλούχηση ενός βιολογικού μορίου, που περιέχεται σε έναν δοκιμαστικό σωλήνα σε κάποιο εργαστήριο. Δεν αποτελούν συναινετικές (consensus) ακολουθίες για έναν πληθυσμό ούτε είναι σειρές γραμμάτων που παράγονται από έναν υπολογιστή. Αυτή η κατάσταση έχει ορισμένες επιπτώσεις στην ερμηνεία των ακολουθιών. Καθεμία από αυτές τις ακολουθίες DNA και RNA σχολιάζεται, προκειμένου να περιγραφεί η ανάλυση των πειραματικών αποτελεσμάτων που υποδηλώνει για ποιον λόγο αρχικά προσδιορίστηκε η συγκεκριμένη ακολουθία. Το μεγαλύτερο ποσοστό των πρωτεϊνικών ακολουθιών που περιλαμβάνονται στις δημόσιες βάσεις δεδομένων δεν προσδιορίστηκε πειραματικά, γεγονός που μπορεί να έχει περαιτέρω επιπτώσεις κατά την εκτέλεση αναλύσεων. Για παράδειγμα, η απόδοση ενός ονόματος στο πρωτεϊνικό προϊόν ή ενός λειτουργικού προσδιορισμού με βάση την υποκειμενική ερμηνεία μιας ανάλυσης ομοιότητας (π.χ. BLAST· βλ. Κεφάλαιο 11) είναι συχνά χρήσιμη, αλλά μπορεί να αποδειχθεί και παραπλανητική. Ως εκ τούτου, οι ακολουθίες DNA, RNA και πρωτεϊνών είναι τα στοιχεία που πρόκειται να αναλυθούν «υπολογιστικά», και αντιπροσωπεύουν το πιο πολύτιμο συστατικό των πρωτογενών βάσεων δεδομένων.

ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ ΝΟΥΚΛΕΟΤΙΔΙΚΩΝ ΑΚΟΛΟΥΘΙΩΝ

Όπως περιγράφηκε προηγουμένως, οι κυριότερες πηγές δεδομένων για νουκλεοτιδικές ακολουθίες είναι οι βάσεις δεδομένων που συμμετέχουν στη συνεργασία International Nucleotide Sequence Database Collaboration: οι DDBJ, EMBL και GenBank – θυμίζουμε ότι νέα ή ενημερωμένα δεδομένα ανταλλάσσονται μεταξύ των τριών αυτών βάσεων μία φορά ανά 24 ώρες. Αυτή η μεταφορά διευκολύνεται χάρη στη χρήση κοινών μορφοποιήσεων δεδομένων

για τα είδη πληροφορίας που περιγράφονται αναλυτικά παρακάτω.

Οι εγγραφές νουκλεοτιδικών ακολουθιών των DDBJ/EMBL/GenBank είναι συχνά η κυριότερη πηγή πληροφοριών αλληλούχησης και βιολογικών δεδομένων, από την οποία παράγονται οι εγγραφές άλλων βάσεων δεδομένων. Επειδή οι βάσεις δεδομένων που εξαρτώνται από την ακρίβεια των εγγραφών της DDBJ/EMBL/GenBank είναι πολλές, παρουσιάζονται ορισμένα σημαντικά ζητήματα που πρέπει να λαμβάνονται υπ' όψιν:

- Εάν σε μια εγγραφή νουκλεϊκού οξέος δεν υποδειχθεί κάποια κωδική ακολουθία, δεν θα δημιουργηθεί η αντίστοιχη εγγραφή στις πρωτεϊνικές βάσεις δεδομένων. Συνεπώς, οι αναζητήσεις για ομοιότητες ακολουθιών στις πρωτεϊνικές βάσεις δεδομένων, που αποτελούν την πιο ευαίσθητη μέθοδο αναζήτησης ομοιοτήτων (Κεφάλαιο 11), ενδέχεται να παραβλέψουν σημαντικές βιολογικές σχέσεις.
- Εάν ένα στοιχείο μιας εγγραφής DDBJ/EMBL/GenBank περιλαμβάνει λανθασμένες πληροφορίες για την πρωτεΐνη, οι πληροφορίες αυτές θα μεταδοθούν σε άλλες βάσεις δεδομένων που προέρχονται απευθείας από την εγγραφή αυτή, ενώ ενδέχεται να διαδοθούν και σε άλλες εγγραφές νουκλεοτιδίων ή πρωτεϊνών με βάση ομοιότητες ακολουθίας.
- Εάν κάποιες σημαντικές πληροφορίες για μια πρωτεΐνη δεν εισαχθούν στο κατάλληλο σημείο μιας εγγραφής ακολουθίας, τα προγράμματα λογισμικού που σχεδιάστηκαν έτσι, ώστε να εξάγουν πληροφορίες από τις εγγραφές αυτές, θα παραβλέψουν πιθανότατα αυτές τις πληροφορίες και, συνεπώς, οι πληροφορίες δεν θα μεταδοθούν στις άλλες βάσεις δεδομένων.

Μορφοποιήσεις βάσεων δεδομένων (Database Formats)

Οι βασικές πληροφορίες είναι κατατεθειμένες στις DDBJ/EMBL/GenBank με μορφή *επίπεδων αρχείων* (*flatfiles*). Η ύπαρξη αντιστοιχίας μεταξύ των διάφορων μορφοποιήσεων ενός «επίπεδου» (flat) αρχείου διευκολύνει την ανταλλαγή δεδομένων μεταξύ των αντίστοιχων βάσεων δεδομένων: στις περισσότερες περιπτώσεις, τα πεδία των διάφορων μορφών ενός αρχείου έχουν μεταξύ τους αντιστοιχία ένα προς ένα. Με την πάροδο των ετών υιοθετήθηκαν και χρησιμοποιήθηκαν εκτεταμένα διάφορες μορφοποιήσεις αρχείων, ενώ άλλες εγκαταλείφθηκαν για διάφορους λόγους. Ο βαθμός επιτυχίας μιας ορισμένης μορφοποίησης εξαρτάται από τη χρησιμότητά της σε μια ποικιλία περιεχομένων, καθώς και οι δυνατότητές της να συμπεριλάβει αποτελεσματικά τις μορφές βιολογικής πληροφορίας που πρέπει να αρχειοθετηθούν και να διατεθούν στην επιστημονική κοινότητα.

Στην απλούστερη μορφή της, μια εγγραφή ακολουθίας μπορεί να αναπαρασταθεί ως μια σειρά νουκλεοτιδίων με μια βασική ετικέτα (tag) ή αναγνωριστικό (identifier). Η πιο δημοφιλής από τις απλές αυτές μορφές είναι η FASTA, η οποία προσφέρει έναν εύκολο τρόπο χειρισμού των πρωτογενών δεδομένων, τόσο από ανθρώπους όσο και από υπολογιστές. Οι εγγραφές νουκλεοτιδικών ακολουθιών σε μορφή FASTA έχουν την παρακάτω μορφή:

```
>U54469.1
CGGTTGCTTGGGTTTTATAACATCAGTCAGTGACAGGCATTTCCAGAGTTGCCCTGTTCAACAATCGATA
GCTGCCTTTGGCCACCAAAATCCCAAACTTAATTAAAGAATTAAATAATTCGAATAATAATTAAGCCCAG
TAACCTACGCAGCTTGAGTGCGTAACCGATATCTAGTATACATTTCGATACATCGAAATCATGGTAGTGT
TGGAGACGGAGAAGGTAAGACGATGATAGACGGCGAGCCGCATGGGTTCGATTTGCGCTGAGCCGTGGCA
GGGAACAACAAAAACAGGGTTGTTGCACAAGAGGGGAGGCGATAGTCGAGCGGAAAAGAGTGCAGTTGGC
GTGGCTACATCATCATTGTGTTCACCGATTATTTTTTGCACAATTGCTTAATATTAATTGTACTTGCACG
CTATTGTCTACGTCATAGCTATCGCTCATCTCTGTCTGTCTCTATCAAGCTATCTCTCTTTCGCGGTCAC
TCGTTCTCTTTTTTCTCTCCTTTCGCATTTGCATACGCATACCACACGTTTTCAGTGTTCTCGCTCTCTC
TCTCTTGTCAAGACATCGCGCGCGTGTGTGTGGGTGTGTCTCTAGCACATATACATAAATAGGAGAGCGG
AGAGACAAATATGGAAAGAATGAAAAAGAGTGAATTACTGCAATTAACCAGTCGCGAACAGTTAAATCAT
ATTTTTGTCGGCCATTGCAGTAAATAAACCGTTGGCTTTCCCTCCTTCACTTTCCACCTCCTTTCTTGAC
```

Εδώ παρουσιάζονται μόνον οι πρώτες γραμμές της ακολουθίας για εξοικονόμηση χώρου. Στην απλούστερη υλοποίηση της μορφής FASTA, ο χαρακτήρας «>» υποδεικνύει την αρχή μιας νέας εγγραφής ακολουθίας: η γραμμή αυτή ονομάζεται *γραμμή ορισμού* (*definition line*, συνήθως «*def line*»). Ένα αναγνωριστικό (σε αυτήν την περίπτωση ο *αριθμός εισαγωγής.αριθμός δημοσίευσης* `U54469.1`) ακολουθείται από την ακολουθία DNA, είτε σε κεφαλαία είτε σε πεζά γράμματα – συνήθως δίνονται 60 χαρακτήρες ανά γραμμή. Οι χρήστες και οι βάσεις δεδομένων μπορούν κατόπιν να προσθέσουν έναν βαθμό πολυπλοκότητας στη μορφοποίηση αυτή. Για παράδειγμα, χωρίς να παραβεί κανέναν από τους παραπάνω κανόνες, μπορεί κανείς να προσθέσει πληροφορίες στη γραμμή ορισμού FASTA, ώστε να αυξηθεί το πληροφοριακό περιεχόμενο αυτής της απλής μορφή, ως εξής:

```
>gb$|$U54469.1$|$DMU54469 Drosophila melanogaster eukaryotic initiation
factor 4E (eIF4E) gene, alternative splice products, complete cds
```

Αυτό το τροποποιημένο αρχείο FASTA περιλαμβάνει τώρα πληροφορίες για τη βάση δεδομένων προέλευσης (`gb` για την GenBank), τον κωδικό πρόσβασης.έκδοσης (accession.version number, εδώ `U54469.1`), ένα αναγνωριστικό όνομα LOCUS (κατά την ονοματολογία της GenBank) ή ένα αναγνωριστικό όνομα καταχώρισης (κατά την ονοματολογία του EMBL) και μια μικρή περιγραφή της βιολογικής οντότητας την οποία αντιπροσωπεύει η ακολουθία αυτή.

Αντιστοίχως, μια εγγραφή πρωτεΐνης αντιπροσωπεύεται ως εξής:

```
>uniprot$|$P48598$|$IF4E_DROME Eukaryotic translation initiation factor 4E
(eIF4E) (eIF-4E) (mRNA cap-binding protein) (eIF-4F 25 kDa subunit)
MQSDFHRMKNFANPKSMFKTSAPSTEQGRPEPPTSAAAPAEAKDVKPKEDPQETGEPAGN
TATTTAPAGDDAVRTEHLYKHPLMNVWTLWYLENDRSKSWEDMQNEITSFDTVEDFWSLY
NHIKPPSEIKLGSDYSLFKKNIRPMWEDAANKQGGRWVITLNKSSKTDLDNLWLDVLLCL
IGEAFDHSDQICGAVINIRGKSNKISIWTADGNNEEAALEIGHKLRDALRLGRNNSLQYQ
LHKDTMVKQGSNVKSIYTL
```

Η εγγραφή αυτή προέρχεται από τη νουκλεοτιδική εγγραφή που δίνεται παραπάνω. Η γραμμή ορισμού έχει την ίδια μορφή με την προηγούμενη περίπτωση, με αναφορά στη βάση δεδομένων προέλευσης (`uniprot`), τον κωδικό πρόσβασης (`P48598`), ένα αναγνωριστικό UniProt (`IF4E_DROME`) και μια μικρή περιγραφή της βιολογικής οντότητας την οποία αντιπροσωπεύει η ακολουθία αυτή.

ΕΠΙΠΕΔΑ ΑΡΧΕΙΑ ΝΟΥΚΛΕΟΤΙΔΙΚΩΝ ΑΛΛΗΛΟΥΧΙΩΝ: ΠΕΡΙΓΡΑΦΗ

Δεδομένου ότι τα επίπεδα αρχεία αποτελούν τη στοιχειώδη μονάδα πληροφορίας στην DDBJ/EMBL/GenBank και διευκολύνουν την ανταλλαγή πληροφοριών μεταξύ αυτών των βάσεων δεδομένων, είναι σημαντικό να διευκρινιστεί το τι αντιπροσωπεύει κάθε πεδίο του επίπεδου αρχείου και τι είδος πληροφορίας παρέχεται σε κάθε τμήμα της εγγραφής. Σήμερα, οι μορφές επίπεδων αρχείων της DDBJ και της GenBank είναι σχεδόν πανομοιότυπες (Παραρτήματα 1.1. και 1.2), ενώ το EMBL χρησιμοποιεί προθέματα τύπου γραμμής (line-type prefixes), τα οποία υποδεικνύουν το είδος της πληροφορίας που παρέχεται σε κάθε γραμμή της εγγραφής (Παράρτημα 1.3). Τα επίπεδα αρχεία χωρίζονται σε τρία βασικά μέρη: την επικεφαλίδα (header) που περιλαμβάνει πληροφορίες [περιγραφείς (descriptors)] οι οποίες σε αφορούν ολόκληρο το αρχείο, τα χαρακτηριστικά (features) που αποτελούν σχόλια επί της εγγραφής και την ίδια τη νουκλεοτιδική ακολουθία. Τα επίπεδα αρχεία των μεγάλων νουκλεοτιδικών βάσεων δεδομένων καταλήγουν στους χαρακτήρες «//» στην τελευταία γραμμή κάθε εγγραφής.

Επικεφαλίδα

Η επικεφαλίδα είναι το τμήμα της εγγραφής που ποικίλλει περισσότερο, ανάλογα με την εκάστοτε βάση δεδομένων. Οι επιμέρους βάσεις δεδομένων δεν είναι υποχρεωμένες να περιλαμβάνουν τις ίδιες πληροφορίες σε αυτό το τμήμα της εγγραφής και πράγματι υπάρχουν μικρές διαφορές μεταξύ των βάσεων. Παρ' όλα αυτά, καταβάλλονται μεγάλες προσπάθειες, ώστε να διασφαλιστεί ότι παρέχονται οι ίδιες πληροφορίες σε όλες τις βάσεις δεδομένων DDBJ/EMBL/GenBank. Η πρώτη γραμμή όλων των επίπεδων αρχείων είναι η γραμμή `LOCUS` (τόπος) στην DDBJ και την GenBank, που αντιστοιχεί στη γραμμή `ID` του EMBL:

DDBJ/GenBank

```
LOCUS       DMU54469                2881 bp    DNA     linear   INV 22-FEB-1998
```

EMBL

```
ID   DM54469    standard; genomic DNA; INV; 2881 BP.
```

Η γραμμή LOCUS/ID αναφέρει ένα αυθαίρετο όνομα (DM54469), που είναι το όνομα τόπου στην ονοματολογία των DDBJ/GenBank και το όνομα καταχώρισης στην ονοματολογία του EMBL. Αυτό το στοιχείο πρέπει να ξεκινά με ένα γράμμα, ενώ οι υπόλοιποι χαρακτήρες μπορούν να είναι είτε γράμματα είτε αριθμοί. Όλα τα γράμματα είναι κεφαλαία και το μήκος του ονόματος δεν πρέπει να υπερβαίνει τους δέκα χαρακτήρες. Παλαιότερα, οι επιμελητές των βάσεων δεδομένων ακολουθιών προσπαθούσαν να δώσουν χρήσιμα ή εύληπτα ονόματα τόπου/καταχώρισης στις εγγραφές, αλλά τα ονόματα αυτά πρέπει να είναι μοναδικά σε μια βάση δεδομένων και, δεδομένου ότι όλα τα ονόματα με κάποιο νόημα έχουν ήδη δοθεί, το όνομα τόπου/καταχώρισης δεν αποτελεί πλέον χρήσιμο στοιχείο μορφής, τόσο, τα ονόματα

αυτά παραμένουν στις εγγραφές, γιατί πολλά πακέτα λογισμικού βασίζονται στην ύπαρξή τους.

Το δεύτερο στοιχείο στη γραμμή τόπου μιας εγγραφής DDBJ/GenBank είναι το μήκος της ακολουθίας – αυτό αντιστοιχεί στο τελευταίο στοιχείο της γραμμής ID μιας καταχώρισης EMBL. Το τρίτο στοιχείο της γραμμής τόπου/ID υποδεικνύει τον τύπο του μορίου, δηλαδή τη βιολογική του φύση. Ο μοριακός τύπος («mol type») είναι συνήθως DNA ή RNA. Σε αυτό το παράδειγμα, ο μοριακός τύπος είναι DNA στην εγγραφή DDBJ/GenBank ή `genomic DNA` στην εγγραφή EMBL.

Το τέταρτο στοιχείο της γραμμής τόπου/ID είναι ο κωδικός κατηγορίας: τρία γράμματα, που είτε έχουν ταξινομική σημασία είτε χρησιμοποιούνται για άλλους σκοπούς ταξινόμησης. Οι κωδικοί αυτοί υπάρχουν για ιστορικούς λόγους από την εποχή κατά την οποία χρησιμοποιούνταν διάφορες κατηγορίες για να χωριστούν τα αρχεία της βάσης δεδομένων σε πιο εύχρηστα τμήματα. Το σημαντικότερο σημείο διαφοράς που θα πρέπει να γνωρίζουν οι χρήστες είναι ο ορισμός που χρησιμοποιεί κάθε βάση δεδομένων για τις διάφορες κατηγορίες στις οποίες διαιρούνται οι οργανισμοί (βλ. Πίνακα 1.1). Για ιστορικούς λόγους, το NCBI αποφάσισε να μην ενημερώσει τον κατάλογο των κατηγοριών των οργανισμών, επειδή δεν πίστευε ότι ο κωδικός τριών γραμμάτων ήταν κατάλληλος για να αντιπροσωπεύσει τη βιοποικιλότητα που υπάρχει στον πλανήτη. Εξαιτίας αυτού, το NCBI δεν δημιούργησε τις κατηγορίες FUN και HUM (βλ. Πίνακα 1.1). Νέες κατηγορίες λειτουργικής βάσης έχουν αποδειχθεί χρήσιμες από την άποψη ότι αντιπροσωπεύουν λειτουργικούς και καθορισμένους τύπους ακολουθίας. Κάποιες από τις πιο σημαντικές λειτουργικές κατηγορίες περιγράφονται στο Πλαίσιο 1.1.

Η ημερομηνία στη γραμμή LOCUS αντιπροσωπεύει την ημερομηνία κατά την οποία δημοσιεύτηκε τελευταία φορά η συγκεκριμένη εγγραφή. Εάν οποιοδήποτε χαρακτηριστικό ή σχόλιο ενημερώθηκε και η εγγραφή δημοσιεύτηκε ξανά, η ημερομηνία θα αντιστοιχεί στην ημερομηνία τελευταίας δημοσίευσης. Πρέπει να σημειωθεί ότι η γραμμή ID του EMBL δεν περιλαμβάνει πληροφορίες ημερομηνίας, καθώς οι πληροφορίες ημερομηνίας παρέχονται σε χωριστές γραμμές: αυτός είναι ένας από τους τύπους γραμμής που είναι πολύ σαφέστερος στη μορφή EMBL απ' ό,τι στις μορφές GenBank/DDBJ.

```
DT   19-MAY-1996 (Rel. 47, Created)
DT   04-MAR-2000 (Rel. 63, Last updated, Version 3)
```

Οι ημερομηνίες που αναφέρονται σε κάθε γραμμή `DT` υποδεικνύουν πότε δημιουργήθηκε η καταχώριση (πρώτη γραμμή) και πότε ενημερώθηκε για τελευταία φορά (δεύτερη γραμμή). Ο αριθμός δημοσίευσης σε κάθε γραμμή υποδεικνύει την πρώτη τριμηνιαία δημοσίευση που έγινε *μετά* από τη δημιουργία ή την τελευταία ενημέρωση της καταχώρισης. Ο αριθμός έκδοσης για την καταχώριση φαίνεται στη δεύτερη γραμμή. Οι αριθμοί έκδοσης επιτρέπουν στους χρήστες να διαπιστώσουν εύκολα εάν βλέπουν την πιο πρόσφατη εγγραφή για μια ορισμένη ακολουθία. Οι αριθμοί έκδοσης αυξάνονται κατά ένα, κάθε φορά που ενημερώνεται η αντίστοιχη καταχώριση, επειδή μια καταχώριση μπορεί να ενημερωθεί πολλές φορές, προτού να εμφανιστεί σε μια τριμηνιαία δημοσίευση. Εάν μια καταχώριση δεν έχει ενημερωθεί

ΠΙΝΑΚΑΣ 1.1 ■ Κατηγορίες οργανισμών που χρησιμοποιούνται στις τρεις κυριότερες βάσεις δεδομένων ακολουθιών DNA.

Διαίρεση		DDBJ	EMBL	GenBank
BCT	Βακτήρια	√		√
FUN	Μύκητες		√	
HUM	Άνθρωπος	√	√	
INV	Ασπόνδυλα	√	√	√
MAM	Άλλα θηλαστικά	√	√	√
ORG	Οργανίδια		√	
PHG	Φάγος	√	√	√
PLN	Φυτό[α]	√	√	√
PRI	Πρωτεύον[β]	√	√	√
PRO	Προκαρυωτικά		√	
ROD	Τρωκτικά	√	√	√
SYN	Συνθετικά και χιμαιρικά	√	√	√
VRL	Ιοί	√	√	√
VRT	Άλλα σπονδυλωτά	√	√	√

[α]Δεν ισχύουν τα ίδια δεδομένα σε όλα· περιλαμβάνει όλες τις FUN ακολουθίες στις DDBJ και GenBank.

[β]Δεν ισχύουν τα ίδια δεδομένα σε όλα· περιλαμβάνει όλες τις HUM ακολουθίες στην GenBank