Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Θέματα Διαχείρισης Δεδομένων για Εφαρμογές Βιοεπιστημών

Διπλωματικη Εργασια

της

ΔΕΣΠΟΙΝΑΣ ΑΙΜ. ΠΕΡΟΥΛΗ

Επιβλέπων: Τιμολέων Σελλής
Καθηγητής Ε.Μ.Π.

Εργαστηριο Συστηματων Βασεων Γνωσεων και Δεδομενων
Αθήνα, Ιούλιος 2006

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Συστημάτων Βάσεων Γνώσεων και Δεδομένων

Θέματα Διαχείρισης Δεδομένων για Εφαρμογές Βιοεπιστημών

Διπλωματικη Εργασια

της

ΔΕΣΠΟΙΝΑΣ ΑΙΜ. ΠΕΡΟΥΛΗ

Επιβλέπων: Τιμολέων Σελλής
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 12η Ιουλίου 2006.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Τιμολέων Σελλής	Ιωάννης Βασιλείου	Νεκτάριος Κοζύρης
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Επίκουρος Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2006

(Υπογραφή)

.....

Δεσποινα Περουλη
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Συστημάτων Βάσεων Γνώσεων και Δεδομένων

Δέσποινα Περούλη, *Θέματα Διαχείρισης Δεδομένων για Εφαρμογές Βιοεπιστημών*, Διπλωματική Εργασία, Εθνικό Μετσόβιο Πολυτεχνείο, Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών, Εργαστήριο Συστημάτων Βάσεων Γνώσεων και Δεδομένων, 13 Ιουλίου 2006.

Σελίδες: 186

Ευχαριστίες

Ειλικρινά αισθάνομαι την ανάγκη να ευχαριστήσω τον επιβλέποντα της διπλωματικής, Καθηγητή Τίμο Σελλή, για την πολύτιμη βοήθεια που ήταν πάντοτε διατεθειμένος να προσφέρει και για το δημιουργικό και φιλικό κλίμα το οποίο καλλιεργεί, όχι μόνο στο εργαστήριο που διευθύνει, μα και στη σχολή γενικότερα. Ιδιαίτερα, επίσης, ευχαριστώ τον Διδάκτορα Θοδωρή Δαλαμάγκα για τη συνεχή καθοδήγηση και τη συνέπειά του, όπως και για την ευχάριστη συνεργασία μας επί ένα σχεδόν ολόκληρο χρόνο. Καθώς η εργασία αυτή ρίχνει σιγά σιγά την αυλαία της πενταετούς προπτυχιακής μου φοίτησης, είμαι ευγνώμων προς τους γονείς και τους αδελφούς μου, αλλά και σε εκείνους που είναι σαν γονείς ή αδέλφια μου, για την ανυπολόγιστη αγάπη τους.

Περίληψη

Τα τελευταία δεκαπέντε χρόνια, σημαντικές εξελίξεις στον ερευνητικό χώρο επιστημών του ευρύτερου τομέα της βιολογίας (αποκωδικοποίηση γονιδιωμάτων, αποτελέσματα σύγχρονων πειραμάτων της γενετικής, της μοριακής βιολογίας) έχουν εγείρει νέες προκλήσεις για τη διαχείριση βάσεων δεδομένων και την πληροφορική, αφού έχουν συσσωρεύσει τεράστιο πλήθος πολυειδών δεδομένων. Η παρούσα διπλωματική εργασία ασχολείται με θέματα γύρω από την αποθήκευση και επεξεργασία αυτών. Αρχικά γίνεται προσπάθεια να προσεγγιστούν με τη ματιά του μηχανικού υπολογιστών απαραίτητες έννοιες σχετικές με τις λειτουργίες που λαμβάνουν χώρα στους ζωντανούς οργανισμούς καθώς και τη δομή των υποκειμένων και αντικειμένων τούτων των ενεργειών. Περιγράφονται, έτσι, μεταξύ των άλλων το κεντρικό δόγμα της μοριακής βιολογίας, η μεταγραφή, η μετάφραση, οι DNA sequencers, τα microarray πειράματα, τα νουκλεϊκά οξέα, οι πρωτεΐνες, το κύτταρο. Αναλύονται, επίσης, η ποικιλομορφία και τα ειδικά χαρακτηριστικά των δεδομένων των βιοεπιστημών, ενώ ενδιαφέρουν και τα διάφορα πρότυπα με τα οποία αποθηκεύονται στις υπάρχουσες βάσεις. Απασχολούν λ.χ. οι ιδιότητες ακολουθιών νουκλεοτιδίων ή αμινοξέων, τρισδιάστατων δομών μακρομορίων, βιολογικών μονοπατιών, όπως και το μοντέλο του NCBI, το BIOML και άλλα XML πρότυπα. Τα ερωτήματα που θέτουν οι αντίστοιχοι επιστήμονες και οι εργασίες που χρειάζεται να εκτελούν αποτελούν επιπλέον αντικείμενο μελέτης. Τέτοιες είναι, για παράδειγμα, η σύγκριση ακολουθιών, η φυλογενετική ανάλυση, η sequence assembly, ο προσδιορισμός δομής από ακολουθία. Από τα προηγούμενα γίνεται εφικτό να εντοπιστούν κύρια προβλήματα (π.χ. προέλευση και ενοποίηση δεδομένων), στα οποία πρέπει να δώσει απάντηση η τεχνολογία των βάσεων δεδομένων και προτείνονται ορισμένες πιθανές λύσεις (επεκτάσεις στην SQL, ανάπτυξη νέου μοντέλου και γλώσσας) για περαιτέρω έρευνα. Τέλος, εξετάζονται το εργαλείο BLAST και το Pathways Database System (PathCase) ως προς το σκοπό και το θεωρητικό υπόβαθρο αλλά και πειραματικά.

Λέξεις Κλειδιά

Βιοεπιστήμες, κεντρικό δόγμα μοριακής βιολογίας, DNA, microarrays, ποικιλομορφία δεδομένων, ακολουθίες, τρισδιάστατες δομές, βιολογικά μονοπάτια, alignment, φυλογενετική ανάλυση, sequence assembly, data provenance and integration, NCBI μοντέλο, BIOML, BLAST, PathCase

Abstract

During the last fifteen years, the remarkable developments in life sciences research (sequenced genomes, results of modern experiments in molecular biology and genetics) have brought the database community and computer science to new challenges, because of the abundance and diversity of the data in such sciences. This diploma thesis deals with storage and management issues about these data types. First of all, basic principles about the functions that take place incide living organisms and the structure of the entities involved are approached, with the eye of a computer engineer. The central dogma of molecular biology, transcription, translation, DNA sequencers, microarray experiments, nucleic acids, proteins and the cell are described among others. Then, the heterogeneity and the special characteristics of life sciences' data are analysed, while various standards in which these data are stored in current databases are also considered. For instance, the properties of nucleotide or amino acid sequences and those of the structure of macromolecules, biological pathways, the NCBI data model, BIOML and other XML standards are discussed. What is more, the queries needed by the related researchers and main processes done by them on data are examined. These processes include alignment, phylogenetic analysis, sequence assembly and determination of 3D structure based on sequence. As a consequence of the preceding study, main problems (like data provenance and integration) which the database community has to handle are detected, while probable solutions that have already been proposed (such as SQL extensions or the introduction of a new data model and language) are mentioned. Finally, BLAST and the Pathways Database System (PathCase) are examined concerning not only their purpose and theoretical background, but also their experimental response.

Keywords

Life sciences, the central dogma of molecular biology, DNA, microarrays, data abundance and diversity, sequences, 3D structures, biological pathways, alignment, phylogenetic analyses, sequence assembly, data provenance and integration, NCBI data model, BIOML, BLAST, PathCase

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

Το κεφάλαιο αυτό έχει σκοπό να καταστήσει σαφές το αντικείμενο της διπλωματικής, αλλά και τον τρόπο με τον οποίο οργανώνεται ο τόμος. Αναφέρεται, επομένως, συνοπτικά το περιεχόμενο κάθε κεφαλαίου που ακολουθεί, ενώ δεν παραλείπεται να διευκρινιστούν και οι λόγοι που οδήγησαν στην ανάπτυξη αυτής της εργασίας. Τέλος, γίνεται λόγος για τη συνεισφορά της στο χώρο των βάσεων δεδομένων.

1.1 Γενικό πλαίσιο ανάπτυξης της εργασίας

Οι τεχνολογίες των βάσεων δεδομένων έχουν προσφέρει αποδοτικές λύσεις στη διαχείριση και επεξεργασία αρκετών ειδών δεδομένων μεγάλου όγκου. Τα τελευταία χρόνια φαίνεται να είναι επίσης πρόσφορο έδαφος για την εφαρμογή τους ο ευρύτερος χώρος των επιστημών της βιολογίας. Η εν λόγω διπλωματική εργασία είναι μια βιβλιογραφική μελέτη, που επικεντρώνεται γύρω από τη μορφή των δεδομένων που υπάρχουν και το είδος των ερωτημάτων που χρειάζονται οι επιστήμες αυτές.

Η κοινότητα των βάσεων έχει αναπτύξει συστήματα για να καλύψει πλήθος αναγκών εφαρμογών, που προέρχονται από διαφορετικούς χώρους. Κλασικά παραδείγματα τέτοιων πεδίων είναι οι εμπορικές συναλλαγές (Σχεσιακά, RDBMS), οι εφαρμογές CAD - computer aided design (Αντικειμενοστρεφή, OODBMS), τα πληροφοριακά συστήματα (Αποθήκες Δεδομένων, Data Warehouses), το διαδίκτυο (Συστήματα XML). Οι ιδιαίτερες απαιτήσεις της διαχείρισης των δεδομένων καθενός από τους παραπάνω χώρους καθοδήγησαν σε μεγάλο βαθμό την ανάπτυξη νέων τρόπων αποθήκευσης και επεξεργασίας τους.

Από την άλλη πλευρά, τις τελευταίες δύο κυρίως δεκαετίες η έρευνα στη μοριακή βιολογία αλλά και στη γενετική, στην εξελικτική βιολογία, στην ιατρική έχει συσσωρεύσει πλούτο δεδομένων. Στα πειραματικά αποτελέσματα πιο παραδοσιακών μεθόδων (NMR, X-ray crystallography) πρέπει να προστεθούν τα αποκωδικοποιημένα γονιδιώματα αρκετών οργανισμών (όπως ανθρώπου, ποντικού, μύγας, βακτηρίων, ιών), καθώς επίσης πορίσματα σχετικών προγραμμάτων υπολογιστών, και έτσι δημιουργείται ένας τεράστιος όγκος στοιχείων προς μελέτη. Μάλιστα αυτά είναι διαφορετικής υφής από την αριθμητική υπόσταση των αντίστοιχων από τις κλασικές εφαρμογές των συστημάτων βάσεων δεδομένων.

1.2 Δομή και περιεχόμενο

Η διπλωματική εργασία στοχεύει στη διερεύνηση των ιδιαιτεροτήτων που έχουν τα δεδομένα και τα ερωτήματα των βιοεπιστημών, σε σχέση με αυτά που έχουν ήδη αντιμετωπίσει τα υπάρχοντα συστήματα βάσεων. Επιχειρεί να ρίξει ένα –όσο το δυνατόν πιο εκτεταμένο σε πλάτος– βλέμμα σε θέματα που άπτονται ταυτόχρονα των βιοεπιστημών και της πληροφορικής. Η έμφαση δίνεται σε εκείνα που μπορεί να απασχολήσουν περισσότερο τον ερευνητή των βάσεων δεδομένων και να τον ωθήσουν, ώστε εκείνος να συνεχίσει μια πιο λεπτομερή μελέτη πάνω σε αυτά.

Ο κορμός της εργασίας αποτελείται από πέντε άνισα μέρη. Αρχικά, γίνεται εισαγωγή σε έννοιες, όρους, θεωρίες και μεθόδους των βιοεπιστημών. Στη συνέχεια, εξετάζονται τα δεδομένα αυτών ως προς τα χαρακτηριστικά και τη μορφή τους, ενώ μελετώνται και αντιπροσωπευτικοί τρόποι αποθήκευσής τους σε υπάρχουσες βάσεις. Κατόπιν, αναλύονται οι πιο βασικές λειτουργίες που χρειάζεται να γίνονται πάνω στα δεδομένα από τους βιοεπιστήμονες ώστε να φανούν τα είδη των ερωτημάτων που είναι χρήσιμο να ασκούνται. Με βάση τα προηγούμενα, τονίζονται προβλήματα που εμφανίζονται για τους ερευνητές των βάσεων δεδομένων και αναφέρονται λύσεις που έχουν προταθεί. Τέλος, ενδεικτικά παρουσιάζονται ένα δημοφιλές πρόγραμμα (BLAST) και ένα σχετικά νέο σύστημα διαχείρισης μιας κατηγορίας βιοδεδομένων (PathCase).

1.3 Βασικές γνώσεις βιολογίας (Κεφάλαια 2, 3)

Η αναγκαιότητα να έρθει ο μηχανικός υπολογιστών σε επαφή με το αντικείμενο των βιοεπιστημών, ώστε να καταλάβει βαθύτερα τα προβλήματα που καλείται εκείνος να επιλύσει, είναι έκδηλη. Είναι μάλιστα χαρακτηριστικό ότι όσο πιο νωρίς επιχειρήσει να εξοικειωθεί με αυτό, τόσο λιγότερες θα είναι οι δυσκολίες που θα αντιμετωπίσει. Αν και μοιάζει σχετικά απλό να μάθει κανείς τα βασικά για τις βιοεπιστήμες, στην ουσία αυτός διαπιστώνει ότι μόνο μια ουσιαστική και αρκετές φορές λεπτομερής κατανόηση των αντίστοιχων θεμάτων θα τον απαλλάξει από νοητικά αδιέξοδα ή παρερμηνείες.

Στο Κεφάλαιο 2 της εργασίας, περιγράφονται οι πιο βασικές βιολογικές οντότητες που χρειάζεται να έχει κανείς υπόψη του. Η παρουσίαση ξεκινά από το υψηλότερο επίπεδο, που είναι αυτό των οργανισμών, και ολοένα αυξάνει την ανάλυση του μεγεθυντικού φακού. Αναφέρονται τα κύρια χαρακτηριστικά των κυττάρων, για να αποκαλυφθούν όχι μόνο η δομή και λειτουργία της ελάχιστης δομικής μονάδας όλων των έμβιων όντων, αλλά και οι απροσδόκητα πολλές ομοιότητες αυτών μεταξύ τους. Η εστίαση είναι μεγαλύτερη σε δύο βιομόρια, που θα απασχολήσουν ιδιαίτερα όλη την υπόλοιπη εργασία, τα νουκλεϊκά οξέα (DNA, RNA) και τις πρωτεΐνες. Επιπλέον, αναφέρονται τα χημικά άτομα που συνθέτουν σε τελική ανάλυση τη ζωή.

Εκτός αυτών, κρίνεται απαραίτητο να θιγούν κρίσιμες λειτουργίες που λαμβάνουν χώρα εντός των οργανισμών σε μοριακό επίπεδο, όπως και κάποιες θεωρίες και μέθοδοι των βιοεπιστημών. Αναλύονται οι μηχανισμοί του διπλασιασμού του γενετικού υλικού και οι φάσεις

που χρειάζονται για την παραγωγή των πρωτεϊνών, στις οποίες μάλιστα εμπλέκεται και ο γενετικός κώδικας. Αναφέρονται, επίσης, το κεντρικό δόγμα της βιολογίας, ο τρόπος της γενετικής έκφρασης και στοιχεία που συμβάλλουν στη γενετική ρύθμιση. Η θεωρία της εξέλιξης απασχολεί αρκετά το υπόλοιπο τμήμα της διπλωματικής εργασίας, για αυτό και δίνονται σημαντικά συμπεράσματά της. Το ίδιο ισχύει και για την πειραματική μέθοδο των microarrays, με την οποία και ολοκληρώνεται το τρίτο κεφάλαιο, ενώ περιγράφονται συνοπτικά οι DNA sequencers και η μέθοδος PCR.

1.4 Κατηγορίες δεδομένων, υπάρχουσες βάσεις, μοντέλα και πρότυπα (Κεφάλαιο 4)

Καθώς έχει σχηματιστεί μία όσο το δυνατόν πιο πλήρης εικόνα για το αντικείμενο έρευνας των βιοεπιστημών, μπορεί να γίνει μια απόπειρα κατηγοριοποίησης των δεδομένων τους ανάλογα με τα ποιοτικά και μορφολογικά χαρακτηριστικά τους. Ταυτόχρονα, πολύτιμη σχετική γνώση αποκτάται από τη μελέτη των βάσεων αυτών των δεδομένων που υπάρχουν ήδη στο χώρο. Από τη σκοπιά του ερευνητή πληροφορικού, έχουν ιδιαίτερο ενδιαφέρον τα μοντέλα και πρότυπα που χρησιμοποιούνται στις βάσεις ή έχουν προταθεί για μελλοντική χρήση. Στις επόμενες τρεις παραγράφους σχολιάζονται περισσότερο τα θέματα αυτά, τα οποία και συνθέτουν το περιεχόμενο του τετάρτου κεφαλαίου.

Η ομαδοποίηση των δεδομένων του ευρύτερου χώρου της βιολογίας πραγματοποιείται με δύο διαφορετικά κριτήρια. Στην πρώτη περίπτωση κατατάσσονται με βάση το είδος τους, δηλαδή το περιεχόμενό τους από τη σκοπιά των βιοεπιστημόνων (λ.χ. χάρτες γονιδίων, βιομονοπάτια, πεδία). Η προσέγγιση αυτή βοηθά στην ομαλή έκβαση από τη συζήτηση που έχει προηγηθεί στα αρχικά κεφάλαια. Με τον δεύτερο τρόπο κατηγοριοποιούνται σύμφωνα με τη μορφή τους, η οποία είναι πιο κοντά στην αντίληψη του μηχανικού υπολογιστών (π.χ. ακολουθίες, γράφοι, εικόνες). Ο τελευταίος αυτός διαχωρισμός εξυπηρετεί και την ανάλυση που ακολουθεί στα επόμενα τμήματα της διπλωματικής εργασίας.

Στις βάσεις που είναι προς το παρόν διαθέσιμες στον τομέα των βιοεπιστημών, τα δεδομένα συνήθως αποθηκεύονται ανάλογα με έναν συνδυασμό από τους παραπάνω τρόπους. Με άλλα λόγια, υπάρχουν κάποιες που αναφέρονται σε ένα μόνο είδος, για παράδειγμα στο DNA, κάποιες που ασχολούνται αποκλειστικά με μία μορφή, όπως τις διαδρομές του μεταβολισμού και κάποιες που συγκεντρώνουν στοιχεία για συγκεκριμένα είδη και μορφές. Άλλες φορές πάλι ακολουθούνται διαφορετικά κριτήρια, σαν τον οργανισμό από τον οποίο προέρχονται τα δεδομένα, ανεξάρτητα από τα χαρακτηριστικά τους. Σίγουρο είναι το γεγονός ότι ορισμένες θεωρούνται οι πρωταρχικές και οι υπόλοιπες αντλούν τα στοιχεία τους από αυτές, σχηματίζοντας έτσι μικρότερες αλλά περισσότερο εξειδικευμένες βάσεις.

Ως προς το μοντέλο αποθήκευσης, είναι αλήθεια ότι δεν υπάρχει ομοφωνία. Οι περισσότερες βάσεις δεδομένων χρησιμοποιούν το δικό τους format. Άλλωστε, κάτι τέτοιο είναι λογικό να συμβαίνει ακριβώς λόγω της ποικιλομορφίας και ιδιοσυγκρασίας των δεδομένων, όπως αναφέρθηκε. Ωστόσο, ξεχωρίζει το μοντέλο του NCBI (National Center for Biotechnology Information), το οποίο εφαρμόζεται σε μικρό αλλά σπουδαίο μερίδιο βάσεων. Εκτός

από την αναλυτική περιγραφή αυτού, αναφέρονται και πρότυπα βασισμένα στην XML, καθώς τα πλεονεκτήματα της αυστηρής δομημένης ιεραρχίας της εξυπηρετούν και στο συγκεκριμένο πεδίο. Ιδιαίτερη έμφαση δίνεται σε ένα από αυτά, το BIOML, που είναι ενδεικτικό.

1.5 Κατηγορίες ερωτημάτων, λειτουργίες πάνω στα δεδομένα (Κεφάλαιο 5)

Τα στοιχεία που παρατίθενται στο τέταρτο κεφάλαιο θεωρείται πως είναι ικανά, για να σχηματίσει ο αναγνώστης μια σφαιρική άποψη γύρω από τα χαρακτηριστικά των βιοδεδομένων και να μπορέσει να προχωρήσει ακόμη ένα βήμα. Το λογικά επόμενο να αναρωτηθεί και εξερευνήσει είναι το είδος των ερωτημάτων που συνήθως ασκούν –ή θα επιθυμούσαν να είναι σε θέση να ασκήσουν– οι ερευνητές του ευρύτερου χώρου της βιολογίας. Η απάντηση σε αυτόν τον προβληματισμό έρχεται μέσα από τη μελέτη των εργασιών που κάνουν οι επιστήμονες πάνω στα δεδομένα.

Επομένως, στόχος σε αυτή τη φάση της εργασίας είναι η επισκόπηση των σημαντικότερων σημείων από τις λειτουργίες στα δεδομένα. Μελετώνται από δύο οπτικές γωνίες αυτές που κρίνονται ως οι πλέον συχνές και χρήσιμες. Εξετάζεται πρώτα το πρόβλημα που υπάρχει έτσι όπως το βλέπει η μεριά των βιοεπιστημόνων, ώστε να γίνει αντιληπτή η πραγματική του αξία για τη μελέτη της βιολογίας. Στη συνέχεια, παρουσιάζεται ο αλγόριθμος που χρησιμοποιείται από τους πληροφορικούς για να λύσουν αυτό το πρόβλημα, στις περιπτώσεις βέβαια που κάτι τέτοιο είναι δυνατό. Εφόσον η μελέτη δεν είναι εξονυχιστική, επιλέγεται να αναλυθεί συνήθως ο πιο ευρέως διαδεδομένος αλγόριθμος. Παράλληλα, αναφέρονται και μοντέλα κατάλληλα για τη λύση αυτών των θεμάτων.

Το πέμπτο, λοιπόν, κεφάλαιο προσπαθεί να δώσει τα κυριότερα σημεία από τις πιο σημαντικές λειτουργίες, περιγράφοντάς τα συνοπτικά αλλά με ακρίβεια. Ανάμεσα σε αυτές είναι η σύγκριση μεταξύ ακολουθιών, που περιλαμβάνει και το alignment, και είναι θεμελιακή για τις περισσότερες από τις υπόλοιπες. Χαρακτηριστική τέτοια περίπτωση είναι η φυλογενετική ανάλυση, που έχει τη βάση της στην εξελικτική θεωρία. Γίνεται, επίσης, λόγος για τον προσδιορισμό σχήματος ή τρισδιάστατης δομής με βάση μόνο την πρωτοταγή δομή του μορίου. Ενδιαφέρων είναι και ο τρόπος με τον οποίο επανασχηματίζεται η αρχική μορφή των ακολουθιών μετά από το σπάσιμό τους σε μικρότερες για τους σκοπούς της αποκωδικοποίησης (genome assembly) και γενικότερα, ό,τι έχει σχέση με την ανάλυση των γονιδιωμάτων.

Το βαθύτερο ζητούμενο, ωστόσο, του πέμπτου κεφαλαίου συνεχίζει να είναι η ανακάλυψη των ερωτημάτων που γίνονται στα δεδομένα των βιοεπιστημών. Για το λόγο αυτό επιχειρείται να καταταχθούν σε διαφορετικές ομάδες. Μεγάλη κατηγορία σχηματίζουν εκείνα που αφορούν similarity queries δεδομένων οποιασδήποτε σχεδόν μορφής. Επιπλέον, συχνά είναι χρήσιμες ερωτήσεις που σχετίζονται με την ανακάλυψη patterns. Ειδικά σύνολα αποτελούν τα ερωτήματα που αφορούν εικόνες αλλά και εκείνα που αναφέρονται σε metadata. Η κατηγοριοποίηση αυτή εφαρμόζεται περισσότερο για να υποστηριχθεί η τακτική σκέψη του αναγνώστη στο θέμα, παρά ως πρόταση τυφλής επιμονής σε αυτήν. Για κάθε ομάδα ερωτημάτων δίνονται πάντα χαρακτηριστικά παραδείγματα.

1.6 Προβλήματα και λύσεις, δύο χαρακτηριστικά συστήματα (Κεφάλαια 6, 7)

Μετά από την ανάλυση των παραπάνω ενοτήτων, εύλογο είναι να έχει δημιουργηθεί η επιθυμία να εντοπιστούν τα προβλήματα που εμφανίζονται στα θέματα που αυτές πραγματεύονται, όπως και οι πιθανές λύσεις. Το έκτο κεφάλαιο έρχεται να αντιμετωπίσει αυτήν την πρόκληση. Πρόθεσή του είναι να φωτίσει τις πιο σοβαρές αδυναμίες που υπάρχουν στα επιμέρους εγχειρήματα, δηλαδή τόσο στην αποθήκευση των δεδομένων όσο και στην επεξεργασία των ερωτημάτων των βιοεπιστημών. Θίγει θέματα σαν την προέλευση (data provenance) και την ενοποίηση (data integration) των δεδομένων, όπως και τη θεώρηση κάποιων τύπων δεδομένων ως πολιτών πρώτης κατηγορίας. Παράλληλα, αναφέρει λύσεις που έχουν προταθεί σε ορισμένες από αυτές τις δυσκολίες από την κοινότητα των βάσεων δεδομένων, αφήνοντας τον αναγνώστη να συμπεράνει την αποτελεσματικότητά τους.

Το έβδομο κεφάλαιο συμπληρώνει την εικόνα παρουσιάζοντας δύο εργαλεία που είναι προϊόντα της βιοπληροφορικής. Πρόκειται για το BLAST (Basic Local Alignment Search Tool) του NCBI και για το PathCase (Pathway Database System) του Case Western Reserve University. Το πρώτο λύνει το πρόβλημα της σύγκρισης δύο ακολουθιών της βιολογίας (νουκλεοτιδίων ή αμινοξέων) και έχει καθιερωθεί να χρησιμοποιείται τα τελευταία δέκα περίπου χρόνια παρά την ύπαρξη αρκετών συναφών. Το δεύτερο είναι ένα σύστημα αποθήκευσης και διαχείρισης βιολογικών μονοπατιών (διαδρομών της ενέργειας, των πρωτεϊνών, των σημάτων στους οργανισμούς). Παρότι είναι σχετικά καινούριο, η μελέτη του παρουσιάζει ιδιαίτερο ενδιαφέρον για τους ερευνητές των βάσεων δεδομένων, καθώς είναι ένα σύστημα ακριβώς στην περιοχή τους. Για τα εργαλεία αυτά αναφέρεται το θεωρητικό τους υπόβαθρο, αλλά δίνονται και παραδείγματα χρήσης τους.

1.7 Συνεισφορά

Με την επιλογή και την κριτική –όπου είναι δυνατόν– ανάπτυξη των παραπάνω θεμάτων, η διπλωματική εργασία πιστεύεται ότι ξεκινά στέρεα την πορεία προς τη ζητούμενη κατεύθυνση. Πιο συγκεκριμένα, θεωρείται ότι η εν λόγω βιβλιογραφική εργασία:

- Εξηγεί τις απαραίτητες έννοιες, διαδικασίες και θεωρίες των επιστημών του χώρου της βιολογίας στον μηχανικό υπολογιστών.
- Διερευνά τα χαρακτηριστικά των δεδομένων των βιοεπιστημών.
- Καταγράφει τις υπάρχουσες βάσεις δεδομένων του χώρου, καθώς επίσης χαρακτηριστικά μοντέλα και πρότυπα που χρησιμοποιούνται ή έχουν προταθεί.
- Προσδιορίζει το είδος των ερωτημάτων που ασκούνται ή είναι επιθυμητό να ασκούνται από τους ερευνητές των βιοεπιστημών.
- Αναλύει χρήσιμες λειτουργίες που γίνονται πάνω στα δεδομένα.

- Αναφέρει προβλήματα και πιθανές λύσεις στην αποθήκευση και επεξεργασία των δεδομένων αυτής της επιστημονικής περιοχής.
- Παρουσιάζει δύο εργαλεία που ξεχωρίζουν στον τομέα αυτό.

Έτσι, διαγράφει τα πρώτα βήματα που ίσως θελήσει να ακολουθήσει κανείς για να φέρει αποτέλεσμα στην εφαρμογή των γνώσεων και τεχνικών της κοινότητας των βάσεων δεδομένων αλλά και την παραγωγή νέων, για τα προβλήματα των ερευνητών των βιοεπιστημών. Η ανάπτυξη έχει καταβληθεί προσπάθεια να είναι όσο το δυνατόν πιο σφαιρική και εκτενής σε πλάτος, ενώ καλεί τον ενδιαφερόμενο αναγνώστη να εμβαθύνει περισσότερο σε ό,τι κρίνει σκόπιμο.

Κεφάλαιο 2

Βιολογικές οντότητες

Το δεύτερο κεφάλαιο έχει στόχο να εξηγήσει σε έναν μηχανικό υπολογιστών τον τρόπο οργάνωσης των ζωντανών οργανισμών. Αναφέρεται η κατηγοριοποίησή τους, καθώς και η δομή και λειτουργία των μονάδων που τους συνθέτουν. Η οργάνωση αυτή φθάνει σε τέσσερα επίπεδα. Ξεκινά από τους οργανισμούς, συνεχίζει στα κύτταρα, προχωρά στα μακρομόρια και καταλήγει στα χημικά άτομα. Από την αναφορά αυτή διακρίνονται σε μεγάλο βαθμό τα κοινά χαρακτηριστικά που εμφανίζουν οι οργανισμοί του πλανήτη και τα οποία είναι συνυφασμένα με τη ζωή.

2.1 Οργανισμοί

Στη γη υπάρχουν περισσότερα από 1,7 εκατομμύρια διαφορετικά είδη οργανισμών. Από αυτά 1000000 ανήκει στα ζώα και 250000 στα φυτά. Η ποικιλομορφία είναι τεράστια και τα ιδιαίτερα χαρακτηριστικά που χρειάζεται να μελετηθούν πολλά. Ωστόσο, σημαντικές και αρκετές ιδιότητες εμφανίζονται σε όλους τους ζωντανούς οργανισμούς, συχνά μάλιστα εκδηλώνονται με τον ίδιο τρόπο. Η ανάπτυξη, ο μεταβολισμός, η αναπαραγωγή, η αναζήτηση θρεπτικών συστατικών, ο θάνατος είναι κάποιες από τις βασικές ιδιότητες.

2.1.1 Ταξινόμηση

Στη σημερινή εποχή οι οργανισμοί δεν κατηγοριοποιούνται με βάση τον τρόπο ζωής ή τις δραστηριότητες τους (ταξινόμηση του Αριστοτέλη), αλλά σύμφωνα με το λεγόμενο *μειξιολογικό κριτήριο*. Τα άτομα που μπορούν να παράγουν γόνιμους απογόνους ανήκουν στην ίδια ομάδα, που ονομάζεται *είδος*. Για εκείνους τους οργανισμούς που αναπαράγονται μονογονικά χρησιμοποιούνται κριτήρια που έχουν να κάνουν με μορφολογικά και βιοχημικά χαρακτηριστικά και είναι λιγότερο αντικειμενικά.

Για παράδειγμα, εφαρμόζοντας το μειξιολογικό κριτήριο, είναι σαφές ότι το άλογο και το γαϊδούρι ανήκουν σε διαφορετικά είδη, αφού το μουλάρι που προκύπτει από τη διασταύρωσή τους είναι στείρο.

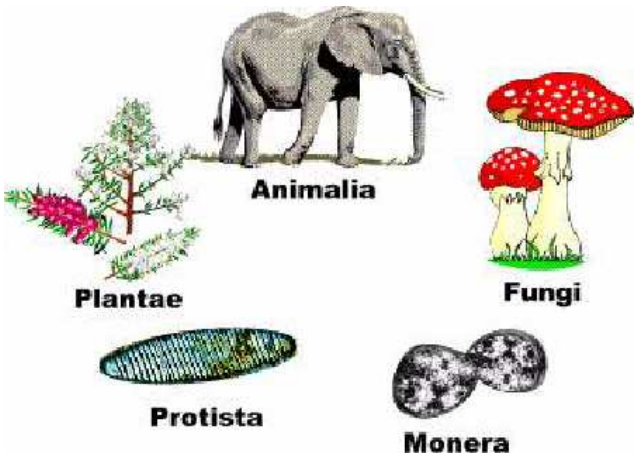
Η ταξινόμηση των οργανισμών δε σταματά στο είδος. Είδη που παρουσιάζουν κοινά χαρακτηριστικά συνιστούν ένα *γένος*. Γένη συγγενικά μεταξύ τους δημιουργούν μια *οικογένεια*.

Πίνακας 2.1: Ταξινόμηση του ανθρώπου

ΒΑΣΙΛΕΙΟ	Ζώα (Animalia)
ΦΥΛΟ	Χορδωτά (Chordota)
ΟΜΟΤΑΞΙΑ	Θηλαστικά (Mammalia)
ΤΑΞΗ	Πρωτεύοντα (Primata)
ΟΙΚΟΓΕΝΕΙΑ	Ανθρωπίδες (Hominidae)
ΓΕΝΟΣ	Άνθρωπος (Homo)
ΕΙΔΟΣ	Άνθρωπος ο σοφός (Homo sapiens)

Αντίστοιχα οι οικογένειες σχηματίζουν τις *τάξεις*, αυτές τις *ομοταξίες* και οι τελευταίες τα *φύλα*. Στον Πίνακα 2.1 φαίνεται η ταξινόμηση για το ανθρώπινο είδος.

Τα πέντε *βασίλεια* της ζωής δημιουργούνται από τα φύλα. Το πρώτο αποτελείται από προκαρυωτικά κύτταρα, κυρίως αυτά των βακτηρίων (Monera), ενώ τα υπόλοιπα τέσσερα αφορούν ευκαρυωτικά κύτταρα. Οι ορισμοί των δύο αυτών τύπων κυττάρων δίνονται στην Ενότητα 2.2.1. Το δεύτερο βασίλειο αφορά κυρίως πρωτόζωα (Protista) και το τρίτο τους μύκητες (Fungi). Το τέταρτο είναι το βασίλειο των φυτών (Plantae) και το πέμπτο αυτό των ζώων (Animalia). Βοηθητικό είναι το Σχήμα 2.1 [1].



Σχήμα 2.1: Τα πέντε βασίλεια της ζωής.

Αξίζει να σημειωθεί ότι οι *ιοί* δεν ανήκουν σε καμία κατηγορία των ζωντανών οργανισμών. Δεν αποτελούνται από κύτταρα, αλλά είναι μικρότεροι σε μέγεθος και η δομή τους είναι πιο απλή. Επιπλέον, δε γίνονται στη δική τους υπόσταση οι μεταβολικές διεργασίες για την παραγωγή ενέργειας, αφού παρασιτούν σε άλλους οργανισμούς. Εκεί μόνο μπορούν να αναπτυχθούν και να αναπαραχθούν, επομένως δεν μπορούν να θεωρηθούν ανεξάρτητοι οργανισμοί.

[1] Πηγή: http://www.palaeos.com/Kingdoms/kingdoms.htm

2.1.2 Πειραματικά μοντέλα

Για τη μελέτη της δομής και της λειτουργίας των οργανισμών χρησιμοποιούνται συνήθως κάποιοι συγκεκριμένοι, πάνω στους οποίους γίνονται τα πειράματα των βιολόγων. Τα χαρακτηριστικά που χρειάζονται, ώστε ένας οργανισμός να θεωρηθεί κατάλληλος για μοντέλο, ποικίλλουν.

Για τα περισσότερα τέτοια χαρακτηριστικά είναι σχετικά εύκολο να κατανοηθεί ο λόγος για τον οποίο απαιτούνται. Η ευκολία με την οποία μπορούν να αναπαραχθούν, αλλά και να διατηρηθούν σε μεγάλες ποσότητες οι οργανισμοί των πειραμάτων, είναι ένα από αυτά. Σημασία έχει, επίσης, κατά πόσο μπορεί να εμφυτευθεί σε αυτούς γενετικό υλικό ξένο, όπως και να εκφραστεί μέσα τους και να μελετηθούν πιθανές μεταλλάξεις. Οι έννοιες αυτές εξηγούνται στο Κεφάλαιο 3. Τέλος, ένας ακόμη παράγοντας είναι το κατά πόσο είναι αποκωδικοποιημένο το γονιδίωμα του οργανισμού, δηλαδή αν είναι γνωστή η ακολουθία των βάσεων των νουκλεοτιδίων που συνθέτουν το γενετικό του υλικό.

Παραδείγματα οργανισμών που χρησιμοποιούνται σε πειράματα υπάρχουν αρκετά. Τα πιο κοινά είναι το βακτήριο Escherichia coli, που είναι μονοκύτταρο, η μαγιά (yeast), η οποία ανήκει στους μύκητες, το φυτό κάρδαμο (thale cress), ένα είδος σκουληκιού (nematode worm), το οποίο ήταν μάλιστα ο πρώτος πολυκύτταρος οργανισμός του οποίου ολοκληρώθηκε η αποκωδικοποίηση του γονιδιώματος, η μύγα (drosophila melanogaster), ένα είδος ψαριού (zebrafish) και βέβαια, το ποντίκι (mus musculus).

Παρότι οι παραπάνω οργανισμοί ακούγεται να είναι πολύ διαφορετικοί μεταξύ τους αλλά και με τον άνθρωπο, είναι πραγματικά αξιοπρόσεκτο το πόσες ομοιότητες μπορούν να βρεθούν μεταξύ τους. Σε αυτές μάλιστα στηρίζονται οι επιστήμονες που ερευνούν την πιθανότητα όλοι οι οργανισμοί να έχουν κοινή καταγωγή. Με τις ομοιότητες αυτές ασχολείται ουσιαστικά όλο το υπόλοιπο κεφάλαιο.

Ένα ιδιαίτερο παράδειγμα που δημιουργεί ερωτηματικά αλλά και ενδιαφέρον αφορά τη λειτουργία που υπαγορεύουν δύο γονίδια διαφορετικών οργανισμών. Αν αφαιρεθεί το γονίδιο eyeless από το γονιδίωμα της μύγας (drosophila melanogaster), τότε αυτή που θα γεννηθεί δε θα έχει μάτια. Αντίστοιχα, αν συμβεί το ίδιο με το γονίδιο aniridia του ανθρώπου, τότε αυτός θα έχει μάτια χωρίς ίριδες. Αν το γονίδιο aniridia αντικαταστήσει το eyeless της μύγας, η μύγα που προκύπτει έχει μάτια.

Αρκετές απορίες που πηγάζουν από πειράματα μένουν αναπάντητες σήμερα. Ένας σοβαρός λόγος για τον οποίο συμβαίνει αυτό είναι ότι δεν υπάρχουν τα κατάλληλα μέσα, για να μελετηθούν οι οργανισμοί. Ως προς το προηγούμενο παράδειγμα, είναι εξαιρετικά δύσκολο χωρίς ηλεκτρονικό υπολογιστή να συγκριθούν οι ακολουθίες που σχηματίζουν τα δύο γονίδια. Ίσως οι ηλεκτρονικοί υπολογιστές καταφέρουν να συνδράμουν πολύ περισσότερο και από το ηλεκτρονικό μικροσκόπιο στο έργο των βιολόγων.

2.2 Κύτταρο

Όλοι οι οργανισμοί αποτελούνται από κύτταρα και αυτά είναι η μικρότερη δομή στη φύση, όπου εμφανίζεται το φαινόμενο της ζωής. Μέχρι το 1665 η ύπαρξή του δεν ήταν απόλυτα βέβαιη για τους επιστήμονες, ενώ χρειάστηκαν δύο αιώνες περίπου για να διατυπωθεί η *κυτταρική θεωρία* (1839) και να λάβει την τελική της μορφή (1885). Σύμφωνα με αυτήν, η θεμελιώδης δομική και λειτουργική μονάδα όλων των οργανισμών είναι το κύτταρο και καθένα προέρχεται από ένα άλλο κύτταρο.

Οι σημαντικότερες ιδιότητές τους είναι κοινές για οποιονδήποτε οργανισμό. Όλα δομούνται από τις ίδιες χημικές ενώσεις και εκδηλώνουν παρόμοιες μεταβολικές διεργασίες. Η μεταφορά ουσιών στο εσωτερικό τους, η αλλαγή θέσης των κυτταρικών δομών, όταν χρειάζεται, και οι πολύπλοκες βιοχημικές διαδικασίες είναι κάποιες από αυτές. Η συνεργασία μεταξύ τους έχει ως τελικό αποτέλεσμα τη λειτουργία των οργανισμών.

2.2.1 Είδη

Διακρίνονται δύο μεγάλες κατηγορίες κυττάρων. Υπάρχουν εκείνα στα οποία το γενετικό τους υλικό περιβάλλεται από μια μεμβράνη και σχηματίζεται έτσι ο *πυρήνας*, όπως και εκείνα στα οποία δε συμβαίνει αυτό. Τα πρώτα ονομάζονται *ευκαρυωτικά* (κάρυο = πυρήνας, ευ = καλώς, ευκαρυωτικός = με καλά σχηματισμένο πυρήνα), ενώ τα δεύτερα *προκαρυωτικά*. Η δομή των ευκαρυωτικών κυττάρων είναι συνθετότερη και με αυτά θα ασχοληθεί η μελέτη των επόμενων ενοτήτων.

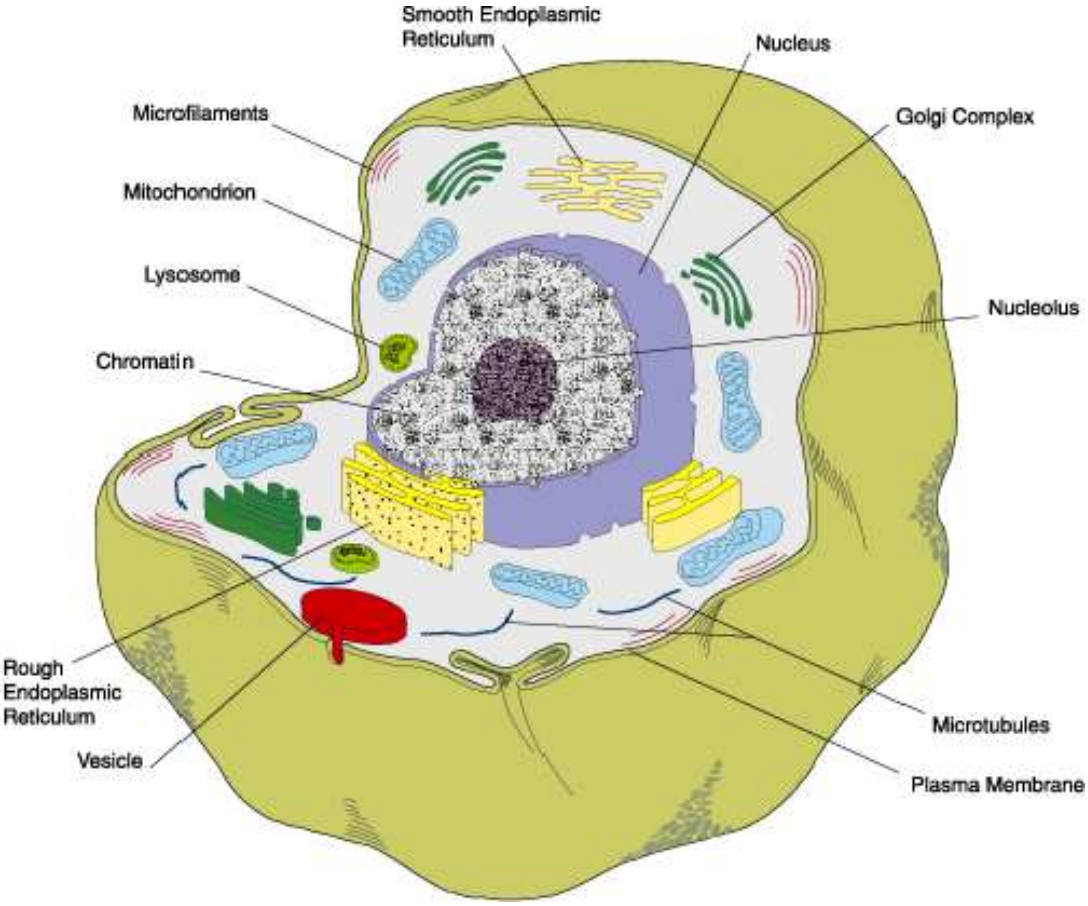
Είναι απαραίτητο να σημειωθεί, ωστόσο, ότι τα είδη των κυττάρων είναι πολλά. Στον άνθρωπο υπάρχουν περίπου 100 διαφορετικά, το καθένα από τα οποία έχει ως στόχο να εξυπηρετήσει διαφορετική λειτουργία (κύτταρα νευρικά, μυικά, κ.ο.κ.) και για αυτό έχει και διαφορετική δομή. Το λεγόμενο *τυπικό κύτταρο*, στο οποίο αναφέρονται οι επόμενες ενότητες, είναι αυτό που συγκεντρώνει όλα τα κοινά γνωρίσματα, εξυπηρετεί τη μελέτη, αλλά ουσιαστικά είναι ανύπαρκτο.

2.2.2 Δομή και λειτουργία

Η δομή ενός τυπικού ζωικού κυττάρου φαίνεται στο Σχήμα 2.2 [2]. Οι διαφορές που έχει σε σχέση με ένα φυτικό κύτταρο δεν είναι πολλές και θα αναφερθούν στη συνέχεια. Αφορούν κυρίως την παρουσία ή μη συγκεκριμένων οργανιδίων.

Μια σημαντική παρατήρηση, που υποβοηθάται από το σχήμα, είναι η έντονη παρουσία μεμβρανών. Αυτές αποτελούνται από φωσφολιπίδια, στερεοειδή και πρωτεΐνες. Εκείνο που χωρίζει το κύτταρο από το εξωτερικό περιβάλλον είναι η *πλασματική μεβράνη*. Ο ρόλος της δεν είναι μόνο να το οριοθετεί, αλλά και να ελέγχει τις ουσίες και τα μηνύματα που αυτό ανταλλάσσει με τον έξω χώρο. Στο εσωτερικό του κυττάρου υπάρχει το *ενδοπλασματικό σύστημα* που περιλαμβάνει τα οργανίδια *ενδοπλασματικό δίκτυο (αδρό και λείο), σύμπλεγμα*

[2] Πηγή: http://www.paternityexperts.com/images/animal%20cell.gif



Σχήμα 2.2: Τυπικό ζωικό κύτταρο.

Golgi, λυσοσώματα, υπεροξειδιοσώματα, κενοτόπια. Η παρουσίαση της λειτουργίας καθενός από αυτά ξεφεύγει από το σκοπό της παρούσας εργασίας.

Το πιο ευδιάκριτο οργανίδιο, αλλά και το κέντρο ελέγχου του κυττάρου, είναι ο *πυρήνας*. Συνήθως υπάρχει ένας ανά κύτταρο, αλλά μπορούν να είναι και περισσότεροι. Η διάμετρός του είναι 5-10μm. Περιβάλλεται και διαχωρίζεται από το *κυτταρόπλασμα* μέσω της *πυρηνικής μεμβράνης*. Στο εσωτερικό του βρίσκεται η μεγαλύτερη ποσότητα του γενετικού υλικού του κυττάρου με τη μορφή της *χρωματίνης*. Επιπλέον, υπάρχουν ένας ή περισσότεροι *πυρηνίσκοι*, στους οποίους γίνεται η σύνθεση του rRNA. Περισσότερα για τη χρωματίνη και το RNA δίνονται στην ενότητα του κεφαλαίου τη σχετική με τα νουκλεϊκά οξέα.

Δύο ακόμη σημαντικά οργανίδια είναι τα *μιτοχόνδρια* και οι *χλωροπλάστες*. Ο ρόλος τους είναι να μετατρέπουν την εξωτερική ενέργεια που λαμβάνει το κύτταρο σε χρησιμοποιήσιμη μορφή, ώστε να καλύψει τις ανάγκες του. Είναι σημαντικό ότι τα οργανίδια αυτά περιέχουν DNA, δηλαδή γενετικό υλικό το οποίο τους επιτρέπει να είναι ανεξάρτητα από αυτό του πυρήνα. Πολλαπλασιάζονται, δηλαδή, από μόνα τους ανάλογα με τις ανάγκες του κυττάρου και συνθέτουν τα ίδια κάποιες από τις πρωτεΐνες που χρειάζονται.

Τα *ριβοσώματα*, που απασχολούν και σε επόμενες ενότητες, είναι μικροί σχηματισμοί στους οποίους γίνεται η πρωτεϊνοσύνθεση. Βρίσκονται πάνω στο αδρό ενδοπλασματικό δίκτυο αλλά και ελεύθερα στο κυτταρόπλασμα, στα μιτοχόνδρια και τους χλωροπλάστες. Οι πρωτεΐνες, μετά τη σύνθεσή τους εκεί, μετακινούνται μέσω των αγωγών του δικτύου, μέσα στους οποίους μπορούν να υποστούν τροποποιήσεις.

Μια σχετικά πρόσφατη ανακάλυψη των βιολόγων λόγω της ηλεκτρονικής μικροσκοπίας είναι ο *κυτταρικός σκελετός*. Πρόκειται για ένα πλέγμα από ινίδια και σωληνίσκους διαφόρων μεγεθών που υποστηρίζουν μηχανικά το κύτταρο. Σε αυτά κυρίως οφείλεται το σχήμα του και η διατήρηση της θέσης κάθε οργανιδίου στο εσωτερικό, ενώ βοηθούν επίσης την κίνηση του κυττάρου.

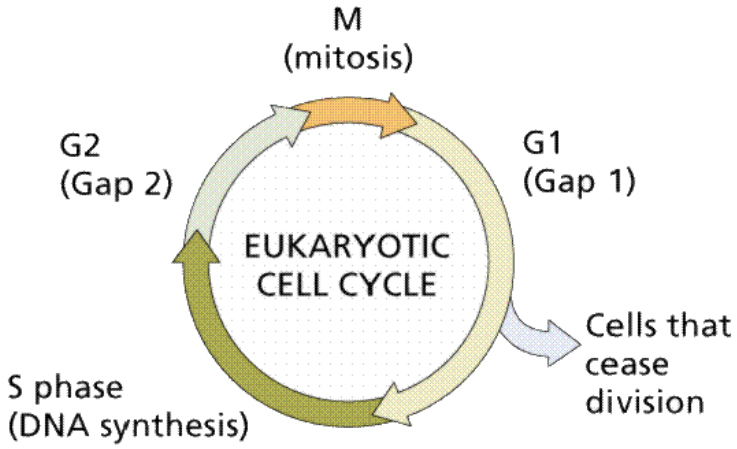
Οι σημαντικότερες διαφορές φυτικών και ζωικών κυττάρων έχουν να κάνουν με οργανίδια. Για παράδειγμα, οι χλωροπλάστες υπάρχουν μόνο στα φυτικά καθώς και το *κυτταρικό τοίχωμα*, που είναι ένα ανθεκτικό εξωτερικό περίβλημα, όπου υπάρχει η κυτταρίνη. Αντίθετα, μόνο στα ζωικά κύτταρα υπάρχει το *κεντροσωμάτιο*, ένα σύμπλεγμα μικροσωληνίσκων που σχετίζεται με τη διαίρεση του κυττάρου.

Τέλος, είναι ενδιαφέρον να επισημανθεί μια γενική παρατήρηση που αφορά τη μορφή του κυττάρου. Το σχήμα του είναι τέτοιο, ώστε να έχει μικρό όγκο και τη μεγαλύτερη δυνατή επιφάνεια. Με αυτόν τον τρόπο ικανοποιεί δύο κρίσιμες απαιτήσεις. Αφενός μπορεί να ανταλλάσσει πολλές ουσίες και μηνύματα με το περιβάλλον του (λόγω της μεγάλης εξωτερικής του επιφάνειας), αφετέρου αυτά μεταβιβάζονται έγκαιρα στο εσωτερικό του (λόγω του μικρού του όγκου).

2.2.3 Κύκλος ζωής

Η χρονική περίοδος ζωής ενός κυττάρου, δηλαδή το διάστημα από τη στιγμή της δημιουργίας του μέχρι αυτήν της διαίρεσής του σε δύο νέα κύτταρα, ονομάζεται κυτταρικός κύκλος. Το

Σχήμα 2.3 [3] δείχνει τις φάσεις της ζωής του.



Σχήμα 2.3: Ο κύκλος ζωής ενός κυττάρου.

Οι φάσεις G1, S, G2, δηλαδή όλες εκτός από τη μίτωση, ονομάζονται *μεσόφαση*. Η G1 είναι η μεγαλύτερη σε διάρκεια, ενώ η S είναι η μικρότερη. Στο στάδιο G1 γίνεται η σύνθεση mRNA, tRNA, ριβοσωμάτων και πρωτεϊνών. Σχετικά με αυτά στοιχεία δίνονται στην ενότητα για τα μακρομόρια. Στο S το γενετικό υλικό διπλασιάζεται. Το στάδιο G2 είναι μεταβατικό πριν τη μίτωση.

Η *μίτωση* είναι εκείνη η περίοδος κατά την οποία το κύτταρο διαιρείται. Χωρίζεται σε τέσσερις περιόδους, που είναι οι: πρόφαση, μετάφαση, ανάφαση, τελόφαση. Είναι τότε που η χρωματίνη παίρνει τη μορφή των χρωμοσωμάτων, όπως θα αναλυθεί στην ενότητα για τα νουκλεϊκά οξέα.

Τέλος, ορισμένα είδη κυττάρων πολυκύτταρων οργανισμών δεν ολοκληρώνουν αυτόν τον κύκλο. Για παράδειγμα, τα νευρικά κύτταρα του ανθρώπου δεν πολλαπλασιάζονται. Αυτό σημαίνει ότι με όσα δημιουργηθούν από την αρχή, με αυτά θα συνεχίσει ο οργανισμός ολόκληρη τη ζωή του.

2.3 Μακρομόρια

Τα μακρομόρια είναι σύνθετες οργανικές ενώσεις μεγάλου μοριακού βάρους, οι οποίες σχηματίζονται με συγκεκριμένο χημικό μηχανισμό και παίζουν πολύ σημαντικό ρόλο στη λειτουργία του κυττάρου. Παραδείγματα είναι οι πρωτεΐνες, τα νουκλεϊκά οξέα, τα λιπίδια και οι υδατάνθρακες.

Τα μακρομόρια είναι πολυμερή, δηλαδή σχηματίζονται από την ένωση πολλών μονομερών μεταξύ τους. Για να δημιουργηθεί ένα διμερές, χάνεται ένα μόριο νερού από δύο μονομερή (μια υδροξυλομάδα από το ένα και ένα άτομο υδρογόνου από το άλλο) και τα τελευταία συνδέονται με ομοιοπολικό δεσμό. Η διαδικασία αυτή ονομάζεται *συμπύκνωση*, ενώ η αντίστροφη είναι

[3] Πηγή: http://www.emc.maricopa.edu/faculty/farabee/BIOBK/cellcycle.gif

η *υδρόλυση*. Ανάλογα παράγονται οι αλυσίδες από περισσότερα μονομερή, μέχρι τελικά να γίνει το μακρομόριο.

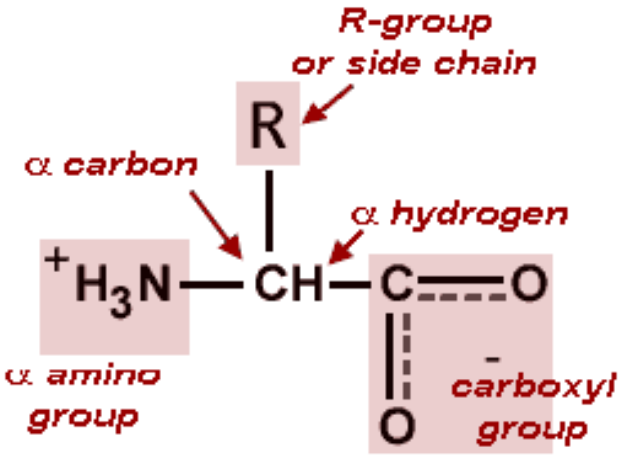
Τα μονομερή που θα αναλυθούν στη συνέχεια της παρούσας εργασίας είναι τα αμινοξέα και τα νουκλεοτίδια. Τα πρώτα είναι οι δομικές μονάδες των πρωτεϊνών και τα δεύτερα εκείνες των νουκλεϊκών οξέων.

2.3.1 Πρωτεΐνες

Από το όνομα που έχει δοθεί σε αυτό το μακρομόριο — όνομα που προέρχεται από τη λέξη πρώτος — υποψιάζεται κανείς και την ιδιαίτερη αξία του για τη ζωή. Είτε χρησιμεύει ως δομικό συστατικό είτε εξυπηρετεί συγκεκριμένη λειτουργία ενός κυττάρου. Θεωρείται ότι είναι το πιο διαδεδομένο και πολυδιάστατο στη μορφή και τη λειτουργία του μακρομόριο. Ακόμη και σε ένα κύτταρο βακτηρίου σαν αυτό της E. coli, το οποίο θεωρείται απλό κύτταρο, υπάρχουν εκατοντάδες διαφορετικές πρωτεΐνες.

Αμινοξέα

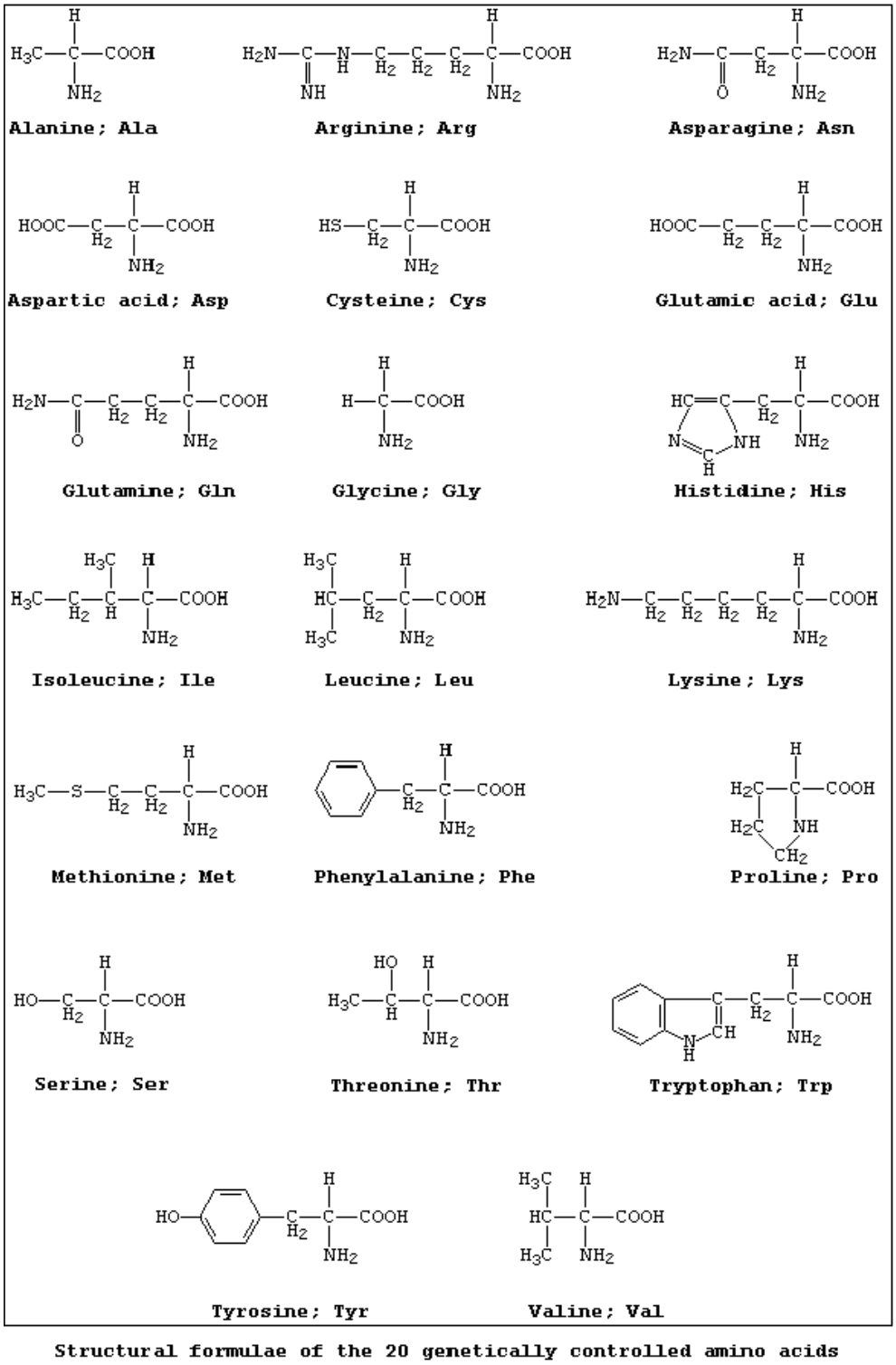
Όπως αναφέρθηκε και στην αρχή της Ενότητας 2.3, η δομική μονάδα των πρωτεϊνών είναι τα αμινοξέα. Παρότι έχουν ανιχνευθεί περισσότερα από 170, πιστεύεται ότι 20 μόνο χρησιμοποιούνται για τη σύνθεση των πρωτεϊνών, αν και τελευταία ο αριθμός αυτός έχει τεθεί υπό αμφισβήτηση και λέγεται ότι τελικά πρέπει να είναι περισσότερα από 20.



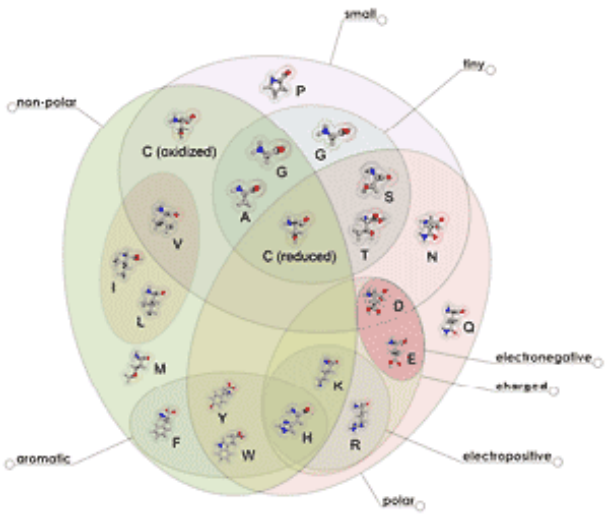
Σχήμα 2.4: Η χημική σύσταση ενός αμινοξέος.

Από την πλευρά της χημικής δομής, τα αμινοξέα διαφέρουν μεταξύ τους μόνο κατά τη λεγόμενη πλευρική ομάδα R. Όπως φαίνεται στο Σχήμα 2.4 [4], κάθε αμινοξύ αποτελείται από ένα άτομο άνθρακα, στο οποίο συνδέονται μία αμινομάδα, μία καρβοξυλομάδα, ένα άτομο υδρογόνου και μία πλευρική ομάδα.

[4] Πηγή: http://www.biology.arizona.edu/biochemistry/problem_sets/aa/Graphics/ChemBasicLabelled.gif



Σχήμα 2.5: Τα 20 αμινοξέα που συνθέτουν τις πρωτεΐνες.



Σχήμα 2.6: Διάγραμμα Venn για τα 20 αμινοξέα.

Τα 20 διαφορετικά αμινοξέα, που προκύπτουν από τις 20 διαφορετικές πλευρικές ομάδες και είναι εκείνα που συμμετέχουν στη σύνθεση των πρωτεϊνών ως δομικές τους μονάδες, φαίνονται στο Σχήμα 2.5 [5]. Επιπλέον, στο σχήμα 2.6 [6] έχει χαραχθεί ένα διάγραμμα Venn ώστε να γίνουν καλύτερα αντιληπτά τα κοινά χαρακτηριστικά που παρουσιάζουν.

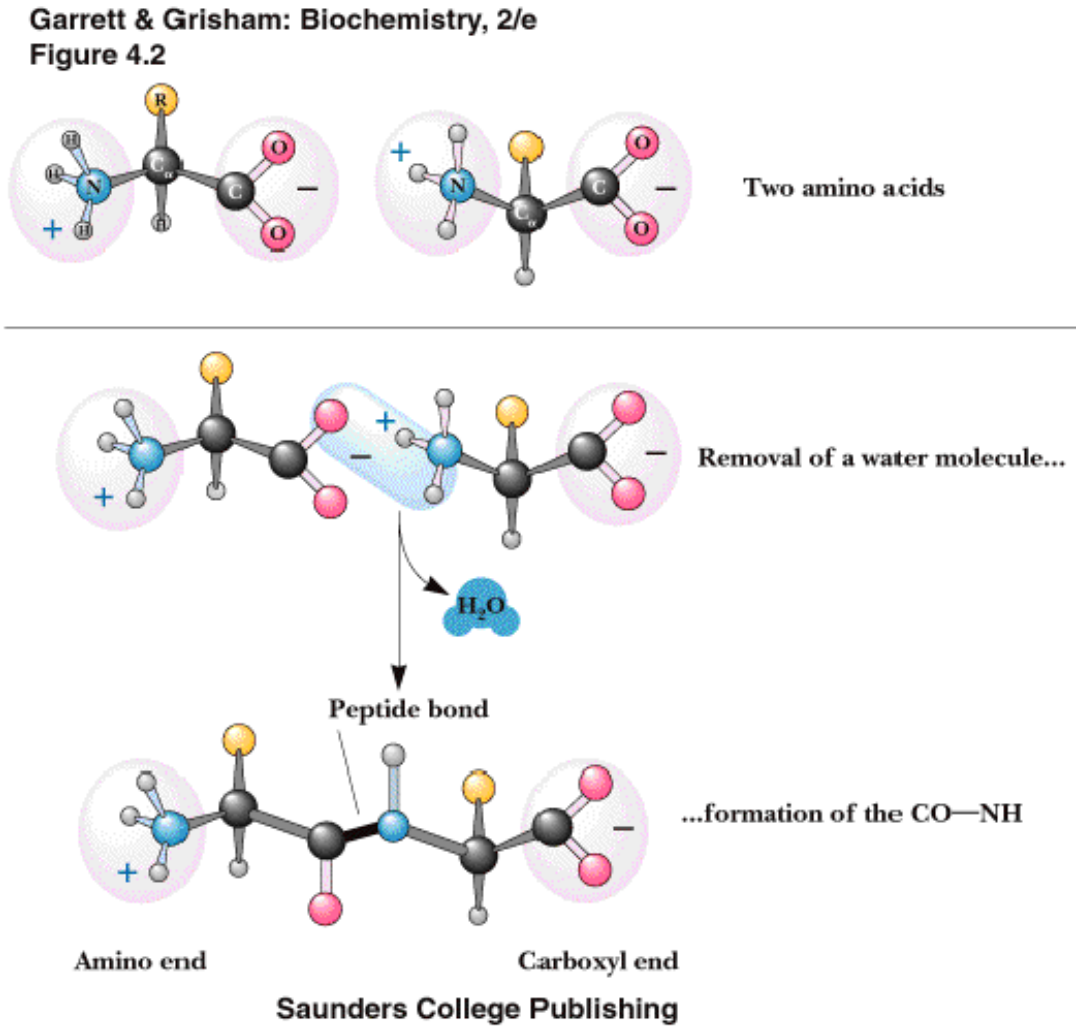
Ο ομοιοπολικός δεσμός που δημιουργείται όταν ενώνονται δύο αμινοξέα ονομάζεται *πεπτιδικός*. Με τη συμπύκνωση, που αναφέρθηκε στην αρχή της Ενότητας 2.3, αποδεσμεύεται ένα μόριο νερού, που προέρχεται από το υδροξύλιο -OH της καρβοξυλομάδας του πρώτου αμινοξέος και το υδρογόνο -H της αμινομάδας του δεύτερου αμινοξέος (βλ. Σχήμα 2.7). Οι δύο αυτές ομάδες ανήκουν στο σταθερό τμήμα ενός αμινοξέος, οπότε είναι ίδιος ο μηχανισμός της ένωσης ανεξάρτητα από τον τύπο (δηλαδή την πλευρική ομάδα) κάθε αμινοξέος.

Διάταξη στο χώρο και λειτουργία

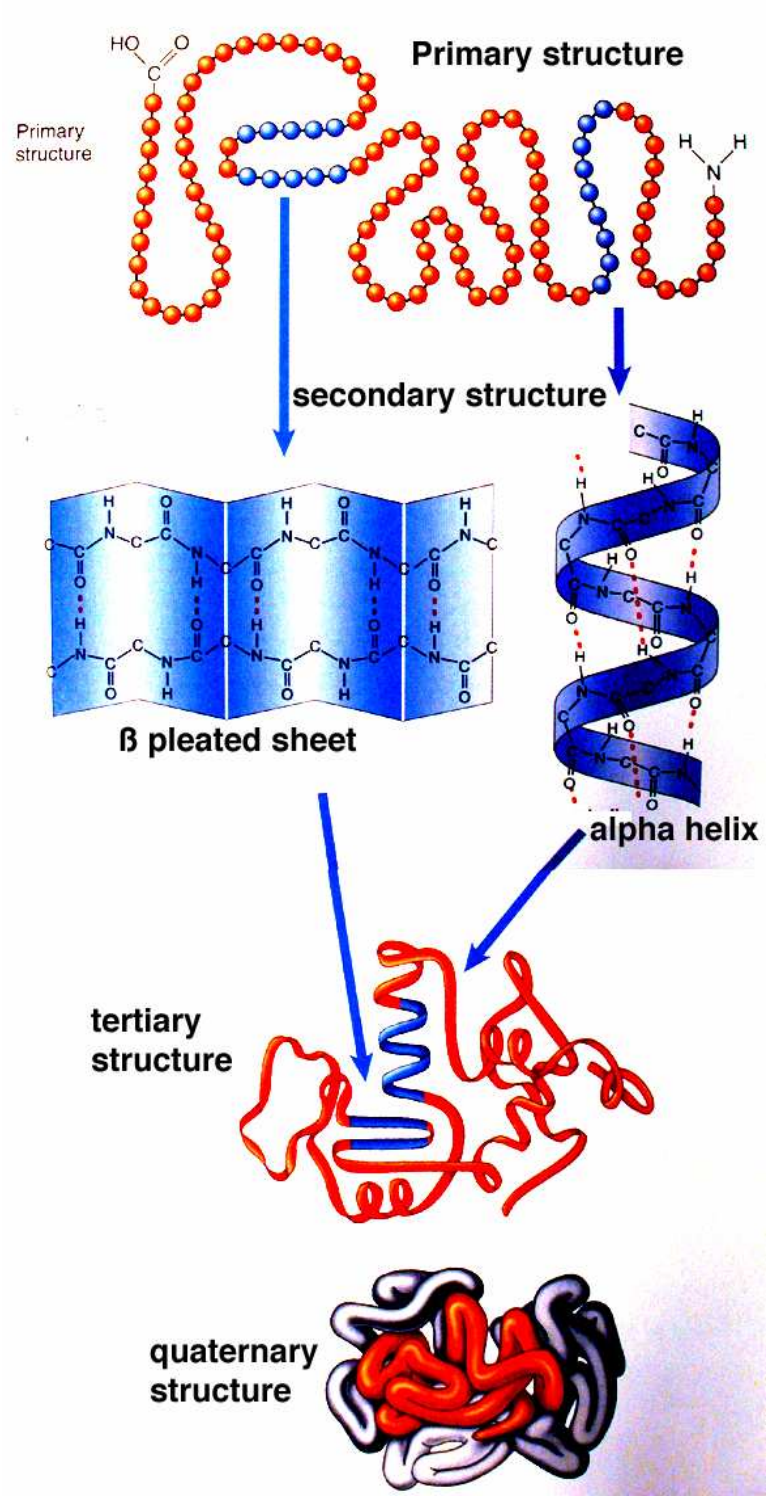
Από όσα έχουν μέχρι στιγμής αναφερθεί πιθανώς να βγάλει κανείς το συμπέρασμα ότι, για να γνωρίζει κανείς τις ιδιότητες μιας πρωτεΐνης, είναι αρκετό να ξέρει την αλληλουχία των αμινοξέων που την αποτελούν. Η αλήθεια, όμως, είναι πως η λειτουργία μιας πρωτεΐνης είναι αποτέλεσμα όχι μόνο της ακολουθίας των αμινοξέων αλλά και της μορφής που έχει αυτή στο χώρο. Ανάλογα με το βιολογικό της ρόλο παίρνει και το κατάλληλο σχήμα.

Διακρίνονται τέσσερα επίπεδα οργάνωσης των πρωτεϊνικών μορίων, τα οποία και φαίνονται στο Σχήμα 2.8 [7]. Η πρωτοταγής δομή είναι η σειρά που έχουν τα αμινοξέα στην πολυπεπτιδική αλυσίδα. Η δευτεροταγής δομή είναι η αναδίπλωση του μορίου στο χώρο, που έχει ως αποτέλεσμα ελικοειδή ή πτυχωτή μορφή. Στο τρίτο επίπεδο, που είναι η τριτοταγής δομή, η πρωτεΐνη αναδιπλώνεται ξανά και λαμβάνει καθορισμένη μορφή. Τέλος, αν η πρωτεΐνη απο-

[5] Πηγή: http://www.faqs.org/nutrition/images/nwaz_02_img0196.jpg
[6] Πηγή: http://www.symmation.com/gallery/images/amino-acids-Venn-Diagram.gif
[7] Πηγή: http://academic.brooklyn.cuny.edu/biology/bio4fv/page/prot_struct4143.JPG



Σχήμα 2.7: Ο σχηματισμός ενός διπεπτιδίου.



Σχήμα 2.8: Η οργάνωση μιας πρωτεΐνης στο χώρο.

τελείται από περισσότερες από μία αλυσίδες, τότε η τεταρτοταγής δομή είναι ο συνδυασμός των επιμέρους πολυπεπτιδικών αλυσίδων σε ένα ενιαίο μόριο.

Χρειάζεται να σημειωθεί ότι η πρωτοταγής δομή, δηλαδή η αλληλουχία των αμινοξέων, επηρεάζει τη διαμόρφωση της πρωτεΐνης στο χώρο, αφού η τελευταία εξαρτάται από τους χημικούς δεσμούς που συνάπτονται ανάμεσα στις πλευρικές ομάδες των αμινοξέων.

Η μεγάλη ποικιλία πρωτεϊνών, λοιπόν, γίνεται αντιληπτό ότι είναι δυνατή λόγω των τόσων διαφορετικών συνδυασμών που μπορούν να προκύψουν με τα 20 διαφορετικά αμινοξέα σε πολυπεπτίδια. Στο ανθρώπινο σώμα υπάρχουν περισσότερες από 30000 διαφορετικές πρωτεΐνες, οι οποίες είναι δομικές, αποτελούν δομικά συστατικά των κυττάρων, ή λειτουργικές, δηλαδή συμμετέχουν στις διάφορες διαδικασίες του.

Τέλος, οι πρωτεΐνες είναι ευαίσθητες στις μεταβολές χαρακτηριστικών του περιβάλλοντός τους, όπως η θερμοκρασία και το pH. Όταν εκτεθούν σε ακραίες συνθήκες, υφίστανται *μετουσίωση*, δηλαδή σπάνε οι δεσμοί ανάμεσα στις πλευρικές ομάδες, καταστρέφεται η τρισδιάστατη δομή τους και τελικά χάνουν τη λειτουργικότητά τους.

Ένζυμα

Τα ένζυμα είναι μία από τις κατηγορίες των λειτουργικών πρωτεϊνών. Η δράση τους έγκειται στην κατάλυση χημικών αντιδράσεων εντός ή εκτός των κυττάρων. Τα ίδια δε συμμετέχουν ως προϊόντα ή αντιδρώντα σε αυτές, οπότε παραμένουν αναλλοίωτα μετά την πραγματοποίησή τους.

Ο ρόλος τους είναι να επιταχύνουν αντιδράσεις που θα μπορούσαν να γίνουν και χωρίς αυτά αλλά πολύ πιο αργά. Οι ανάγκες των κυττάρων συνήθως είναι άμεσες, για αυτό και η παρουσία των ενζύμων είναι εντελώς απαραίτητη. Η ταχύτητα των αντιδράσεων με αυτά μπορεί να αυξηθεί ακόμη και 100000000 φορές.

Εκτός αυτού, τα ένζυμα μειώνουν την ενέργεια ενεργοποίησης που είναι αναγκαία για να γίνει μία αντίδραση. Χωρίς αυτά, η ενέργεια που θα χρειαζόταν για να ξεκινήσει η αντίδραση, είτε εξώθερμη είτε ενδόθερμη, θα ήταν απαγορευτική για την επιβίωση του κυττάρου, ιδιαίτερα αν αυτή παρεχόταν θερμικά.

Είναι σημαντικό να τονισθεί ο υψηλός βαθμός εξειδίκευσης που εμφανίζουν. Τις περισσότερες φορές ένα ένζυμο καταλύει μία και μόνο αντίδραση. Το γεγονός αυτό έχει ιδιαίτερη αξία για τις μελέτες της δράσης ορισμένων φαρμάκων. Αν για παράδειγμα ένα φάρμακο αναστέλλει τη δράση ενός ενζύμου και είναι γνωστό ότι κανένο άλλο ένζυμο δεν μπορεί να επιταχύνει την ίδια αντίδραση, τότε έχει διακοπεί η πραγματοποίηση της αντίδρασης.

2.3.2 Νουκλεϊκά οξέα

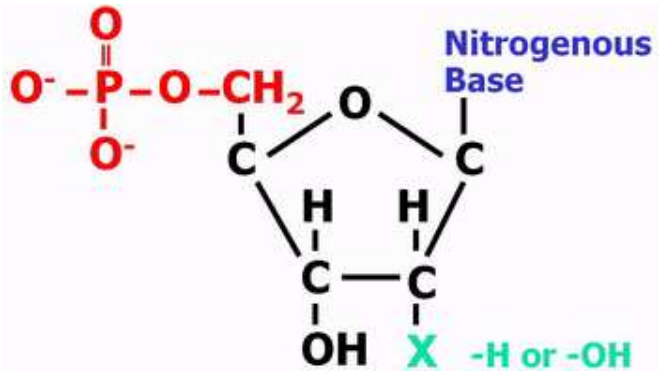
Τα δύο είδη νουκλεϊκών οξέων, το DNA και το RNA, εξάπτουν το ενδιαφέρον όχι μόνο των βιολόγων αλλά και των υπόλοιπων ανθρώπων. Ο λόγος είναι το γεγονός ότι σχετίζονται πολύ στενά με τη γενετική πληροφορία, δηλαδή τα στοιχεία εκείνα που έχουν να κάνουν με την κληρονομικότητα και τα ιδιαίτερα χαρακτηριστικά κάθε ατόμου.

Σε αντίθεση με τις πρωτεΐνες, τα νουκλεϊκά οξέα βρίσκονται μόνο εντός του κυττάρου.

Στα ευκαρυωτικά κύτταρα η μεγαλύτερη ποσότητα του DNA βρίσκεται στον πυρήνα, αλλά υπάρχει μικρό μέρος του και σε δύο άλλα οργανίδια του κυττάρου, τα μιτοχόνδρια και τους χλωροπλάστες. Σε πολλά βακτήρια πολύ μικρό ποσοστό του DNA τους περιέχεται στα πλασμίδια, που είναι ανεξάρτητα από το κεντρικό μόριο DNA. Τέλος, το RNA μπορεί να εντοπιστεί και εντός και εκτός του πυρήνα.

Νουκλεοτίδια

Η δομική μονάδα των νουκλεϊκών οξέων είναι τα νουκλεοτίδια. Η χημική τους σύσταση φαίνεται στο Σχήμα 2.9 [8]. Αποτελούνται από μία πεντόζη, δηλαδή ένα σάκχαρο με πέντε άτομα άνθρακα, ένα μόριο φωσφορικού οξέος και μία αζωτούχα βάση. Τα νουκλεοτίδια του DNA έχουν την πεντόζη δεσοξυριβόζη, ενώ εκείνα του RNA περιέχουν την πεντόζη ριβόζη. Για το λόγο αυτό το DNA λέγεται δεσοξυριβονουκλεϊνικό, ενώ το RNA ριβονουκλεϊνικό οξύ.



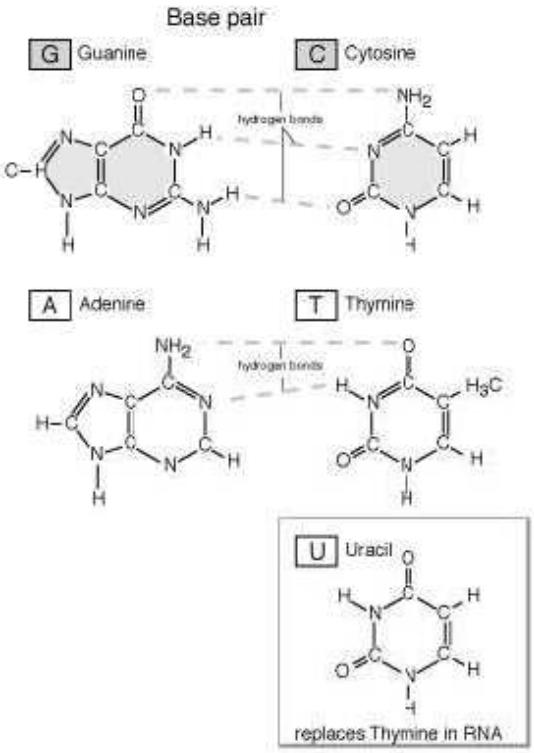
Σχήμα 2.9: Η χημική σύσταση ενός νουκλεοτιδίου.

Όπως τα αμινοξέα, έτσι και τα νουκλεοτίδια έχουν ένα σταθερό και ένα μεταβλητό μέρος. Αναφερόμενοι στον ίδιο τύπο νουκλεϊκού οξέος, δηλαδή είτε DNA είτε RNA, τα νουκλεοτίδια διαφέρουν μόνο στην αζωτούχα βάση. Για αυτόν ακριβώς το λόγο συνηθίζεται να αναφέρονται τα νουκλεϊκά οξέα ως ακολουθίες βάσεων και όχι νουκλεοτιδίων. Για κάθε είδος νουκλεϊκού οξέος υπάρχουν τέσσερις διαφορετικές βάσεις, το οποίο σημαίνει ότι υπάρχουν μόνο τέσσερα διαφορετικά νουκλεοτίδια, σε αντίθεση με τα είκοσι διαφορετικά αμινοξέα των πρωτεϊνών.

Οι βάσεις που συναντά κανείς στα νουκλεοτίδια του DNA είναι οι: αδενίνη (A), γουανίνη (G), κυτοσίνη (C) και θυμίνη (T). Στο RNA μπορεί να συναντήσει όλες τις προηγούμενες εκτός από τη θυμίνη, αντί της οποίας υπάρχει η ουρακίλη (U). Η χημική σύνθεση των βάσεων αυτών έχει χαραχθεί στο Σχήμα 2.10 [9].

Ένα σημαντικό χαρακτηριστικό αυτών των βάσεων είναι η λεγόμενη *συμπληρωματικότητα*. Ανάμεσα σε συγκεκριμένα ζευγάρια από αυτές μπορούν να σχηματιστούν δεσμοί υδρογόνου, με άλλα λόγια είναι δυνατό να αναπτυχθούν ελκτικές δυνάμεις, όταν τα αντίστοιχα μόρια βρεθούν σε ικανά κοντινή απόσταση. Τα ζεύγη των βάσεων είναι A-T, A-U, G-C. Μάλιστα,

[8] Πηγή: http://graphics.csie.ntu.edu.tw/~zick/bio/diff/nucleotide.jpg
[9] Πηγή: http://www.accessexcellence.org/RC/VL/GG/images/base_pair.gif



Σχήμα 2.10: Τα ζεύγη των συμπληρωματικών αζωτούχων βάσεων.

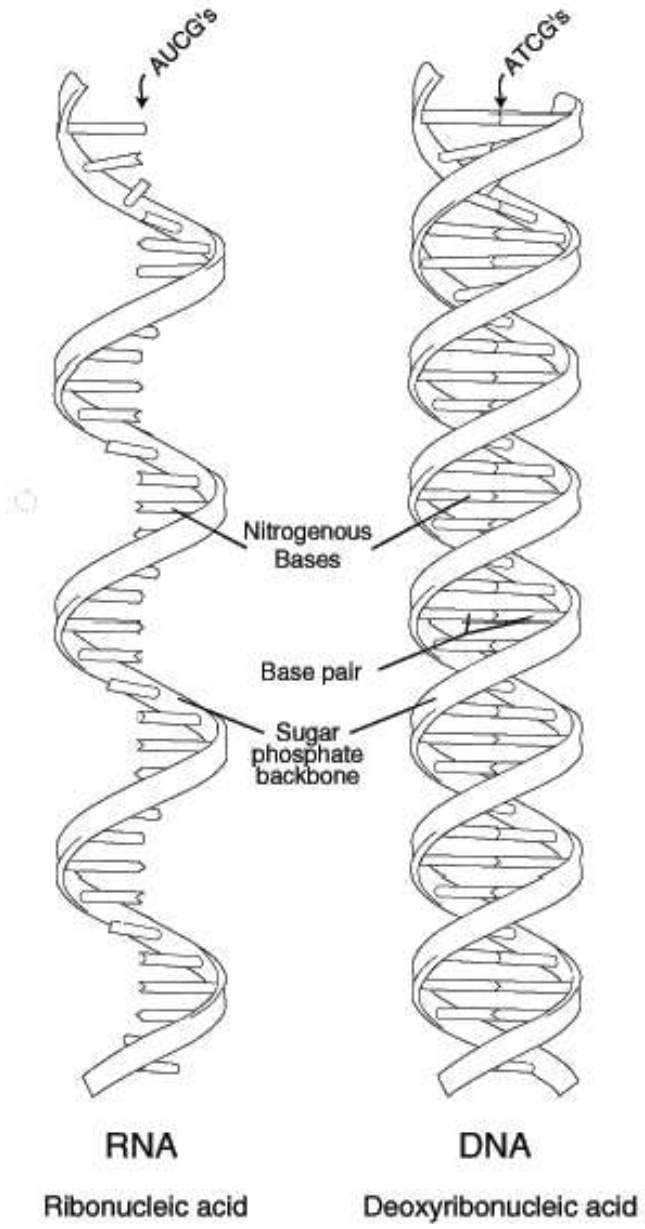
στο Σχήμα 2.10 διακρίνονται οι δύο δεσμοί υδρογόνου μεταξύ αδενίνης και θυμίνης/ουρακίλης και οι τρεις δεσμοί υδρογόνου μεταξύ γουανίνης και κυτοσίνης. Η ιδιότητα της συμπληρωματικότητας είναι πολύ σημαντική για τη δομή του DNA, όπως θα εξηγηθεί στην ενότητα για τη δομή που ακολουθεί.

Τα νουκλεοτίδια συνδέονται μεταξύ τους με ομοιοπολικό δεσμό, για να σχηματίσουν το μακρομόριο του DNA ή του RNA. Αυτό γίνεται με τη διαδικασία της συμπύκνωσης, που εξηγήθηκε στην αρχή της Ενότητας 2.3. Ο τρόπος σύνδεσής τους, επομένως, είναι ακριβώς ανάλογος με αυτόν των αμινοξέων. Ο ομοιοπολικός δεσμός που σχηματίζεται ονομάζεται *φωσφοδιεστερικός*.

Δομή

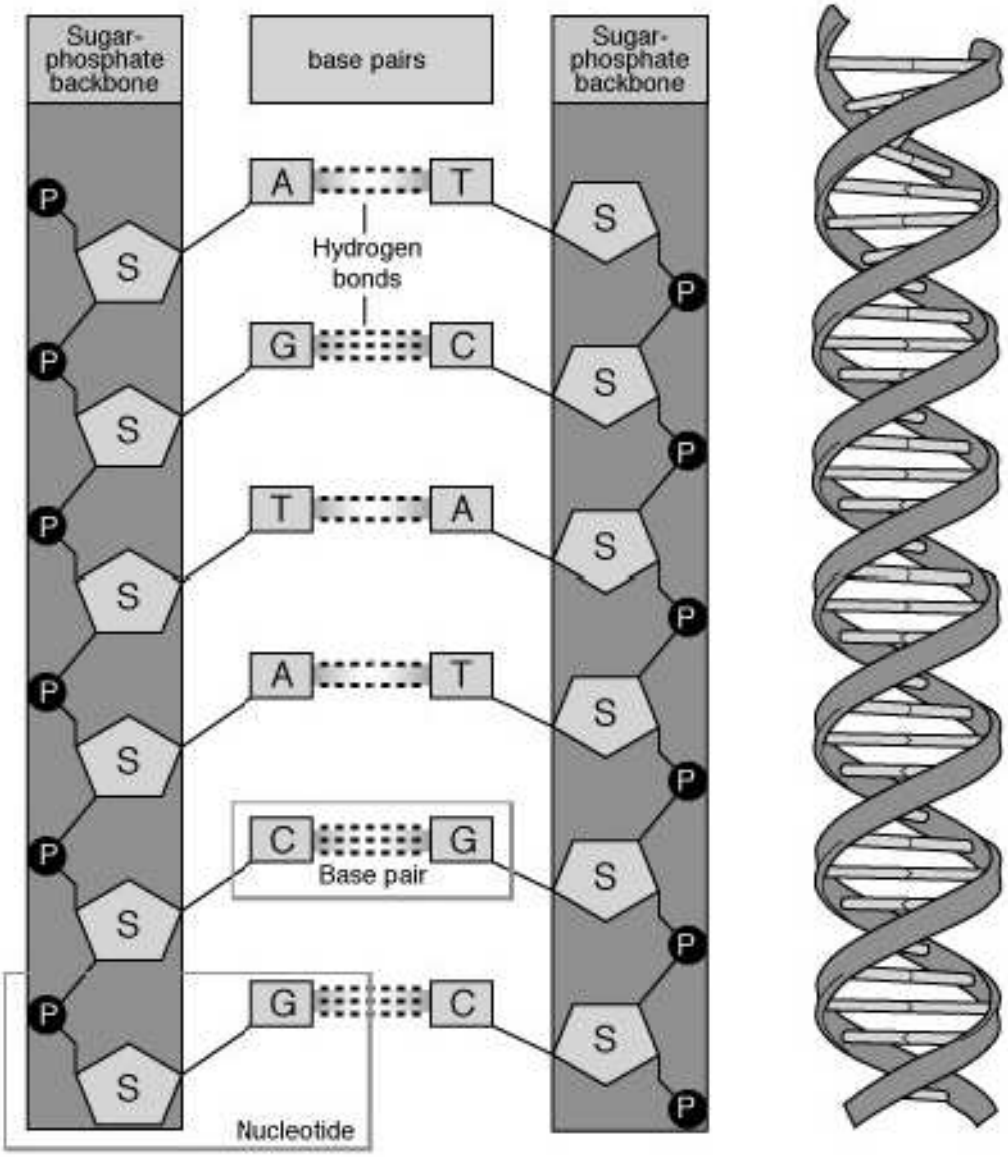
Εκτός από τη σύνθεση των νουκλεοτιδίων τους, το DNA και το RNA παρουσιάζουν μία ακόμη διαφορά στη δομή τους. Ενώ το RNA αποτελείται από μία μόνο αλυσίδα νουκλεοτιδίων, το DNA είναι δίκλωνο. Η διαφορά αυτή γίνεται καλύτερα αντιληπτή παρατηρώντας το Σχήμα 2.11 [10]. Κάποιες φορές, ωστόσο, η μονόκλωνη έλικα του RNA αναδιπλώνεται σε ορισμένα σημεία. Στα επόμενα η προσοχή θα εστιαστεί στη διάταξη του DNA, καθώς είναι περισσότερο πολύπλοκη και παρουσιάζει μεγαλύτερο ενδιαφέρον.

[10] Πηγή: http://www.accessexcellence.org/RC/VL/GG/rna.html



Σχήμα 2.11: Οι έλικες των δύο ειδών νουκλεϊκών οξέων.

Το μοντέλο για τη δομή του DNA παρουσιάστηκε το 1953 από τους J. Watson και F. Crick, εργασία για την οποία τιμήθηκαν το 1962 με το βραβείο Nobel στον τομέα της ιατρικής. Ιστορικά αξίζει να σημειωθεί πως το DNA πρωτοεντοπίστηκε στον πυρήνα των κυττάρων το 1869, ενώ μέχρι το 1944 δεν υπήρχαν σοβαρές ενδείξεις για το γεγονός ότι αποτελεί το γενετικό υλικό των οργανισμών. Για το βιολογικό του ρόλο θα γίνει αναφορά στην επόμενη ενότητα.



Σχήμα 2.12: Η δομή του DNA.

Τα στοιχεία που συνιστούν το λεγόμενο μοντέλο της διπλής έλικας του DNA έχουν ήδη σποραδικά αναφερθεί και συγκεντρώνονται στο Σχήμα 2.12 [11]. Το μόριο του DNA

[11] Πηγή: http://www.accessexcellence.org/RC/VL/GG/dna2.html

αποτελείται από δύο πολυνουκλεοτιδικές αλυσίδες, που σχηματίζουν στο χώρο διπλή έλικα. Ο άξονας κάθε αλυσίδας σχηματίζεται από τη δεσοξυριβόζη και τη φωσφορική ομάδα κάθε νουκλεοτιδίου, ενώ προς το εσωτερικό της έλικας προεξέχουν κάθετα οι αζωτούχες βάσεις. Χάρη στην ιδιότητα της συμπληρωματικότητας των βάσεων, που έχει αναφερθεί στη συζήτηση για τα νουκλεοτίδια, ανάμεσα στους δύο κλώνους ασκούνται ελκτικές δυνάμεις οι οποίες τους συγκρατούν. Επιπλέον, από τη στιγμή που είναι γνωστή η αλληλουχία των βάσεων της μιας αλυσίδας είναι δεδομένη και η αλληλουχία στην άλλη, αφού οι δύο κλώνοι είναι συμπληρωματικοί.

Βιολογικός ρόλος

Το 1952 αποδείχθηκε και τυπικά, δηλαδή πειραματικά (από τους Hershey, Chase), ότι το DNA είναι το γενετικό υλικό των οργανισμών. Εξαίρεση στον κανόνα αυτό αποτελούν κάποιοι ιοί, οι οποίοι έχουν RNA ως γενετικό υλικό.

Το γενετικό υλικό είναι εκείνο στο οποίο είναι γραμμένες όλες οι γενετικές πληροφορίες, δηλαδή όλα τα χαρακτηριστικά που πρέπει να εκφραστούν σε οποιοδήποτε κύτταρο του οργανισμού. Αυτό ρυθμίζει τη λειτουργία των κυττάρων, επομένως καθορίζει όλες τις ιδιότητες ενός οργανισμού.

Τα μόρια του DNA έχουν, λοιπόν, τριπλό ρόλο. Αποθηκεύουν τη γενετική πληροφορία, καθοδηγούν μέσω αυτής τις εργασίες των κυττάρων και τη μεταβιβάζουν αναλλοίωτη από γενιά σε γενιά. Το τελευταίο γεγονός δεν εμποδίζει τη δημιουργία γενετικής ποικιλομορφίας. Ο τρόπος με τον οποίο ελέγχεται η δραστηριότητα των κυττάρων είναι μέσω της παραγωγής των πρωτεϊνών, θέμα που θα εξεταστεί σε μεγαλύτερη λεπτομέρεια στο τρίτο κεφάλαιο.

Ο βιολογικός ρόλος του RNA, πέρα από γενετικό υλικό των ιών, εξαρτάται από τον τύπο του. Οι τέσσερις τύποι RNA αναφέρονται στην επόμενη παράγραφο, όπου και γίνεται λόγος για τη σημασία τους στη ζωή των κυττάρων.

Είδη του RNA

Υπάρχουν τέσσερα είδη μορίων RNA. Το αγγελιοφόρο RNA (mRNA) είναι εκείνο που συντίθεται από τμήμα του DNA, με σκοπό να παραχθεί ένας τύπος πρωτεΐνης. Ο ρόλος του είναι να μεταφέρει την πληροφορία, που καθορίζει την παραγωγή των πρωτεϊνών, από τον πυρήνα του κυττάρου, όπου βρίσκεται το DNA, στα ριβοσώματα, όπου γίνεται η πρωτεϊνοσύνθεση. Το μεταφορικό RNA (tRNA) μεταφέρει τα αμινοξέα στα ριβοσώματα για να γίνει η σύνθεση των πρωτεϊνών. Το τρίτο είδος είναι το μικρό πυρηνικό RNA (snRNA), το οποίο συνδέεται με πρωτεΐνες και σχηματίζει σωματίδια που καταλύουν μία συγκεκριμένη διαδικασία, την ωρίμανση του mRNA. Περισσότερα για αυτούς τους τύπους RNA θα δοθούν στην ενότητα για την παραγωγή των πρωτεϊνών του Κεφαλαίου 3. Τέλος, το ριβοσωμικό RNA (rRNA) αποτελεί μαζί με πρωτεΐνες δομικό συστατικό των ριβοσωμάτων.

Το γονιδίωμα και η οργάνωσή του

Ο όρος γονιδίωμα αναφέρεται στο σύνολο του γενετικού υλικού ενός κυττάρου. Όπως αναφέρθηκε, στην πλειοψηφία των περιπτώσεων το γενετικό υλικό των οργανισμών είναι DNA. Για να μεταβεί κανείς ομαλά από το μοντέλο της διπλής έλικας του DNA στην οργάνωση του γονιδιώματος, χρειάζεται πρώτα να περιγραφούν άλλοι όροι, όπως αυτοί της χρωματίνης και των χρωμοσωμάτων.

Το δίκλωνο μόριο του DNA συσπειρώνεται πάρα πολύ στο χώρο και παίρνει είτε κυκλική μορφή (βακτήρια, μιτοχόνδρια, χλωροπλάστες) είτε ευθεία (ευκαρυωτικά κύτταρα). Στα επόμενα αναφέρεται η οργάνωσή του στα ευκαρυωτικά κύτταρα.

Ο μεγάλος βαθμός συμπύκνωσης του μορίου του DNA φαίνεται στο Σχήμα 2.13 [12]. Η δίκλωνη αλυσίδα τυλίγεται γύρω από πρωτεΐνες (τις ιστόνες) και έτσι σχηματίζεται η *χρωματίνη*. Αυτή, δηλαδή, είναι μία νουκλεοπρωτεΐνη, η οποία αποτελείται από DNA, RNA και πρωτεΐνες. Ανάλογα με το στάδιο ζωής του κυττάρου, η χρωματίνη είναι περισσότερο ή λιγότερο συμπυκνωμένη. Όταν το κύτταρο δεν είναι στη φάση διαίρεσής του, τότε έχει τη μορφή πλέγματος, που είναι το λεγόμενο *δίκτυο χρωματίνης*. Αντίθετα, στη φάση της μίτωσης η συμπίεση είναι πολύ μεγάλη και η μορφή που λαμβάνει η χρωματίνη είναι αυτή των *χρωμοσωμάτων*.

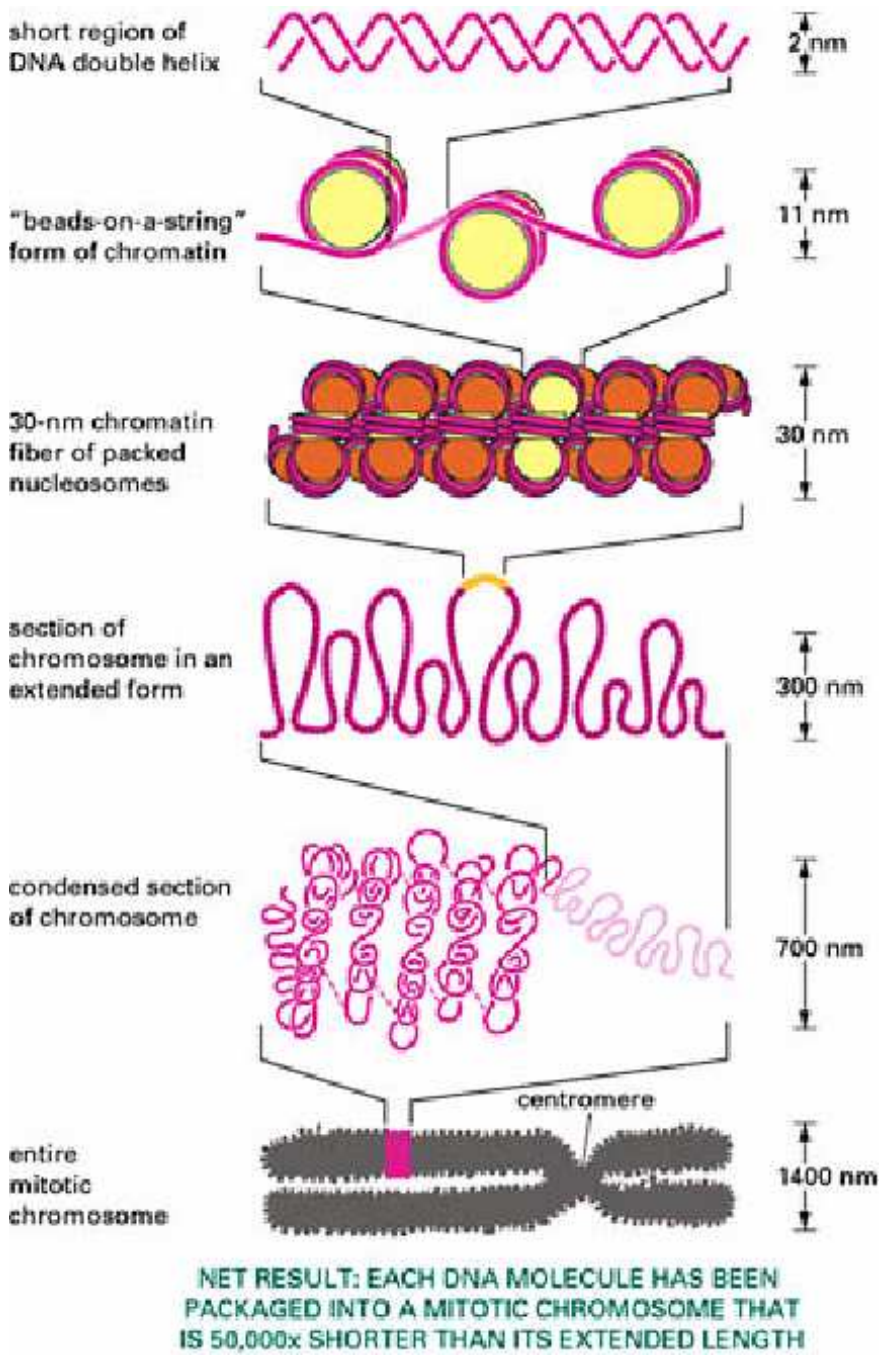
Στο Σχήμα 2.14 [13] διακρίνονται οι ονομασίες για τα τμήματα ενός χρωμοσώματος καθώς και η θέση του σε ένα ευκαρυωτικό κύτταρο. Το χρωμόσωμα αποτελείται από τις δύο αδελφές *χρωματίδες*, οι οποίες συνδέονται σε ένα σημείο, το *κεντρομερίδιο*. Ο αριθμός των χρωμοσωμάτων που υπάρχουν στα διάφορα φυτικά και ζωικά κύτταρα είναι αυστηρά καθορισμένος. Για παράδειγμα, τα κύτταρα του ανθρώπου έχουν 23 ζεύγη χρωμοσωμάτων (σε κάθε ζευγάρι το ένα χρωμόσωμα προέρχεται από τον πατέρα και το άλλο από τη μητέρα), με εξαίρεση τους *γαμέτες*, που είναι τα κύτταρα για την αναπαραγωγή και έχουν 23 απλά χρωμοσώματα.

Ανάλογα με το αν τα χρωμοσώματα ενός κυττάρου βρίσκονται σε ζευγάρια ή όχι, τα κύτταρα χαρακτηρίζονται ως *διπλοειδή* ή *απλοειδή*, αντίστοιχα. Τα χρωμοσώματα που εμφανίζονται στο ίδιο ζευγάρι ονομάζονται *ομόλογα*. Αυτό δε σημαίνει ότι το ένα είναι αντίγραφο του άλλου, αλλά πως έχουν ίδιο σχήμα και μέγεθος και περιέχουν γονίδια που ελέγχουν τα ίδια χαρακτηριστικά και βρίσκονται στην ίδια θέση (*γονιδιακός τόπος*) σε καθένα από τα χρωμοσώματα. Ωστόσο, είναι πιθανό να ορίζουν με διαφορετικό τρόπο τα χαρακτηριστικά αυτά.

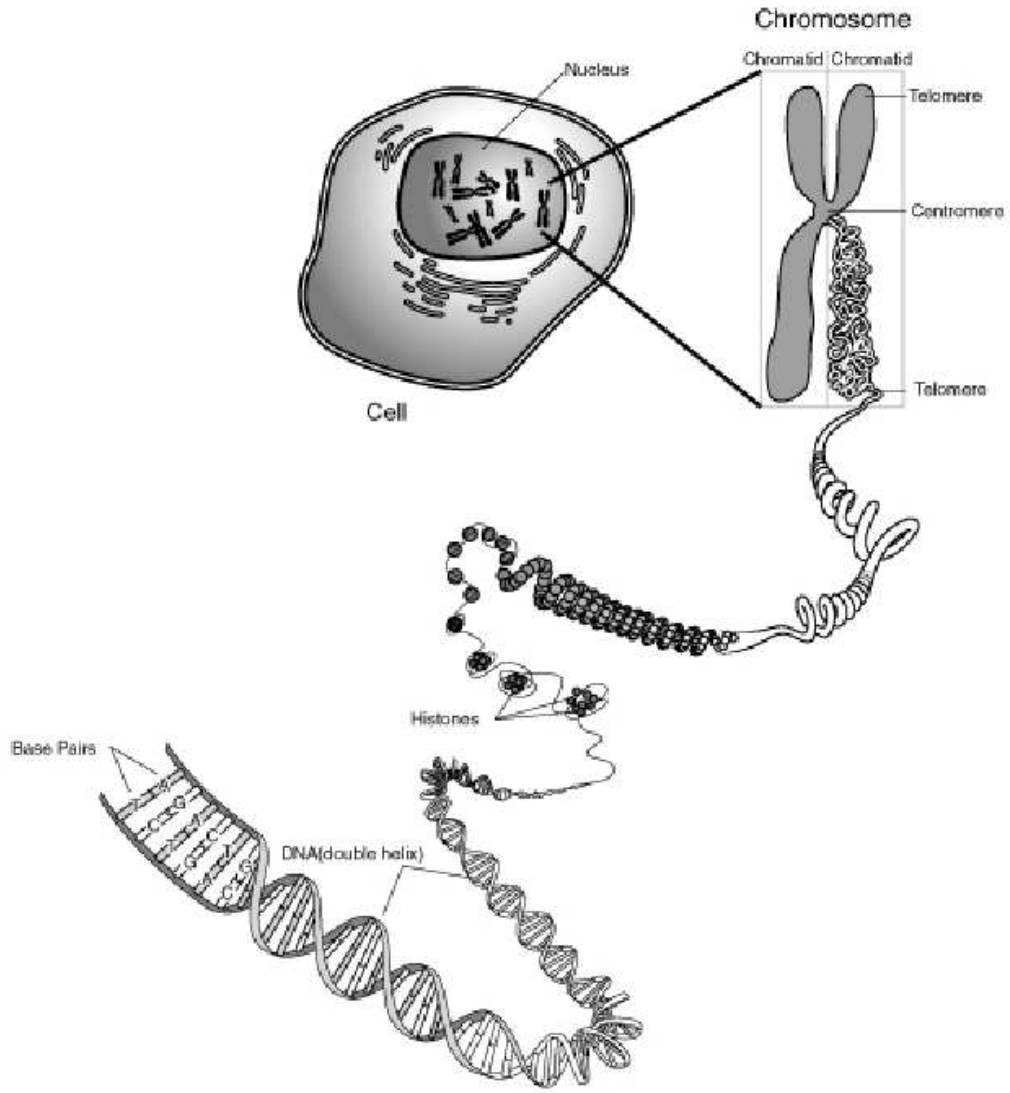
Γονίδια

Από το γενετικό υλικό ενός οργανισμού ένα μικρό μόνο μέρος αποτελεί τα γονίδια. Αυτά είναι τα τμήματα του γονιδιώματος που μεταγράφονται σε mRNA. Η προηγούμενη πρόταση θα γίνει σαφής όταν περιγραφεί η διαδικασία σύνθεσης των πρωτεϊνών στο Κεφάλαιο 3. Συνοπτικά μπορεί να ειπωθεί από τώρα ότι γονίδιο είναι τμήμα του γενετικού υλικού που είναι υπεύθυνο είτε για την παραγωγή πρωτεΐνης είτε για την παραγωγή τριών από τα τέσσερα είδη RNA

[12] Πηγή: http://amazingbeauty.org/nature/Fig_8.10-250.jpg
[13] Πηγή: http://www.mtsinai.on.ca/pdmg/images/chromosome.jpg

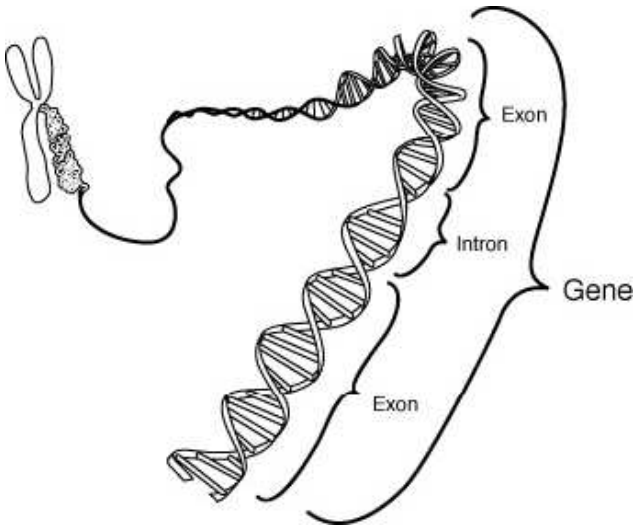


Σχήμα 2.13: Η συμπύκνωση του μορίου του DNA.



Σχήμα 2.14: Η οργάνωση του DNA σε χρωμοσώματα.

(tRNA, rRNA, snRNA).



Σχήμα 2.15: Γονίδιο.

Η εικόνα ενός γονιδίου δίνεται στο Σχήμα 2.15 [14]. Είναι σημαντικό να παρατηρηθεί ότι τα περισσότερα γονίδια των ευκαρυωτικών κυττάρων είναι ασυνεχή. Με άλλα λόγια, οι αλληλουχίες των βάσεων ενός γονιδίου, που καθορίζουν τις αλληλουχίες αμινοξέων ή νουκλεοτιδίων του παραγόμενου προϊόντος (μιας πρωτεΐνης ή ενός t/r/snRNA), διακόπτονται από ενδιάμεσες αλληλουχίες βάσεων. Οι πρώτες ονομάζονται *εξώνια (exons)*, ενώ οι δεύτερες *εσώνια (introns)*.

Αριθμητικά στοιχεία

Το γενετικό υλικό των προκαρυωτικών κυττάρων, δηλαδή το δίκλωνο κυκλικό μόριο DNA, έχει μήκος περίπου 1mm και αναδιπλώνεται τόσο, ώστε τελικά να έχει μήκος μέσα στο κύτταρο 1μm. Το συνολικό DNA σε κάθε διπλοειδές κύτταρο του ανθρώπου έχει μήκος περίπου 2m. Η συμπύκνωσή του γίνεται σε τέτοιο βαθμό, που φθάνει να χωρά στον πυρήνα, που έχει διάμετρο περίπου 10μm. Στην πρώτη περίπτωση η συμπύκνωση γίνεται 1000 φορές, ενώ στη δεύτερη 200000 φορές.

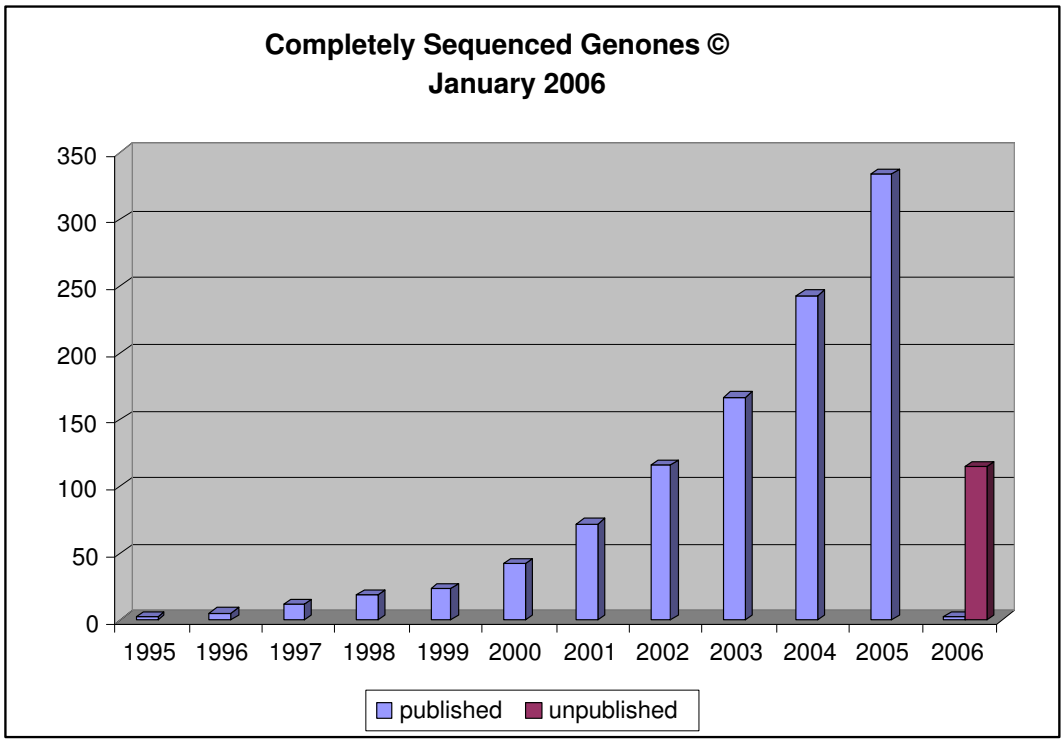
Για έναν μηχανικό υπολογιστών ίσως είναι πιο εύκολο να συνειδητοποιήσει τα παραπάνω νούμερα, αν δει την αναλογία με τη συμπίεση αρχείων σε υπολογιστές. Στην περίπτωση των προκαρυωτικών κυττάρων είναι σα να συμπιέζεται ένα αρχείο του 1MB και να γίνεται 1KB, ενώ για το ανθρώπινο κύτταρο το ανάλογο είναι να συμπιεστεί ένα αρχείο των 200MB σε 1KB [15].

Αριθμητικά στοιχεία για το μήκος του DNA και το μέγεθος της συμπύκνωσής του υπάρχουν επίσης στο Σχήμα 2.13. Σε αυτά μπορεί να προστεθεί και το γεγονός ότι το βήμα της

[14] Πηγή: http://www.accessexcellence.org/RC/VL/GG/images/exon.gif

[15] Οι υπολογισμοί είναι προσεγγιστικοί. Για παράδειγμα το μήκος του DNA σε ένα διπλοειδές ανθρώπινο κύτταρο είναι 1,8m και 1KB = 1024B, ενώ εδώ θεωρήθηκε το πρώτο στα 2m και ότι 1KB = 1000B.

έλικας του DNA είναι 3,4nm.



Σχήμα 2.16: Πλήθος οργανισμών για τους οποίους έχει ολοκληρωθεί η αποκωδικοποίηση του γενετικού τους υλικού μέχρι και τον Ιανουάριο του 2006.

Στο Σχήμα 2.16 [16] φαίνεται η προοδευτική πορεία των τελευταίων ετών στην ολοκλήρωση της αποκωδικοποίησης του γενετικού υλικού διαφόρων οργανισμών. Επιπλέον, στο Σχήμα 2.17 [17] φαίνονται λίγοι από αυτούς τους οργανισμούς, καθώς και το μέγεθος του γενετικού υλικού τους. Είναι χαρακτηριστικό ότι αν τυπωνόταν όλο το ανθρώπινο γονιδίωμα, θα χρειαζόντουσαν περίπου 200 τόμοι σε μέγεθος τηλεφωνικού καταλόγου, δηλαδή των 1000 σελίδων ο καθένας, ενώ η ανάγνωσή του υπολογίζεται ότι θα απαιτούσε περίπου 26 χρόνια εργασίας.

2.4 Χημικά άτομα

Αν και στη φύση υπάρχουν περισσότερα από 92 χημικά στοιχεία, μόνο 27 συναντώνται σε ζωντανούς οργανισμούς. Μάλιστα 4 από αυτά (άνθρακας, υδρογόνο, οξυγόνο, άζωτο) είναι τα επικρατέστερα σε ποσοστό 96%.

Η προτίμηση σε αυτά τα τέσσερα στοιχεία δεν είναι τυχαία. Ένας από τους λόγους που θεωρείται ότι ξεχωρίζουν είναι το γεγονός ότι συμβάλλουν στη σταθερότητα και την ποικιλομορφία, χαρακτηριστικά συνυφασμένα με τη ζωή. Πιο συγκεκριμένα, τα στοιχεία αυτά μπορούν να σχηματίζουν ομοιοπολικούς δεσμούς, οι οποίοι είναι σταθεροί δεσμοί, ενώ με

[16] Πηγή: www.genomesonline.org
[17] Πηγή: http://en.wikipedia.org/wiki/Genome

Organism	Genome size (base pairs)
Virus, Phage -X174;	5386 - First sequenced genome
Virus, Phage	5×10^4
Archaeum, *Nanoarchaeum equitans*	5×10^5 - Smallest non-viral genome Dec, 2005
Bacterium, *Buchnera aphidicola*	6×10^5
Bacterium, *Wigglesworthia glossinidia*	7×10^5
Bacterium, *Escherichia coli*	4×10^6
Amoeba, *Amoeba dubia*	6.7×10^{11} - Largest known genome Dec, 2005
Plant, *Fritillary assyrica*	1.3×10^{11}
Fungus, *Saccharomyces cerevisiae*	2×10^7
Nematode, *Caenorhabditis elegans*	8×10^7
Insect, *Drosophila melanogaster*	2×10^8
Mammal, *Homo sapiens*	3×10^9

Σχήμα 2.17: Ορισμένοι οργανισμοί και το μέγεθος του γονιδιώματός τους.

εξαίρεση το υδρογόνο μπορούν να συνδεθούν με περισσότερα του ενός άτομα δημιουργώντας πολλούς διαφορετικούς συνδυασμούς. Εξάλλου, για αυτό η χημεία που ασχολείται με τις ενώσεις του άνθρακα λέγεται οργανική.

Το υπόλοιπο 4% αποτελείται από φώσφορο, θείο, νάτριο, κάλιο, ασβέστιο, μαγνήσιο και χλώριο καθώς και από ιχνοστοιχεία.

2.4.1 Η παρουσία του νερού

Το υδατικό περιβάλλον μέσα στα κύτταρα αλλά και έξω από αυτά είναι δεδομένο για όλους τους οργανισμούς. Από αυτό τα κύτταρα –είτε πολυκύτταρων οργανισμών όπως ο άνθρωπος, είτε μονοκύτταρων σαν την αμοιβάδα– αντλούν τα απαραίτητα συστατικά για την επιβίωσή τους και εκκρίνουν τα παράγωγα του μεταβολισμού τους. Ακόμα και στο εσωτερικό τους, όμως, το 80% των συστατικών τους αποτελείται από νερό. Έτσι, οι ουσίες διευκολύνονται στη μεταφορά τους καθώς και στις αντιδράσεις μεταξύ τους. Σε ορισμένες μάλιστα συμμετέχει και το ίδιο το νερό, όπως η υδρόλυση, που αναφέρθηκε στην αρχή της Ενότητας 2.3.

Κεφάλαιο 3

Βιολογικές διαδικασίες, θεωρίες και μέθοδοι

Το τρίτο κεφάλαιο αποτελεί συμπλήρωμα του δεύτερου. Αναλύει θέματα που έχουν σχέση με τις διαδικασίες στις οποίες συμμετέχουν οι βιολογικές οντότητες, θεωρίες για αυτές, καθώς επίσης τεχνικές μέτρησης χαρακτηριστικών τους. Θεωρείται και αυτό απαραίτητο για την κατανόηση των βασικών αρχών των βιοεπιστημών.

Οι διαδικασίες που περιγράφονται έχουν σχέση με το γενετικό υλικό. Είναι ο διπλασιασμός του, η μεταγραφή, η μετάφραση. Αναφέρονται, επιπλέον, περισσότερο λεπτομερή θέματα, σαν την ωρίμανση και το ρόλο των ενζύμων σε κάθε φάση. Αναγκαστικά εξηγείται και ο γενετικός κώδικας, αφού χρειάζεται για τη μετάφραση των νουκλεοτιδίων σε αμινοξέα.

Οι θεωρίες που απασχολούν είναι κρίσιμες για την κατανόηση αρκετών ζητημάτων. Ενδιαφέρει να γίνει γνωστό το κεντρικό δόγμα της βιολογίας, του οποίου το όνομα δηλώνει ότι έχει κύρια θέση στη μοριακή βιολογία. Εξηγούνται η έννοια της γενετικής έκφρασης και ο τρόπος που γίνεται η ρύθμιση της δραστηριότητας των γονιδίων. Η εξέλιξη είναι και αυτή μία από τις θεωρίες που συζητώνται.

Τέλος, είναι χρήσιμο για τον μηχανικό υπολογιστών που έχει σκοπό να ασχοληθεί με τα θέματα των βιοεπιστημών να καταλάβει για ποιο λόγο πολλά από τα δεδομένα έχουν μια συγκεκριμένη μορφή. Περιγράφονται, λοιπόν, σημαντικές πειραματικές μέθοδοι, όπως είναι τα microarrays, η PCR και οι DNA sequencers.

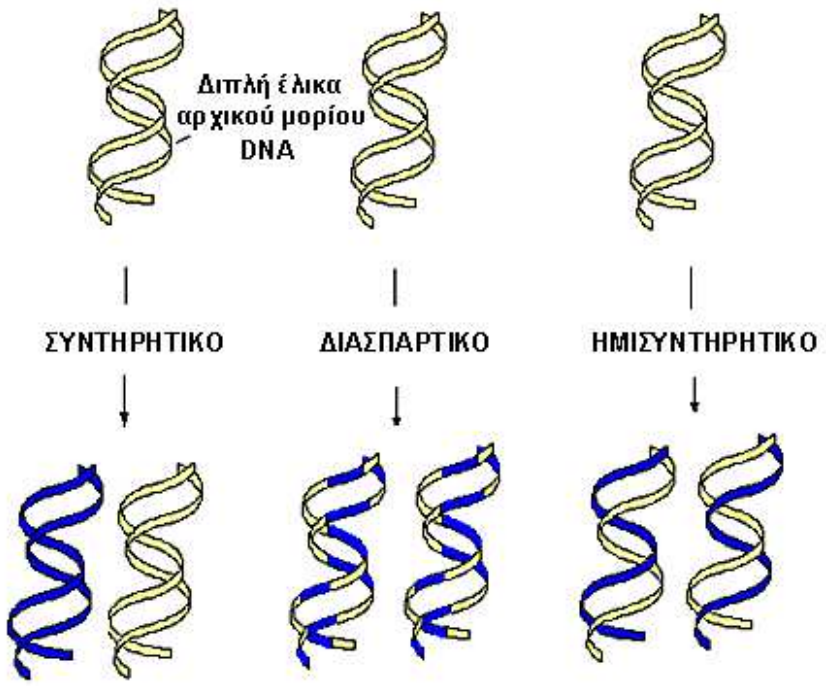
3.1 Σημαντικές διαδικασίες στις οποίες συμμετέχει το γενετικό υλικό

Στην ενότητα αυτή περιγράφονται με σημαντική λεπτομέρεια οι διαδικασίες της αντιγραφής του DNA και της παραγωγής των πρωτεϊνών. Τμήματα της δεύτερης είναι η μεταγραφή, η ωρίμανση και η μετάφραση. Η μεταγραφή, επιπλέον, αποτελεί στάδιο και της παραγωγής των περισσότερων ειδών του RNA. Τέλος, αναφέρεται ο ρόλος του γενετικού κώδικα.

3.1.1 Αντιγραφή του DNA

Όπως αναφέρθηκε στο Κεφάλαιο 2 σχετικά με τον κύκλο ζωής ενός κυττάρου, ο διπλασιασμός του DNA συμβαίνει λίγο πριν το κύτταρο χωριστεί σε δύο νέα. Κάθε θυγατρικό παίρνει ποιοτικά και ποσοτικά το ίδιο ακριβώς γενετικό υλικό με το μητρικό του. Η διαδικασία αυτή εξασφαλίζει τη μεταφορά της γενετικής πληροφορίας από το αρχικό στα δύο καινούρια και επαγωγικά, από έναν οργανισμό στον απόγονό του.

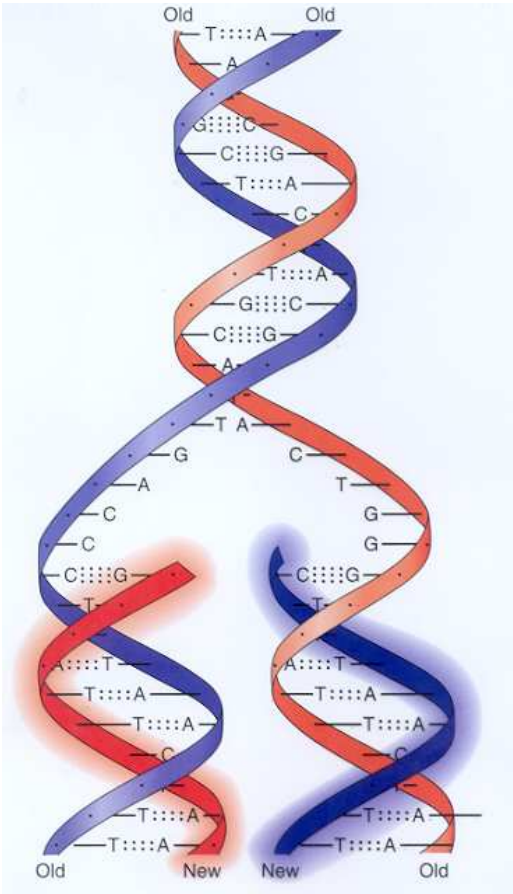
Πιθανά μοντέλα αυτοδιπλασιασμού του DNA φαίνονται στο Σχήμα 3.1. Πειραματικά αποδείχθηκε (από τους Meselson, Stahl το 1958) ότι ο μηχανισμός αντιγραφής του DNA είναι *ημισυντηρητικός*. Αυτό σημαίνει ότι καθένα από τα δύο θυγατρικά μόρια που προκύπτουν αποτελούνται από τη μία αλυσίδα του δίκλωνου μητρικού μορίου και από μία νέα. Η καινούρια, βέβαια, αλυσίδα είναι πανομοιότυπη με την αντίστοιχη μητρική.



Σχήμα 3.1: Πιθανά μοντέλα διπλασιασμού του DNA.

Η διαδικασία της αντιγραφής στα προκαρυωτικά κύτταρα έχει μία σημαντική διαφορά σε σχέση με τα ευκαρυωτικά, όσον αφορά τα σημεία έναρξης. Επειδή η πολυπλοκότητα της οργάνωσης αλλά και το μέγεθος του γενετικού υλικού στα βακτήρια είναι πολύ μικρότερα από αυτά των ανώτερων οργανισμών, η αντιγραφή του DNA αρκεί να ξεκινά από ένα μόνο σημείο του μορίου. Αντίθετα, στα ευκαρυωτικά κύτταρα υπάρχουν περισσότερες θέσεις έναρξης, ώστε η διαδικασία να ολοκληρώνεται και πάλι γρήγορα. Η αντιγραφή του DNA έχει, πάντως, μελετηθεί πολύ περισσότερο στα προκαρυωτικά και συγκεκριμένα, στο βακτήριο Escherichia coli.

Ωστόσο, τα επιμέρους βήματα που ακολουθούνται για την αντιγραφή του μορίου παρουσιάζουν σημαντικές ομοιότητες και στα δύο είδη κυττάρων και γίνονται αντιληπτά μέσω του Σχήματος 3.2. Αρχικά, σπάνε οι δεσμοί υδρογόνου που συγκρατούν τις δύο αλυσίδες μαζί, ώστε να ξετυλιχθεί στο σημείο εκείνο η έλικα. Στη συνέχεια αντιγράφονται και οι δύο αλυσίδες ταυτόχρονα σύμφωνα με την αρχή συμπληρωματικότητας των βάσεων (βλ. Κεφάλαιο 2). Έτσι, σταδιακά δημιουργούνται οι δύο θυγατρικές αλυσίδες και το μητρικό μόριο αντικαθιστάται από δύο καινούρια, τα οποία είναι πανομοιότυπα τόσο μεταξύ τους όσο και με το αρχικό. Χρειάζεται να σημειωθεί ότι σε όλα τα στάδια της αντιγραφής είναι ιδιαίτερα κρίσιμος ο ρόλος ενζύμων.



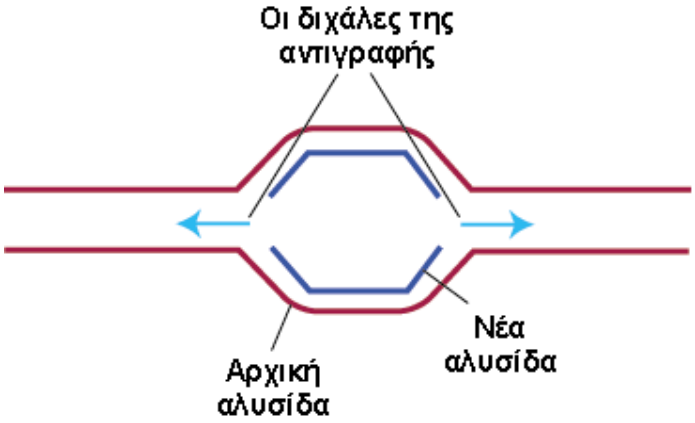
Σχήμα 3.2: Σπάσιμο δεσμών αρχικής έλικας, σχηματισμός νέων συμπληρωματικών αλυσίδων.

Η δράση των ενζύμων

Στην παρούσα ενότητα δίνονται στοιχεία που χρειάζονται για την κατανόηση σε μεγαλύτερο βάθος του μηχανισμού της αντιγραφής του DNA. Τα στοιχεία αυτά σχετίζονται με τις εργασίες που εκτελούν τα ένζυμα και συμπληρώνουν όσα προηγουμένως αναφέρθηκαν.

Τα ένζυμα που συμμετέχουν στη διαδικασία είναι οι DNA πολυμεράσες και ελικάσες, η DNA δεσμάση, κάποια επιδιορθωτικά και ένα σύμπλοκο ενζύμων που λέγεται πριμόσωμα. Οι

πολυμεράσες είναι εκείνες που επιμηκύνουν τις θυγατρικές αλυσίδες τοποθετώντας τα συμπληρωματικά νουκλεοτίδια απέναντι από τη μητρική αλυσίδα, ενώ μπορούν επίσης να διορθώνουν λάθη που συμβαίνουν σε αυτήν την τοποθέτηση. Οι *ελικάσες* έχουν τη δυνατότητα να σπάνε τους δεσμούς υδρογόνου ανάμεσα στους δύο κλώνους, ώστε να δημιουργηθεί η λεγόμενη *θηλιά* εκεί όπου ξεκινά η διχάλα (fork) και ξετυλίγεται το μόριο (Σχήμα 3.3). Τα *επιδιορθωτικά* ένζυμα διορθώνουν όσα λάθη δεν καταφέρνουν να διορθώσουν οι DNA πολυμεράσες και τελικά η πιθανότητα εμφάνισης σφάλματος στους ευκαρυωτικούς οργανισμούς περιορίζεται στο 1 στα 10^{10} νουκλεοτίδια. Οι ρόλοι του πριμοσώματος και της DNA δεσμάσης εξηγούνται στα επόμενα.



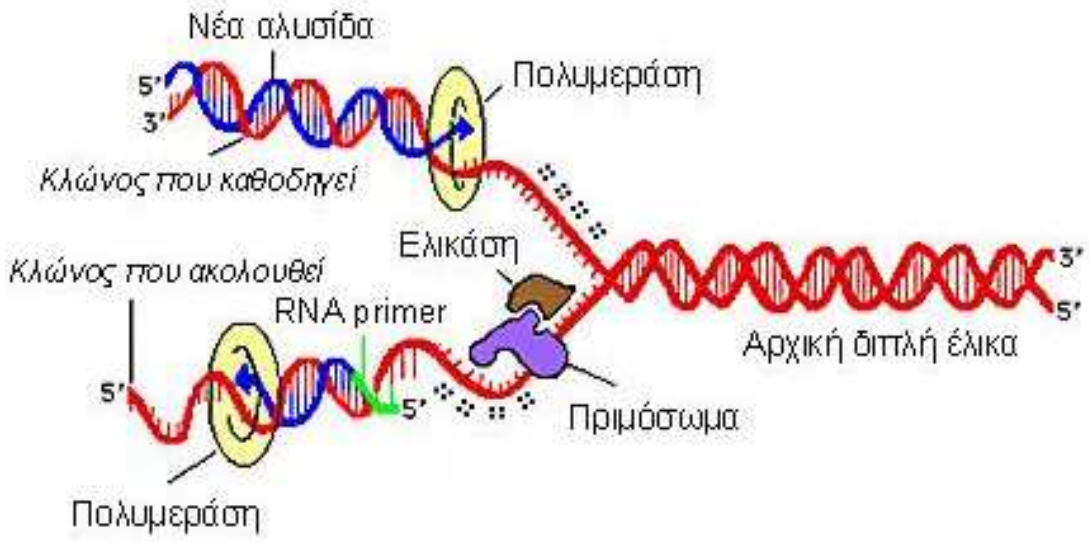
Σχήμα 3.3: Η θηλιά της αντιγραφής του DNA.

Οι DNA πολυμεράσες δεν έχουν την ικανότητα να ξεκινήσουν την αντιγραφή στα σημεία όπου έχουν αρχίσει να ξετυλίγουν την έλικα οι ελικάσες. Αντίθετα, επιμηκύνουν τα λεγόμενα *πρωταρχικά τμήματα (primers)*. Αυτά είναι μικρά τμήματα RNA, συμπληρωματικά των μητρικών αλυσίδων, τα οποία συντίθενται από το *πριμόσωμα*, που αναφέρθηκε στην προηγούμενη παράγραφο.

Εδώ χρειάζεται να εξηγηθεί ένα λεπτό σημείο σχετικά με τον προσανατολισμό της διπλής έλικας του μορίου του DNA. Από τη δομή του μορίου, που αναφέρθηκε στο Κεφάλαιο 2, είναι γνωστό ότι ο φωσφοδιεστερικός δεσμός σχηματίζεται μεταξύ του υδροξυλίου του 3′ άνθρακα της πεντόζης του ενός νουκλεοτιδίου και της φωσφορικής ομάδας που είναι συνδεδεμένη στον 5′ άνθρακα της πεντόζης του επόμενου νουκλεοτιδίου. Έτσι, το πρώτο νουκλεοτίδιο της αλυσίδας έχει μια ελεύθερη φωσφορική ομάδα στον 5′ άνθρακα, ενώ το τελευταίο της αλυσίδας έχει ελεύθερο το υδροξύλιο που ανήκει στον 3′ άνθρακα. Για το λόγο αυτό λέγεται ότι η αλυσίδα έχει *προσανατολισμό 5′–>3′*. Τον προσανατολισμό αυτό έχουν οι δύο αλυσίδες του μορίου του DNA. Ωστόσο, είναι *αντιπαράλληλες*, δηλαδή το άκρο 3′ της μιας είναι απέναντι από το άκρο 5′ της άλλης.

Τα ένζυμα εκτελούν τις εργασίες τους πάνω σε μία αλυσίδα του DNA ακολουθώντας τη φορά 3′–>5′. Αυτό σημαίνει ότι κατά τη διάρκεια της αντιγραφής η σύνθεση του DNA στις θυγατρικές αλυσίδες είναι συνεχής στη μία και ασυνεχής στην άλλη. Τα τμήματα της

ασυνεχούς αλυσίδας συνδέονται μεταξύ τους μέσω της δράσης της DNA *δεσμάσης*. Επιπλέον, το ένζυμο αυτό συνδέει όλα τα κομμάτια των νέων αλυσίδων που προκύπτουν από τις διάφορες θέσεις έναρξης της αντιγραφής. Τα παραπάνω φαίνονται στο Σχήμα 3.4 έχει δημιουργηθεί με βάση το αντίστοιχο στο [24].



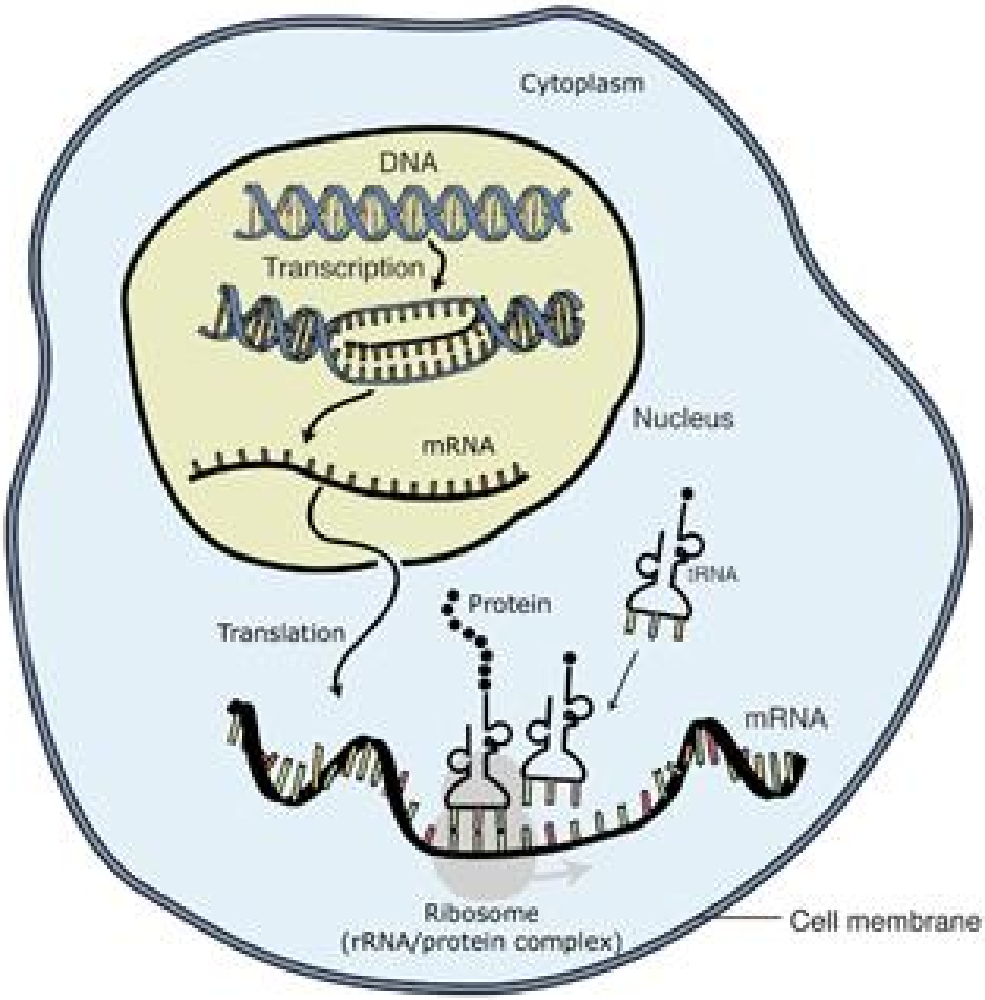
Σχήμα 3.4: Η συνεργασία των ενζύμων.

3.1.2 Παραγωγή πρωτεϊνών

Οι οδηγίες για το πώς, πότε και ποιες πρωτεΐνες πρέπει να παραχθούν σε ένα κύτταρο υπάρχουν στον πυρήνα του, στο DNA. Για να συντεθούν οι πρωτεΐνες με βάση αυτές τις οδηγίες ακολουθούνται οι πορείες της μεταγραφής και της μετάφρασης. Η πορεία που ακολουθείται φαίνεται συνοπτικά στο Σχήμα 3.5, ενώ στις ακόλουθες παραγράφους περιγράφεται βήμα προς βήμα.

Μεταγραφή

Για να δημιουργηθούν οι πρωτεΐνες στα ριβοσώματα του κυττάρου σύμφωνα με τις οδηγίες του γενετικού υλικού οι οποίες βρίσκονται στον πυρήνα του, οι γενετικές πληροφορίες μεταφέρονται μέσω του mRNA στο κυτταρόπλασμα, δηλαδή εκτός του πυρήνα. Όπως αναφέρθηκε στο Κεφάλαιο 2, αυτός είναι ο λόγος που το συγκεκριμένο είδος RNA ονομάζεται *αγγελιοφόρο*. Η διαδικασία με την οποία παράγεται το mRNA λέγεται *μεταγραφή*, όνομα του οποίου ο λόγος χρήσης θα γίνει σαφής στη συνέχεια.



Σχήμα 3.5: Μεταγραφή, μετάφραση και ο ρόλος του RNA.

Χρειάζεται να γίνει μια σύντομη παρατήρηση, που είναι ουσιαστικά υπενθύμιση από το Κεφάλαιο 2. Οι πληροφορίες που είναι αναγκαίες για τη σύνθεση μιας πρωτεΐνης βρίσκονται σε ένα μικρό τμήμα του DNA. Έτσι, το mRNA που δημιουργείται έχει πολύ μικρότερο μέγεθος σε σχέση με το DNA του πυρήνα, καθώς περιέχει μόνο τα αναγκαία στοιχεία. Επιπλέον, σε αντίθεση με το DNA που είναι μία δίκλωνη έλικα στην οποία ισχύει η συμπληρωματικότητα των βάσεων, το mRNA αποτελείται από μία μόνο αλυσίδα. Με άλλα λόγια το mRNA είναι το ελάχιστο δυνατό κινητό αντίγραφο του γονιδίου που είναι υπεύθυνο για την παραγωγή μιας πρωτεΐνης.

Η μεταγραφή λαμβάνει χώρα μέσα στον πυρήνα. Στο σημείο που βρίσκεται το γονίδιο που φέρει τη ζητούμενη πληροφορία για τη σύνθεση της πρωτεΐνης, η δίκλωνη έλικα του DNA ξετυλίγεται. Απέναντι από τα δεσοξυριβονουκλεοτίδια της μίας αλυσίδας του, η οποία ονομάζεται *μεταγραφόμενη*, τοποθετούνται τα συμπληρωματικά ριβονουκλεοτίδια σχηματίζοντας έτσι την αλυσίδα του mRNA. Με αυτόν τον τρόπο το mRNA που προκύπτει έχει τις ίδιες βάσεις — με εξαίρεση την ουρακίλη στη θέση της αδενίνης — με την άλλη αλυσίδα του DNA η οποία ονομάζεται *κωδική*.

Ο όρος, λοιπόν, μεταγραφή (transcription), χρησιμοποιείται επειδή η γενετική πληροφορία που είναι καταγεγραμμένη στη γλώσσα του DNA μεταγράφεται στη γλώσσα του RNA. Οι βάσεις θυμίνη (T), γουανίνη (G), κυτοσίνη (C) διατηρούνται και στο mRNA, ενώ τη θέση της αδενίνης (A) καταλαμβάνει η ουρακίλη (U).

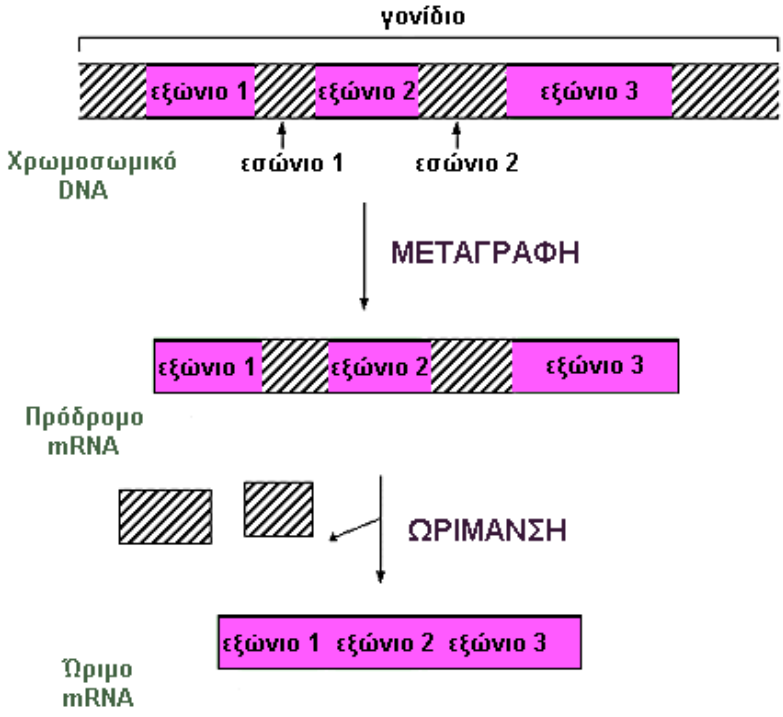
Όπως στην αντιγραφή του DNA έτσι και στη μεταγραφή του συμμετέχουν ένζυμα. Τα κυριότερα από αυτά είναι οι RNA *πολυμεράσες*, οι οποίες και συνθέτουν την αλυσίδα του mRNA. Χρειάζεται να σημειωθεί ότι κάποιες περιοχές του DNA ονομάζονται *υποκινητές (promoters)* και κάποιες άλλες *αλληλουχίες λήξης της μεταγραφής (terminators)*. Όπως δηλώνουν και τα ονόματά τους, οι πρώτες σηματοδοτούν την αρχή της περιοχής από την οποία ξεκινά η μεταγραφή, ενώ οι δεύτερες το σημείο στο οποίο τελειώνει. Αυτές μπορούν να τις αναγνωρίσουν οι RNA *πολυμεράσες*.

Ωρίμανση

Όπως αναφέρθηκε στο Κεφάλαιο 2 για τα γονίδια, στους ευκαρυωτικούς οργανισμούς αυτά αποτελούνται από εσώνια και εξώνια. Το mRNA που δημιουργείται από τη διαδικασία που περιγράφηκε ως τώρα ονομάζεται *πρόδρομο*. Είναι εκείνο το οποίο περιέχει και τις ακολουθίες που μεταφράζονται σε αμινοξέα (εξώνια) και εκείνες που δεν μεταφράζονται (εσώνια).

Προτού, όμως, βγει από τον πυρήνα, το mRNA απαλλάσσεται από τις περιττές για τη μετάφραση αλληλουχίες. Η διαδικασία αυτή ονομάζεται *ωρίμανση* (Σχήμα 3.6). Σε αυτήν συμμετέχει το μικρό πυρηνικό RNA (snRNA), που έχει αναφερθεί και στο Κεφάλαιο 2. Έτσι, προκύπτει το ώριμο (mature) RNA, το οποίο έχει μόνο εξώνια και δύο μόνο αμετάφραστες περιοχές, μία σε κάθε άκρο του, τις 5′ και 3′ αμετάφραστες περιοχές (για την ονομασία των άκρων βλέπε την ενότητα 3.1.1 για τη δράση των ενζύμων).

Στους προκαρυωτικούς οργανισμούς, αντίθετα, δεν υπάρχει το στάδιο της ωρίμανσης. Μάλιστα, από τη στιγμή που δεν υπάρχει καν πυρηνική μεμβράνη, η μετάφραση αρχίζει προτού



Σχήμα 3.6: Η διαδικασία της ωρίμανσης του mRNA.

τελειώσει η μεταγραφή του RNA.

Γενετικός κώδικας

Όταν δημιουργείται RNA από DNA κατά τη μεταγραφή είναι εύλογο να θεωρήσει κανείς ότι κάθε βάση του DNA (A, T, G, C) αντιστοιχίζεται σε μια συμπληρωματική της του mRNA (U, A, C, G). Ωστόσο, δεν μπορεί να υποτεθεί κάτι ανάλογο κατά τη μετάφραση.

Κατά την τελευταία ένα λεξιλόγιο τεσσάρων γραμμάτων μεταφράζεται σε ένα λεξιλόγιο τουλάχιστον είκοσι. Είκοσι θεωρείται ότι είναι ο διαφορετικός αριθμός αμινοξέων που υπάρχουν σε πρωτεΐνες, αν και αυτός ο αριθμός έχει αρχίσει να τίθεται υπό αμφισβήτηση. Η μετάφραση αυτή γίνεται με βάση τον *γενετικό κώδικα* (Σχήμα 3.7 [1]). Σύμφωνα με αυτό μια τριάδα νουκλεοτιδίων κωδικοποιεί ένα αμινοξύ. Η τριάδα αυτή των βάσεων ονομάζεται *κωδικόνιο (codon)*.

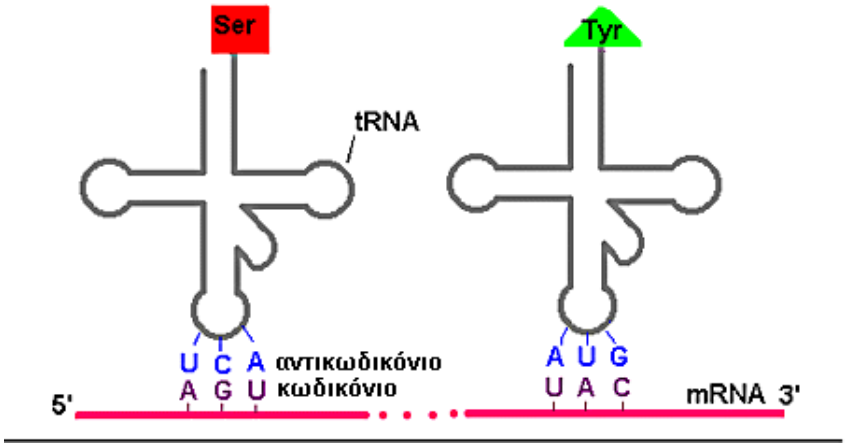
Διαφορετικά κωδικόνια είναι δυνατό να αντιστοιχούν στο ίδιο αμινοξύ. Αυτό είναι λογικό, αν σκεφθεί κανείς ότι οι δυνατοί συνδυασμοί τεσσάρων βάσεων ανά τρεις είναι 64 (4^3). Ο αριθμός αυτός υπερκαλύπτει τα είκοσι ζητούμενα αμινοξέα.

Τα περισσότερα χαρακτηριστικά του γενετικού κώδικα έχουν ήδη αναφερθεί. Είναι κώδικας τριπλέτας (αντιστοιχεί κωδικόνια σε αμινοξέα), συνεχής (δεν παραλείπεται νουκλεοτίδιο του mRNA), μη επικαλυπτόμενος (ένα νουκλεοτίδιο μπορεί να ανήκει σε ένα μόνο κωδικό-

[1] Πηγή: http://campus.queens.edu/faculty/jannr/Genetics/images/codon.jpg

First base	Second base: U	C	A	G	Third base
U	UUU, UUC — Phenylalanine F UUA, UUG — Leucine L	UCU, UCC, UCA, UCG — Serine S	UAU, UAC — Tyrosine Y UAA Stop codon UAG Stop codon	UGU, UGC — Cysteine C UGA Stop codon UGG Tryptophan W	U C A G
C	CUU, CUC, CUA, CUG — Leucine L	CCU, CCC, CCA, CCG — Proline P	CAU, CAC — Histidine H CAA, CAG — Glutamine Q	CGU, CGC, CGA, CGG — Arginine R	U C A G
A	AUU, AUC, AUA — Isoleucine I AUG Methionine start codon M	ACU, ACC, ACA, ACG — Threonine T	AAU, AAC — Asparagine N AAA, AAG — Lysine K	AGU, AGC — Serine S AGA, AGG — Arginine R	U C A G
G	GUU, GUC, GUA, GUG — Valine V	GCU, GCC, GCA, GCG — Alanine A	GAU, GAC — Aspartic acid D GAA, GAG — Glutamic acid E	GGU, GGC, GGA, GGG — Glycine G	U C A G

Σχήμα 3.7: Ο γενετικός κώδικας.



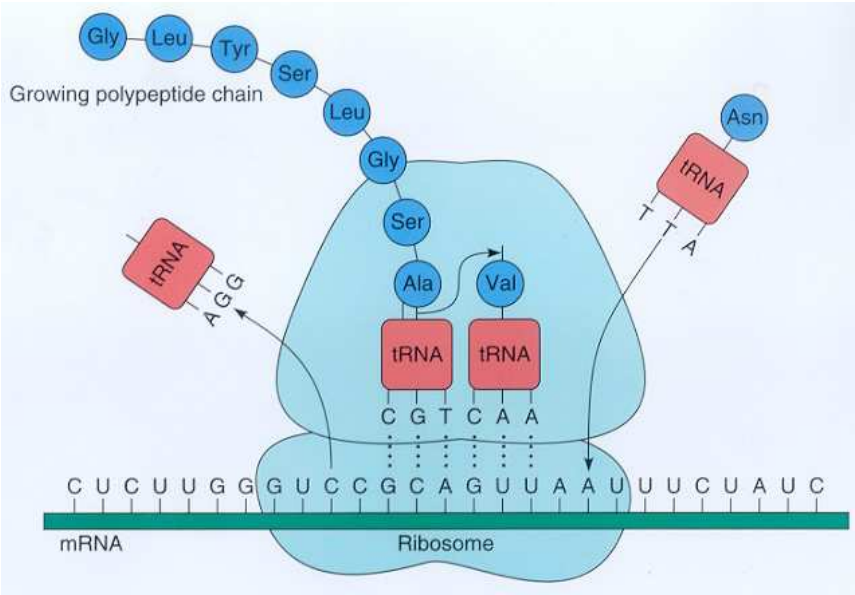
Σχήμα 3.8: Κωδικόνια και αντικωδικόνια.

νιο), σχεδόν καθολικός (ισχύει σε όλους τους οργανισμούς), εκφυλισμένος (στα περισσότερα αμινοξέα αντιστοιχούν παραπάνω από 1 κωδικόνια), έχει κωδικόνια έναρξης (AUG) και λήξης (UAG, UGA, UAA).

Η καθολικότητα του γενετικού κώδικα έχει μία προφανή αλλά και εκπληκτική συνέπεια. Το mRNA από οποιονδήποτε οργανισμό μπορεί να μεταφραστεί σε φυτά, ζώα ή βακτήρια — σε εργαστηριακό περιβάλλον — και να παράγει την ίδια πρωτεΐνη.

Σημειώνεται, τέλος, μια λεπτομέρεια που φαίνεται στο Σχήμα 3.8. Ο γενετικός κώδικας αναφέρεται στα νουκλεοτίδια του mRNA. Το tRNA που φέρνει τα αμινοξέα για την πρωτεϊνοσύνθεση φέρει τα λεγόμενα *αντικωδικόνια (anticodons)*. Αυτά είναι τα συμπληρωματικά των κωδικονίων. Έτσι, αν το αμινοξύ Τυροσίνη χαρακτηρίζεται από την τριπλέτα UAC στον γενετικό κώδικα, δηλαδή στην αλληλουχία του mRNA, τότε το tRNA θα έχει το αντικωδικόνιο UAC.

Μετάφραση



Σχήμα 3.9: Η μετάφραση του mRNA.

Καθώς είναι γνωστές οι ιδιότητες του γενετικού κώδικα, η διαδικασία της μετάφρασης δεν παρουσιάζει κάποια δυσκολία στην κατανόηση (Σχήμα 3.9 [2]). Το ώριμο mRNA μεταφέρεται στο κυτταρόπλασμα, από τον πυρήνα όπου ήταν. Εκεί συνδέεται με ριβόσωμα, που αρχίζει να το διαβάζει. Όταν συναντήσει το κωδικόνιο έναρξης, το πρώτο tRNA φέρνει στο ριβόσωμα το αμινοξύ μεθειονίνη. Το αμινοξύ αυτό είναι πάντοτε το πρώτο κάθε πρωτεΐνης. Το ριβόσωμα διαβάζει διαδοχικά όλη την ακολουθία του mRNA και τα tRNA φέρνουν τα κατάλληλα αμινοξέα. Κάποτε φθάνει σε κάποιο από τα κωδικόνια λήξης, οπότε και η παραγωγή της πρωτεΐνης έχει ολοκληρωθεί.

[2] Πηγή: http://library.tedankara.k12.tr/chemistry/vol1/biochem/trans100.htm

3.2 Βασικές θεωρίες

Στοιχεία για τέσσερις βασικές έννοιες δίνονται στην παρούσα ενότητα. Με βάση όσα έχουν ήδη περιγραφεί, μπορούν να εξηγηθούν το κεντρικό δόγμα της μοριακής βιολογίας, το περιεχόμενο της γενετικής έκφρασης και πώς γίνεται η γονιδιακή ρύθμιση. Επίσης, καταγράφονται τα βασικά σημεία που συνθέτουν τη θεωρία της εξέλιξης.

3.2.1 Το κεντρικό δόγμα της μοριακής βιολογίας

Το κεντρικό δόγμα της βιολογίας διατυπώθηκε, αν και όχι στην πλήρη μορφή του, από τον F. Crick το 1958 και περιγράφεται στο Σχήμα 3.10. Το περιεχόμενό του είναι ότι τα νουκλεϊκά οξέα διπλασιάζονται μόνα τους και παράγουν τις πρωτεΐνες. Η συμπλήρωση στο εν λόγω σχήμα, αρκετά μεταγενέστερα από τον F. Crick, είναι ένα βέλος και από το RNA προς το DNA. Έχει βρεθεί, δηλαδή, ότι μερικοί ιοί έχουν RNA ως γενετικό υλικό και η αντίστροφη μεταγραφάση (ένζυμο) το χρησιμοποιεί για να συνθέσει DNA. Επομένως, το κεντρικό δόγμα της βιολογίας εκφράζεται σήμερα σύμφωνα με το Σχήμα 3.11.

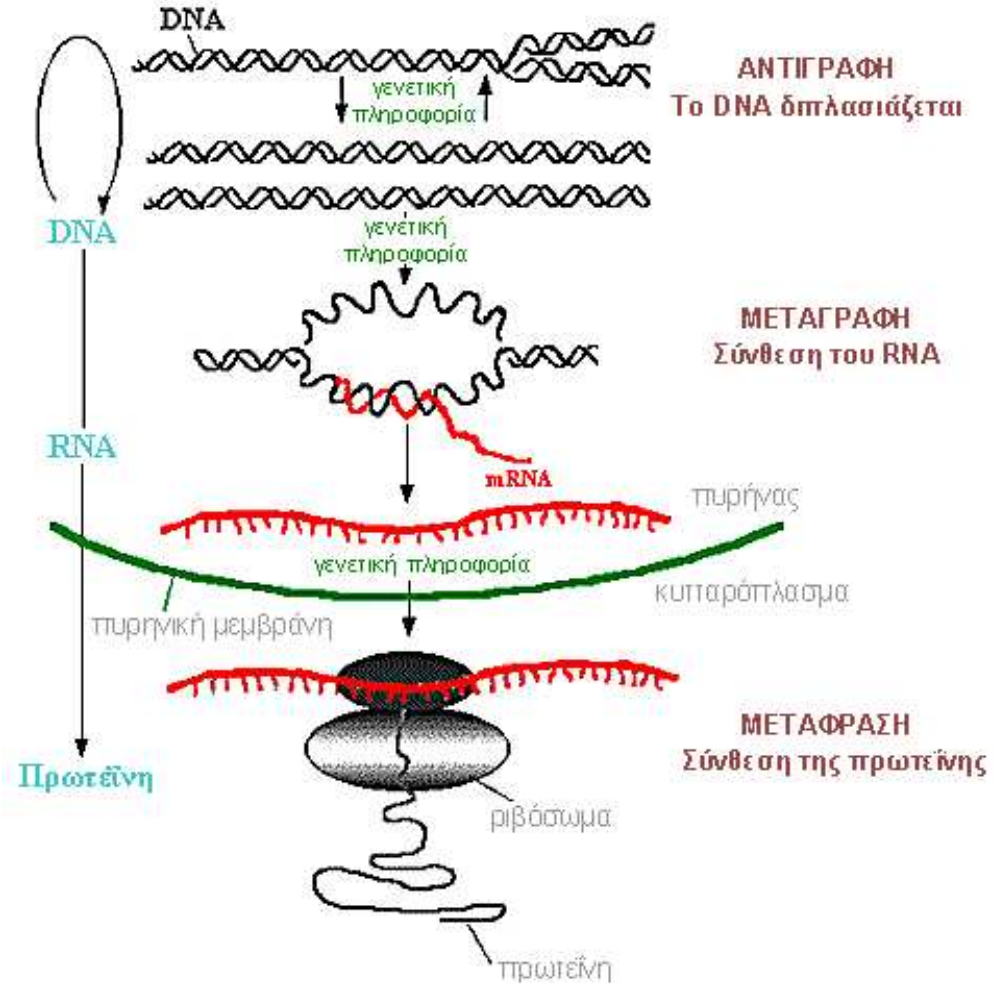
3.2.2 Γενετική έκφραση

Ο όρος της γενετικής έκφρασης (gene expression) αναφέρεται στη διαδικασία με την οποία η γενετική πληροφορία σταδιακά μετατρέπεται στις δομές και λειτουργίες ενός κυττάρου. Πιο απλά είναι τα διαδοχικά βήματα μέσω των οποίων από ένα γονίδιο - δηλαδή μια ακολουθία DNA στη συντριπτική πλειοψηφία των περιπτώσεων - δημιουργείται το κατάλληλο είδος RNA και στη συνέχεια ενδέχεται να γίνεται η σύνθεση της αντίστοιχης πρωτεΐνης, έτσι ώστε να μπορούν να λάβουν χώρα οι απαραίτητες λειτουργίες ενός κυττάρου. Ένα γονίδιο λέγεται ότι εκφράζεται σε δύο περιπτώσεις, οι οποίες εξετάζονται στη συνέχεια.

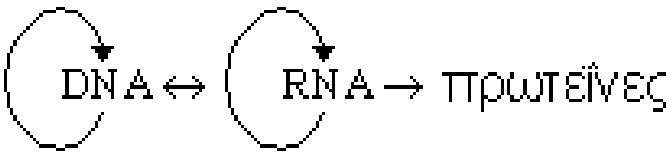
Η πρώτη αφορά γονίδια που είναι υπεύθυνα για τη σύνθεση κάποιας πρωτεΐνης. Το DNA βρίσκεται στον πυρήνα ενός κυττάρου στην περίπτωση των ευκαρυωτικών οργανισμών. Οι πρωτεΐνες συντίθενται στα ριβοσώματα που βρίσκονται μέσα στο κύτταρο, αλλά εκτός του πυρήνα (στο ενδοκυττάριο υγρό ή στο ενδοπλασματικό δίκτυο). Αυτό σημαίνει ότι με κάποιο τρόπο οι πληροφορίες του γενετικού υλικού πρέπει να μεταφερθούν στα ριβοσώματα. Η εργασία αυτή επιτελείται από το mRNA (messenger RNA).

Για το σκοπό αυτό μέσα στον πυρήνα γίνεται μεταγραφή (transcription) του γονιδίου που έχει τις πληροφορίες για την παραγωγή μιας πρωτεΐνης, δηλαδή από συγκεκριμένο τμήμα της δίκλωνης έλικας DNA παράγεται η αντίστοιχη μονόκλωνη αλυσίδα mRNA. Στη συνέχεια το mRNA κατευθύνεται στα ριβοσώματα όπου με βάση αυτό συντίθεται η πρωτεΐνη. Η διαδικασία αυτή περιγράφεται με τον όρο μετάφραση (translation) της γενετικής πληροφορίας. Τα τελευταία χρόνια πιστεύεται ότι δεν υπάρχει ένα προς μία αντιστοιχία γονιδίων και πρωτεΐνών. Με άλλα λόγια θεωρείται ότι από ένα γονίδιο μπορούν να παραχθούν περισσότερες της μιας πρωτεΐνες, αν και το θέμα αυτό είναι ακόμη αντικείμενο έρευνας.

Η δεύτερη κατηγορία γονιδίων που εκφράζονται περιλαμβάνει εκείνα που είναι υπεύθυνα για την παραγωγή των δύο άλλων ειδών RNA, τα οποία δεν μεταφράζονται. Αυτά είναι το



Σχήμα 3.10: Το κεντρικό δόγμα της μοριακής βιολογίας.



Σχήμα 3.11: Το κεντρικό δόγμα της μοριακής βιολογίας στην πλήρη μορφή του.

tRNA (transfer RNA) και το rRNA (ribosomal RNA). Το πρώτο μεταφέρει στα ριβοσώματα τα αμινοξέα που χρειάζονται για να δημιουργηθεί το πρωτεϊνικό μόριο, ενώ το δεύτερο αποτελεί δομικό συστατικό των ριβοσωμάτων.

Σχεδόν όλα τα παραπάνω έχουν αναφερθεί κατά τόπους στα Κεφάλαια 2 και 3. Σε αυτήν, όμως, την ενότητα αφενός δίνονται συγκεντρωτικά, αφετέρου εξηγείται μια έννοια που συναντάται αρκετά συχνά, αυτή της γενετικής έκφρασης.

3.2.3 Γονιδιακή ρύθμιση

Είναι εύλογο ύστερα από όλες τις προηγούμενες αναλύσεις να δημιουργηθούν κάποια πιο φιλοσοφικά ερωτήματα. Σε αυτά ανήκουν απορίες όπως, πώς καθορίζεται σε ποια χρονική στιγμή και ποσότητα θα παραχθεί μια πρωτεΐνη, αν ένα γονίδιο θα είναι ενεργό σε κάποιο κύτταρο και ανενεργό σε κάποιο άλλο ή πώς γίνεται η προσαρμογή ενός οργανισμού στο περιβάλλον του, δηλαδή η ενεργοποίηση γονιδίων που ως κάποια στιγμή ήταν στον ίδιο οργανισμό ανενεργά και το αντίστροφο.

Θέματα σαν τα παραπάνω είναι το αντικείμενο της ρύθμισης της γονιδιακής έκφρασης. Θεωρείται ένας από τους πιο δύσκολους κλάδους της μοριακής βιολογίας, ενώ αναμένονται πολύτιμα συμπεράσματα από την έρευνα σε αυτόν. Μέχρι στιγμής υπάρχουν αρκετά στοιχεία τα οποία, όμως, δεν είναι αρκετά για να σχηματίζουν ολοκληρωμένη εικόνα.

Στους ευκαρυωτικούς οργανισμούς, η έκφραση ρυθμίζεται σε τέσσερα επίπεδα. Κατά τη μεταγραφή, μόνο όταν προσδεθεί στον υποκινητή ενός γονιδίου ο ορθός συνδυασμός μεταγραφικών παραγόντων (ειδική κατηγορία πρωτεϊνών), ξεκινά η RNA πολυμεράση τη διαδικασία. Κατά τη μετάφραση, ο χρόνος ζωής ενός mRNA στο κυτταρόπλασμα δεν είναι ο ίδιος για όλα τα γονίδια από τα οποία προέρχεται, ενώ ποικίλλει και η ικανότητα πρόσδεσής του στα ριβοσώματα. Επίσης, παράγοντες ρύθμισης μετά τη μεταγραφή είναι η ταχύτητα με την οποία γίνεται η ωρίμανση και η είσοδος του mRNA στον πυρήνα. Αντίστοιχα, μετά τη μετάφραση μπορεί η πρωτεΐνη να χρειάζεται τροποποιήσεις πριν γίνει βιολογικά ενεργή.

Οι προκαρυωτικοί οργανισμοί έχουν και εκείνοι ανάλογους τρόπους ρύθμισης της γενετικής τους έκφρασης. Πιο κοινός είναι η οργάνωση των γονιδίων σε ομάδες, που ονομάζονται *οπερόνια*, ώστε η μεταγραφή και η μετάφρασή τους να γίνεται ή για όλα και ταυτόχρονα ή για κανένα.

3.2.4 Εξέλιξη

Στο ερώτημα πώς έχει δημιουργηθεί η ποικιλομορφία των ειδών στον πλανήτη, δύο είναι οι κυρίαρχες θεωρίες που δίνουν απάντηση. Σύμφωνα με την πρώτη, οι διάφοροι οργανισμοί εμφανίστηκαν ο ένας ανεξάρτητα από τον άλλο και από τότε ο καθένας διατήρησε τα χαρακτηριστικά του. Η δεύτερη, πάλι, υποστηρίζει ότι αυτή η ποικιλομορφία οφείλεται στην εξελικτική πορεία. Με άλλα λόγια, οι πρόγονοι των οργανισμών είναι κοινοί, αλλά με το πέρασμα των χρόνων τα χαρακτηριστικά τους διαφοροποιήθηκαν.

Η μελέτη του πώς έγινε αυτή η διαφοροποίηση, καθώς επίσης ποιοι είναι οι πρόγονοι των ειδών αποτελεί το αντικείμενο της εξέλιξης (evolution). Στη συνέχεια παρουσιάζονται

δύο θεωρίες που είναι από τις πιο βασικές στον κλάδο αυτό της βιολογίας. Θυμίζεται ότι η ορθότητα των απόψεων αυτών δεν είναι αποδεδειγμένη, για αυτό άλλωστε λέγονται θεωρίες. Στηρίζουν, όμως, τα πορίσματά τους σε σημαντικές ενδείξεις.

Ο U.B. Lamarck ήταν ο πρώτος που διατύπωσε στις αρχές του 19ου αιώνα αμφιβολίες για τη σταθερότητα των χαρακτηριστικών των οργανισμών. Υποστήριξε ότι δύο είναι οι αρχές με βάση τις οποίες μεταβάλλονται οι ιδιότητές τους. Η μία είναι *η αρχή της χρήσης και αχρησίας*, η οποία σταδιακά αποβάλλει από τους οργανισμούς όποια όργανα τείνουν να μη χρησιμοποιούνται και δυναμώνει όσα κάνουν το αντίθετο. Η δεύτερη είναι *η αρχή της κληρονόμησης των επίκτητων χαρακτηριστικών*. Σύμφωνα με αυτήν ό,τι αποκτά ένας οργανισμός κατά τη διάρκεια της ζωής του κληρονομείται και στους απογόνους του.

Η θεωρία του Δαρβίνου (C. Darwin) είναι η δεύτερη της οποίας η συμβολή στην πρόοδο της εξελικτικής βιολογίας είναι μεγάλη. Είναι γνωστή και ως *θεωρία της φυσικής επιλογής*. Συνοπτικά υποστηρίζει ότι από τη στιγμή που οι ανάγκες των πληθυσμών της γης είναι μεγαλύτερες από εκείνες που μπορούν να καλυφθούν, η φύση αναγκάζει τα άτομα σε έναν αγώνα επιβίωσης στον οποίο επικρατούν εκείνοι που μπορούν να προσαρμοστούν καλύτερα στο περιβάλλον.

Από τις βασικές διαφορές των δύο θεωριών ξεχωρίζουν δύο. Η πρώτη είναι ότι ο Lamarck μίλησε για άτομα, ενώ ο Δαρβίνος για πληθυσμούς. Η δεύτερη αφορά το ρόλο της φύσης. Ουσιαστικά η θεωρία του Lamarck υποστηρίζει ότι η φύση *προκαλεί* τα άτομα να αλλάξουν, ενώ σύμφωνα με τη θεωρία του Δαρβίνου απλά *επιλέγει* εκείνους που είναι ήδη διαφορετικοί. Για παράδειγμα, ο μακρύς λαιμός της καμηλοπάρδαλης οφείλεται κατά τον ένα στη συνεχή προσπάθεια των ζώων να φθάσουν την τροφή ψηλά στα δέντρα, ενώ κατά τον άλλο κάποτε υπήρχαν και καμηλοπαρδάλεις με κοντούς λαιμούς οι οποίες, όμως, δεν έφταναν την τροφή και έτσι δεν κατάφεραν να επιβιώσουν.

Σήμερα, η θεωρία που διαμορφώνεται και κατέχει την προεξέχουσα θέση είναι η λεγόμενη *συνθετική*. Βασίζεται στη θεωρία του Δαρβίνου, αλλά λαμβάνει υπόψη της και δεδομένα από επιστήμες, όπως η Παλαιοντολογία και η Γενετική. Σε γενικές γραμμές υποστηρίζει πως μονάδα της εξελικτικής διαδικασίας είναι ο πληθυσμός και αυτή διαμορφώνεται σύμφωνα με τον παράγοντα της ποικιλομορφίας (κύριο ρόλο στον οποίο παίζουν οι μεταλλάξεις), της φυσικής επιλογής και της γενετικής απομόνωσης.

Τέλος, η πορεία της εξέλιξης των οργανισμών μπορεί να παρασταθεί με ένα δέντρο που λέγεται *φυλογενετικό*. Στη βάση του τοποθετείται το κοινό προγονικό είδος και στα κλαδιά του τα σύγχρονα είδη. Σε ενδιάμεσους κόμβους συνήθως μπαίνουν είδη που δεν υπάρχουν πια. Περισσότερα για αυτό δίνονται στο Κεφάλαιο 5.

3.3 Τεχνικές και πειραματικές μέθοδοι

Στην επαφή με τις βιοεπιστήμες συχνά συναντά κανείς την αναφορά σε πειράματα και διαδικασίες τεχνικού περιεχομένου. Κάποιες από αυτές αξίζουν ιδιαίτερα της προσοχής. Έχουν επιλεχθεί να αναλυθούν οι DNA sequencers, η μέθοδος PCR και η πειραματική διαδικασία των microarrays. Στην αρχή της ενότητας παρατίθενται ορισμένοι ορισμοί που, επίσης, είναι

πιθανό να αναρωτηθεί κανείς τι σημαίνουν.

3.3.1 Όροι

Για να περιγραφούν οι συνθήκες στις οποίες πραγματοποιείται μια διαδικασία, συνήθως πειραματική, συχνά χρησιμοποιούνται οι όροι που διευκρινίζονται στη συνέχεια.

in vivo

Η βιολογική διαδικασία γίνεται σε ζωντανό οργανισμό.

in vitro

Η βιολογική διαδικασία γίνεται σε δοκιμαστικό σωλήνα.

in silico

Όρος που πρωτοχρησιμοποιήθηκε το 1989 και ακολουθεί τη λογική των παραπάνω. Περιγράφει βιολογική διαδικασία που προσομοιώνεται σε ηλεκτρονικό υπολογιστή.

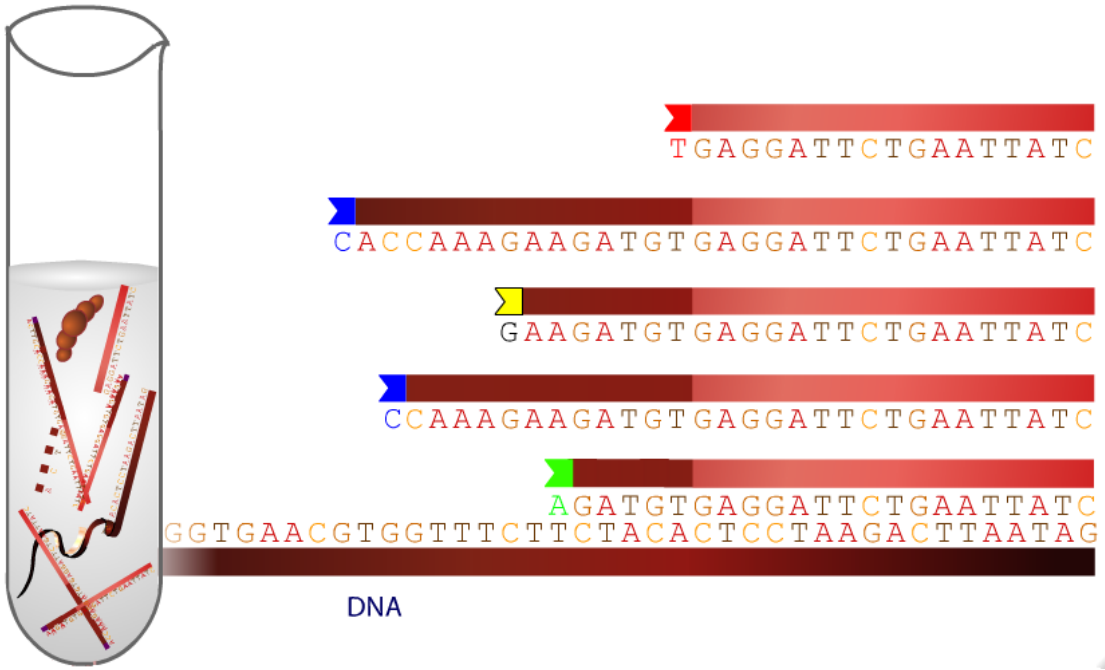
3.3.2 DNA sequencer

Είναι ενδιαφέρον και χρήσιμο να γνωρίζει ο πληροφορικός τα προγράμματα με τα οποία γίνεται γνωστή μια ακολουθία DNA. Αυτά λέγονται sequencers και εμπλέκονται στην πειραματική διαδικασία με την οποία διαβάζονται ακολουθίες. Τα σχήματα της παραγράφου έχουν πηγή τους το http://www.dnalc.org.

Αρχικά τοποθετείται στο δοκιμαστικό σωλήνα το μόριο του οποίου είναι άγνωστη η ακολουθία, καθώς επίσης ένα μικρό τμήμα ακολουθίας νουκλεοτιδίων που είναι συμπληρωματικό στη μία αλυσίδα του και το ένζυμο DNA πολυμεράση. Τα παραπάνω βρίσκονται σε πολλά αντίγραφα μέσα στο σωλήνα. Υπάρχουν, επιπλέον, τα συνηθισμένα νουκλεοτίδια και κάποια άλλου τύπου νουκλεοτίδια, τα οποία διαφέρουν ελάχιστα στη χημική σύσταση από τα κλασικά νουκλεοτίδια. Οι κύριες διαφορές τους είναι ότι μπορούν να συνδεθούν μόνο από τη μια πλευρά με άλλο νουκλεοτίδιο και το γεγονός ότι εκπέμπουν μήκος κύματος (διαφορετικό για καθένα από τα τέσσερα νουκλεοτίδια) κατάλληλα ανιχνεύσιμο.

Αφού σπάσει η δίκλωνη έλικα των μορίων υπό εξέταση, τα μικρά συμπληρωματικά τμήματα προσδένονται στο αντίστοιχο σημείο της μιας εκ των δύο αλυσίδων και η DNA πολυμεράση συνδέει στο μικρό τμήμα νουκλεοτίδια. Η επιμήκυνση αυτή σταματά, όταν συνδέει κάποιο από τα διαφορετικά νουκλεοτίδια που αναφέρθηκαν, αφού δε γίνεται δίπλα σε αυτό να προσδεθεί και δεύτερο, όπως εξηγήθηκε. Από τη στιγμή που υπάρχουν πολλά (δισεκατομμύρια) αντίτυπα, παράγονται και ανάλογα επιμηκυμμένα τμήματα. Η εικόνα σε αυτό το στάδιο είναι αυτή του Σχήματος 3.12.

Στη συνέχεια εμπλέκεται ο DNA sequencer. Τα επιμηκυμμένα τμήματα εισάγονται στο μηχάνημα και κατατάσσονται από το μικρότερο στο μεγαλύτερο. Στη συνέχεια περνούν διαδοχικά από ανιχνευτή ικανό να συλλάβει το μήκος κύματος των διαφορετικών ως προς τη



Σχήμα 3.12: Οι ακολουθίες που παράγονται από το πείραμα και διαβάζει ο DNA sequencer.

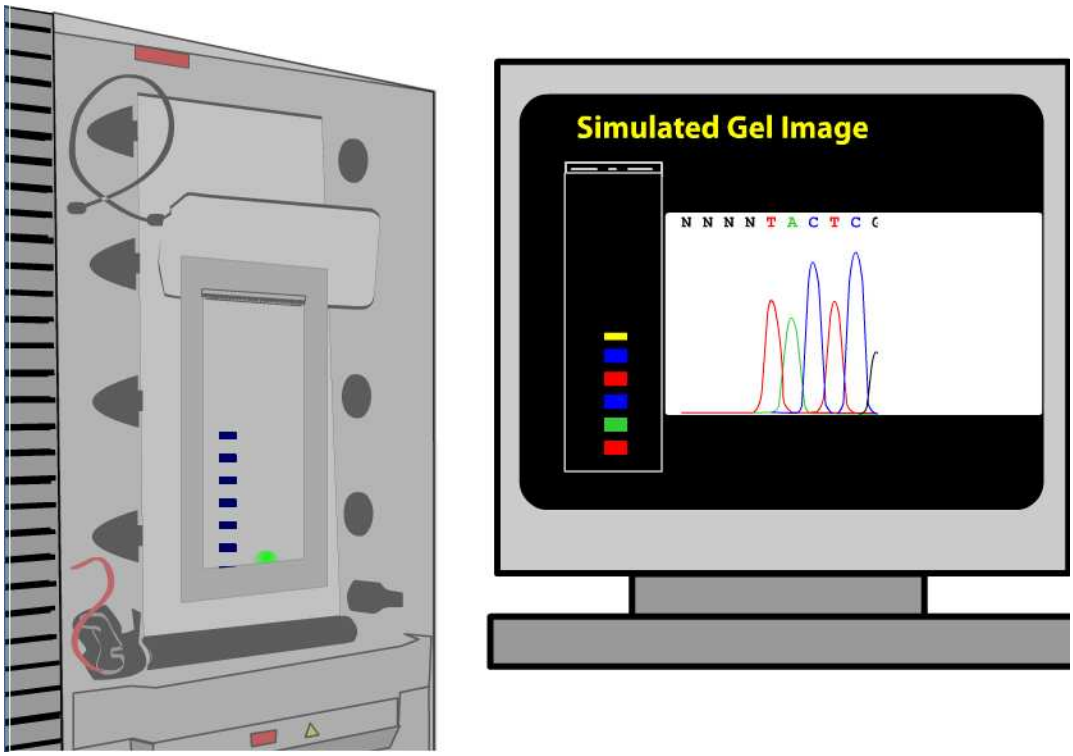
χημική σύσταση νουκλεοτιδίων που βρίσκονται στο τέλος κάθε ακολουθίας. Έτσι, παράγει ένα διάγραμμα με τα μήκη κύματος, όπως φαίνεται στο Σχήμα 3.13.

Εφόσον τα τμήματα έχουν καταταχθεί, περνούν από το μηχάνημα κατά αύξουσα σειρά μεγέθους και έτσι σχηματίζεται η ζητούμενη αρχική ακολουθία των γραμμάτων (Σχήμα 3.14). Ο τεράστιος αριθμός των μορίων DNA και των μικρών αλυσίδων, που είναι συμπληρωματικές στη μια του αλυσίδα, εξασφαλίζει με βάση τις πιθανότητες ότι για κάθε σημείο της αλυσίδας του θα υπάρχει επιμηκυμμένο τμήμα που να σταματά (λόγω των νουκλεοτιδίων διαφορετικού τύπου) εκεί.

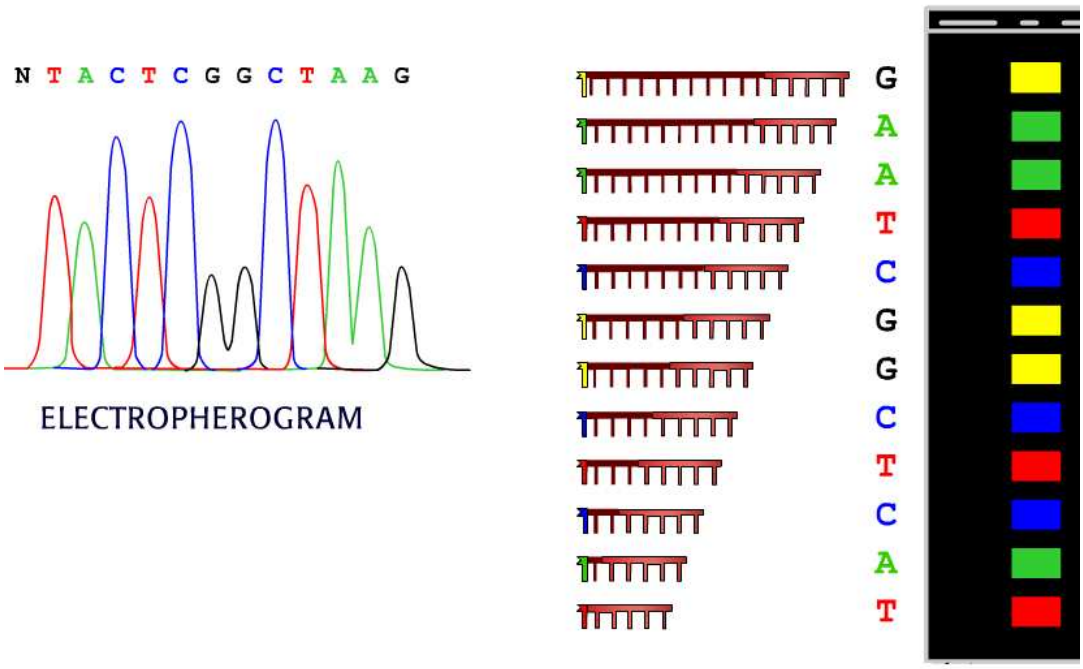
3.3.3 PCR

Η αλυσιδωτή αντίδραση πολυμεράσης ή αλλιώς PCR (Protein Chain Reaction) είναι μια τεχνική που ανακαλύφθηκε το 1985 στα πλαίσια της τεχνολογίας του ανασυνδυασμένου DNA.

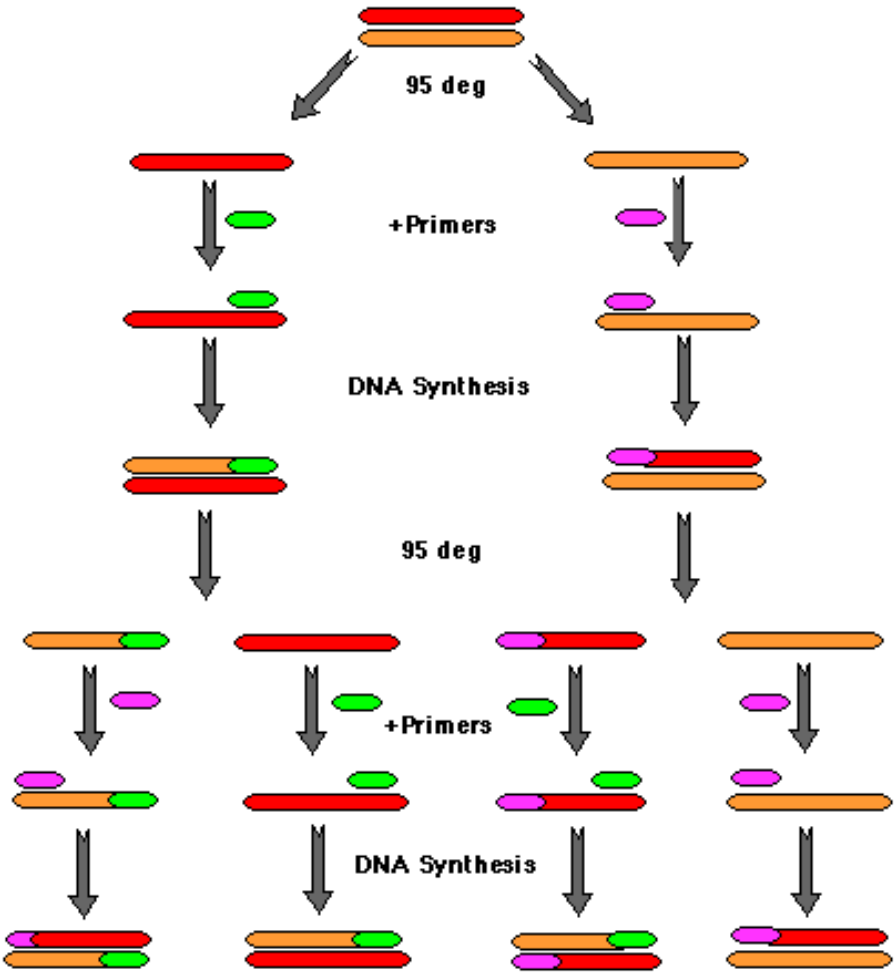
Επιτρέπει την επιλεκτική αντιγραφή, εκατομμύρια φορές, ειδικών αλληλουχιών DNA από ένα σύνθετο τμήμα μορίων DNA χωρίς τη μεσολάβηση ζωντανού κυττάρου. Συγκεκριμένα, χρειάζεται να σπάσει το δίκλωνο μόριο σε δύο αλυσίδες και το ένζυμο DNA πολυμεράση να επιμηκύνει κάποια συμπληρωματικά τμήματα (primers) που τοποθετούνται στην κάθε αλυσίδα. Μετά από επανάληψη n φορές της διαδικασίας, προκύπτουν 2^n πανομοιότυπα με το αρχικό μόρια. Στο Σχήμα 3.15 φαίνεται η κατάσταση για δύο κύκλους.



Σχήμα 3.13: Η διαδικασία της ανάγνωσης από τον DNA sequencer.



Σχήμα 3.14: Τα τελικά αποτελέσματα του DNA sequencer.



Σχήμα 3.15: Η μέθοδος PCR για δύο κύκλους.

3.3.4 Microarrays

Η τεχνολογία των microarrays αναπτύχθηκε τη δεκαετία του '90 και χρησιμοποιείται για πολλούς διαφορετικούς λόγους. Για το λόγο αυτό υπάρχουν και αρκετά διαφορετικά είδη πειραμάτων (DNA/protein/tissue/transfection/antibody/chemical compound microarrays). Η λογική, ωστόσο, στην οποία βασίζονται αυτά είναι η ίδια. Ίδιος, επίσης, είναι και ο βαθύτερος σκοπός: η εύρεση της γενετικής έκφρασης. Στα ακόλουθα περιγράφεται ένα παράδειγμα για DNA microarray, στο οποίο παρουσιάζονται τα βασικά σημεία της τεχνικής που ακολουθείται.

Όπως έχει αναφερθεί (Ενότητα 3.2.2), ένα γονίδιο θεωρείται ότι εκφράζεται σε ένα κύτταρο σε δύο περιπτώσεις, κοινό χαρακτηριστικό των οποίων είναι η μεταγραφή του γονιδίου σε mRNA. Επιπλέον, είναι γνωστό (Κεφάλαιο 2) πως δεν εκφράζονται όλα τα γονίδια σε όλα τα κύτταρα, αλλά ανάλογα με τον τύπο του κυττάρου ενεργοποιείται η μεταγραφή των κατάλληλων γονιδίων.

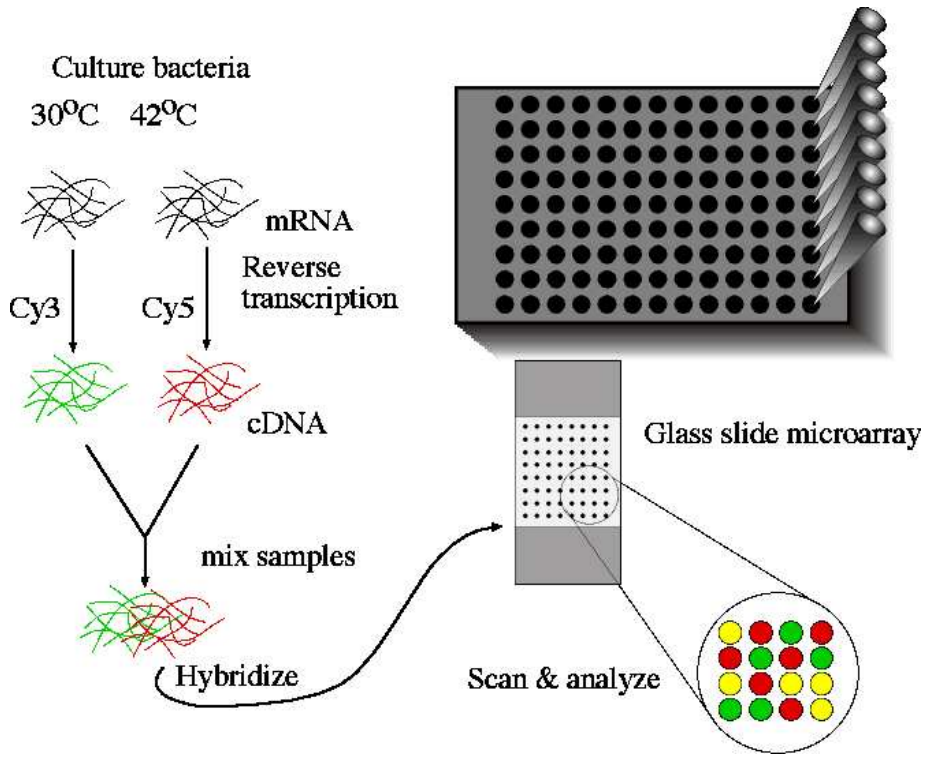
Ένας τρόπος για να βρεθεί, επομένως, ποια γονίδια εκφράζονται (δηλαδή μεταγράφονται) σε ένα κύτταρο είναι να βρεθούν όλα τα διαφορετικά mRNAs του. Όταν αυτό γίνει, μπορεί να λάβει χώρα στο εργαστήριο μια διαδικασία γνωστή ως *αντίστροφη μεταγραφή*: από το μονόκλωνο mRNA δημιουργείται, επίσης μονόκλωνο, DNA με τον αντίστροφο τρόπο από αυτόν με τον οποίο γίνεται η μεταγραφή. Το DNA που σχηματίζεται ονομάζεται cDNA (complementary DNA). Από τη στιγμή, μάλιστα, που έχει μία μόνο αλυσίδα, αν βρεθεί η συμπληρωματική του και υπάρχουν οι κατάλληλες περιβαλλοντικές συνθήκες, θα ενωθεί με αυτήν και θα γίνει δίκλωνο. Το γεγονός αυτό εκμεταλλεύεται η μέθοδος των πειραμάτων των microarrays.

Έστω ότι είναι διαθέσιμα τα mRNAs δύο κυττάρων ίδιου τύπου (π.χ. και τα δύο νευρικά), εκ των οποίων το ένα είναι υγιές, αλλά το άλλο ασθενές. Ζητούμενο είναι να προσδιοριστούν τα γονίδια (ένα ή περισσότερα) στα οποία οφείλεται η ασθενής κατάσταση του δεύτερου κυττάρου, αν υποθέσουμε βέβαια ότι η αιτιολογία είναι γενετικής φύσεως. Εκτελείται, τότε, ένα microarray πείραμα σαν αυτό του Σχήματος 3.16.

Σε κάθε τετράγωνο του πίνακα τοποθετείται cDNA από ένα γονίδιο από εκείνα που ενδιαφέρει να εξεταστούν. Έτσι, η αντιστοιχία γονιδίων και τετραγώνων είναι ένα προς ένα. Επιπλέον, δημιουργούνται όλα τα δυνατά cDNAs των δύο κυττάρων και χρωματίζονται ανάλογα με το κύτταρο προέλευσής τους. Αφού αναμιχθούν μεταξύ τους, ρίχνονται στον πίνακα. Όλες αυτές βέβαια οι διαδικασίες απαιτούν τις κατάλληλες χημικές συνθήκες, που εδώ δεν αναφέρονται, καθώς δεν εξυπηρετούν επιθυμητό σκοπό της παρούσας ανάλυσης.

Η έκφραση του κάθε γονιδίου στο κάθε κύτταρο γίνεται φανερή από το χρώμα που παίρνει το τετράγωνο που καταλαμβάνει αυτό στον πίνακα. Επειδή, όπως αναφέρθηκε, είναι μονόκλωνο, μπορεί να κολλήσει με τη συμπληρωματική του αλυσίδα. Έτσι, αν αυτή υπάρχει, είτε από το ένα κύτταρο προερχόμενη είτε από το άλλο, το μόριο θα γίνει δίκλωνο. Αυτό συνεπάγεται ότι το τετράγωνό του θα χρωματιστεί με το χρώμα που αντιστοιχεί στο εν λόγω κύτταρο.

Υπάρχει, βέβαια, η πιθανότητα το γονίδιο να εκφράζεται και στα δύο κύτταρα ή σε κανένα. Στην πρώτη περίπτωση το χρώμα του τετραγώνου του θα είναι ο συνδυασμός των χρωμάτων που αντιστοιχούν σε κάθε κύτταρο, ενώ στη δεύτερη θα διατηρήσει το αρχικό. Ένας τέτοιος

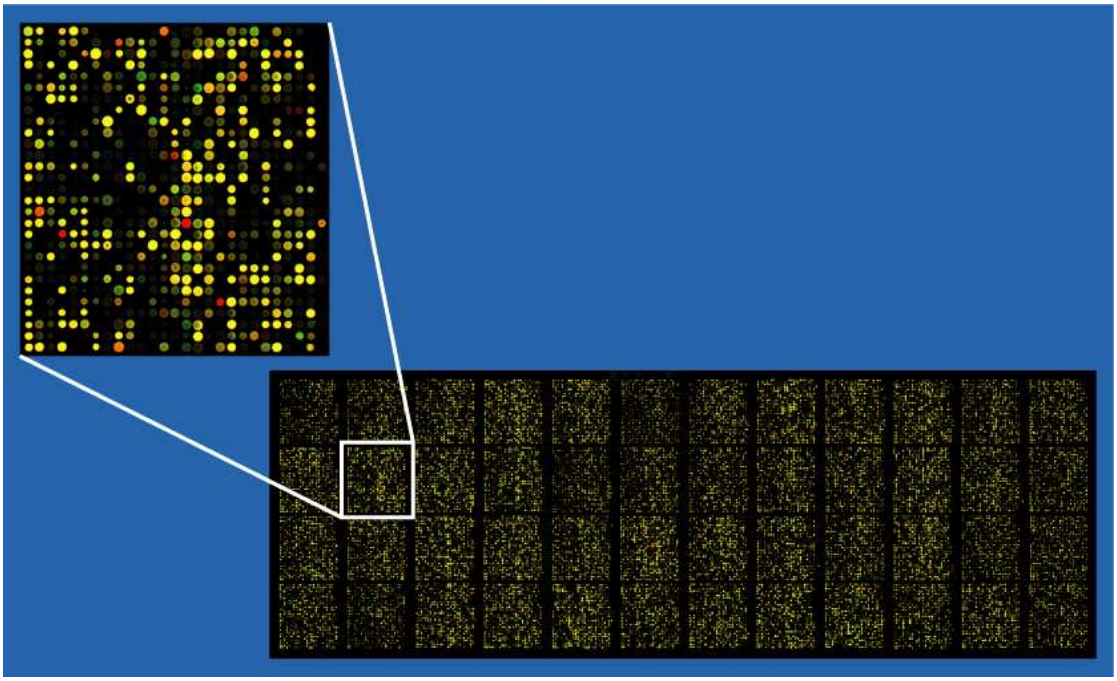


Σχήμα 3.16: Πείραμα microarray.

τελικός πίνακας φαίνεται στο Σχήμα 3.17. Τα χρώματα για τα δύο κύτταρα ήταν κόκκινο και πράσινο, οπότε ο συνδυασμός τους έβγαλε το κίτρινο.

Με την τεχνολογία των microarrays είναι δυνατό να τοποθετηθούν στον πίνακα χιλιάδες γονίδια. Επιπλέον, όπως ήδη ειπώθηκε, υπάρχουν αρκετές παραλλαγές του τι τύπου κύτταρα εξετάζονται, αλλά όλες βασίζονται στη διαδικασία που περιγράφηκε.

Πέρα από την κατανόηση της μεθόδου, σημασία για το μηχανικό υπολογιστών έχει κυρίως η μορφή των πειραματικών αποτελεσμάτων, τα λεγόμενα gene expression data, που αναφέρονται στο Κεφάλαιο 4. Πρόκειται για έναν αλγεβρικό πίνακα με τόσες θέσεις όσες έχει και αυτός των πειραμάτων. Η τιμή κάθε στοιχείου του προκύπτει ως ο λογάριθμος του λόγου της ποσότητας του cDNA, που βρίσκεται μέσα στο αντίστοιχο τετράγωνο του πίνακα του πειράματος και προέρχεται από το ένα κύτταρο-δείγμα, προς την αντίστοιχη που προέρχεται από το δεύτερο δείγμα.



Σχήμα 3.17: Αποτέλεσμα του πειράματος microarray.

Κεφάλαιο 4

Δεδομένα

Τα δεδομένα των βιοεπιστημών τα οποία χρειάζεται να αποθηκεύονται και να τίθενται υπό επεξεργασία έχουν δύο κύρια χαρακτηριστικά, που είναι πηγή σημαντικών δυσκολιών για αυτές τις εργασίες. Το πρώτο είναι η τεράστια ποσότητά τους και το δεύτερο η ποικιλομορφία τους. Ένας επιπλέον παράγοντας που είναι απαραίτητο να λαμβάνεται υπόψη είναι το γεγονός ότι η έρευνα στη βιολογία βρίσκεται σε συνεχή πρόοδο, με αποτέλεσμα όχι μόνο να αυξάνονται τα δεδομένα αλλά και να ανακαλύπτονται νέοι τύποι τους.

Είναι κοινή διαπίστωση ότι η έρευνα των επιστημόνων έχει ανάγκη τη μελέτη πολύ μεγάλων ποσοτήτων πληροφοριών. Υπολογίζεται ότι ο όγκος δεδομένων της μοριακής βιολογίας αυξάνει με εκθετικό ρυθμό. Η ανακάλυψη των γονιδιωμάτων του ανθρώπου (3 δισεκατομμύρια βάσεις) αλλά και άλλων οργανισμών, όπως του ποντικού (2,6 δισεκατομμύρια βάσεις), της μύγας (137 εκατομμύρια βάσεις) και της μαγιάς (12,1 εκατομμύρια βάσεις) είναι ενδεικτική της επανάστασης που έχει συντελεστεί τα τελευταία είκοσι περίπου χρόνια όσον αφορά τα ποσοτικά μεγέθη. (βλ. Αριθμητικά Στοιχεία, Κεφάλαιο 2)

Ωστόσο, η πολυπλοκότητα των δεδομένων αυτών θεωρείται ακόμη μεγαλύτερο πρόβλημα. Πρόκειται για διαφορετικούς τύπους δεδομένων, που ταυτόχρονα συσχετίζονται μεταξύ τους. Τα είδη είναι πολλά, το ίδιο και οι μορφές με τις οποίες εμφανίζονται, όπως θα γίνει φανερό και στη συνέχεια του κεφαλαίου. Σε αυτά θα πρέπει να προστεθεί το γεγονός ότι καινούριοι τύποι εμφανίζονται με σημαντική συχνότητα.

Η ιδιοσυγκρασία των βιοδεδομένων καθιστά δύσκολη την αναπαράστασή τους με ένα μοντέλο. Προς το παρόν υπάρχουν λίγα μοντέλα που ανταποκρίνονται στις απαιτήσεις, τα οποία μάλιστα αφορούν συγκεκριμένο τύπο δεδομένων και δεν εφαρμόζονται σε όλο το πλήθος των διαθέσιμων βάσεων. Χρειάζεται, επίσης, να σημειωθεί ότι το όποιο μοντέλο υιοθετείται φαίνεται αναγκαίο να είναι αρκετά ευέλικτο, καθώς οι εξελίξεις στη βιολογία τρέχουν με γρήγορους ρυθμούς.

Στη συνέχεια του κεφαλαίου γίνεται λόγος για τους τύπους δεδομένων που ενδιαφέρουν, για τα πρότυπα, μοντέλα και formats που χρησιμοποιούνται ως επί το πλείστον καθώς και για τις υπάρχουσες βάσεις δεδομένων.

4.1 Είδη

Στην ενότητα αυτή αναφέρονται οι τύποι δεδομένων τους οποίους μελετούν οι ερευνητές της βιολογίας και των συγγενών επιστημών. Καταγράφονται, δηλαδή, τα δεδομένα που χρειάζονται οι επιστήμονες να αποθηκεύουν και να επεξεργάζονται. Για κάθε είδος δίνεται μια σύντομη περιγραφή, ενώ όπου κρίνεται απαραίτητη βαθύτερη κατανόηση ο αναγνώστης μπορεί να ανατρέξει στις αντίστοιχες παραγράφους από τα δύο προηγούμενα κεφάλαια της παρούσας εργασίας.

4.1.1 Νουκλεϊκά οξέα

Στη μεγάλη πλειοψηφία των περιπτώσεων το ενδιαφέρον είναι επικεντρωμένο γύρω από το DNA και όχι το RNA. Σε αυτήν την κατηγορία ανήκουν τα γονιδιώματα των οργανισμών, τα πλασμίδια των προκαρυωτικών κυττάρων (κυκλικά μόρια DNA), το γενετικό υλικό που υπάρχει σε οργανίδια όπως τα μιτοχόνδρια αλλά και το RNA των ριβοσωμάτων. Η ομάδα αυτή αφορά τις γενετικές πληροφορίες των έμβιων όντων και η έρευνά της είναι εμφανώς υψίστης σημασίας.

4.1.2 ESTs

Τα ESTs (expressed sequence tags) δημιουργούνται από βιβλιοθήκες cDNA (complementary DNA). Αυτό σημαίνει ότι αποτελούνται μόνο από τμήματα DNA τα οποία μεταγράφονται και ταυτόχρονα δεν περιέχουν εσώνια (introns). Για το λόγο αυτό προσφέρουν σημαντικές πληροφορίες σχετικές με την έκφραση των γονιδίων. Το μέγεθός τους είναι μικρό, περίπου 200 με 500 νουκλεοτίδια, και μπορούν να ανιχνευθούν με τη μέθοδο PCR. Ανήκουν στη γενικότερη κατηγορία των STSs (sequence tagged sites), που είναι τμήματα DNA με γνωστή αλληλουχία νουκλεοτιδίων και μοναδικότητα στο ανθρώπινο γονιδίωμα.

4.1.3 Επίπεδα γονιδιακής έκφρασης

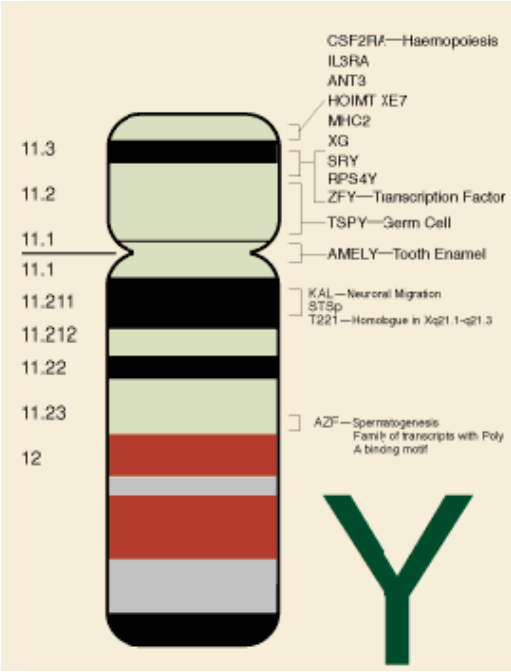
Ιδιαίτερο ενδιαφέρον έχουν στοιχεία που δίνουν πληροφορίες για την έκφραση των γονιδίων (gene expression). Αυτά έχουν να κάνουν με τη χρονική στιγμή και συχνότητα με την οποία δημιουργείται mRNA από ένα γονίδιο σε ένα είδος κυττάρου. Στην περίπτωση που το γονίδιο είναι υπεύθυνο για την δημιουργία κάποιας πρωτεΐνης τότε μπορούν να συλλεχθούν και πληροφορίες για το χρόνο και το ρυθμό παραγωγής αυτής. Όλα αυτά τα στοιχεία προέρχονται τις περισσότερες φορές από τα πειράματα των microarrays, ενώ ήδη αναφέρθηκε ότι και τα ESTs προσφέρουν πληροφορίες για τη γονιδιακή έκφραση.

Τουλάχιστον τρία είναι τα επίπεδα στα οποία μπορεί να γίνει ανάλυση των χαρακτηριστικών της έκφρασης των γονιδίων. Στο πρώτο εξετάζεται η συμπεριφορά ενός μόνο γονιδίου σε διαφορετικό σημείο της ζωής ενός κυττάρου ή σε διαφορετικού τύπου κύτταρα ή σε διαφορετικές περιβαλλοντικές συνθήκες. Στο δεύτερο ερευνώνται οι σχέσεις που μπορεί να έχουν κάποια γονίδια μεταξύ τους, ώστε να δημιουργηθούν ομάδες (clusters) από αυτά. Για παράδειγμα, κατά πόσον η έκφρασή τους ενεργοποιείται από τους ίδιους παράγοντες. Το τρίτο

επίπεδο αφορά συσχετίσεις γονιδίων και πρωτεϊνών.

4.1.4 Χάρτες γονιδίων

Είναι σημαντικό για τους επιστήμονες να γνωρίζουν την ακριβή θέση ενός γονιδίου, δηλαδή σε ποιο συγκεκριμένο σημείο ενός γνωστού χρωμοσώματος βρίσκεται. Με τις πληροφορίες αυτές σχηματίζουν τους λεγόμενους *φυσικούς χάρτες (physical maps)*, παράδειγμα των οποίων φαίνεται στο Σχήμα 4.1 [1]. Ανάλογα με την ανάλυσή τους και τη μέθοδο δημιουργίας τους διακρίνονται σε τέσσερα είδη (chromosomal, cDNA, macrorestriction, contig).

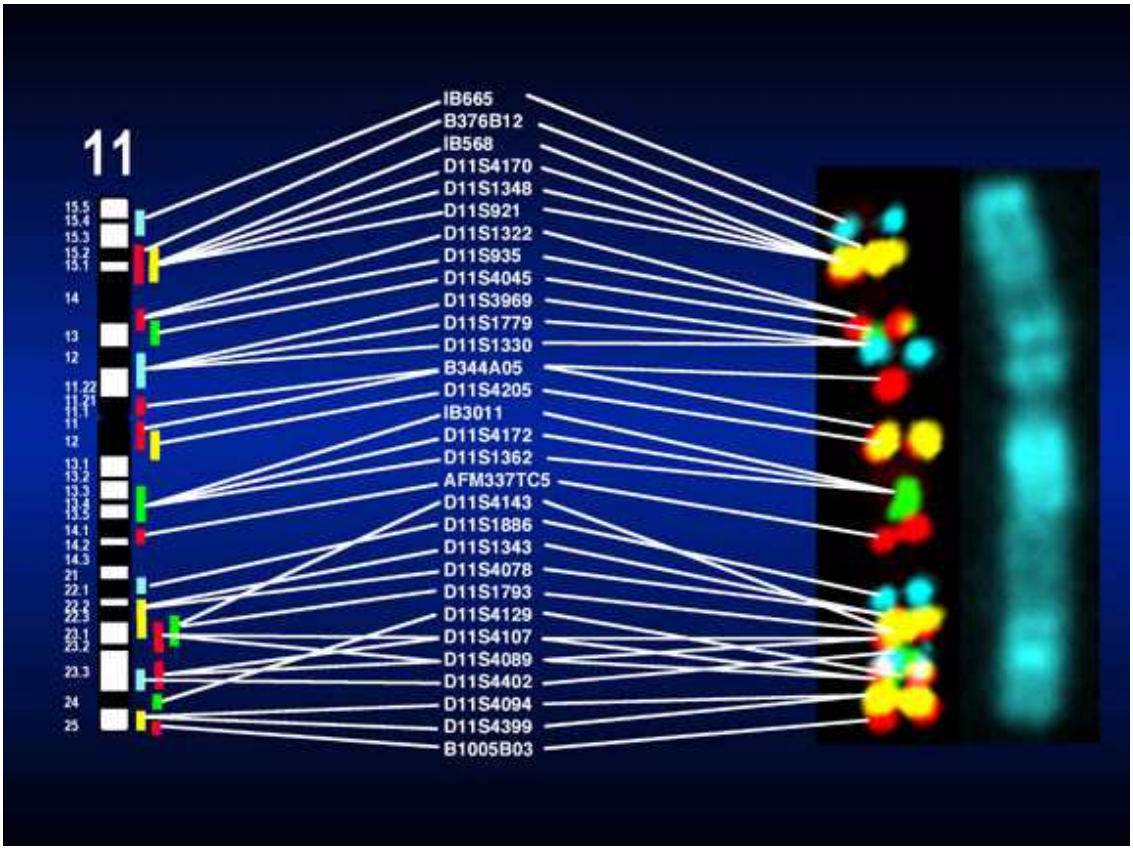


Σχήμα 4.1: Φυσικός χάρτης των γονιδίων του χρωμοσώματος Υ.

Ένας ακόμη σπουδαίος τύπος διαγράμματος είναι οι *γενετικοί χάρτες* (genetic maps). Η απόσταση μεταξύ δύο γονιδίων μετριέται σε cM (centimorgans, από τον Αμερικανό γενετιστή Thomas Hunt Morgan) και είναι ανάλογη της πιθανότητας που έχουν δύο γονίδια του ίδιου χρωμοσώματος να κληροδοτηθούν και τα δύο στον απόγονο. Οι χάρτες αυτοί έχουν μεγάλη αξία για τη μελέτη κληρονομικών ασθενειών, ακόμη και όταν δεν είναι γνωστή η ακριβής θέση ενός γονιδίου αλλά μόνο η περιοχή στην οποία βρίσκεται (Σχήμα 4.2 [2]).

Η δημιουργία τόσο των φυσικών όσο και των γενετικών χαρτών είναι σε εξέλιξη για την πλειοψηφία των οργανισμών με αποκωδικοποιημένο γονιδίωμα. Όσον αφορά τη σχέση που έχουν μεταξύ τους, απόσταση 1cM σε ένα γενετικό χάρτη αντιστοιχεί περίπου σε απόσταση ενός εκατομμυρίου βάσεων σε ένα φυσικό χάρτη.

[1] Πηγή: http://www.ucl.ac.uk/tcga/ScienceSpectra-pages/pics/14-bradman_fig_1.gif
[2] Πηγή: http://www.csmc.edu/csri/korenberg/images/papers/figure2.jpg



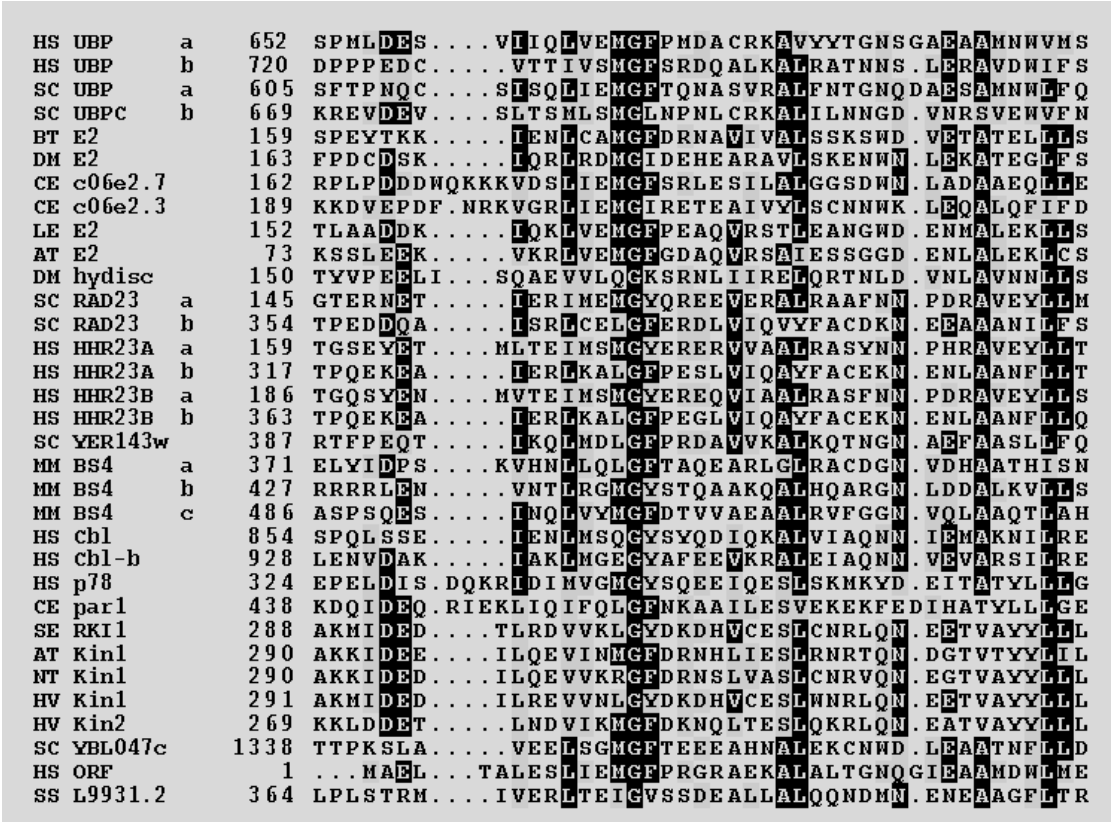
Σχήμα 4.2: Γενετικός χάρτης για το χρωμόσωμα 11 του ανθρώπου.

4.1.5 Πρωτεΐνες

Ένα μεγάλο τμήμα της έρευνας στη μοριακή βιολογία αφορά τις πρωτεΐνες. Η βιολογική τους αξία δικαιολογεί το γεγονός αυτό και έχει τονισθεί στα δύο προηγούμενα κεφάλαια. Ξεχωριστό ενδιαφέρον παρουσιάζουν τα ένζυμα, ενώ και οι πρωτεΐνες που συμβάλλουν στη μεταγραφή του DNA μελετώνται ιδιαίτερα. Για την τελευταία κατηγορία γίνεται λόγος και στην επόμενη παράγραφο.

4.1.6 Motifs, transcription factors

Αρκετές φορές είναι ζητούμενη η ανακάλυψη στατιστικά συχνά επαναλαμβανόμενων τμημάτων (patterns) σε νουκλεϊκά οξέα ή πρωτεΐνες. Αυτά είναι γνωστά με την ονομασία motifs και διακρίνονται σε sequence motifs, αν αφορούν ακολουθίες και structural motifs, όταν αναφέρονται σε τρισδιάστατες δομές. Παράδειγμα sequence motif δίνεται στο Σχήμα 4.3 [3].



Σχήμα 4.3: Sequence motifs.

Μια ειδική κατηγορία στην οποία έχει εφαρμογή η ανακάλυψη motifs είναι αυτή των παραγόντων της μεταγραφής. Ο όρος transcription factors αναφέρεται σε πρωτεΐνες που συμμετέχουν στη μεταγραφή του DNA. Οι παράγοντες αυτοί μάλιστα διακρίνονται σε τρεις κατηγορίες (general, upstream, inducible).

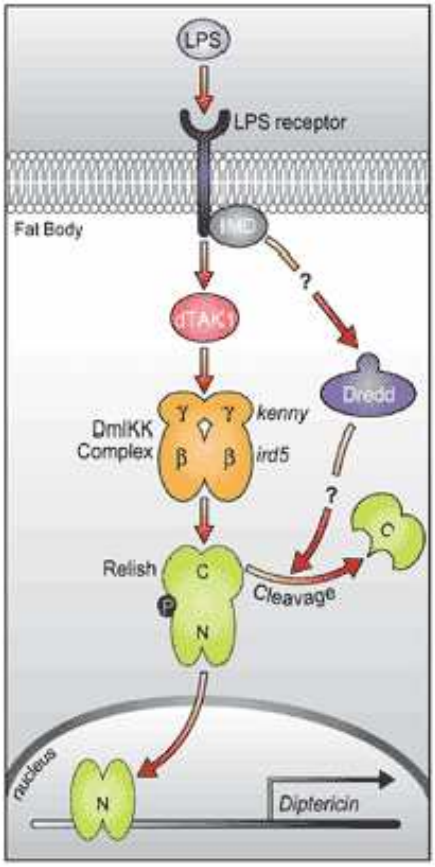
[3] Πηγή: http://www.isrec.isb-sib.ch/profile/isrec96/uba_aln_big.gif

4.1.7 Βιομονοπάτια

Τα βιομονοπάτια (biological pathways) περιλαμβάνουν τριών ειδών διαδρομές μέσα σε έναν οργανισμό. Η πρώτη ομάδα αφορά το μεταβολισμό και τις βιοχημικές αντιδράσεις που λαμβάνουν χώρα (metabolical, biochemical). Η δεύτερη ασχολείται με την πορεία που ακολουθείται για τη μεταγραφή του DNA και τη σύνθεση των πρωτεΐνών (transcription, protein synthesis). Η τρίτη κατηγορία είναι εκείνη που αποτελείται από τις διαδρομές των σημάτων (signaling).

Στο δεύτερο είδος ανήκουν και τα λεγόμενα δίκτυα γονιδίων (gene networks). Σε αυτά τοποθετούνται τα γονίδια ως κόμβοι που δέχονται ως είσοδο πρωτεΐνες, οι οποίες είναι παράγοντες της μεταγραφής, και έχουν έξοδο τα επίπεδα της γενετικής έκφρασης.

Συνοπτικά, μπορεί να ειπωθεί ότι τα βιομονοπάτια αφορούν την παραγωγή και τη διαδρομή της ενέργειας, των πρωτεϊνικών μορίων και των ηλεκτρικών σημάτων σε έναν οργανισμό. Στο Σχήμα 4.4 [4] φαίνεται ενδεικτικά η διαδρομή που ακολουθεί ένα σήμα σε κύτταρο.



Σχήμα 4.4: Παράδειγμα μονοπατιού σημάτων (Drosophila antibacterial).

[4] Πηγή: http://www.umassmed.edu/infdis/graphics/silverman1.gif

4.1.8 Πεδία

Φυσικά μεγέθη, βαθμωτά ή διανυσματικά, δίνουν αξιοσημείωτες πληροφορίες για τη συμπεριφορά του κυττάρου και του οργανισμού ως συνόλου. Χαρακτηριστικά παραδείγματα είναι η διαφορά δυναμικού κατά μήκος της επιφάνειας ενός κυττάρου, η ροή ασβεστίου αλλά και πρωτεϊνών από και προς το κύτταρο, η κλινική απόκριση σε φάρμακα.

4.1.9 Μαθηματικό μοντέλο, περιορισμοί

Παρά τη χρησιμότητα, έχει δοθεί μικρή έμφαση στη συστηματική αναπαράσταση, αποθήκευση και επεξεργασία του μαθηματικού ή στατιστικού μοντέλου που χρησιμοποιείται για την εξαγωγή μετρήσεων και συμπερασμάτων. Η προσοχή έχει στραφεί αντίθετα στα δεδομένα εισόδου ή εξόδου αυτού.

Τα μαθηματικά μοντέλα που χρησιμοποιούνται στις επιστήμες τις σχετικές με τη βιολογία εισάγουν για τα δεδομένα και περιορισμούς που ξεφεύγουν από το επίπεδο των λογικών κανόνων. Πρόκειται για ισότητες ή ανισότητες που σχετίζονται κυρίως με αρχές διατήρησης (μάζας, ορμής, ενέργειας) της φυσικής και της χημείας. Έχουν τη μορφή είτε τοπικού είτε γενικού περιορισμού. Για παράδειγμα, αν ενδιαφέρει η μάζα αντιδρώντων και προϊόντων σε μία αντίδραση, τότε είναι τοπικός ο περιορισμός, ενώ αν η αναφορά γίνεται σε έναν κύκλο αντιδράσεων, ο περιορισμός είναι γενικός.

4.1.10 Άρθρα και σημειώσεις

Είναι σχεδόν βέβαιο ότι ένας επιστήμονας που ενδιαφέρεται για κάποιο συγκεκριμένο θέμα θα επιθυμεί να ενημερωθεί για τις σχετικές με αυτό δημοσιεύσεις. Πλέον σχεδόν όλα τα περιοδικά έχουν ηλεκτρονικές εκδόσεις.

Επιπλέον, βοηθητικές γνώσεις αποκτώνται και από την ανάγνωση σχολίων και σημειώσεων που συνοδεύουν μια εγγραφή σε βάση δεδομένων. Ένα τέτοιο παράδειγμα είναι τα διαφορετικά ονόματα με τα οποία μπορεί να συναντήσει κανείς την ίδια πρωτεΐνη.

4.2 Μορφές

Είναι χρήσιμο να γνωρίζει ο ενδιαφερόμενος μηχανικός υπολογιστών τους τύπους δεδομένων από τη σκοπιά της βιολογίας, ώστε να μπορεί να κατανοεί το αντικείμενο και τις απαιτήσεις του. Ίσως είναι, όμως, ακόμη σημαντικότερο να ξέρει τη μορφή που έχουν αυτά τα δεδομένα ανεξάρτητα από τη βιολογική τους σημασία, για να μπορεί να βρει κατάλληλους τρόπους αποθήκευσης και επεξεργασίας. Στα επόμενα καταγράφονται οι διαφορετικές μορφές δεδομένων που ήδη χρησιμοποιούνται.

4.2.1 Ακολουθίες

Η πλειονότητα των βάσεων που υπάρχουν αυτή τη στιγμή έχουν δεδομένα αυτής της μορφής. Ως ακολουθίες αποθηκεύονται τόσο νουκλεϊκά οξέα (αλληλουχίες νουκλεοτιδίων) όσο και

πρωτεΐνες (αλληλουχίες αμινοξέων).

Στην περίπτωση του DNA και του RNA το λεξιλόγιο είναι τέσσερα γράμματα: A (DNA) ή U (RNA), T, G, C. Όπως έχει αναφερθεί στο Κεφάλαιο 2, αυτά αντιπροσωπεύουν τις τέσσερις αζωτούχες βάσεις αδενίνη (μόνο στο DNA), ουρακίλη (μόνο στο RNA), θυμίνη, γουανίνη, κυτοσίνη αντίστοιχα. Αν και το DNA είναι δίκλωνο, λόγω της αρχής της συμπληρωματικότητας των βάσεων, αρκεί η ακολουθία της μιας μόνο αλυσίδας για να είναι γνωστή και αυτή της άλλης. Το RNA αποτελείται ούτως ή άλλως από μία μόνο αλυσίδα.

Για τις πρωτεΐνες το λεξιλόγιο είναι μεγαλύτερο. Αποτελείται από τα είκοσι αμινοξέα για τα οποία έχει πάλι γίνει λόγος στο Κεφάλαιο 2. Τελευταία υπάρχουν ενδείξεις ότι ο αριθμός των διαφορετικών αμινοξέων που δομούν τις πρωτεΐνες είναι μεγαλύτερος. Ωστόσο, στις υπάρχουσες βάσεις δεδομένων δε συναντώνται άλλα πέρα από αυτά τα είκοσι.

Στην κατηγορία αυτή πρέπει να ενταχθούν και τα ESTs όπως και τα sequence motifs, αφού σε τελική ανάλυση αποτελούνται από νουκλεοτίδια ή αμινοξέα.

Παράδειγμα ακολουθίας νουκλεοτιδίων του DNA: ATTAAACACTGTAATCTTAAGT
Παράδειγμα ακολουθίας αμινοξέων: RQVPDARLLKSMSYQEAMELSY

4.2.2 Τρισδιάστατες δομές

Η διάταξη ενός μακρομορίου στο χώρο δίνει σημαντικές πληροφορίες για τη λειτουργία του. Έχει, λοιπόν, νόημα να μπορεί να αποθηκεύεται και να τίθεται υπό επεξεργασία η τρισδιάστατη δομή ενός μορίου. (Σχήμα 4.5 [5]) Λόγω των δυσκολιών, ωστόσο, που υπάρχουν σε αυτές τις δύο εργασίες ο αριθμός των βάσεων με τρισδιάστατα στοιχεία που υπάρχουν προς το παρόν είναι περιορισμένος.

Το κύριο βιομόριο του οποίου η διάταξη στο χώρο ενδιαφέρει είναι οι πρωτεΐνες. Οι πληροφορίες που παρέχονται από την τριτοταγή και τεταρτοταγή δομή τους χάνονται, όταν η πρωτεΐνη αποθηκεύεται απλά σα μια ακολουθία αμινοξέων. Όπως αναφέρθηκε, από το σχήμα της συμπεραίνονται αρκετά και για τη λειτουργία της.

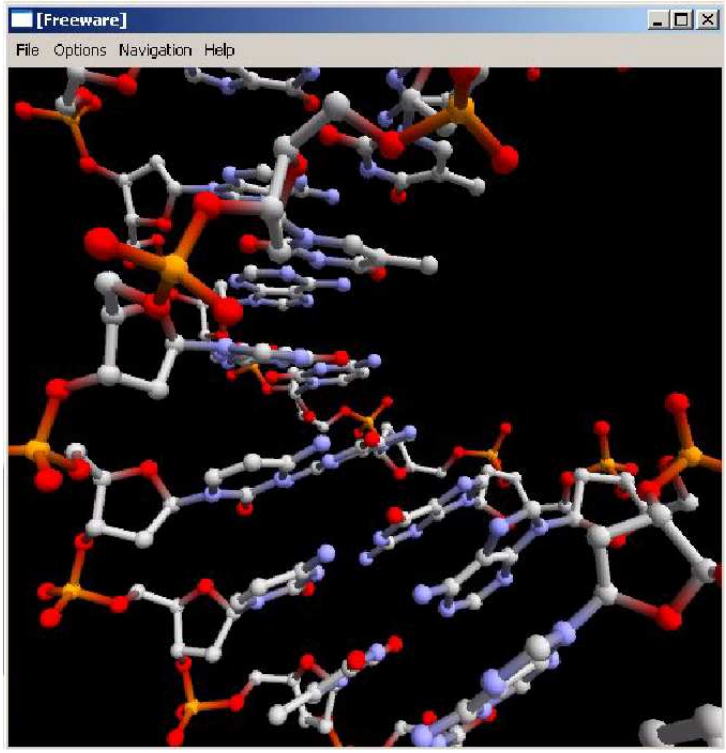
Σχετικά με τα νουκλεϊκά οξέα, μεγαλύτερο όφελος υπάρχει από την αποθήκευση της τριασδιάστατης δομής του RNA παρά από αυτήν του DNA. Αυτό ισχύει επειδή είναι γνωστό ότι το DNA είναι μια δίκλωνη έλικα. Αντίθετα, για το RNA δεν είναι εύκολα προβλέψιμα τα σημεία στα οποία αναδιπλώνεται.

Τέλος, στην ομάδα αυτή ανήκουν και τα structural motifs.

4.2.3 Πίνακες

Η μορφή που έχουν τα δεδομένα εξαρτάται όχι μόνο από τη φύση τους αλλά και από την πειραματική μέθοδο από την οποία προέκυψαν. Αρκετά στοιχεία προέρχονται από τα πειράματα των microarrays. Τα αποτελέσματα με αυτή τη μέθοδο είναι μεγάλοι και αραιοί πίνακες πραγματικών αριθμών. Περισσότερες λεπτομέρειες για τους microarrays έχουν δοθεί στο Κεφάλαιο 3.

[5] Πηγή: http://www.geocities.com/pdbviewer/Viewer1.jpg



Σχήμα 4.5: Τρισδιάστατη δομή μακρομορίου.

4.2.4 Γράφοι

Χρησιμοποιούνται για την αναπαράσταση των βιομονοπατιών αλλά και των γονιδιακών χαρτών. Μεγάλη ποικιλία τύπων γράφων έχει εφαρμογή σε αυτά τα δεδομένα: κατευθυνόμενοι ή μη, φωλιασμένοι ή όχι, δέντρα. Ένα παράδειγμα φαίνεται στο Σχήμα 4.6 [6].

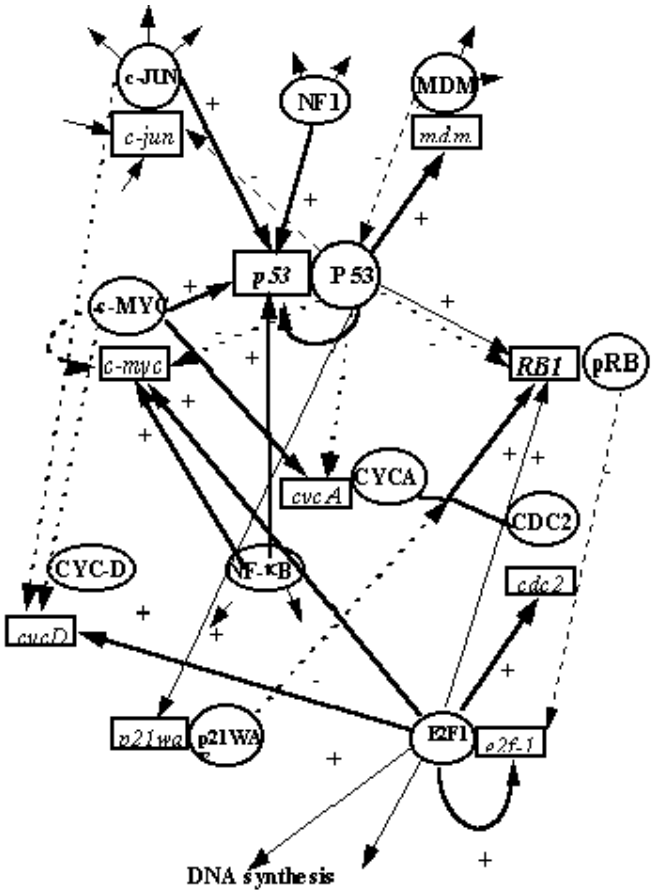
4.2.5 Ψηφιακές εικόνες

Κάποιες πειραματικές τεχνικές (π.χ. AFM - atomic force microscopy, fluorescence microscopy) προσφέρουν αποτελέσματα σε αυτή τη μορφή. Θεωρείται ότι η συμβολή των εικόνων υψηλής ανάλυσης στην κατανόηση πολύπλοκων συστημάτων, όπως είναι το νευρικό του ανθρώπου, θα είναι μεγάλη. (Σχήμα 4.7 [7])

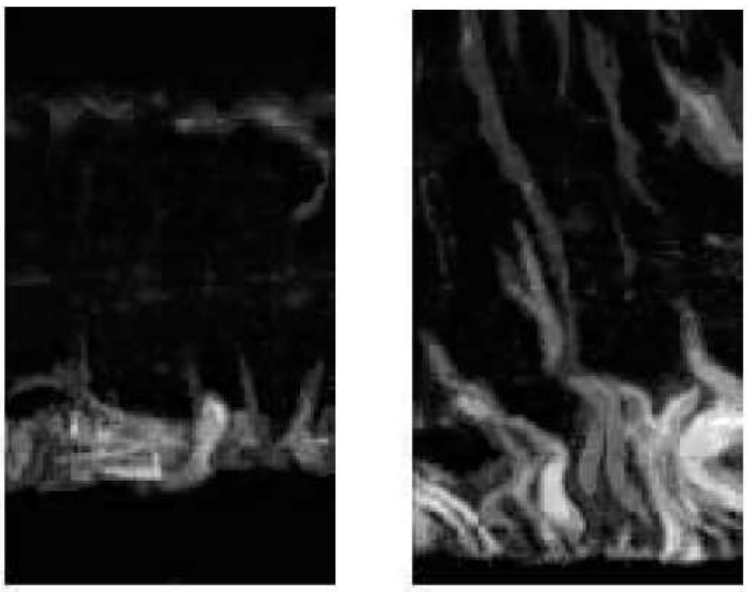
4.2.6 Κείμενο

Η χρήση κειμένου εξυπηρετεί πληροφορίες όπως είναι τα επιστημονικά άρθρα και οι σημειώσεις. Η αξία αυτών στη μελέτη των ερευνητών επισημάνθηκε στην ενότητα 4.1.10.

[6] Πηγή: http://wwwicg.bionet.nsc.ru/SRCG/TRRD/images/gene_network.gif
[7] Πηγή: [55]



Σχήμα 4.6: Παράδειγμα ενός gene network.



Σχήμα 4.7: Εικόνες πρωτεΐνης σε κανονικό και τραυματισμένο αμφιβληστροειδή χιτώνα.

4.3 Μοντέλα και πρότυπα

Ιδιαίτερα κρίσιμη για την εφαρμογή κάθε λειτουργίας που επιθυμεί ο χρήστης μιας βάσης βιοδεδομένων να κάνει, είναι η επιλογή του μοντέλου δεδομένων και των προτύπων. Αν και δεν υπάρχει ένα ενιαίο μοντέλο ή πρότυπο που να χρησιμοποιείται σε όλες τις διαθέσιμες βάσεις, στην ενότητα αυτή συζητώνται κάποια που ξεχωρίζουν. Αυτά είναι το μοντέλο δεδομένων του NCBI και ορισμένα XML πρότυπα με έμφαση στο BIOML.

4.3.1 Το μοντέλο δεδομένων του NCBI

Το National Center for Biotechnology Information [36] είναι ένας κρατικός ερευνητικός οργανισμός των Ηνωμένων Πολιτειών Αμερικής που δραστηριοποείται τα τελευταία είκοσι περίπου χρόνια σε θέματα σχετικά με τις επιστήμες υγείας και ειδικότερα τη μοριακή βιολογία. Ανάμεσα στις άλλες υπηρεσίες που παρέχονται ελεύθερα μέσω της ιστοσελίδας του είναι και η χρήση εργαλείων που βοηθούν στην αξιοποίηση των δεδομένων πολλών σχετικών βάσεων δεδομένων.

Το μοντέλο δεδομένων που χρησιμοποιείται, ώστε να είναι δυνατά τόσο ο συνδυασμός πολλών πληροφοριών από διαφορετικές βάσεις όσο και η αποδοτική επεξεργασία αυτών έχει ενδιαφέρον να μελετηθεί. Επιπλεόν, η εστίαση σε αυτό είναι εύλογη, επειδή το NCBI αποτελεί βασική και συχνά χρησιμοποιούμενη πηγή για τους ερευνητές των βιοεπιστημών. Στις βάσεις δεδομένων στις οποίες αναφέρεται ανήκουν και οι DDBJ, EMBL, GenBank, SWISS-PROT, PIR, PRF, PDB, αναλυτικότερα στοιχεία για τις οποίες δίνονται στην ενότητα 4.4.

Στην παρούσα ενότητα δε θα γίνει λόγος για τα εργαλεία του NCBI που χρησιμοποιούν αυτό το μοντέλο και πραγματοποιούν τις απαραίτητες λειτουργίες (αναζήτηση μιας εγγραφής,

προσθήκη, σύγκριση ακολουθιών, μορφή εμφάνισης αποτελέσματος κτλ) όπως είναι τα Entrez, LocusLink, Sequin, BLAST.

Γενικά χαρακτηριστικά

Το μοντέλο δεδομένων του NCBI έχει σχεδιαστεί πρωτίστως για την αποθήκευση και επεξεργασία δεδομένων σε μορφή ακολουθίας. Τα είδη με τα οποία ασχολείται από άποψη βιολογίας είναι τα νουκλεϊκά οξέα (DNA, RNA) και οι πρωτεΐνες. Ωστόσο, παρέχει πληροφορίες και για δημοσιεύσεις που σχετίζονται με αυτά, καθώς και για φυσικούς και γενετικούς χάρτες.

Ο σχεδιασμός του μοντέλου έχει γίνει, ώστε αυτό να αντέξει σε βάθος χρόνου. Έμφαση, δηλαδή, έχει δοθεί στη σταθερότητα και ευελιξία του. Για το λόγο αυτό υπάρχουν στοιχεία που αποτελούν τον κορμό του και άλλα που πιο εύκολα μπορούν να αλλάξουν. Με άλλα λόγια, έχει ληφθεί υπόψη ότι η έρευνα στο πεδίο της βιολογίας είναι πιθανό να επιφέρει σημαντικές αλλαγές στις υπάρχουσες θεωρίες και έτσι έχει γίνει προσπάθεια το μοντέλο να στηριχθεί κατά το δύνατόν μόνο στις πολύ βασικές βιολογικές παρατηρήσεις.

Επιπλέον, έχει δοθεί ιδιαίτερη προσοχή στο γεγονός ότι η χρήση του από τους σχετικούς επιστήμονες έχει στόχο την ανακάλυψη νέας γνώσης. Δεν είναι μόνο η άντληση των δεδομένων από τις βάσεις που ενδιαφέρει αλλά και ο δημιουργικός συνδυασμός τους. Αυτός επιτυγχάνεται, όταν πληροφορίες που τοποθετήθηκαν σε δύο βάσεις δεδομένων και αφορούν αντικείμενα που σχετίζονται μεταξύ τους εμφανίζονται συνδεδεμένες στον χρήστη που αναζητά τη μία ή την άλλη από αυτές. Παράλληλα, για την ανακάλυψη γνώσης έχει σημασία να μπορούν να εφαρμοστούν κατάλληλοι υπολογισμοί πάνω στα δεδομένα.

Αξίζει να σημειωθούν από τώρα δύο σημεία στα οποία πρωτοτυπεί το εν λόγω μοντέλο και τα οποία θα αναλυθούν στα επόμενα. Το πρώτο αφορά το γεγονός ότι τόσο η ακολουθία του DNA όσο και η ακολουθία της παραγόμενης από αυτό πρωτεΐνης αντιμετωπίζονται ισάξια και οι δύο ως πολίτες πρώτης κατηγορίας του μοντέλου. Στα περισσότερα άλλα μοντέλα τα χαρακτηριστικά της μιας ακολουθίας δίνονται αναφορικά με την άλλη, μετατοπίζοντας έτσι το βάρος προς την πρώτη ή τη δεύτερη ακολουθία. Το δεύτερο στοιχείο αφορά τις πληροφορίες που σχετίζονται με μια ακολουθία. Αυτές είναι δυνατό να ανταλλάσσονται και τίθενται υπό επεξεργασία ξεχωριστά από την ίδια την ακολουθία.

Τέλος, το μοντέλο του NCBI ακολουθεί το πρότυπο ASN.1 (Abstract Syntax Notation 1). Αυτό είναι ένα ISO πρότυπο χαμηλού επιπέδου, από το οποίο υψηλού επιπέδου formats μπορούν να δημιουργηθούν με διάφορα εργαλεία, ώστε να είναι τα αποτελέσματα ευανάγνωστα από τον άνθρωπο. Η ενότητα αυτή για το μοντέλο του NCBI δεν ασχολείται με αυτά τα εργαλεία, όπως έχει ήδη αναφερθεί, παρά μόνο με τα χαρακτηριστικά του.

Bioseq

Η βιολογική ακολουθία είναι ο βασικός τύπος δεδομένων του μοντέλου. Περιέχει τις πληροφορίες για την ακολουθία ενός μορίου νουκλεϊκού οξέος ή πρωτεΐνης. Έχει τουλάχιστον έναν Seq-id, ενώ μπορεί να έχει Seq-annot και Seq-descr, στοιχεία για τα οποία θα γίνει αναφορά στις επόμενες ενότητες.

Όλα τα είδη Bioseq, τα οποία περιγράφονται στη συνέχεια, ορίζουν τις ακολουθίες τους με βάση ένα κοινό γραμμικό σύστημα ακεραίων. Έτσι, το μήκος τους χαρακτηρίζεται πάλι από έναν ακέραιο.

Raw. Ένα string που αποτελείται από αζωτούχες βάσεις ή αμινοξέα. Το μήκος του είναι γνωστό και ίσο με τον αριθμό αυτών των δομικών μονάδων. Αυτός ο τύπος είναι ο πιο απλός αλλά και πιο αναμενόμενος, όταν έχει κανείς στο νου του τη μορφή μιας ακολουθίας.

Segmented. Περιέχει μόνο τους Seq-id των Raw Bioseq από τις οποίες αποτελείται. Ο τύπος αυτός είναι χρήσιμος για να περιγράψει μεγάλες ακολουθίες ενός γονιδιώματος στις οποίες δεν είναι γνωστά όλα τα μέρη. Χαρακτηριστικό τέτοιο παράδειγμα είναι ένα γονίδιο στο οποίο είναι γνωστά τα εξώνια αλλά όχι τα εσώνια. (βλ. Κεφάλαιο 2)

Virtual. Χρησιμεύει στην περιγραφή μορίου του οποίου δεν είναι γνωστή η ακολουθία των δομικών μονάδων από τις οποίες αποτελείται αλλά μόνο ο τύπος του (π.χ. πρωτεΐνη, DNA). Η θέση, το μήκος και η τοπολογία του (π.χ. γραμμική, κυκλική κτλ) μπορεί να είναι επίσης γνωστά. Τα εσώνια ενός γονιδίου είναι το πιο συχνό παράδειγμα τέτοιας περίπτωσης.

Delta. Ο τύπος αυτός μοιάζει με τον Segmented με τη διαφορά ότι ολόκληρη η ακολουθία, άλλα τμήματα της οποίας είναι γνωστά και άλλα άγνωστα, περιγράφεται από το ίδιο Seq-id. Χρησιμοποιείται για τις λεγόμενες ακολουθίες HTGS (High Throughput Genome Sequences), των οποίων η παραγωγή από τα ερευνητικά κέντρα βρίσκεται σε εξέλιξη, δεν έχει ολοκληρωθεί.

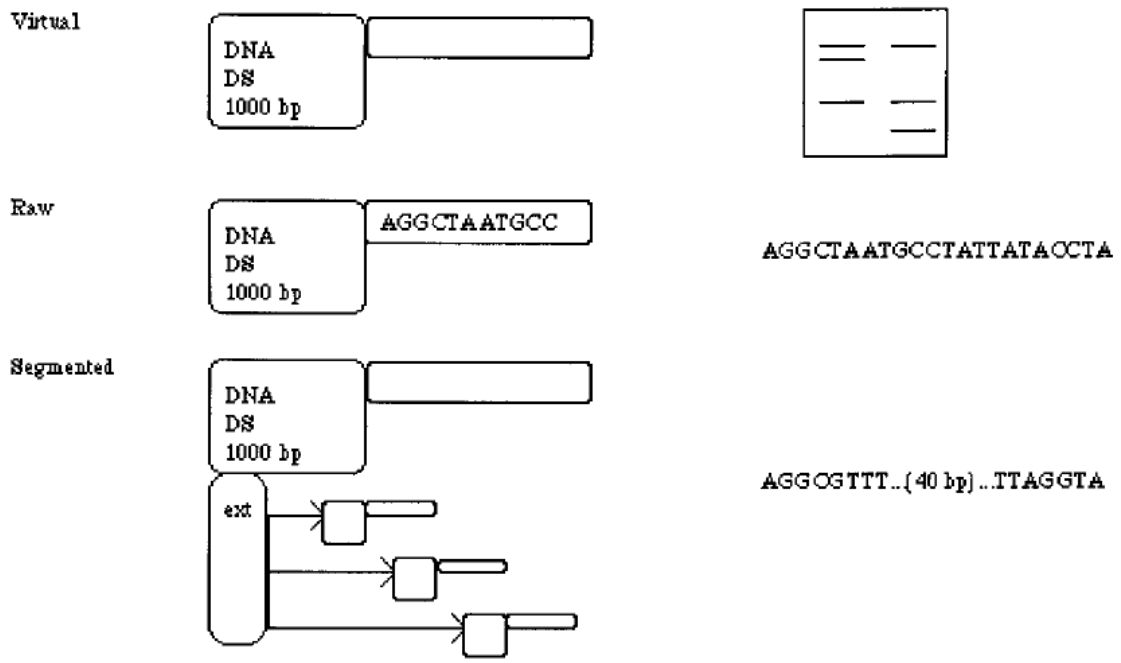
Map. Όπως δηλώνει και το όνομά του, ανταποκρίνεται στις ανάγκες αποθήκευσης και επεξεργασίας των φυσικών και γενετικών χαρτών. Ο τύπος αυτός έχει ομοιότητα με τον Virtual, όμως η πληροφορία που ενδιαφέρει σε αυτόν είναι τα επιμέρους χαρακτηριστικά των γονιδίων.

Τα παραπάνω είδη βιολογικών ακολουθιών φαίνονται στα Σχήμα 4.8 και Σχήμα 4.9. [8]

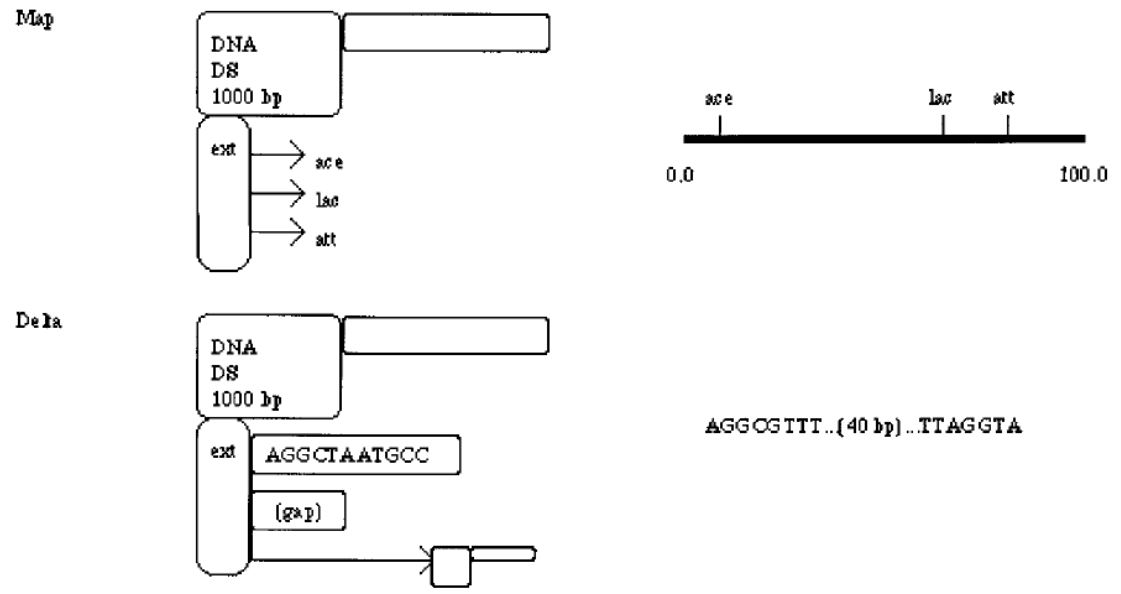
Bioseq-set

Αρκετές φορές είναι χρήσιμο να αντιμετωπίζονται κάποιες ακολουθίες ως ομάδα και όχι ξεχωριστά. Για το λόγο αυτό εξυπηρετεί ο τύπος Bioseq-set που είναι ένα σύνολο από Bioseq. Επιπλέον, κάποιες πληροφορίες (αυτές των Seq-descr) αποθηκεύονται μόνο μία φορά και αναφέρονται σε όλες τις ακολουθίες του σετ. Τα κυριότερα είδη των ομάδων που υπάρχουν φανερώνουν τις περιπτώσεις που είναι όντως βολικό να χρησιμοποιηθεί αυτός ο τύπος και περιγράφονται συνοπτικά στη συνέχεια.

[8] Πηγή: [4]



Σχήμα 4.8: Virtual, Raw, Segmented Bioseq.



Σχήμα 4.9: Delta, Map Bioseq.

Nuc-prot. Η πιο συνηθισμένη περίπτωση στην οποία είναι εύλογη η χρήση Bioseq-set είναι μια ακολουθία νουκλεοτιδίων που ανήκει σε γονίδιο και μία ή περισσότερες ακολουθίες αμινοξέων που παράγονται από αυτήν. Το πιο αξιοσημείωτο είναι το γεγονός ότι αντιμετωπίζονται όλες οι ακολουθίες ισάξια, χωρίς να έχει κάποια τον κύριο ρόλο και οι υπόλοιπες να είναι προσβάσιμες από αυτήν μέσω κάποιου συνδέσμου.

Population and Phylogenetic Studies. Για τη μελέτη ενός πληθυσμού συγκρίνονται οι ακολουθίες διαφόρων ατόμων του ίδιου είδους για το ίδιο γονίδιο, ενώ στη φυλογενετική μελέτη η σύγκριση γίνεται ανάμεσα σε άτομα διαφορετικών ειδών. Είναι εμφανές ότι και στις δύο περιπτώσεις έχει νόημα να κρατώνται αυτές οι ακολουθίες ως ομάδες και όχι κάθε μια ξεχωριστά.

Parts. Ο τύπος αυτός χρησιμεύει στην αποθήκευση των ακολουθιών που αποτελούν μία Segmented Bioseq, σύμφωνα με όσα αναφέρθηκαν στην προηγούμενη ενότητα για τα είδη των Bioseq.

Seg. Από όσα έχουν μέχρι στιγμής αναφερθεί είναι εύλογο να αποθηκεύονται μία Segmented Bioseq και ένα Parts Bioseq-set σε μία ομάδα. Αυτή είναι τύπου Seg Bioseq-set.

Equiv. Αυτός ο τύπος χρησιμοποιείται για να αποθηκεύονται μαζί ισοδύναμες Bioseqs. Για παράδειγμα, διάφοροι γενετικοί χάρτες για χρωμοσώματα ή φυσικοί χάρτες για το ίδιο χρωμόσωμα, που έχουν, όμως, δημιουργηθεί με διαφορετική μέθοδο.

Seq-id

Ο τρόπος με τον οποίο μια εγγραφή ξεχωρίζει από μία άλλη είναι μέσω των αναγνωριστικών. Αυτά είναι περισσότερα από ένα. Ο λόγος για αυτήν την ποικιλία είναι το γεγονός ότι οι ακολουθίες με τις οποίες ασχολείται το μοντέλο του NCBI προέρχονται από πολλές διαφορετικές πηγές-βάσεις, καθεμιά από τις οποίες είτε χρησιμοποιεί τα δικά της αναγνωριστικά είτε αντιστοιχίζει διαφορετική εγγραφή σε ένα αναγνωριστικό από αυτήν που αντιστοιχίζει μια άλλη βάση στο ίδιο. Οι κυριότεροι τύποι Seq-id περιγράφονται στα επόμενα.

Locus. Το αναγνωριστικό αυτό υπάρχει στις παλαιότερες από τις βάσεις και για λόγους συμβατότητας διατηρείται. Η ονομασία του προέρχεται από τον όρο locus, που στη γενετική αναφέρεται στη θέση ενός γονιδίου σε ένα χρωμόσωμα. Πρόκειται για ένα όνομα που είχε στόχο να φανερώνει το βιολογικό ρόλο της ακολουθίας στην οποία αντιστοιχεί και να χρησιμοποιηθεί ως κλειδί, αλλά παρουσιάστηκαν δύο σημαντικά προβλήματα. Αφενός λόγω της γρήγορης αύξησης του αριθμού των ακολουθιών των βάσεων κατέστη εξαιρετικά δύσκολη η απόδοση δηλωτικών ονομάτων, αφετέρου διαπιστώθηκε ότι είχαν δοθεί λάθος ονόματα, που έπρεπε να διορθωθούν και να επιφέρουν επιπλέον δυσκολίες, όταν οι ερευνητές ανακάλυπταν ότι δεν ήταν τελικά η βιολογική σημασία της ακολουθίας εκείνη που αρχικά θεωρούσαν.

Accession. Ο αριθμός αυτός είναι το κλειδί μιας εγγραφής σε σημαντικό τμήμα από τις βάσεις που χρησιμοποιεί το NCBI (DDBJ/EMBL/GenBank). Δεν αντικατοπτρίζει το βιολογικό ρόλο της ακολουθίας στην οποία αντιστοιχεί, αλλά –σύμφωνα με την τελευταία τακτική– αποτελείται από δύο κεφαλαία λατινικά γράμματα και έξι ψηφία. Ακόμη και αυτός ο τρόπος ορισμού κλειδιού, ωστόσο, εμφανίζει προβλήματα και δεν είναι απόλυτα σταθερός. Χαρακτηριστικό παράδειγμα είναι η κατάσταση στην οποία μια εγγραφή ανανεώνεται με την προσθήκη αζωτούχων βάσεων σε αυτήν και οι χρήστες θεωρούν ότι πρόκειται για άλλη εγγραφή από αυτήν που αρχικά ήξεραν και αναζητούσαν. Για το λόγο αυτό κάποιες φορές εισήχθησαν και δευτερεύοντες accession numbers, που όμως δεν έφεραν πολύ καλύτερα αποτελέσματα.

Accession.Version. Για να αντιμετωπιστούν τα προβλήματα από το αναγνωριστικό accession number, πρόσφατα άρχισε η χρήση αυτού του τύπου αναγνωριστικού. Σε αυτόν δίνεται επιπλέον και η έκδοση στην οποία βρίσκεται η εγγραφή.

gi. Ο αριθμός αυτός, που αντιστοιχεί στο GenInfo Identifier, λειτουργεί σαν κλειδί για το μοντέλο του NCBI. Παρότι οι χρήστες στις αναζητήσεις τους συνήθως χρησιμοποιούν τους accession numbers, εσωτερικά στο σύστημα του NCBI όλοι οι υπολογισμοί και οι διεργασίες γίνονται με βάση τους gi numbers. Αυτός ο τύπος αναγνωριστικού καθιερώθηκε ώστε η κάθε ακολουθία να έχει ένα μοναδικό αριθμό ανεξάρτητο από την πηγή προέλευσής της, κάτι το οποίο δεν ισχύει για τους accession numbers.

Reference. Ο τύπος αυτός αναγνωριστικού έχει τη μορφή του accession.version, αλλά προσδίδει φυσικό νόημα στην ακολουθία. Αυτό επιτυγχάνεται προσθέτοντας μπροστά από τον αριθμό δύο γράμματα του λατινικού αλφαβήτου. Τα πιο συχνά χρησιμοποιούμενα είναι NC_ για τα χρωμοσώματα, NM_ για τα mRNAs και NP_ για τις πρωτεΐνες.

Seq-descr

Ένας περιγραφητής μπορεί να χαρακτηρίζει μία bioseq ή ένα bioseq-set (βλ. και παράγραφο για Seq-id). Οι πληροφορίες που δίνει για αυτά αφορούν στοιχεία σχετικά με το βιολογικό περιεχόμενο και την προέλευση της ακολουθίας ή της ομάδας των ακολουθιών. Για παράδειγμα, σε ποιον οργανισμό ανήκει, με ποια τεχνική αποκτήθηκε. Η χρησιμότητά του τονίζεται ακόμη περισσότερο από το γεγονός ότι ιδιαίτερα στην περίπτωση των bioseq-sets αποφεύγεται η αποθήκευση πλεονάζουσας πληροφορίας. Ένας μόνο περιγραφητής χρησιμοποιείται για όλη την ομάδα ακολουθιών, καθώς σε όλα τα μέλη της αντιστοιχούν λογικά οι πληροφορίες του. Δύο είναι τα είδη των seq-descr, τα οποία και σχολιάζονται στα αμέσως επόμενα.

BioSource. Περιέχει στοιχεία σχετικά με την πηγή προέλευσης, όπως από ποιον οργανισμό προέρχεται η ακολουθία, το ιστορικό της στο NCBI, τη θέση της μέσα στο κύτταρο (π.χ. αν είναι DNA που ανήκει στον πυρήνα ή στα μιτοχόνδρια).

MolInfo. Αφορά πληροφορίες για τον τύπο του μορίου που αποθηκεύεται, τη μέθοδο με την οποία αποκτήθηκε η ακολουθία των δομικών του μονάδων και το βαθμό στον οποίο αυτή η ακολουθία είναι πλήρης.

Seq-annot

Σε αντίθεση με τους seq-descrs, πολλές seq-annots μπορούν να συνοδεύουν μια bioseq ή ένα bioseq-set. Πρόκειται για σημειώσεις σχετικές με την ακολουθία ή την ομάδα ακολουθιών. Το περιεχόμενο και η μορφή τους γίνονται φανερά από τα είδη των seq-annots. Σημαντικό είναι ότι αυτές οι σημειώσεις μπορούν να διακινηθούν (π.χ. μεταξύ επιστημόνων) ανεξάρτητα από τις ίδιες τις ακολουθίες στις οποίες αναφέρονται.

Seq-feat. Είναι ένα σύνολο ιδιοτήτων καθεμιά από τις οποίες χαρακτηρίζει συγκεκριμένο τμήμα ή σημείο μιας ακολουθίας. Οι ιδιότητες αυτές έχουν προκαθορισμένα δηλωτικά ονόματα και είναι αρκετές. Παραδείγματα αυτών είναι η ιδιότητα CDS (Coding Regions), η Gene και η Bond. Συγκεκριμένα η CDS είναι ο τρόπος με τον οποίο μια περιοχή του DNA μεταφράζεται σε πρωτεΐνη περιλαμβάνοντας πιθανές διαφοροποιήσεις από την τυπική διαδικασία της μετάφρασης. Από την άλλη, η Gene φανερώνει τμήμα της ακολουθίας που αντιστοιχεί σε γονίδιο και ταυτόχρονα είναι πιθανό να προσφέρει αναφορές σε γενετικούς χάρτες για περισσότερες πληροφορίες. Τέλος, η Bond χαρακτηρίζει το δεσμό ανάμεσα σε δύο μόρια, λόγου χάρη πρωτεΐνης.

Seq-align. Χρησιμοποιείται για να αποθηκεύεται ένα alignment, δηλαδή μία ευθυγράμμιση μεταξύ δύο ακολουθιών. (Περισσότερα για το alignment στο Κεφάλαιο 5.) Κρατώνται σε μορφή λίστας συντεταγμένων τα ζευγάρια που ευθυγραμμίζονται από τις δύο ή περισσότερες ακολουθίες όπως και το μήκος τους. Όταν παρεμβάλλεται κενό, τότε μπαίνει η τιμή -1. Στην περίπτωση που το alignment δεν είναι συνεχές αλλά διακοπτόμενο, τότε κρατώνται περισσότερες λίστες.

Seq-graph. Αυτός ο τύπος σημειώσεων περιλαμβάνει τους γράφους. Εξυπηρετεί την αποθήκευση στοιχείων κυρίως για φυσικά μεγέθη, όπως η διαφορά δυναμικού κατά μήκος τμήματος ακολουθίας ή η υδροφοβικότητα.

Δημοσιεύσεις

Όσο πλήρη και αν είναι τα στοιχεία που παρέχονται από τις βάσεις δεδομένων είναι πολλές φορές αναγκαίο να ανατρέξει κανείς σε σχετικές δημοσιεύσεις. Ο ρόλος των δημοσιεύσεων είναι διττός. Αφενός μπορεί να συνδέουν δύο εγγραφές από διαφορετικές ή την ίδια βάση δεδομένων, αφετέρου μπορούν να λειτουργήσουν ως σημεία έναρξης μιας αναζήτησης για περισσότερες πληροφορίες στο σχετικό θέμα.

Είναι εύλογο, επομένως, να υπάρχει πρόσβαση σε όλα τα στοιχεία μιας δημοσίευσης. Αυτά είναι το κύριο άρθρο, τα στοιχεία που σχετίζονται με τους συγγραφείς αλλά και οι πατέντες. Επιπλέον, δεν πρέπει να αμεληθούν οι ηλεκτρονικές καταχωρήσεις δεδομένων, που προηγούνται των δημοσιεύσεων ή αντικαθιστούν πλήρως αυτές.

Το μοντέλο του NCBI προσφέρει δύο τρόπους για την καταχώρηση των πληροφοριών των σχετικών με τους συγγραφείς. Είτε μια δομή με ξεχωριστά πεδία για κάθε χαρακτηριστικό (π.χ. όνομα, επίθετο κτλ) είτε ένα μόνο string, στο οποίο να μπαίνουν όλα. Αν και η

πρώτη μορφή είναι πιο καλά δομημένη, η δεύτερη χρησιμεύει, επειδή στις διάφορες βάσεις δεν ακολουθείται το ίδιο format και είναι δύσκολη η αντιστοίχιση του κάθε πεδίου.

Για τα άρθρα το μοντέλο διαθέτει ξεχωριστό format για τα βασικά είδη πηγών (βιβλία, περιοδικά, κτλ), ενώ διαφοροποιείται ακόμη περισσότερο στην περίπτωση πατέντας. Αυτό οφείλεται στο γεγονός ότι η τελευταία αποτελεί περισσότερο νομικό παρά ερευνητικό κείμενο.

Τέλος, ξεχωριστή μέριμνα λαμβάνεται για τις ηλεκτρονικές καταχωρήσεις δεδομένων σε μια βάση που πιθανώς συνοδεύουν μια δημοσίευση. Είναι φανερό ότι ενδιαφέρουν ειδικά τα ονόματα των ερευνητών, τα οποία μάλιστα θα αντιστοιχούν σε εκείνα των συγγραφέων της ανάλογης δημοσίευσης.

Παράδειγμα

Παρατίθεται μια εγγραφή της βάσης GenBank για να γίνουν περισσότερο κατανοητά όσα θεωρητικά αναφέρθηκαν. Το συγκεκριμένο παράδειγμα αφορά το mRNA του γονιδίου SOD1 του ανθρώπου.

```
LOCUS       NM_000454      981 bp    mRNA   linear PRI 12-JUN-2006
DEFINITION  Homo sapiens superoxide dismutase 1, soluble
            (amyotrophic lateral sclerosis 1 (adult)) (SOD1), mRNA.
ACCESSION   NM_000454
VERSION     NM_000454.4  GI:48762945
KEYWORDS .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 981)
  AUTHORS   Muller,F.L., Song,W., Liu,Y., Chaudhuri,A., Pieke-Dahl,S.,
            Strong,R., Huang,T.T., Epstein,C.J., Roberts,L.J. II, Csete,M.,
            Faulkner,J.A. and Van Remmen,H.
  TITLE     Absence of CuZn superoxide dismutase leads to elevated oxidative
            stress and acceleration of age-dependent skeletal muscle atrophy
  JOURNAL   Free Radic. Biol. Med. 40 (11), 1993-2004 (2006)
   PUBMED   16716900
  REMARK    GeneRIF: Sod1 knockout mice have a dramatic accleration of age
            related muscle mass loss and a 20% reduction in body weight.
REFERENCE   2  (bases 1 to 981)
  AUTHORS   Stewart,H.G., Mackenzie,I.R., Eisen,A., Brannstrom,T.,
            Marklund,S.L. and Andersen,P.M.
```

```
  TITLE     Clinicopathological phenotype of ALS with a novel G72C SOD1 gene
            mutation mimicking a myopathy
  JOURNAL   Muscle Nerve 33 (5), 701-706 (2006)
   PUBMED   16435343
  REMARK    GeneRIF: the phenotypic spectrum of ALS associated with SOD1(G72C)
            mutations to include presenting features that mimic a myopathy.

.............................................................................

REFERENCE   155 (bases 85 to 644)
  AUTHORS   Sherman,L., Dafni,N., Lieman-Hurwitz,J. and Groner,Y.
  TITLE     Nucleotide sequence and expression of human chromosome 21-encoded
            superoxide dismutase mRNA
  JOURNAL   Proc. Natl. Acad. Sci. U.S.A. 80 (18), 5465-5469 (1983)
   PUBMED   6577438
REFERENCE   156 (bases 1 to 981)
  AUTHORS   Philip,T., Fraisse,J., Sinet,P.M., Lauras,B., Robert,J.M. and
            Freycon,F.
  TITLE     Confirmation of the assignment of the human SODS gene to chromosome
            21q22
  JOURNAL   Cytogenet. Cell Genet. 22 (1-6), 521-523 (1978)
   PUBMED   752535
COMMENT     REVIEWED REFSEQ: This record has been curated by NCBI
            staff. The reference sequence was derived from W17182.1,
            AV756797.1, BC001034.1, X02317.1 and CA448539.1.
            On Jun 16, 2004 this sequence version replaced gi:40255241.

            Summary: The protein encoded by this gene binds copper and zinc
            ions and is one of two isozymes responsible for destroying free
            superoxide radicals in the body. The encoded isozyme is a soluble
            cytoplasmic protein, acting as a homodimer to convert
            naturally-occuring but harmful superoxide radicals to molecular
            oxygen and hydrogen peroxide. The other isozyme is a mitochondrial
            protein. Mutations in this gene have been implicated as causes of
            familial amyotrophic lateral sclerosis. Rare transcript variants
            have been reported for this gene.
            COMPLETENESS: complete on the 3' end.
FEATURES             Location/Qualifiers
     source          1..981
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
```

```
                /db_xref="taxon:9606"
                /chromosome="21"
                /map="21q22.11"
gene            1..981
                /gene="SOD1"
                /note="synonyms: ALS, SOD, ALS1, IPOA, homodimer"
                /db_xref="GeneID:6647"
                /db_xref="HGNC:11179"
                /db_xref="MIM:147450"
STS             40..222
                /gene="SOD1"
                /standard_name="GDB:374780"
                /db_xref="UniSTS:156962"
CDS             149..613
                /gene="SOD1"
                /EC_number="1.15.1.1"
                /go_component="cytoplasm"
                /go_function="antioxidant activity; copper ion binding;
                copper, zinc superoxide dismutase activity [pmid 6316150];
                metal ion binding; oxidoreductase activity; zinc ion
                binding"
                /go_process="nervous system development [pmid 8351519];
                response to oxidative stress; superoxide metabolism"
                /note="Cu/Zn superoxide dismutase; indophenoloxidase A;
                SOD, soluble; superoxide dismutase, cystolic; superoxide
                dismutase (aa 120-154); Cu /Zn superoxide dismutase"
                /codon_start=1
                /product="superoxide dismutase 1, soluble"
                /protein_id="NP_000445.1"
                /db_xref="GI:4507149"
                /db_xref="GeneID:6647"
                /db_xref="HGNC:11179"
                /db_xref="MIM:147450"
                /translation="MATKAVCVLKGDGPVQGIINFEQKESNGPVKVWGSIKGLTEGLH
                GFHVHEFGDNTAGCTSAGPHFNPLSRKHGGPKDEERHVGDLGNVTADKDGVADVSIED
                SVISLSGDHCIIGRTLVVHEKADDLGKGGNEESTKTGNAGSRLACGVIGIAQ"
STS             158..403
                /gene="SOD1"
                /standard_name="Sod1"
                /db_xref="UniSTS:144477"
STS             327..427
```

```
                     /gene="SOD1"
                     /standard_name="MARC_26673-26674:1036188363:1"
                     /db_xref="UniSTS:269054"
     STS             612..864
                     /gene="SOD1"
                     /standard_name="SHGC-87564"
                     /db_xref="UniSTS:30768"
     STS             618..918
                     /gene="SOD1"
                     /standard_name="GDB:185171"
                     /db_xref="UniSTS:155426"
     polyA_signal    691..696
                     /gene="SOD1"
     STS             695..886
                     /gene="SOD1"
                     /standard_name="G20764"
                     /db_xref="UniSTS:41750"
     polyA_site      712
                     /gene="SOD1"
                     /experiment="experimental evidence, no additional details
                     recorded"
     STS             730..882
                     /gene="SOD1"
                     /standard_name="G34694"
                     /db_xref="UniSTS:78956"
     polyA_signal    940..945
                     /gene="SOD1"
     polyA_site      961
                     /gene="SOD1"
                     /experiment="experimental evidence, no additional details
                     recorded"
     polyA_site      966
                     /gene="SOD1"
ORIGIN
        1 gtttggggcc agagtgggcg aggcgcggag gtctggccta taaagtagtc gcggagacgg
       61 ggtgctggtt tgcgtcgtag tctcctgcag cgtctggggt ttccgttgca gtcctcggaa
      121 ccaggacctc ggcgtggcct agcgagttat ggcgacgaag gccgtgtgcg tgctgaaggg
      181 cgacggccca gtgcagggca tcatcaattt cgagcagaag gaaagtaatg gaccagtgaa
      241 ggtgtgggga agcattaaag gactgactga aggcctgcat ggattccatg ttcatgagtt
      301 tggagataat acagcaggct gtaccagtgc aggtcctcac tttaatcctc tatccagaaa
      361 acacggtggg ccaaaggatg aagagaggca tgttggagac ttgggcaatg tgactgctga
```

```
     421 caaagatggt gtggccgatg tgtctattga agattctgtg atctcactct caggagacca
     481 ttgcatcatt ggccgcacac tggtggtcca tgaaaaagca gatgacttgg gcaaaggtgg
     541 aaatgaagaa agtacaaaga caggaaacgc tggaagtcgt ttggcttgtg gtgtaattgg
     601 gatcgcccaa taaacattcc cttggatgta gtctgaggcc ccttaactca tctgttatcc
     661 tgctagctgt agaaatgtat cctgataaac attaaacact gtaatcttaa aagtgtaatt
     721 gtgtgacttt ttcagagttg ctttaaagta cctgtagtga gaaactgatt tatgatcact
     781 tggaagattt gtatagtttt ataaaactca gttaaaatgt ctgtttcaat gacctgtatt
     841 ttgccagact taaatcacag atgggtatta aacttgtcag aatttctttg tcattcaagc
     901 ctgtgaataa aaaccctgta tggcacttat tatgaggcta ttaaaagaat ccaaattcaa
     961 actaaaaaaa aaaaaaaaaa a
//
```

4.3.2 XML πρότυπα

Η αυστηρά δομημένη και ιεραρχική μορφή των XML αρχείων έχει επηρεάσει και την κοινότητα που ασχολείται με τη μοντελοποίηση των βιοδεδομένων. Στην ενότητα αυτή περιγράφεται με αρκετές λεπτομέρειες ένα από τα αρκετά XML πρότυπα, το BIOML, το οποίο δίνει μια αντιπροσωπευτική γεύση. Αναφέρονται, επίσης, στοιχεία και για άλλα σημαντικά πρότυπα βασισμένα στην XML. Μια εξαιρετική πηγή αναφοράς αυτών, στην οποία μπορούν να αναζητηθούν περισσότερες πληροφορίες είναι η ιστοσελίδα του Paul Gordon (Institute for Marine Biosciences, NRCC, Halifax, NS, Canada) [40].

BIOML

Η BIOML (BIOpolymer Markup Language) [26] έχει αναπτυχθεί από τις Proteometrics, LLC & Proteometrics Canada, Ltd. Στόχος της είναι να περιγράψει τις πληροφορίες που παρέχονται από πειράματα και αφορούν κυρίως γονίδια και πρωτεΐνες.

Έχει σχεδιαστεί με τέτοιο τρόπο, ώστε να εκπληρώνει έξι στόχους. Εκτός από τον προφανή, που είναι η πιστή αναπαράσταση των γονιδίων και πρωτεϊνών, χρειάζεται επιπλέον να είναι επεκτάσιμη, ώστε να συμφωνεί με το πρότυπο της XML. Παράλληλα, είναι αναγκαία η εύκολη κατανόησή της από τους ανθρώπους συνδέοντας λογικά αλλά και με σαφήνεια τα στοιχεία (elements). Μάλιστα, οι πληροφορίες είναι φωλιασμένες στη δενδρική μορφή της ιεραρχίας των elements. Τέλος, πρέπει να υποστηρίζει δεδομένα που δεν είναι ASCII, καθώς και τη μετατροπή άλλων σχετικών τύπων αρχείου σε BIOML.

Στους Πίνακες 4.1-4.14 περιγράφονται τα elements και τα attributes της γλώσσας, ώστε να αποκτηθεί μια πρώτη επαφή με αυτήν.

Πίνακας 4.1: Elements που αντιστοιχούν σε υψηλού επιπέδου βιολογικές οντότητες.

ELEMENTS	ATTRIBUTES	ΠΕΡΙΓΡΑΦΗ ΛΕΙΤΟΥΡΓΙΑΣ
chromosome	number	Περικλείει ένα χρωμόσωμα
sts_domain	start end	Περικλείει μια περιοχή ενός χρωμοσώματος που περιορίζεται από δύο Sequence Tagged Sites (STSs)
locus	start end	Περικλείει την περιγραφή για μια θέση
clone	-	Περικλείει την περιγραφή για έναν κλώνο
plasmid	-	Περικλείει την περιγραφή για ένα πλασμίδιο

Πίνακας 4.2: Σχετικά με το DNA και το RNA.

ELEMENTS	ATTRIBUTES	ΠΕΡΙΓΡΑΦΗ ΛΕΙΤΟΥΡΓΙΑΣ
dna/rna	start end	Περικλείει ένα νουκλεϊκό οξύ (DNA/RNA)
promotor	start end	Περικλείει μια περιοχή έναρξης της μεταγραφής
gene	comp	Περικλείει ένα γονίδιο
exon	start end type	Περικλείει ένα εξώνιο
intron	start end	Περικλείει ένα εσώνιο
ddomain/rdomain	start end type	Περικλείει μια περιοχή (DNA/RNA)
da/ra	type	Ένα νουκλεοτίδιο του (DNA/RNA)
dmod/rmod	atbr>type occ	Μια μετάλλαξη που αφορά το (DNA/RNA)
dvariant/rvariant	at occ type	Μια αλλαγή του (DNA/RNA) σε ένα συγκεκριμένο σημείο
dstart/rstart	at	Ένα κωδικόνιο αρχής
dstop/rstop	at	Ένα κωδικόνιο λήξης

Πίνακας 4.3: Για τις πρωτεΐνες.

ELEMENTS	ATTRIBUTES	ΠΕΡΙΓΡΑΦΗ ΛΕΙΤΟΥΡΓΙΑΣ
protein	comp	Περικλείει μια πρωτεΐνη
subunit	comp	Περικλείει μια αλυσίδα ενός πρωτεϊνικού μορίου
homolog	-	Ένας άλλος οργανισμός στον οποίο υπάρχει η ίδια πρωτεΐνη

Πίνακας 4.4: Για τα πεπτίδια.

ELEMENTS	ATTRIBUTES	ΠΕΡΙΓΡΑΦΗ ΛΕΙΤΟΥΡΓΙΑΣ
peptide	start end	Περικλείει ένα πεπτίδιο
domain	start end id type	Χαρακτηρίζει ένα τμήμα ενός πεπτιδίου

Πίνακας 4.5: Για τα αμινοξέα.

ELEMENTS	ATTRIBUTES	ΠΕΡΙΓΡΑΦΗ ΛΕΙΤΟΥΡΓΙΑΣ
aa	at type to	Ένα αμινοξύ (Το attribute "to" έχει νόημα μόνο για type="C"
amod	at type occ	Μια μετάλλαξη ενός αμινοξέος
alink	at type to occ	Μια σύνδεση μεταξύ δύο αμινοξέων
avariant	at type occ	Μια αλλαγή αμινοξέος σε κάποιο σημείο

Πίνακας 4.6: Γενικού σκοπού.

ELEMENTS	ATTRIBUTES	ΠΕΡΙΓΡΑΦΗ ΛΕΙΤΟΥΡΓΙΑΣ
name	-	Το όνομα του στοιχείου μέσα στο οποίο περικλείεται
alt_name	order	Εναλλακτικό όνομα για το στοιχείο στο οποίο περικλείεται
note	id order	Σημείωση που περιγράφει το στοιχείο στο οποίο περικλείεται
comment	-	Σημείωση που περιγράφει τον BIOML κώδικα. Θα πρέπει να αγνοείται από τον εκάστοτε φυλλομετρητή
copyright	-	Τα πνευματικά δικαιώματα για ένα BIOML αρχείο

Πίνακας 4.7: Σχετικά με πληροφορίες για τους οργανισμούς.

ELEMENTS	ATTRIBUTES	ΠΕΡΙΓΡΑΦΗ ΛΕΙΤΟΥΡΓΙΑΣ
organism	id type	Περικλείει τα χαρακτηριστικά ενός οργανισμού
species	id	Περικλείει την τάξη και το είδος του οργανισμού
common_name	-	Το καθιερωμένο όνομα του οργανισμού
alt_common_name	-	Εναλλακτική ονομασία για τον οργανισμό
taxon	id type	Τα χαρακτηριστικά μιας σχετικής ιεραρχίας οργανισμών

Πίνακας 4.8: Για την τοποθεσία.

ELEMENTS	ATTRIBUTES	ΠΕΡΙΓΡΑΦΗ ΛΕΙΤΟΥΡΓΙΑΣ
tissue	id type	Περικλείει τα χαρακτηριστικά ενός τύπου ιστού
cell	id type	Περικλείει τα χαρακτηριστικά ενός τύπου κυττάρου
organelle	id type	Περικλείει τα χαρακτηριστικά ενός οργανιδίου
particle	id type	Περικλείει τα χαρακτηριστικά ενός τύπου μορίου

Πίνακας 4.9: Για τη βιβλιογραφία.

ELEMENTS	ATTRIBUTES	ΠΕΡΙΓΡΑΦΗ ΛΕΙΤΟΥΡΓΙΑΣ
reference	id	Περικλείει μια αναφορά σε βιβλιογραφία
author	-	Το όνομα ενός από τους συγγραφείς π.χ. 'Beavis RC'
title	-	Ο τίτλος της αναφοράς
journal	-	Το όνομα του περιοδικού στο οποίο δημοσιεύθηκε το άρθρο
book_title	-	Τίτλος του βιβλίου που περιέχει το άρθρο
editor	-	Ο εκδότης του αντίστοιχου βιβλίου
volume	-	Τεύχος περιοδικού ή τόμος βιβλίου
pages	-	Ο αριθμός σελίδων

Πίνακας 4.10: Για αναφορές σε βάσεις δεδομένων.

ELEMENTS	ATTRIBUTES	ΠΕΡΙΓΡΑΦΗ ΛΕΙΤΟΥΡΓΙΑΣ
db_entry	id name entry format query	Αναφορά στην εγγραφή μιας βάσης δεδομένων

Πίνακας 4.11: Για πόρους.

ELEMENTS	ATTRIBUTES	ΠΕΡΙΓΡΑΦΗ ΛΕΙΤΟΥΡΓΙΑΣ
file	format URL	Ένα element που δείχνει σε αρχείο
query	format query query_string	Ένα element που δίνει το ερώτημα σε έναν εξυπηρετητή

Πίνακας 4.12: Για δυαδικά δεδομένα.

ELEMENTS	ATTRIBUTES	ΠΕΡΙΓΡΑΦΗ ΛΕΙΤΟΥΡΓΙΑΣ
binary	format length	Ένα element που περικλείει πληροφορίες σε δυαδική μορφή
data	format length	Ένα element που περικλείει πληροφορίες σε κάποια τυποποιημένη μορφή

Πίνακας 4.13: Για φόρμες.

ELEMENTS	ATTRIBUTES	ΠΕΡΙΓΡΑΦΗ ΛΕΙΤΟΥΡΓΙΑΣ
form	type action	Ένα element που περικλείει μία φόρμα
input	type name value width	Ένα element που επιτρέπει την εισαγωγή πληροφοριών από το χρήστη
text	-	Ένα element που περικλείει κείμενο που θα εισαχθεί στη φόρμα

Πίνακας 4.14: Καθολικά attributes.

ELEMENTS	ATTRIBUTES	ΠΕΡΙΓΡΑΦΗ ΛΕΙΤΟΥΡΓΙΑΣ
ALL	label	Παρέχει στον φυλλομετρητή ένα αναγνωριστικό κειμένου
ALL	state	Παρέχει στον φυλλομετρητή πληροφορίες για την εμφάνιση ενός element
ALL	id	Παρέχει στον φυλλομετρητή ένα αναγνωριστικό νούμερο

Ένα παράδειγμα στο οποίο χρησιμοποιούνται αρκετά από τα elements και τα attributes που έχουν αναφερθεί στους προηγούμενους πίνακες δίνεται στη συνέχεια. Το αρχείο αυτό[9] περιγράφει το γονίδιο που είναι υπεύθυνο για την παραγωγή της ινσουλίνης καθώς και την ίδια την πρωτεΐνη που παράγεται.

```
<?xml version="1.0"?>
<!DOCTYPE bioml SYSTEM "bioml.dtd">
<bioml>
<note>
    The following is a valid BIOML file describing the gene
    and the gene product that becomes human insulin.
</note> <organism> <species>Homo sapiens</species> <chromosome
number="11">
    <locus label="HUMINS locus">
    <gene label="Insulin gene">
        <dna start="1" end="4992" label="Complete HUMINS sequence">
            <ddomain start="1" end="2185" label="flanking domain"/>
            <ddomain start="1340" end="1823" label="polymorphic domain"/>
            <ddomain start="2424" end="2495" label="Signal peptide"/>
            <ddomain start="2496" end="2585" label="Chain B"/>
            <ddomain start="2586" end="2610" label="Chain C(1)"/>
            <ddomain start="3397" end="3476" label="Chain C(2)"/>
            <ddomain start="3477" end="3539" label="Chain A"/>
            <exon start="2186" end="2227" label="Exon 1"/>
            <intron start="2228" end="2406" label="Intron 1"/>
            <exon start="2407" end="2610" label="Exon 2"/>
            <intron start="2611" end="3396" label="Intron 2"/>
            <exon start="3397" end="3615" label="Exon 3"/>
            <ddomain  start="3615" end="4992" label="flanking domain"/>
<comment>
    The browser will ignore any symbol that cannot be a nucleotide
    residue, so the numbers can remain in place to aid the author.
</comment>
        1 ctcgaggggc ctagacattg ccctccagag agagcaccca acaccctcca ggcttgaccg
       61 gccagggtgt ccccttccta ccttggagag agcagcccca gggcatcctg caggggggtgc
      121 tgggacacca gctggccttc aaggtctctg cctccctcca gccaccccac tacacgctgc
      181 tgggatcctg gatctcagct ccctggccga caacactggc aaactcctac tcatccacga
      241 aggccctcct gggcatggtg gtccttccca gcctggcagt ctgttcctca cacaccttgt
```

[9]Πηγή: http://www.bioml.com/insulin3.htm

```
      301 tagtgcccag cccctgaggt tgcagctggg ggtgtctctg aagggctgtg agcccccagg
      361 aagccctggg gaagtgcctg ccttgcctcc ccccggccct gccagcgcct ggctctgccc
      421 tcctacctgg gctccccccca tccagcctcc ctccctacac actcctctca aggaggcacc
      481 catgtcctct ccagctgccg ggcctcagag cactgtggcg tcctggggca gccaccgcat
      541 gtcctgctgt ggcatggctc agggtggaaa gggcggaagg gaggggtcct gcagatagct

.....................................................................

     4501 agtgacaagg tcgttgtggc tccaggtcct tgggggtcct gacacagagc ctcttctgca
     4561 gcacccctga ggacagggtg ctccgctggg cacccagcct agtgggcaga cgagaaccta
     4621 ggggctgcct gggcctactg tggcctggga ggtcagcggg tgaccctagc taccctgtgg
     4681 ctgggccagt ctgcctgcca cccaggccaa accaatctgc acctttcctg agagctccac
     4741 ccagggctgg gctggggatg gctgggcctg gggctggcat gggctgtggc tgcagaccac
     4801 tgccagcttg ggcctcgagg ccaggagctc accctccagc tgccccgcct ccagagtggg
     4861 ggccagggct gggcaggcgg gtggacggcc ggacactggc cccggaagag gagggaggcg
     4921 gtggctggga tcggcagcag ccgtccatgg gaacacccag ccggccccac tcgcacgggt
     4981 agagacaggc gc
            </dna>
        </gene>
    </locus>
</chromosome> <protein comp="6xS[1]">
    <name>Insulin</name>
    <subunit id="1" comp="1xP[1]D[3]+1xP[1]D[7]">
        <peptide id="1" start="1" end="110">
            <db_entry entry="INS_HUMAN" format="SWISSPROT"/>
            <db_entry entry="IPHU" format="PIR"/>
            <domain id="1" type="signal" start="1" end="24"/>
            <domain id="2" type="helix" start="33" end="46"/>
            <domain id="3" type="mature" start="25" end="54">
                <name>Chain B</name>
                <aa type="C" at="31" to="96"/>
                <aa type="H" at="34">
                    <avariant at="34" type="D"/></aa>
                <aa type="C" at="43" to="109"/>
                <aa type="F" at="48">
                    <avariant type="S" at="48"/></aa>
                <aa type="F" at="49">
                    <avariant at="49" type="L"/></aa>
            </domain>
            <domain id="4" type="propeptide" start="55" end="89">
                <name>Chain C</name>
```

```
                <aa type="R" at="89">
                    <avariant at="89" type="H"/>
                    <avariant at="89" type="L"/></aa>
                </domain>
            <domain id="5" type="helix" start="91" end="95"/>
            <domain id="6" type="helix" start="102" end="108"/>
            <domain id="7" type="mature" start="90" end="110">
                <name>Chain A</name>
                <aa type="V" at="92">
                    <avariant type="L"/></aa>
                <aa type="C" at="95" to="100"/>
                <aa type="C" at="96" to="31"/>
                <aa type="C" at="100" to="95"/>
                <aa type="C" at="109" to="43"/>
            </domain>
            MALWMRLLPL LALLALWGPD PAAAFVNQHL CGSHLVEALY LVCGERGFFY
            TPKTRREAED LQVGQVELGG GPGAGSLQPL ALEGSLQKRG IVEQCCTSIC
            SLYQLENYCN
        </peptide>
    </subunit>
</protein>
</organism>
</bioml>
```

Άλλα

Όπως αναφέρθηκε και στην εισαγωγή του κεφαλαίου, υπάρχουν πολλά πρότυπα, χωρίς κάποιο από αυτά να κατέχει τον πρωτεύοντα ρόλο. Αυτό συμβαίνει όχι μόνο επειδή αντιμετωπίζουν διαφορετικά προβλήματα, αλλά και επειδή δεν έχει επικρατήσει η χρήση κάποιου. Στα επόμενα γίνεται μια σύντομη περιγραφή για αρκετά πρότυπα βασισμένα στην XML.

BSML. Η Bioinformatic Sequence Markup Language [27] ανήκει στην LabBook, Inc. και δημιουργήθηκε με χρηματοδότηση του National Human Genome Research Institute (NHGRI). Στόχος της είναι η περιγραφή των ακολουθιών και των σημειώσεων που χρειάζεται να τις συνοδεύουν. Ακολουθεί διαφορετική λογική από τη BIOML, που περιγράφηκε με περισσότερες λεπτομέρειες, αλλά εκπληρώνει τον ίδιο σκοπό.

GAME. Πρόκειται για ακόμη ένα πρότυπο που ασχολείται με την κατάλληλη μοντελοποίηση των ακολουθιών. Το Genome Annotation Markup Elements [29] έχει αναπτυχθεί από το πανεπιστήμιο του UC Berkeley. Δίνει έμφαση στην εύκολη ανταλλαγή πληροφοριών για αυτές τις ακολουθίες μεταξύ των ερευνητών.

PSDML. Το πρότυπο αυτό (Protein Sequence Database Markup Language) χρησιμοποιείται για να αποθηκεύονται πληροφορίες που αφορούν πρωτεΐνες σε μορφή ακολουθίας στη βάση δεδομένων PIR (Protein Information Resource) [22]. Είναι σχεδιασμένο με τρόπο που επιτρέπει την αποδοτική σύγκριση αυτών.

MAGE-ML. Η MAGE-ML (Microarray Gene Expression Markup Language) [35] έχει στόχο την περιγραφή και διακίνηση πληροφοριών σχετικών με τα microarray πειράματα (βλ. Κεφάλαιο 3). Αναπτύχθηκε από το Object Management Group και αποτελεί τη συνέχεια παλαιότερης γλώσσας MAML.

BlastXML. Μία από τις μορφές στις οποίες το πρόγραμμα BLAST, για το οποίο γίνεται λόγος στο Κεφάλαιο 7 της παρούσας εργασίας, μπορεί να παρουσιάσει το αποτέλεσμά του είναι και σε αρχείο μιας επέκτασης της XML.[10] Προκαταβολικά αναφέρεται ότι το πρόγραμμα αυτό ασχολείται με το alignment (ευθυγράμμιση) ακολουθιών.

4.4 Υπάρχουσες βάσεις δεδομένων

Σήμερα υπάρχουν περισσότερες από χίλιες διαφορετικές βάσεις δεδομένων σχετικές με τις βιοεπιστήμες. Καθεμιά από αυτές χρησιμοποιεί το δικό της format για να αποθηκεύει τις πληροφορίες, αν και κάθε εγγραφή τους έχει τη μορφή αρχείου κειμένου. Κάποιες βάσεις δεδομένων έχουν ήδη τουλάχιστον εικοσαετή ιστορία (EMBL 1980, SWISS-PROT 1986). Η κατηγοριοποίησή τους, ωστόσο, είναι πιο λογικό να γίνει σύμφωνα με το περιεχόμενό τους παρά με τα χρόνια ζωής τους.

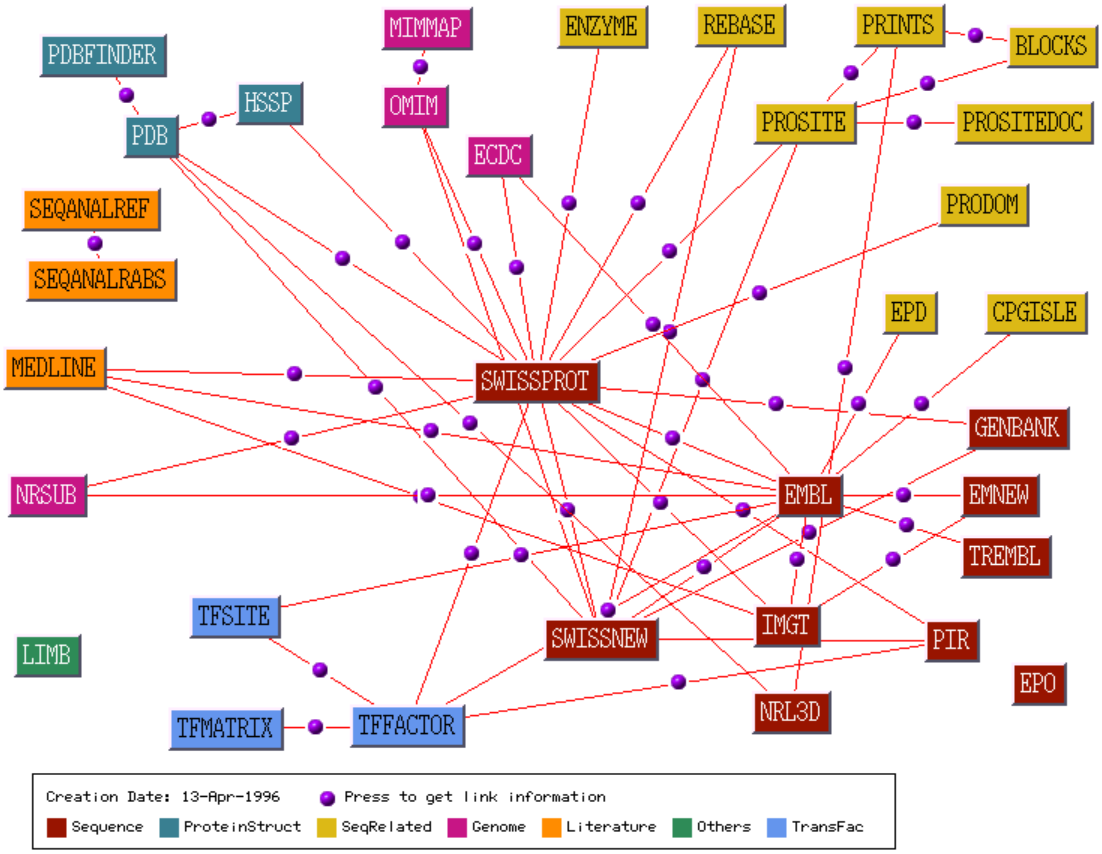
Χρήσιμοι κατάλογοι βάσεων βιοδεδομένων και πληροφορίες για αυτές μπορούν να αναζητηθούν μεταξύ των άλλων στα National Institutes of Health (NIH) [20] και European Bioinformatics Institute (EBI) [17].

Στο σχήμα 4.10 [11] φαίνεται μια ταξινόμηση αρκετών βάσεων των βιοεπιστημών με βάση το είδος των δεδομένων τους. Στα επόμενα θα επιχειρηθεί μια σύντομη περιγραφή των πιο σημαντικών από αυτές ανά κατηγορία.

4.4.1 Ακολουθιών νουκλεϊκών οξέων

Στον τομέα αυτο κυριαρχεί η συνεργασία μεταξύ των τριών μεγάλων οργανισμών National Center for Biotechnology Information (NCBI), USA, European Bioinformatics Institute, UK, National Institute of Genetics, Japan. Αυτοί έχουν καταφέρει να ανταλλάσουν καθημερινά δεδομένα, ώστε οι βάσεις GenBank, EMBL Data Library, DNA Data Bank of Japan (DDBJ) να έχουν το ίδιο περιεχόμενο, απλά σε διαφορετικό format.

[10] http://www.visualgenomics.ca/gordonp/xml/BlastXML/dtd/blastxml.dtd
[11] http://www.ii.uib.no/bio/seminars/sem97db/net.gif



Σχήμα 4.10: Ενδεικτική ταξινόμηση διαφόρων βάσεων δεδομένων.

Από αυτές τις τρεις δημόσιες βάσεις αντλούν το περιεχόμενό τους οι περισσότερες υπόλοιπες που έχουν ανάλογο είδος δεδομένων. Συνήθως προσφέρουν κάποια επιπλέον προγράμματα, κάνοντας ίσως τις μικρότερες βάσεις περισσότερο λειτουργικές.

Είναι σημαντικό να δοθεί έμφαση στο ρυθμό αύξησης του μεγέθους της GenBank[12]. Σήμερα διπλασιάζει την ποσότητα πληροφοριών που διαθέτει κάθε έξι με οκτώ μήνες, ενώ αυτός ο ρυθμός ολοένα αυξάνεται. Μάλιστα, προς το παρόν περιέχει περίπου επτά εκατομμύρια εγγραφές, που αντιστοιχούν σε εννιά δισεκατομμύρια νουκλεοτίδια.

Παράδειγμα του format μιας εγγραφής της βάσης GenBank έχει δοθεί στην ενότητα 4.3.1. Εκεί υπάρχουν αρκετά στοιχεία για να κατανοηθεί και το μοντέλο στο οποίο στηρίζεται.

4.4.2 Γονιδιωμάτων

Οι ακολουθίες των νουκλεοτιδίων των γονιδιωμάτων των διαφόρων οργανισμών υπάρχουν στην GenBank[13]. Μέχρι τον Οκτώβριο του 2000 ήταν διαθέσιμα τα μερικώς ή πλήρως αποκωδικοποιημένα γονιδιώματα περίπου 900 ειδών. Ωστόσο, έχουν δημιουργηθεί και ξεχωριστές βάσεις δεδομένων για τον κάθε οργανισμό. Οι βάσεις των γονιδιωμάτων εκτός από τις ακολουθίες και τις σημειώσεις για αυτές περιέχουν επιπλέον στοιχεία, όπως μπορεί να είναι οι γενετικοί χάρτες ή βοηθητικές πηγές πληροφοριών.

4.4.3 Ακολουθιών πρωτεϊνών

Σε αυτήν την κατηγορία οι βάσεις δεδομένων στις οποίες επικεντρώνεται το μεγαλύτερο ενδιαφέρον είναι και πάλι τρεις. Ωστόσο, δεν υπάρχει μεταξύ τους ανάλογος συνασπισμός όπως αυτός των GenBank, EMBL, DDBJ που αναφέρθηκε στην ενότητα 4.4.1.

Η πρώτη από αυτές είναι ξανά η GenBank[14], η οποία περιλαμβάνει και ακολουθίες αμινοξέων εκτός από νουκλεοτιδίων. Το format των εγγραφών είναι βέβαια το ίδιο με αυτό που δόθηκε στην ενότητα 4.4.1.

Η δεύτερη ιδιαίτερα σημαντική βάση είναι η PIR-International [22]. Είναι προϊόν της συνεργασίας των National Biomedical Research Foundation (Georgetown University, Washington, DC, USA), Munich Information Center for Protein Sequences (MIPS, Munich, Germany), Japan International Protein Information Database (Tsukuba, Japan). Ουσιαστικά αποτελείται από επτά μικρότερες βάσεις δεδομένων (PIR-PSD, iProClass, ASDB, P/R-NREF, NRL3D, ALN, RESID).

Με αρκετές μικρότερες βάσεις δεδομένων σχετίζεται και η τρίτη πολύ σημαντική βάση, η SWISS-PROT [16]. Σε αυτήν υπάρχουν πληροφορίες από τις ερευνητικές ομάδες των Swiss Institute of Bioinformatics, European Bioinformatics Institute, UK. Δύο από αυτές τις μικρότερες είναι η Enzyme, βάση για τα ένζυμα και η PROSITE, που αναφέρεται σε motifs. Το format μιας εγγραφής της Enzyme [15] φαίνεται στο ακόλουθο παράδειγμα, για την κατανόηση του οποίου παρέχονται επίσης τα αναγνωριστικά και η σημασία τους.

[12] http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Nucleotide
[13] http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Genome
[14] http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Protein

Πίνακας 4.15: Τα σύμβολα μιας εγγραφής της Enzyme

ID	Το αναγνωριστικό (Identification)
DE	Το επίσημο όνομα (Description)
AN	Εναλλακτικά ονόματα (Alternate)
CA	Καταλυτική δράση (Catalytic activity)
CF	Συμπαράγοντες (Cofactors)
CC	Σχόλια (Comments)
DI	Ασθένειες που σχετίζονται με το ένζυμο (Disease)
PR	Αντιστοιχία στη βάση PROSITE
DR	Αντιστοιχία στη βάση Swiss-Prot
//	Γραμμή τερματισμού

```
ID   1.14.17.3
DE   Peptidylglycine monooxygenase.
AN   Peptidyl alpha-amidating enzyme.
AN   PAM.
AN   Peptidyl-glycine alpha-amidating monooxygenase.
AN   Peptidylglycine 2-hydroxylase.
CA   Peptidylglycine + ascorbate + O(2) = peptidyl(2-hydroxyglycine)+
CA   dehydroascorbate + H(2)O.
CF   Copper.
CC   -!- Peptidylglycines with a neutral amino acid residue in the
CC       penultimate position are the best substrates for the enzyme.
CC   -!- The enzyme also catalyzes the dismutation of the product to
CC       glyoxylate and the corresponding desglycine peptide amide.
CC   -!- Involved in the final step of biosynthesis of alpha-
CC   melanotropin and related biologically active peptides.
PR   PROSITE; PDOC00080;
DR   P08478, AMD1_XENLA ;  P12890, AMD2_XENLA ;  P10731, AMD_BOVIN  ;
DR   P19021, AMD_HUMAN  ;  P97467, AMD_MOUSE  ;  P14925, AMD_RAT    ;
//
```

Τα ID, // είναι απαραίτητο να υπάρχουν στην αρχή και στο τέλος κάθε εγγραφής, αντίστοιχα. Όλα τα υπόλοιπα μπορούν να εμφανίζονται περισσότερες από μία φορές και η σημασία τους φαίνεται στον πίνακα 4.15.

4.4.4 Τρισδιάστατων δομών μακρομορίων

Σε αυτό το πεδίο οι βάσεις που υπάρχουν είναι ιδιαίτερα περιορισμένες λόγω της δυσκολίας της φύσης των δεδομένων. Η πιο γνωστή από αυτές είναι η PDB (Protein Data Bank) [38] και –όπως δηλώνει το όνομά τησ– ασχολείται με τις πρωτεΐνες. Τα περιεχόμενά της προέρχονται κυρίως από τα πειράματα με ακτίνες X.

Σε αντίθεση με τις βάσεις που περιέχουν ακολουθίες, ο αριθμός των εγγραφών της είναι περιορισμένος. Είναι αποθηκευμένες περίπου 45000 πρωτεΐνες. Το μικρό μέγεθός της, ωστόσο, δεν είναι το μόνο πρόβλημα, καθώς θεωρείται σχετικά δυσνόητος ο τρόπος με τον οποίο καταχωρούνται τα δεδομένα για την τρισδιάστατη απεικόνιση.

4.4.5 Δεδομένων σχετικών με τη γενετική έκφραση (gene expression data)

Πιστεύεται ότι οι βάσεις με αυτό το περιεχόμενο θα αυξηθούν σημαντικά τα επόμενα χρόνια. Αυτή τη στιγμή δεν υπάρχουν, όμως, πολλές διαθέσιμες. Η πλέον οργανωμένη είναι και πάλι η συλλογή dbEST μέσα στη GenBank, ενώ αξιόλογες πηγές είναι και οι ιστοτόποι των National Human Genome Research Initiative's Microarray Project [30], Stanford Genome Resources [18]. Επίσης, formats και πρότυπα αποθήκευσης για τα δεδομένα αυτά είναι ακόμη υπό ανάπτυξη με σημαίνοντα το ρόλο του European Bioinformatics Institute.

4.4.6 Βιομονοπατιών (biological pathways)

Στη συγκεκριμένη κατηγορία δεσπόζει η KEGG (Kyoto Encyclopedia of Genes and Genomes) [31]. Παρότι έχει συλλέξει και δεδομένα για γονίδια και τα προϊόντα τους, ξεχωρίζει για τις πληροφορίες σχετικά με τα μεταβολικά μονοπάτια αλλά και τις συνδέσεις μεταξύ των πληροφοριών διαφορετικών ειδών δεδομένων. Είναι και αυτή διαθέσιμη προς αναζήτηση στον κάθε ενδιαφερόμενο μέσω του παγκόσμιου ιστού.

Μία ακόμη βάση δεδομένων που διακρίνεται σε αυτόν τον τομέα είναι η WIT (What Is There?) [34], που αναπτύχθηκε από τα Argonne National Labs. Περιλαμβάνει μεταβολικά μονοπάτια για οργανισμούς με αποκωδικοποιημένο γονιδίωμα.

Υπάρχουν, τέλος, αρκετές μικρότερες και εξειδικευμένες βάσεις, όπως η Ecocyc pathway DB [25], με περιεχόμενο σχετικό με το βακτήριο Escherichia coli.

4.4.7 Δημοσιεύσεων, άρθρων, βιβλιογραφικού υλικού

Η πλέον καθιερωμένη βάση δεδομένων είναι η PubMed[15]. Αυτή η βιβλιογραφική βάση περιέχει περιλήψεις δημοσιεύσεων και επιστημονικών άρθρων των βιοεπιστημών, ενώ περιλαμβάνει την MEDLINE. Η τελευταία είναι βάση που στηρίζει το περιεχόμενό της στη US National Library of Medicine. Για την αξιοποίηση των πληροφοριών της PubMed μπορούν να χρησιμοποιηθούν όλα τα εργαλεία και προγράμματα που προσφέρει ελεύθερα στο Internet το NCBI, το οποίο και τη διατηρεί.

[15] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubmed

Κεφάλαιο 5

Ερωτήματα

Το παρόν κεφάλαιο στοχεύει να αντιμετωπίσει ζητήματα που έχουν σχέση με τα ερωτήματα τα οποία χρειάζεται να θέτουν οι ερευνητές των βιοεπιστημών. Προσπαθεί να δώσει τη γενική εικόνα του περιβάλλοντος από το οποίο τα queries πηγάζουν. Για το λόγο αυτό ασχολείται όχι μόνο με τα ερωτήματα αυτά καθεαυτά αλλά και με τις εργασίες εκείνες που τα καθιστούν απαραίτητα.

Η πρώτη ενότητα εποπτεύει το χώρο των ερωτημάτων. Παραθέτει χαρακτηριστικά τους παραδείγματα, για να έχει ο αναγνώστης καλύτερα σχηματισμένη άποψη για το είδος τους. Ταυτόχρονα, επιχειρεί να τα κατηγοριοποιήσει, ώστε να γίνουν πιο συγκεκριμένα τα κοινά χαρακτηριστικά που κάποια παρουσιάζουν. Φιλοδοξεί, όμως, να γίνει και εφαλτήριο για να οδηγηθεί κανείς με αληθινή περιέργεια στα υπόλοιπα θέματα του κεφαλαίου, αναζητώντας το λόγο για τον οποίο χρειάζονται τα ερωτήματα αυτά.

Η δεύτερη ενότητα είναι και η κυρίαρχη, τουλάχιστον ποσοτικά. Περιγράφει τις λειτουργίες που γίνονται πάνω στα δεδομένα στα οποία αναφέρθηκε το προηγούμενο κεφάλαιο. Αναφέρει, παράλληλα, τους κυριότερους αλγορίθμους με τους οποίους οι βιοπληροφορικοί έχουν αντιμετωπίσει τις εργασίες αυτές και έχουν τις περισσότερες φορές καταφέρει να δώσουν αποδοτική λύση. Παραδείγματα πλασιώνουν πάντοτε την ανάλυση, ώστε να διαλύονται οι περισσότερες πιθανές απορίες.

Τέλος, η τρίτη ενότητα έχει ως θέμα της μοντέλα που χρησιμοποιούνται για την αντιμετώπιση των προβλημάτων που έχουν ήδη αναφερθεί. Πρόκειται για μοντέλα αναπαράστασης που είναι καθιερωμένα σε άλλους χώρους, αλλά βρίσκουν εφαρμογή και στις βιοεπιστήμες. Η μελέτη δεν προχωρά σε ζητήματα της θεωρίας τους, όμως προσφέρει παραδείγματα του τρόπου με τον οποίο μπορούν να είναι χρήσιμα για συγκεκριμένες λειτουργίες.

5.1 Ενδεικτική κατηγοριοποίηση

Στην ενότητα αυτή διατυπώνονται κάποια ερωτήματα που είναι συνηθισμένο ή χρήσιμο να τίθενται από τους ερευνητές των βιοεπιστημών. Σκοπός της είναι να δώσει αφενός ένα κατά το δυνατόν αντιπροσωπευτικό δείγμα τέτοιων ερωτήσεων και αφετέρου ένα έναυσμα για τη μελέτη του υπόλοιπου κεφαλαίου, στο οποίο εξετάζονται σε μεγαλύτερη λεπτομέρεια οι λει-

τουργίες που καθιστούν απαραίτητες αυτές τις ερωτήσεις.

Είναι γεγονός ότι υπάρχει μεγάλη ποικιλία ερωτημάτων που μπορούν να τεθούν, ακριβώς λόγω της ποικιλίας και του πλήθους των δεδομένων των βιοεπιστημών (Κεφάλαιο 4). Επιπλέον, χρειάζεται βαθιά γνώση του αντικειμένου της βιολογίας και των συναφών κλάδων, για να εντοπιστεί το μεγαλύτερο ποσοστό εκείνων που περισσότερο ενδιαφέρουν.

Στη συνέχεια επιχειρείται μια πρώτη προσέγγιση ανακάλυψης της εν λόγω ομάδας ερωτημάτων. Για κάθε κατηγορία ερωτημάτων δίνονται χαρακτηριστικά παραδείγματα. Πολύτιμοι οδηγοί στάθηκαν τα [45], [55], [13].

5.1.1 Ομοιότητας

Η κατηγορία αυτή αφορά μεγάλο μέρος δεδομένων (όπως ακολουθίες, γράφοι, τρισδιάστατες δομές) και καλύπτει ένα πολύ σημαντικό τμήμα των ερωτημάτων. Πρόκειται για τη σύγκριση δύο ή περισσότερων δεδομένων ίδιας μορφής.

Παράδειγμα α. Να βρεθούν οι ακολουθίες (νουκλεοτιδίων ή αμινοξέων) που είναι όμοιες με δοθείσα ακολουθία (των αντίστοιχων δομικών μονάδων) ή τμήμα αυτής. Το συγκεκριμένο πρόβλημα σε ανάλογη διατύπωση συναντιέται σε αρκετές εφαρμογές της πληροφορικής, όπως η επεξεργασία κειμένου.

Παράδειγμα β. Να βρεθούν οι τρισδιάστατες πρωτεϊνικές δομές που είναι όμοιες με δοθείσα. Ουσιαστικά είναι η γενίκευση του προηγούμενου παραδείγματος στις τρεις διαστάσεις.

Παράδειγμα γ. Να βρεθούν οι τρισδιάστατες πρωτεϊνικές δομές που έχουν ομοιότητες με πρωτεΐνη, της οποίας είναι γνωστή μόνο η ακολουθία. Το πρόβλημα αυτό δεν ανάγεται σε εκείνο του παραδείγματος α, δηλαδή στη σύγκριση των αντίστοιχων ακολουθιών. Αυτό επειδή, αν και πρωτεΐνες με παρόμοιες ακολουθίες έχουν και παρόμοιες τρισδιάστατες δομές, το αντίστροφο δεν ισχύει.

Παράδειγμα δ. Να βρεθούν ακολουθίες αμινοξέων που αντιστοιχούν σε πρωτεϊνικές δομές οι οποίες παρουσιάζουν ομοιότητες με δοθείσα δομή πρωτεΐνης. Η λύση της σύγκρισης με όλες τις τρισδιάστατες δομές, ώστε από τις όμοιες να βρεθούν οι ζητούμενες ακολουθίες, δεν είναι πλήρης. Ο λόγος είναι το γεγονός ότι στις υπάρχουσες βάσεις δεδομένων βρίσκονται πολύ περισσότερες ακολουθίες από δομές, με αποτέλεσμα ακολουθίες που πιθανώς ανήκουν στη λύση να μη γίνεται να προσπελαστούν.

Παράδειγμα ε. Να βρεθούν τα μεταβολικά μονοπάτια ενός οργανισμού των οποίων οι γράφοι είναι ισομορφικοί με άλλον που δίνεται. Το συγκεκριμένο πρόβλημα παρουσιάζει σημαντικές δυσκολίες για τα υπάρχοντα συστήματα βάσεων δεδομένων ως προς την επεξεργασία του ερωτήματος.

Από τα προβλήματα των παραπάνω παραδειγμάτων, εκείνα των α και β αντιμετωπίζονται πλέον αποδοτικά. Τα υπόλοιπα είναι αντικείμενο έρευνας. Επίσης, χρειάζεται να σημειωθεί ότι τα προβλήματα περιπλέκονται ακόμη περισσότερο, όταν τα δεδομένα βρίσκονται σε διαφορετικές βάσεις.

5.1.2 Για patterns

Η συγκεκριμένη ομάδα ερωτημάτων αναφέρεται κυρίως στα motifs (Κεφάλαιο 4). Αφορά την ανακάλυψη patterns, την εύρεσή τους σε συγκεκριμένο δείγμα αλλά και τη σύγκριση μεταξύ τους.

Παράδειγμα α. Να βρεθεί αν υπάρχει κάποιο pattern στις ακολουθίες νουκλεοτιδίων του ίδιου γονιδίου σε διαφορετικά είδη (όπως στον homo sapiens και στον homo universalis).

Παράδειγμα β. Να βρεθούν όλα τα patterns που εμφανίζονται σε δοθείσα ακολουθία.

Παράδειγμα γ. Να βρεθούν τυχόν ομοιότητες στα patterns που είναι γνωστό ότι εμφανίζονται σε συγκεκριμένη τριασδιάστατη δομή.

5.1.3 Για metadata

Ερωτήματα που αφορούν δευτερεύουσες πληροφορίες σε συνάρτηση με ένα δεδομένο είναι επίσης χρήσιμα. Αυτές περιλαμβάνουν τις συνθήκες των πειραμάτων, την ημερομηνία τέλεσής τους και φυσικά, την επιστημονική ομάδα που είναι υπεύθυνη για αυτά.

Παράδειγμα. Να βρεθούν όλα τα δεδομένα που προέκυψαν για τα διάφορα δείγματα από το microarray πείραμα από το οποίο προέκυψε και δοθέν δεδομένο.

5.1.4 Για εικόνες

Αυτά τα ερωτήματα μπορούν να χωριστούν σε τρεις υποκατηγορίες. Είναι τα spatial, που έχουν να κάνουν με χωρικές πληροφορίες που μπορούν να εξαχθούν από τις εικόνες. Υπάρχουν επίσης τα semantic, τα οποία αφορούν δεδομένα που εξάγονται για υψηλού επιπέδου αντικείμενα, όπως ο τύπος κυττάρου. Τέλος, τα spatio-temporal αναφέρονται σε παροδικές αλλαγές χωρικών χαρακτηριστικών όπως είναι λ.χ. η ανάπτυξη του κυτταρικού σκελετού.

Παράδειγμα α. Να βρεθούν όλες οι εικόνες στις οποίες η χωρική κατανομή μιας πρωτεΐνης είναι η ίδια με αυτήν σε δοθείσα εικόνα.

Παράδειγμα β. Να βρεθούν όλες οι εικόνες κυττάρων του αμφιβληστροειδούς στις οποίες τα κύτταρα έχουν συγκεκριμένο σχήμα.

Παράδειγμα γ. Να βρεθεί το σύνολο των εικόνων στις οποίες η μεταβολή μιας πρωτεΐνης σε ένα κύτταρο είναι ανάλογη με τη μεταβολή άλλης συγκεκριμένης πρωτεΐνης.

5.2 Λειτουργίες

Στην ενότητα αυτή δίνονται στοιχεία για εργασίες που ενδιαφέρει τους ερευνητές των βιοεπιστημών να εκτελούν. Περιγράφεται κάθε φορά ποια είναι αυτή η εργασία και ποιος αλγόριθμος μπορεί να χρησιμοποιηθεί από τους βιοπληροφορικούς για να γίνει. Παρατίθεται, επιπλέον, παράδειγμα, για να είναι κατανοητό το πρόβλημα και η διαδικασία επίλυσής του, όπου υπάρχει. Η ανάλυση σκοπεύει να εισάγει τον αναγνώστη σε καθένα από τα θέματα που καλύπτει.

Οι λειτουργίες που απασχολούν την υπόλοιπη ενότητα καλύπτουν ένα αρκετά μεγάλο φάσμα εργασιών αλλά και δυσκολιών. Πρόκειται για τη σύγκριση δύο ή περισσότερων ακολουθιών (alignment), τη φυλογενετική ανάλυση, τη sequence assembly, την ανάλυση γονιδιωμάτων, τον προσδιορισμό δευτεροταγούς δομής μακρομορίου, καθώς και την εύρεση της τριτοταγούς δομής του.

5.2.1 Σύγκριση ακολουθιών (alignment)

Η σύγκριση δύο ή περισσότερων ακολουθιών είναι μια πολύ βασική λειτουργία. Η ομοιότητά τους μπορεί να υποδηλώνει κοινή προέλευση σύμφωνα με την εξελικτική θεωρία, κοινή τρισδιάστατη δομή, ίδιο λειτουργικό ρόλο ή απλά ένα τυχαίο γεγονός. Η σύγκριση αυτών των ακολουθιών έχει στενή σχέση με την έννοια του alignment.

Ο όρος *alignment* αναφέρεται στη διαδικασία ευθυγράμμισης δύο ακολουθιών με το ένα στοιχείο της μίας κάτω από το αντίστοιχο της άλλης, κατά τέτοιο τρόπο ώστε να επιτυγχάνεται η μέγιστη δυνατή ταυτοποίηση. Με άλλα λόγια επιθυμητό είναι να βρίσκονται τελικά όσο το δυνατό περισσότερα κατακόρυφα ζευγάρια ίδιων στοιχείων. Το alignment αυτό αναφέρεται ως *pairwise* σε αντίθεση με τα *multiple alignments*, στα οποία συγκρίνονται περισσότερες των δύο ακολουθιών.

Επιπλέον, η παράταξη των στοιχείων των ακολουθιών μπορεί να ενδιαφέρει να γίνει σε τοπικό ή σε ολικό επίπεδο. Στην περίπτωση της τοπικής αντιστοίχισης *(local alignment)*, τμήματα μόνο των δύο ακολουθιών ευθυγραμμίζονται, ενώ στην ολική *(global alignment)* η ευθυγράμμιση αφορά ολόκληρες τις ακολουθίες.

Ένα θέμα που προκύπτει είναι κατά πόσο επιτρέπεται ή όχι στο alignment η χρήση κενών. Στις περιπτώσεις που αφορούν τη βιολογία η χρήση κενών είναι επιτρεπτή. Έτσι, η ακολουθία a b c d μπορεί να ευθυγραμμιστεί με την bd κατά τον βέλτιστο τρόπο σύμφωνα με τον Πίνακα 5.1. Στο global alignment υπάρχουν συνήθως πολλά μικρά διασκορπισμένα κενά, ενώ στο local οι περιοχές με τα κενά είναι μεγαλύτερες.

Πίνακας 5.1: Alignment με χρήση κενών.

a b c d
- b - d

Απαραίτητο είναι να προσδιοριστεί ένα μέτρο με το οποίο ένα alignment να θεωρείται βέλτιστο. Σε περιπτώσεις προβλημάτων με τα παραπάνω χαρακτηριστικά αυτό δεν είναι προφανές. Με το μάτι εύκολα προσδιορίζεται ότι το τρίτο από τα ζεύγη του Πίνακα 5.2 είναι το καλύτερο alignment. Ωστόσο, σε μεγάλες ακολουθίες, που ενδιαφέρουν τοπικά alignment, χρειάζεται ένα αντικειμενικό κριτήριο για την εύρεση του βέλτιστου.

Στα επόμενα είναι χρήσιμο να δοθούν δύο ορισμοί και να παρουσιαστούν δύο είδη πινάκων. Οι ορισμοί ξεδιαλύνουν πότε δύο ακολουθίες είναι όμοιες και πότε ομόλογες, ενώ οι πίνακες υποβοηθούν την εκτέλεση του alignment αλλά και την αξιολόγησή του.

Πίνακας 5.2: Πιθανά alignments δύο ακολουθιών.

g c t g a a c g c t a t a a t c
g c t g a - a - - c g - - c t - a t a a t c
g c t g - a a - c g - c t a t a a t c -

Όταν δύο ακολουθίες κρίνονται ως προς την *ομοιότητα (similarity)*, παρατηρούνται ή μετρώνται στοιχεία που είναι ίδια σε αυτές, ανεξάρτητα από την πηγή προέλευσης. Από την άλλη, όταν δύο ακολουθίες είναι *ομόλογες (homologous)*, αυτό σημαίνει ότι ανήκουν σε οργανισμούς που έχουν κάποιο κοινό πρόγονο. Οι ομόλογες ακολουθίες εξάγονται από παρατηρήσεις σε όμοιες ακολουθίες.

Στη συνέχεια δίνεται ένας τρόπος σύγκρισης δύο ακολουθιών, που στηρίζεται κυρίως στην εποπτική παρατήρηση. Δοθέντων αυτών κατασκευάζεται ένας πίνακας του οποίου οι γραμμές αντιστοιχούν στα στοιχεία της μιας ακολουθίας και οι στήλες στα στοιχεία της άλλης. Οι θέσεις μέσα στον πίνακα είναι αρχικά 0 ή κενές. Μία θέση γεμίζει μόνο όταν το στοιχείο της γραμμής της είναι το ίδιο με το στοιχείο της στήλης της. Τότε το περιεχόμενο της θέσης σε αυτήν την περίπτωση γίνεται 1 (θεωρώντας ότι αρχικά ήταν 0) ή γεμίζει με το στοιχείο της γραμμής/στήλης της (θεωρώντας ότι αρχικά ήταν κενή). Ο πίνακας αυτός είναι γνωστός ως *dot matrix*.

Στα Σχήματα 5.1, 5.2 [1] δίνονται δύο παραδείγματα τέτοιων πινάκων. Ο πρώτος αντιστοιχεί στην επαναλαμβανόμενη ακολουθία ABRACADABRACADABRA και τον εαυτό της, ενώ ο δεύτερος στην καρκινική ακολουθία MAX I STAY AWAY AT SIX AM και τον εαυτό της. Εύκολα παρατηρείται ότι όταν οι ακολουθίες που συγκρίνονται είναι οι ίδιες, η διαγώνιος είναι γεμάτη. Επίσης, όταν η ακολουθία αποτελείται από επαναλαμβανόμενες υπακολουθίες της - κάτι που συμβαίνει αρκετά συχνά στις ακολουθίες της βιολογίας - μικρότερες διαγώνιοι είναι ξανά γεμάτες. Έτσι, με τον αλγόριθμο κατασκευής του dot matrix παρέχεται ένας εποπτικός τρόπος για να διαπιστωθεί η ομοιότητα δύο ακολουθιών.

Ωστόσο, ο εν λόγω τρόπος παρουσιάζει τέσσερα μειονεκτήματα για τις εφαρμογές της βιολογίας. Πρώτον, έχει χρονική και χωρική πολυπλοκότητα μ*ν (όπου μ, ν τα μήκη των δύο ακολουθιών που συγκρίνονται), που για μεγάλα μ και ν είναι απαγορευτικό μέγεθος. Δεύτερον, για να διαπιστωθεί αν η διαγώνιος τελικά γεμίζει, κατασκευάζεται ολόκληρος ο πίνακας. Τρίτον, όταν το λεξιλόγιο είναι μικρό (π.χ. 4 γράμματα για τα νουκλεοτίδια), τότε μπορεί να εμφανιστεί θόρυβος, δηλαδή πολλές μικρές διαγώνιοι όπου όλες θα πρέπει να μελετηθούν, για να βρεθεί η βέλτιστη. Τέταρτον, δεν παρέχει έναν άμεσο τρόπο εύρεσης της βέλτιστης διαγωνίου (στην περίπτωση που δεν είναι γεμάτη η μέγιστη διαγώνιος), με άλλα

[1] Πηγή: [45]

	A	B	R	A	C	A	D	A	B	R	A	C	A	D	A	B	R	A
A	A			A		A		A			A		A		A			A
B		B							B							B		
R			R							R							R	
A	A			A		A		A			A		A		A			A
C					C							C						
A	A			A		A		A			A		A		A			A
D							D							D				
A	A			A		A		A			A		A		A			A
B		B							B							B		
R			R							R							R	
A	A			A		A		A			A		A		A			A
C					C							C						
A	A			A		A		A			A		A		A			A
D							D							D				
A	A			A		A		A			A		A		A			A
B		B							B							B		
R			R							R							R	
A	A			A		A		A			A		A		A			A

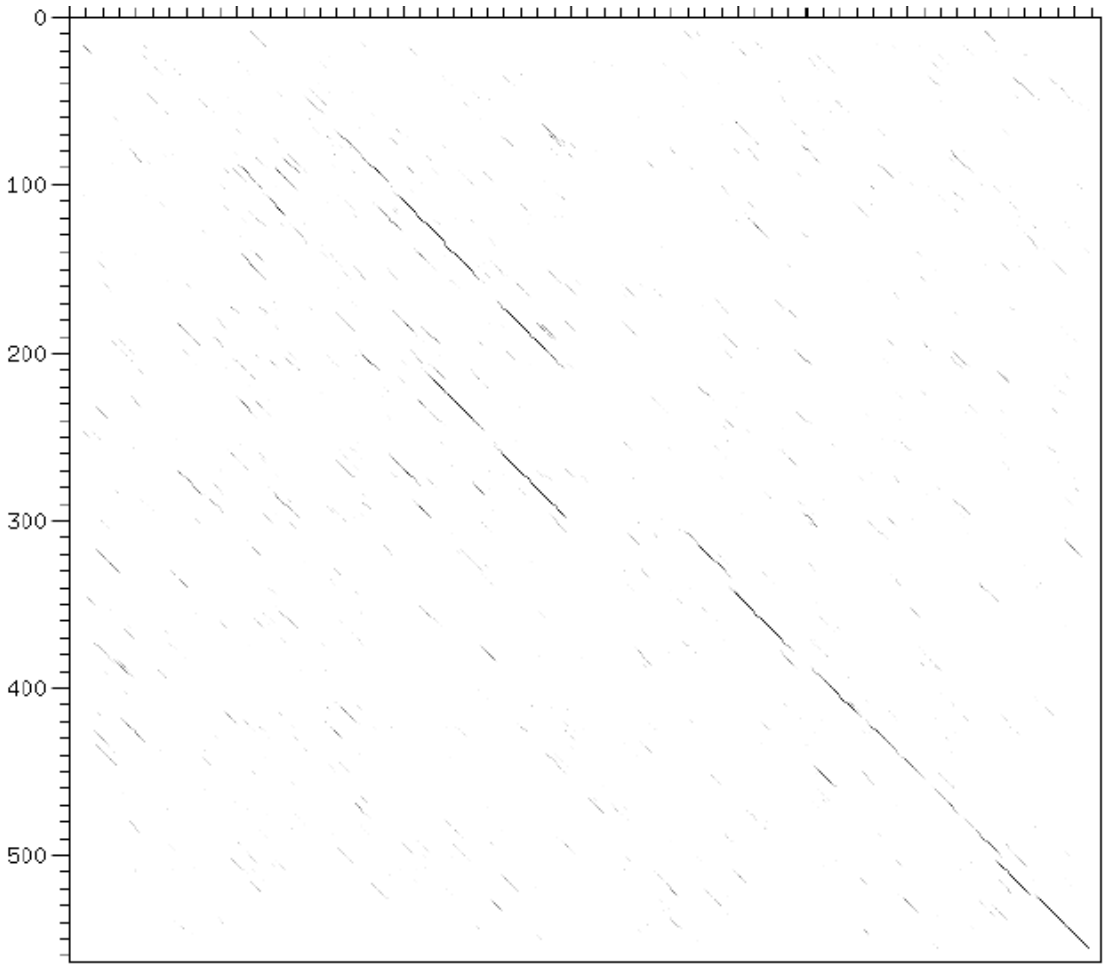
Σχήμα 5.1: Dotplot για την επαναλαμβανόμενη ακολουθία ABRACADABRACADABRA.

	M	A	X	I	S	T	A	Y	A	W	A	Y	A	T	S	I	X	A	M
M	M																		M
A		A					A		A		A		A					A	
X			X														X		
I				I												I			
S					S										S				
T						T								T					
A		A					A		A		A		A					A	
Y								Y				Y							
A		A					A		A		A		A					A	
W										W									
A		A					A		A		A		A					A	
Y								Y				Y							
A		A					A		A		A		A					A	
T						T								T					
S					S										S				
I				I												I			
X			X														X		
A		A					A		A		A		A					A	
M	M																		M

Σχήμα 5.2: Dotplot για την καρκινική ακολουθία MAX I STAY AWAY AT SIX AM.

λόγια δεν δίνει ακριβές μέτρο ομοιότητας μεταξύ των δύο ακολουθιών.

Τα παραπάνω, όμως, δε σημαίνουν ότι δε χρησιμοποιείται στις βιοεπιστήμες και μάλιστα με επιτυχία. Αρκετά προγράμματα που εκτελούν alignments ακολουθιών (όπως το BLAST, βλ. Κεφάλαιο 7) ακολουθούν αλγορίθμους που βασίζονται σε dot matrix, αλλά βελτιωμένους. Στο Σχήμα 5.3 [2] φαίνεται ο πίνακας που σχηματίζεται από τη σύγκριση δύο ακολουθιών αμινοξέων.



Σχήμα 5.3: Dotplot για τις ακολουθίες αμινοξέων human coagulation factor XII (F12; SWISS-PROT P00748), tissue plasminogen activator (PLAT; SWISS-PROT P00750).

Όπως έχει αναφερθεί, σημαντικό είναι να υπάρχει ένα μέτρο ικανό να δηλώνει το βαθμό ομοιότητας δύο ακολουθιών. Με άλλα λόγια, είναι αναγκαίο κάθε alignment με κάποιο τρόπο να βαθμολογείται. Υπάρχουν διάφορες τακτικές, κάποιες από τις οποίες δίνουν τιμές για κάθε ταίριασμα, διαφορά ή κενό και έτσι προκύπτει ένα τελικό σκορ. Από τις πιο διαδεδομένες, ωστόσο, τεχνικές είναι αυτές που χρησιμοποιούν τους λεγόμενους *πίνακες αντικατάστασης (substitution matrices)*.

[2] Πηγή: [4]

Σε αυτό το είδος πινάκων ανήκουν οι ονομαζόμενοι πίνακες PAM (Percent Accepted Mutation) και BLOSUM (BLOcks SUbstitution Matrix). Αυτοί οι συχνά χρησιμοποιούμενοι πίνακες είναι τριγωνικοί 20x20 - όσα και τα διαφορετικά αμινοξέα που συνθέτουν πρωτεΐνες - και έχουν κατασκευαστεί από τους βιολόγους (βλ. Σχήμα 5.4) [3]. Παρέχουν τιμές βαρών που δείχνουν την πιθανότητα αντικατάστασης ενός αμινοξέος από άλλο - λόγω μετάλλαξης - με την πάροδο του χρόνου. Από το γεγονός αυτό προέρχεται και η ονομασία τους. Ο δε όρος BLOCKS στον πίνακα BLOSUM προέρχεται από την ομώνυμη βάση δεδομένων, που περιέχει ακολουθίες πρωτεϊνών που είναι ευθυγραμμισμένες (aligned).

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	12															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4

Σχήμα 5.4: Ο πίνακας αντικατάστασης PAM250.

Οι πίνακες αντικατάστασης στηρίζονται στην παρατήρηση ότι δεν είναι όλες οι αλλαγές μεταξύ των αμινοξέων ισότιμες. Υπάρχουν αμινοξέα που έχουν παρεμφερή χημική σύσταση και αν το ένα αντικαταστήσει το άλλο σε μια πολυπεπτιδική αλυσίδα, το αποτέλεσμα μπορεί

[3] Πηγή: [4]

να μη διαφέρει σημαντικά. Το γεγονός αυτό αξιοποιείται για την αξιολόγηση ενός alignment, δίνοντας μεγαλύτερο σκορ σε ένα ζευγάρι διαφορετικών αλλά χημικά συγγενών αμινοξέων από ότι σε ένα άλλο τυχαίο ζευγάρι.

Ουσιαστικά πρόκειται για ομάδες πινάκων PAM ή BLOSUM και όχι για έναν πίνακα. Η μονάδα μέτρησης ενός PAM αντιστοιχεί σε δύο ακολουθίες που διαφέρουν μόνο κατά 1%. Την περίπτωση να διαφέρουν στο 25% καλύπτει ο πίνακας PAM30, ενώ ο πίνακας PAM250 ορίζεται να αναλογεί σε ακολουθίες που διαφέρουν στο 80% των δομικών τους μονάδων. Οι πίνακες BLOSUM βασίζονται στην ίδια λογική, είναι όμως πιο σύγχρονοι και για την κατασκευή τους έχουν ληφθεί υπόψη περισσότερα στοιχεία που προέρχονται από τη θεωρία της εξέλιξης.

Ανάλογοι πίνακες μπορούν να σχηματιστούν και για την περίπτωση των νουκλεοτιδίων. Μάλιστα, είναι πολύ πιο συχνές οι αντικαταστάσεις μεταξύ των A, G ή των T, C παρά οι αντίστροφοι συνδυασμοί. Ένα πολύ απλό παράδειγμα πίνακα αντικατάστασης θα μπορούσε να είναι αυτό του Πίνακα 5.3.

Πίνακας 5.3: Πιθανός πίνακας αντικατάστασης για νουκλεοτίδια.

	a	t	g	c
a	20	10	5	5
t	10	20	5	5
g	5	5	20	10
c	5	5	10	20

Εκτός από τους αλγόριθμους που στηρίζονται σε dot matrix υπάρχουν και άλλοι που βασίζονται σε άλλες τεχνικές. Για παράδειγμα, η μέθοδος του δυναμικού προγραμματισμού προσφέρει αρκετές δυνατότητες, όμως η πολυπλοκότητα των αλγορίθμων που προκύπτουν είναι απαγορευτική για πολύ μεγάλες ακολουθίες.

5.2.2 Εξελικτική φυλογενετική ανάλυση

Όπως υποδηλώνει το όνομά της, η συγκεκριμένη ανάλυση ασχολείται με τη μελέτη της εξελικτικής πορείας των ειδών στο πέρασμα του χρόνου. Οι θεωρίες της εξέλιξης προσπαθούν να αξιοποιήσουν παρατηρήσεις που βασίζονται στη σύγκριση των οργανισμών και να βγάλουν συμπεράσματα για την πιθανή ύπαρξη κοινού προγόνου. Οι παρατηρήσεις αυτές προέρχονται συνήθως από τη σύγκριση ακολουθιών, κάποιες φορές όμως και από πιο γενικά χαρακτηριστικά λ.χ. την ύπαρξη ή όχι σπονδυλικής στήλης. Τα αποτελέσματα που προκύπτουν παρουσιάζονται με τη μορφή διαφόρων τύπων δένδρων.

Μία από τις βασικές αρχές στις οποίες στηρίζεται η φυλογενετική ανάλυση (phylogenetic analysis) είναι το γεγονός ότι συμβαίνουν μεταλλάξεις σε όλους τους οργανισμούς ανά συγκεκριμένα χρονικά διαστήματα στο πέρασμα των αιώνων (βλ. Εξέλιξη, Κεφάλαιο 3). Θεωρείται πως η σχέση μεταξύ του χρόνου και του αριθμού των αλλαγών που συμβαίνουν είναι γραμ-

μική και προβλέψιμη. Ωστόσο, ο ρυθμός με τον οποίο γίνονται αυτές οι αλλαγές διαφέρει όχι μόνο από οργανισμό σε οργανισμό αλλά και στα ίδια τα μακρομόρια του ίδιου οργανισμού μεταξύ τους (π.χ. νουκλεϊκά οξέα και πρωτεΐνες). Τα παραπάνω φανερώνουν το μέγεθος του προβλήματος και των δυσκολιών του.

Η φυλογενετική ανάλυση ακολουθεί τέσσερα βήματα, εκ των οποίων τα δύο πρώτα σχετίζονται με όσα αναφέρθηκαν στην ενότητα 5.2.1 και τα δύο τελευταία με το δέντρο που κατασκευάζεται. Οι αρχικές δύο εργασίες είναι η πραγματοποίηση του pairwise/multiple alignment και η αξιολόγηση αυτών με την επιλογή των κατάλληλων μοντέλων αντικατάστασης. Στη συνέχεια χρειάζεται να δημιουργηθεί το φυλογενετικό δέντρο και να εκτιμηθούν τα αποτελέσματά του.

Παράδειγμα τέτοιου δέντρου φαίνεται στο Σχήμα 5.5 [4]. Αυτό αφορά ομάδες πρωτεϊνών που είναι παράγοντες της μεταγραφής και έχουν συγκεκριμένη δομή και λειτουργία (homeodomain transcriptors). Το μήκος των κλαδιών συνηθίζεται να είναι ανάλογο με την πάροδο του χρόνου.

Υπάρχουν δύο βασικές μέθοδοι για την κατασκευή των δέντρων. Η πρώτη είναι το *ιεραρχικό clustering* και η δεύτερη η *cladistic ανάλυση*. Η επικρατούσα μέθοδος είναι η δεύτερη, καθώς έχει δειχθεί σε αρκετές περιπτώσεις ότι ανεξάρτητα στοιχεία (π.χ. παλαιοντολογικά ευρήματα) συμφωνούν με τα πορίσματα των cladistics περισσότερο. Κατά συνέπεια, στη συνέχεια θα αναφερθούν τεχνικές που ακολουθούν την ανάλυση cladistic. Δύο clustering αλγόριθμοι είναι ο UPGMA (Unweighted Pair Group Method with Arithmetic Mean) και ο neighbor-joining.

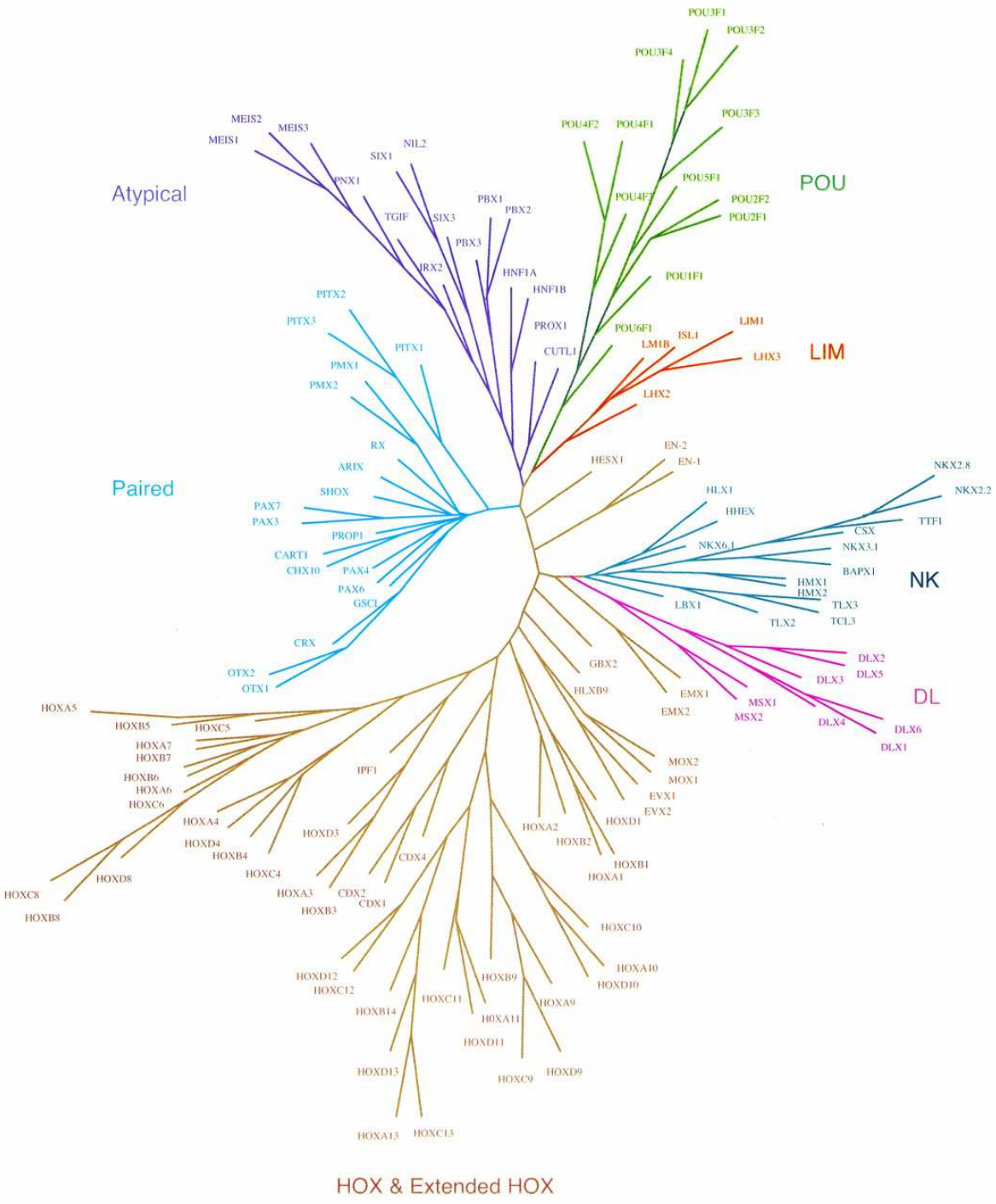
Εδώ χρειάζεται να σημειωθεί ότι ακόμη και μελέτες που χρησιμοποιούν την ίδια μέθοδο για τη φυλογενετική ανάλυση είναι δυνατόν να κατασκευάζουν διαφορετικά δέντρα. Το γεγονός αυτό υπογραμμίζει τη σπουδαία σημασία που έχει για το τελικό αποτέλεσμα η ποιότητα των αρχικών δεδομένων που προέρχονται από το alignment.

Στο Σχήμα 5.6 φαίνονται τα βασικά στοιχεία που αποτελούν ένα δέντρο που έχει προκύψει από cladistics. Ο όρος taxon αναφέρεται σε οποιοδήποτε ομάδα οργανισμών που έχει κάποιο όνομα, ενώ κλάδος - λέξη από την οποία προέρχεται και το όνομα της μεθόδου cladistics - είναι ένας μονοφυλετικός taxon.

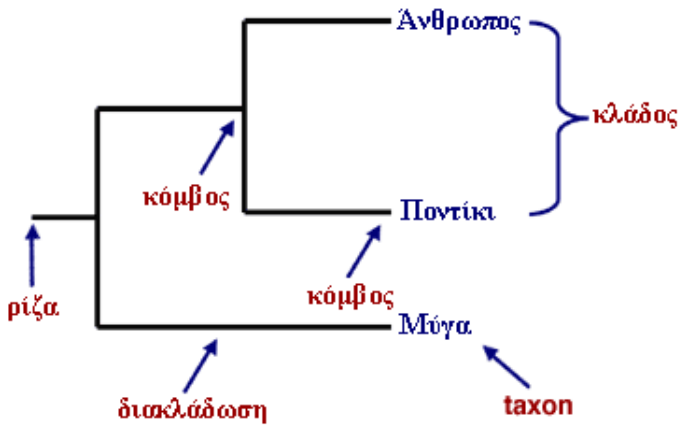
Τρεις είναι οι συνηθέστερα χρησιμοποιούμενες τεχνικές για τη δημιουργία του ζητούμενου δέντρου, οι οποίες μάλιστα εφαρμόζουν τη μέθοδο cladistics. Αυτές είναι οι maximum likelihood, maximum parsimony, neighbor-joining. Οι δύο πρώτες ανήκουν στη λεγόμενη κατηγορία τεχνικών που βασίζονται στους χαρακτήρες (character-based), σε αντίθεση με αυτές που βασίζονται στην απόσταση (distance-based). Η διαφορά των δύο κατηγοριών έγκειται στο γεγονός ότι η δεύτερη δημιουργεί το δέντρο με βάση μόνο τιμές ενδεικτικές των αποστάσεων μεταξύ των συγκρινόμενων ακολουθιών, ενώ η πρώτη λαμβάνει υπόψη της τις διαφορές μεταξύ των χαρακτήρων τους.

Η τεχνική maximum parsimony, για παράδειγμα, κατασκευάζει το δέντρο με τον ελάχιστο αριθμό αλλαγών. Αν οι ομόλογες ακολουθίες είναι οι ATCG, ATGG, TCCA, TTCA, τότε το βέλτιστο δέντρο είναι αυτό του Σχήματος 5.7. Όλα τα άλλα φυλογενετικά δέντρα χρειάζονται

[4] Πηγή: http://webpages.marshall.edu/ harrah5/big%20tree.jpg

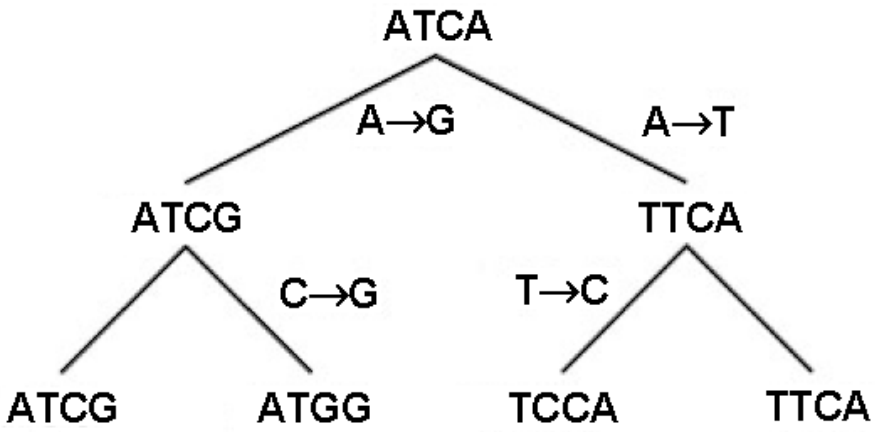


Σχήμα 5.5: Φυλογενετικό δέντρο για τους homeodomain transcriptors.



Σχήμα 5.6: Τα στοιχεία ενός φυλογενετικού δέντρου.

περισσότερες από τέσσερις αντικαταστάσεις γραμμάτων για να φτιαχθούν.



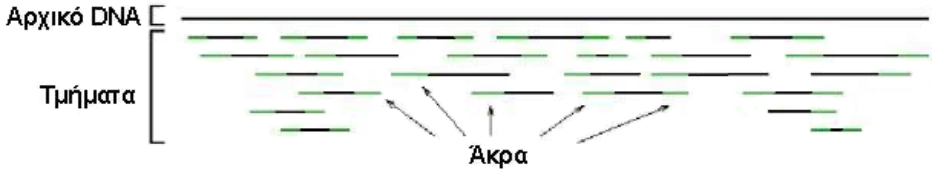
Σχήμα 5.7: Φυλογενετικό δέντρο με βάση την maximum parsimony.

Από την άλλη, η τεχνική maximum likelihood χρησιμοποιεί πιθανότητες. Υπολογίζει την πιθανότητα να συμβεί μία αλλαγή ενός γράμματος για κάθε γράμμα της ακολουθίας και με βάση αυτές βρίσκει την πιθανότητα να προκύψει μια ακολουθία παιδί από μία πρόγονό της. Η διαδικασία αυτή γίνεται για όλο το δέντρο. Από όλα τα δυνατά φυλογενετικά δέντρα βέλτιστο θεωρείται εκείνο με τη μεγαλύτερη πιθανότητα να συμβεί στην πραγματικότητα.

5.2.3 Sequence & genome assembly, ανάλυση γονιδιωμάτων

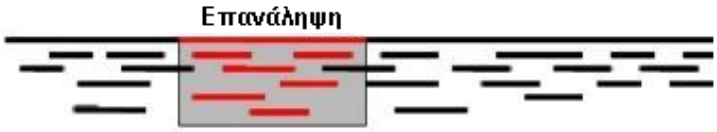
Η sequence assembly λύνει πρόβλημα που πηγάζει από τον τρόπο που γίνεται η αποκωδικοποίηση των γονιδιωμάτων. Για να γίνει εφικτή η ανάγνωση των ακολουθιών των γονιδιωμάτων, έχει χρησιμοποιηθεί η μέθοδος *shotgun*. Με αυτήν το γονιδίωμα σπάει σε πάρα πολλά κομμά-

τια μήκους όχι μεγαλύτερου από 900 νουκλεοτίδια το καθένα (Σχήμα 5.8 [5]). Αυτό συμβαίνει, επειδή τα προγράμματα που διαβάζουν τις ακολουθίες, οι *sequencers*, δεν μπορούν να δεχθούν μεγαλύτερες ως είσοδο.



Σχήμα 5.8: Οι ακολουθίες που δημιουργούνται από τη μέθοδο shotgun.

Το πρόβλημα που προκύπτει έγκειται στην επανασύνδεση αυτών των κομματιών, ώστε να προκύψει η αρχική ακολουθία. Οι δυσκολίες που εμφανίζονται για να γίνει αυτή η εργασία, η sequence assembly, οφείλονται σε διάφορους παράγοντες. Ο πιο σημαντικός είναι το γεγονός ότι υπάρχουν πολλές μικρές επαναλαμβανόμενες ακολουθίες (μήκους μέχρι και χιλιάδων νουκλεοτιδίων) κατά μήκος της αρχικής ακολουθίας (Σχήμα 5.9). Είναι πιθανό τότε να τύχει τα κομμάτια που προκύπτουν να είναι ακριβώς ίδια μεταξύ τους.



Σχήμα 5.9: Επανάληψη στο γονιδίωμα και τμήματα που δημιουργούνται.

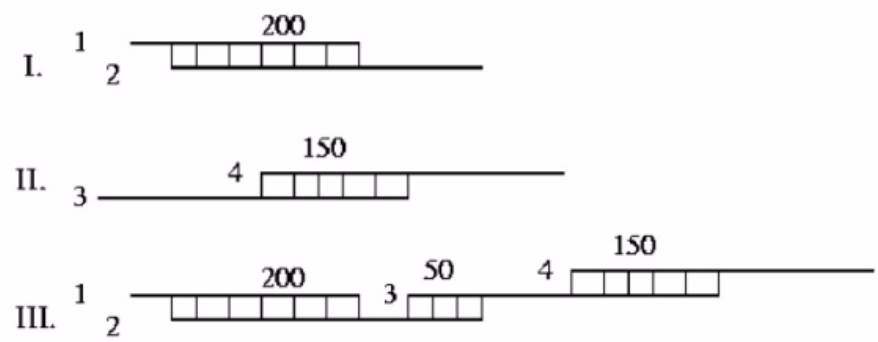
Η εργασία της επανασύνδεσης αναφέρεται και ως *genome assembly*, όταν στόχος είναι η επαναδημιουργία όχι απλά μιας ακολουθίας αλλά ολόκληρου του χρωμοσώματτος. Σε αυτήν την περίπτωση τα μικρότερα κομμάτια μπορεί σε πλήθος να φθάνουν ακόμη και τα 1000.

Για να γίνει περισσότερο κατανοητή η διαδικασία της επανένωσης των επιμέρους τμημάτων, παρατίθεται ένας greedy αλγόριθμος για την πραγματοποίησή της. Αρχικά, γίνονται όλα τα δυνατά alignments μεταξύ των μικρών τμημάτων και στη συνέχεια, επιλέγονται τα δύο που έχουν τη μεγαλύτερη επικάλυψη. Η επικάλυψη αυτή αφορά τα άκρα των δύο τμημάτων. Μάλιστα μια περιοχή συνεχόμενων αλληλεπικαλύψεων ονομάζεται *contig*. Αυτά τα δύο κομμάτια ενώνονται και η διαδικασία του alignment και της συνένωσης επαναλαμβάνονται, μέχρι να προκύψει τελικά ένα μόνο τμήμα, δηλαδή μία ενιαία ακολουθία.

Στο Σχήμα 5.10 φαίνεται ένα απλό παράδειγμα χρήσης του παραπάνω αλγορίθμου για τη συνένωση τεσσάρων μόνο τμημάτων. Από αυτά το κομμάτι 1 επικαλύπτεται με το 2 σε 200 νουκλεοτίδια, το 3 με το 4 σε 150 και το κομμάτι 2 με το 3 σε 50 νουκλεοτίδια. Στη φάση III του εν λόγω σχήματος είναι πλέον ευδιάκριτο ποια είναι η αρχική ακολουθία. Ο αλγόριθμος αυτός είναι αρκετά απλός και δε θα έβγαζε σωστά αποτελέσματα, αν στην αρχική ακολουθία

[5] Τα Σχήματα 5.8 - 5.11 έχουν γίνει με βάση την πηγή http://www.cbcb.umd.edu/research/assembly_primer.shtml

υπήρχαν επαναλήψεις, όπως αναφέρθηκε στα προηγούμενα.



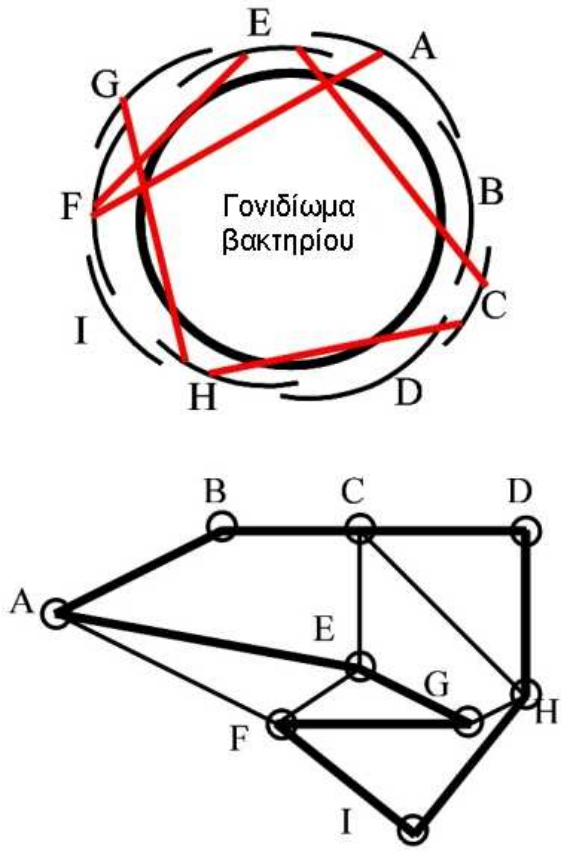
Σχήμα 5.10: Τα βήματα ενός απλού greedy αλγορίθμου για sequence assembly τεσσάρων τμημάτων ακολουθίας.

Η διαδικασία sequence assembly πραγματοποιείται από προγράμματα ηλεκτρονικών υπολογιστών. Αν και υπάρχουν αρκετά και μάλιστα ελεύθερα στο διαδίκτυο, τα πιο διάσημα είναι το Celera Assembler και το Arachne. Το πρώτο αναπτύχθηκε το διάστημα 1998-2002 από την εταιρεία Celera Genomics, η οποία πήρε μέρος και στην αποκωδικοποίηση του γονιδιώματος του ανθρώπου. Ένας από τους πρωτοπόρους της ήταν ο Gene Myers, ο οποίος τώρα βρίσκεται στο UC Berkeley. Το Arachne ξεκίνησε στο MIT το 2000 ως διδακτορική θέση του Serafim Batzoglou ο οποίος είναι σήμερα στο Stanford University. Τα παραπάνω προγράμματα είναι σύγχρονα και αντικατέστησαν προηγούμενα, όπως τα Phrap, TIGR Assmbler, λόγω των αυηξημένων απαιτήσεων για την αποκωδικοποίηση μεγάλων γονιδιωμάτων.

Τα περισσότερα προγράμματα που εκτελούν sequence assembly σχηματίζουν ένα γράφο Hamilton. Κόμβοι του είναι τα κομμάτια που είναι επιθυμητό να επανασυνδεθούν, ενώ οι ακμές του σχηματίζονται ανάμεσα στα κομμάτια με αλληλοεπικαλυπτόμενες υπακολουθίες. Στο Σχήμα 5.11 φαίνεται ένα παράδειγμα τέτοιου γράφου. Οι ακμές που είναι λιγότερο έντονες δείχνουν τις λανθασμένες ακμές που θα υπεισέρχονταν, αν επαναλαμβανόμενα μέρη της αρχικής ολόκληρης ακολουθίας θεωρούνταν αλληλεπικαλύψεις μεταξύ των επιμέρους κομματιών. Αυτά φαίνονται επίσης με ευθείες κόκκινες γραμμές και στον κύκλο που αντιπροσωπεύει το γονιδίωμα ενός βακτηρίου (το γονιδίωμά του είναι ένα κυκλικό μόριο DNA, βλ. Κεφάλαιο 2).

Η αποκωδικοποίηση των γονιδιωμάτων αρκετών οργανισμών τις τελευταίες δύο δεκαετίες (βλ. Κεφάλαιο 2) προσφέρει μεταξύ των άλλων την πρόκληση της ανάλυσης αυτών. Αναγκάζει, δηλαδή, την έρευνα να στραφεί όχι μόνο στη μελέτη του ρόλου μεμονωμένων γονιδίων σε συγκεκριμένες βιολογικές διαδικασίες αλλά και στη διερεύνηση συσχετίσεων μεταξύ γονιδίων, στον ακριβή προσδιορισμό εκείνων που είναι υπεύθυνα για την παραγωγή πρωτεϊνών και στην εξακρίβωση της δράσης αυτών των πρωτεϊνικών μορίων.

Είναι αλήθεια ότι δεν είναι γνωστή ή προβλέψιμη η λειτουργία για περίπου το ένα τρίτο των γονιδίων ενός οργανισμού από εκείνους με αποκωδικοποιημένο γονιδίωμα. Ακόμα και για τα υπόλοιπα δύο τρίτα συχνά οι γνώσεις περί του ρόλου τους είναι γενικές και αφηρημένες.



Σχήμα 5.11: Ο γράφος Hamilton για εννιά κομμάτια ακολουθιών ενός γονιδιώματος.

Χαρακτηριστικό είναι το παράδειγμα του βακτηρίου Escherichia coli K12 το οποίο θεωρείται ο περισσότερο μελετημένος οργανισμός και για τον οποίο τουλάχιστον το 40% των γονιδίων του έχει άγνωστη δραστηριότητα.

Η συγκριτική μελέτη των γονιδίων των διαφόρων οργανισμών είναι δυνατό να προσφέρει σημαντική βοήθεια στη διαλεύκανση του ρόλου αυτών. Πειραματικά αλλά και από τη θεωρία της εξέλιξης έχει υποστηριχθεί ότι γονίδια που είναι υπεύθυνα για το ίδιο χαρακτηριστικό σε διαφορετικούς οργανισμούς έχουν αρκετές φορές παρεμφερή συμπεριφορά. Έτσι, γνώσεις που έχουν αποκτηθεί σε έναν πιο απλό και καλύτερα μελετημένο οργανισμό μπορούν κατάλληλα να μεταφερθούν σε έναν ανώτερο.

Οι διαφορές που κάνουν την ανάλυση των γονιδιωμάτων πολύ πιο αποδοτική σε σχέση με παλαιότερα είναι κυρίως δύο. Από τη μια πλευρά, είναι διαθέσιμες οι ακολουθίες για τα γονίδια, οπότε είναι πιο πλούσιο και συγκεκριμένο το υπό μελέτη υλικό. Από την άλλη, η πληροφορική διαθέτει στις βιοεπιστήμες πολλά εξαιρετικά χρήσιμα υπολογιστικά εργαλεία, για να πραγματοποιηθούν οι επιθυμητές έρευνες.

5.2.4 Προσδιορισμός σχήματος ή τρισδιάστατης δομής από ακολουθία

Δύο εργασίες που έχουν επίσης σπουδαία αξία για τους ερευνητές των βιοεπιστημών είναι η εύρεση της δευτεροταγούς ή τριτοταγούς δομής μορίων για τα οποία είναι γνωστή η ακολουθία των δομικών τους στοιχείων. Πιο συγκεκριμένα, ενδιαφέρονται για τον προσδιορισμό του σχήματος του RNA και της δομής στο χώρο των πρωτεϊνικών μορίων δοθείσης της ακολουθίας νουκλεοτιδίων ή αμινοξέων αντίστοιχα. Στη συνέχεια εξετάζεται κάθε τέτοια περίπτωση ξεχωριστά.

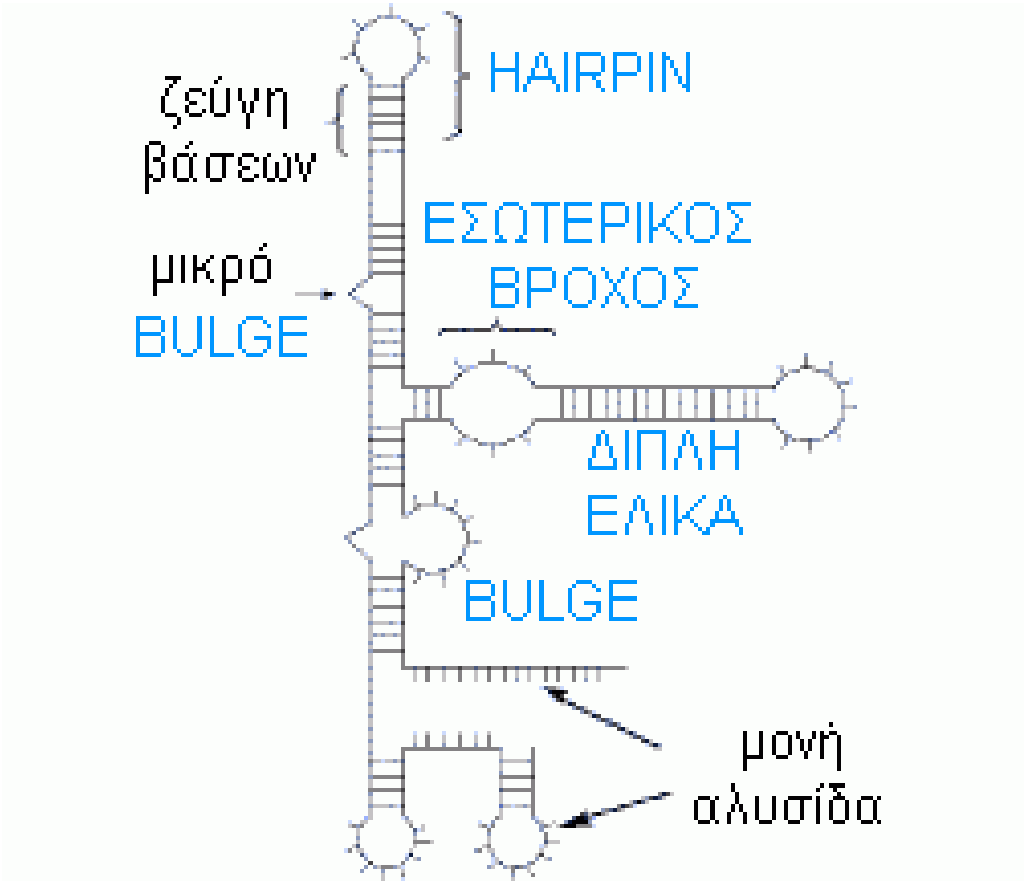
Όπως έχει αναφερθεί (Κεφάλαιο 2), το RNA –σε αντίθεση με το DNA– αποτελείται από μία μόνο πολυνουκλεοτιδική αλυσίδα. Η ακολουθία των δομικών της λίθων είναι η πρωτοταγής της δομή. Δευτεροταγής είναι η δομή που σχηματίζεται από τις αναδιπλώσεις του μορίου. Με άλλα λόγια, το μονόκλωνο μόριο του RNA δεν έχει το σχήμα μιας ευθείας αλλά ένα αρκετά πιο πολύπλοκο.

Στο Σχήμα 5.12 [6] φαίνεται ένα παράδειγμα τέτοιου σχήματος. Συχνά παρατηρείται ότι σε ορισμένα σημεία του RNA δημιουργούνται κάποιοι βρόχοι ή σχηματισμοί οι οποίοι είναι structural motifs (Κεφάλαιο 4). Αυτά μπορεί να περιλαμβάνουν εσωτερικούς βρόχους (internal loops), τμήματα διπλής έλικας, εξογκώματα (bulges) ή hairpins, όπως δείχνει το εν λόγω σχήμα.

Είναι ενδιαφέρον να δει κανείς σε ποια σημεία και κάτω από ποιες συνθήκες εμφανίζονται τα παραπάνω ή άλλα παρεμφερή motifs και γενικότερα ποιοι παράγοντες επηρεάζουν τη δευτεροταγή δομή του RNA. Η ιδιότητα της συμπληρωματικότητας των αζωτούχων βάσεων (Κεφάλαιο 2) και η ενέργεια των δεσμών που σχηματίζονται ανάμεσα σε αυτές και καθορίζει τη σταθερότητα της έλξης τους είναι οι πιο καθοριστικοί.

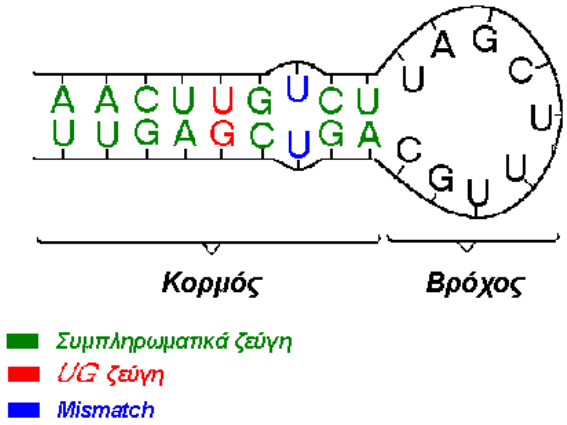
Στο Σχήμα 5.13 [7] φαίνεται ένα παράδειγμα του τρόπου με τον οποίο η συμπληρωματικότητα των βάσεων κατευθύνει τη δευτεροταγή δομή. Κατά μήκος της αλυσίδας του RNA

[6] Πηγή: http://darwin.nmsu.edu/ molb470/fall2003/Projects/samara/2ryrnan.gif
[7] Πηγή: http://www.cs.man.ac.uk/ gowrishv/beta-release/manual/node98.html



Σχήμα 5.12: Structural motifs του RNA.

υπάρχουν περιοχές ανάμεσα στις οποίες ασκούνται ελκτικές δυνάμεις, αν αυτή αναδιπλωθεί στο κατάλληλο σημείο. Εκτός από τα ζεύγη βάσεων που πρότειναν στο μοντέλο τους για τη δομή των DNA, RNA οι Watson, Crick (A-U, G-C, βλ. Κεφάλαιο 2), δυνάμεις είναι δυνατό να ασκούνται και ανάμεσα στις U-G. Ωστόσο, αυτή η εξαίρεση δε θα απασχολήσει άλλο την ανάλυση στα επόμενα.



Σχήμα 5.13: Ένας τρόπος σχηματισμού motif.

Υπάρχει, όμως, ένα ακόμη λεπτό ζήτημα. Είναι δυνατόν να υπάρχουν περισσότερα από ένα ζευγάρια περιοχών τα οποία να είναι συμπληρωματικά μεταξύ τους. Αυτό σημαίνει ότι θα υπάρχουν πολλές δυνατές αναδιπλώσεις του μορίου. Μάλιστα, δεν είναι απίθανο η αναδίπλωσή του σε κάποιο να αποκλείει την ταυτόχρονη αναδίπλωσή του και σε κάποιο άλλο. Σε αυτήν την περίπτωση τελικά επιλέγεται από τη φύση να σχηματιστεί ο πιο σταθερός δεσμός.

Από τα παραπάνω συμπεραίνεται ότι ένα κατάλληλο μοντέλο πρόβλεψης της δευτεροταγούς δομής χρειάζεται να λαμβάνει υπόψη και τους δύο αυτούς παράγοντες. Θα πρέπει, δηλαδή, να βρίσκει όλες τις συμπληρωματικές περιοχές κατά μήκος της ακολουθίας των βάσεων του RNA και από αυτές να επιλέγει εκείνες που δημιουργούν τους πιο ισχυρούς χημικούς δεσμούς (δεσμοί υδρογόνου, Κεφάλαιο 2).

Υπάρχουν διάφορες προσεγγίσεις για την επίλυση αυτού του προβλήματος. Εδώ θα αναφερθεί η πιο δημοφιλής, που είναι η μοντελοποίηση και λύση η οποία βασίζεται στις τυπικές γραμματικές και πιο συγκεκριμένα, στις context-free και stochastic context-free.

Οι κανόνες της context-free γραμματικής του Σχήματος 5.14 είναι ικανοί να παράγουν την ακολουθία μιας αλυσίδας RNA. Τα τερματικά σύμβολα είναι οι τέσσερις αζωτούχες βάσεις. Ο πρώτος μάλιστα κανόνας είναι εκείνος που επιτρέπει την παραγωγή συμπληρωματικών περιοχών βάσεων, ενώ μέσω των υπολοίπων σχηματίζεται και το εναπομείναν τμήμα της αλυσίδας. Εξάλλου, οι context-free γραμματικές είναι ιδανικές για να εκφράσουν παλινδρομικές ακολουθίες.

Εκείνο που μένει ακόμη να ειπωθεί είναι ο τρόπος με τον οποίο επιλέγεται ένας κανόνας παραγωγής σε σχέση με έναν άλλο, έτσι ώστε να δημιουργηθεί το πιο σταθερό μοντέλο. Επαναλαμβάνεται ότι από άποψη βιολογίας αυτό έχει να κάνει με την ενέργεια του κάθε

```
S -> aSu | uSa | gSc | cSg

S -> aS | uS | gS | cS
S -> Sa | Su | Sg | Sc

S -> ε

S -> aSu -> agScu -> aguSacu
  -> agugSacu -> aguguSacu
  -> aguguacu
```

Σχήμα 5.14: Κανόνες context-free γραμματικής για τη δευτεροταγή δομή του RNA και ενδεικτικό παράδειγμα.

δεσμού.

Έχουν προταθεί αρκετοί αλγόριθμοι και μέθοδοι για να λύσουν το συγκεκριμένο πρόβλημα. Κάποιοι στηρίζονται στον δυναμικό προγραμματισμό και πετυχαίνουν μάλιστα κυβική πολυπλοκότητα, την καλύτερη ως τώρα δυνατή. Ιδιαίτερα ενδιαφέρουσα είναι η προσέγγιση με τις stochastic context-free (SCFG) γραμματικές. Συνοπτικά μπορεί να λεχθεί ότι αυτές αντιστοιχίζουν στους κανόνες παραγωγής πιθανότητες και με βάση αυτές επιλέγονται εκείνοι που θα εκτελεστούν.

Το πρόβλημα του προσδιορισμού της τρισδιάστατης πρωτεϊνικής δομής από την πρωτοταγή, δηλαδή την ακολουθία των αμινοξέων, είναι περισσότερο πολύπλοκο. Στο θέμα αυτό δεν θεωρείται πως υπάρχει ένας κυρίαρχος τρόπος λύσης. Εξάλλου, πρόκειται περισσότερο για προσεγγίσεις της λύσης παρά για σίγουρους τρόπους αντιμετώπισης λόγω της δυσκολίας του. Η αξιοπιστία αυτών φθάνει μέχρι περίπου 70%, γεγονός που δείχνει ότι υπάρχει αρκετή δουλειά ακόμη να γίνει στο χώρο. Ανάμεσα μάλιστα στις ερευνητικές ομάδες που ασχολούνται με το συγκεκριμένο πρόβλημα είναι και εκείνη της IBM, που δημιούργησε το δημοφιλές πρόγραμμα-νικητή στο σκάκι.

Κάθε χρόνο διεξάγεται ένας διαγωνισμός ανάμεσα σε εκείνους που ευελπιστούν να δώσουν απάντηση στην εύρεση της δομής των πρωτεϊνικών μορίων στο χώρο. Αυτός ονομάζεται CASP (Critical Assessment of Techniques for Protein Structure Prediction). Οι διαγωνιζόμενοι καλούνται να προβλέψουν τη δομή πρωτεΐνης από την ακολουθία της. Η πρωτεΐνη αυτή είναι κάποια, της οποίας η δομή έχει πρόσφατα ανακαλυφθεί πειραματικά, αλλά δεν έχει προλάβει να ανακοινωθεί στο ευρύ κοινό.

Οι βασικοί λόγοι για τους οποίους είναι σημαντικό να μπορεί να προβλεφθεί η τριτοταγής ή τεταρτοταγής δομή (Κεφάλαιο 2) του εν λόγω μορίου μέσω προγράμματος υπολογιστή είναι

δύο. Ο κύριος είναι το γεγονός ότι η εύρεση με άλλες μεθόδους είναι πολύ απαιτητική. Πειραματικές τεχνικές, όπως η X-ray crystallography, προσδιορίζουν με ακρίβεια το ζητούμενο, αλλά χρειάζονται χρονοβόρα και εξαιρετικά δύσκολη διαδικασία. Χαρακτηριστικό είναι ότι ενώ η βάση SWISS-PROT έχει περίπου 87000 εγγραφές ακολουθιών πρωτεϊνών, η PDB δεν έχει ούτε 13000 εγγραφές για την αντίστοιχη δομή τους στο χώρο (Κεφάλαιο 4). Ο δεύτερος λόγος έχει τονισθεί και άλλες φορές στην εργασία. Πιστεύεται ότι η πρόβλεψη αυτή αξίζει τον κόπο, επειδή η διάταξη στο χώρο προσδιορίζει σε μεγάλο βαθμό και τη λειτουργία της πρωτεΐνης.

Οι παράγοντες που επηρεάζουν τη δομή των πρωτεϊνών είναι αρκετοί. Ένας από αυτούς σχετίζεται με τη σχέση των αμινοξέων με το νερό. Υπάρχουν εκείνα που είναι *υδρόφοβα* και άλλα που είναι *υδρόφιλα*. Από τις ονομασίες αντιλαμβάνεται κανείς ότι τα πρώτα δεν έλκονται από το υδάτινο περιβάλλον, ενώ τα δεύτερα το αντίθετο. Έτσι, όταν μια πρωτεΐνη βρεθεί σε χώρο με έντονη την παρουσία του νερού, τα υδρόφοβα τμήματα τείνουν να "κρυφθούν" προς το εσωτερικό της, ενώ τα υδρόφιλα έρχονται σε επαφή με το νερό.

Για παράδειγμα, το Σχήμα 5.15 [8] έχει πληροφορίες που αφορούν την πρωτεΐνη hen egg white lysozyme. Στο διάγραμμα έχει σχεδιαστεί η υδροφοβικότητα κάθε αμινοξέος της αλυσίδας. Ελάχιστο εμφανίζεται στα 17, 44, 70, 93, 117. Η δεύτερη εικόνα είναι η διάταξη της πρωτεΐνης στο χώρο. Με πιο έντονη γραμμή σκιαγραφούνται οι περιοχές που αντιστοιχούν σε ελάχιστο του πρώτου διαγράμματος (τοπικό ή ολικό). Παρατηρείται ότι το μόριο στρίβει σχεδόν σε κάθε ένα από τα σημεία που υπάρχει ελάχιστο. Αυτό δεν αποτελεί έναν απαραβίαστο κανόνα, αλλά είναι μια χρήσιμη πληροφορία για την πρόβλεψη της τριτοταγούς δομής.

Υπάρχουν αρκετές μέθοδοι για την πρόβλεψη της τριτοταγούς δομής. Αρκετές προσπάθειες έχουν γίνει να προσδιοριστεί με βάση φυσικές ιδιότητες. Να υπολογιστεί, δηλαδή, μία συνάρτηση ενέργειας της οποίας η ελαχιστοποίηση θα δίνει τη λύση για το ποια είναι η διάταξη της πρωτεΐνης. Ωστόσο, αυτή η προσέγγιση δεν έχει δώσει προς το παρόν ικανοποιητικά αποτελέσματα είτε λόγω ακατάλληλης συνάρτησης είτε λόγω των αλγορίθμων που χρησιμοποιούνται για την εύρεση του ολικού ελαχίστου και οι οποίοι παγιδεύονται σε τοπικά ελάχιστα.

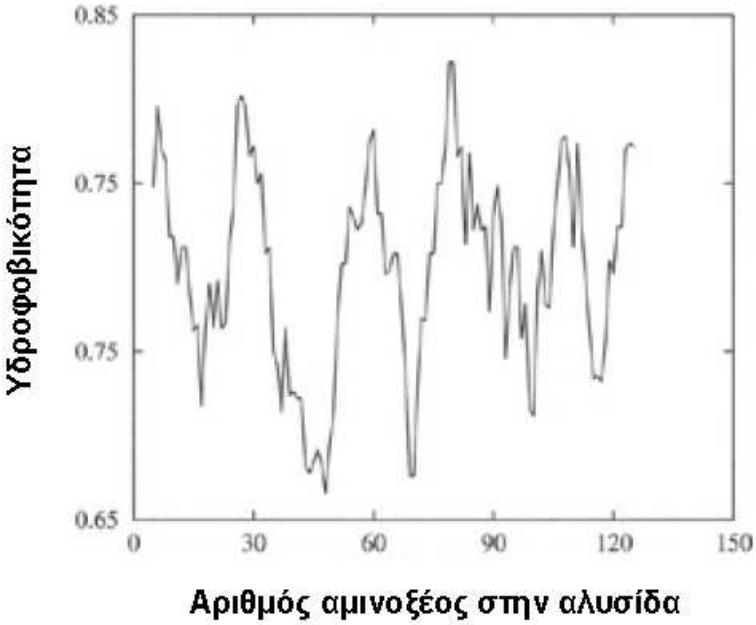
Μία κατηγορία μεθόδων στηρίζεται περισσότερο στην εμπειρική γνώση. Στόχος είναι να εξαχθούν συμπεράσματα για την άγνωστη δομή με βάση άλλη ή άλλες που είναι γνωστές και οι ακολουθίες τους έχουν σημαντικές ομοιότητες με την ακολουθία της ζητούμενης. Στο Σχήμα 5.16 [9] φαίνεται ένα διάγραμμα που υποστηρίζει την εφαρμογή αυτής της λογικής. Όπως είναι και το αναμενόμενο, αν μοιάζουν αρκετά οι ακολουθίες της άγνωστης και της γνωστής δομής, παρεμφερής θα είναι και η διάταξη των ατόμων στις τρεις διαστάσεις.

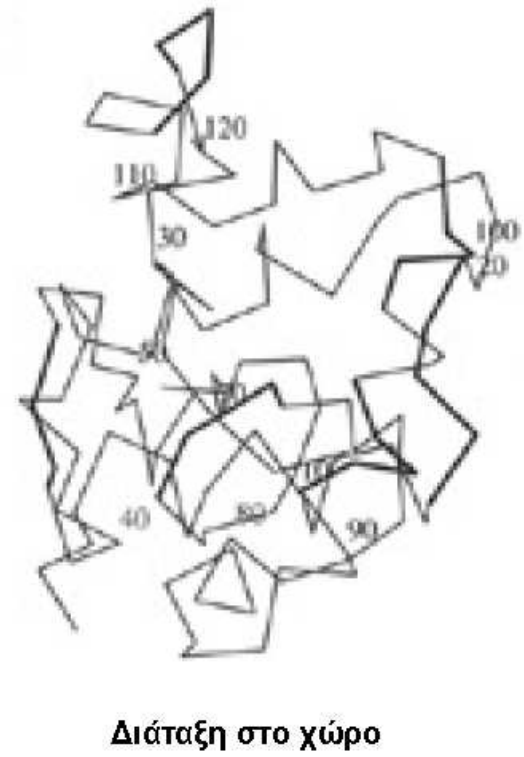
Τα αποτελέσματα της homology modelling, τεχνικής που βασίζεται στη μέθοδο που μόλις περιγράφηκε, είναι συγκρίσιμα με αυτά που παράγουν πειραματικές τεχνικές χαμηλής ανάλυσης. Στο Σχήμα 5.17 [10] εικονίζεται με κόκκινο χρώμα το μοντέλο για τη διάταξη μιας πρωτεΐνης και με γκρι η αληθινή της διάταξη.

[8] Πηγή: [45]

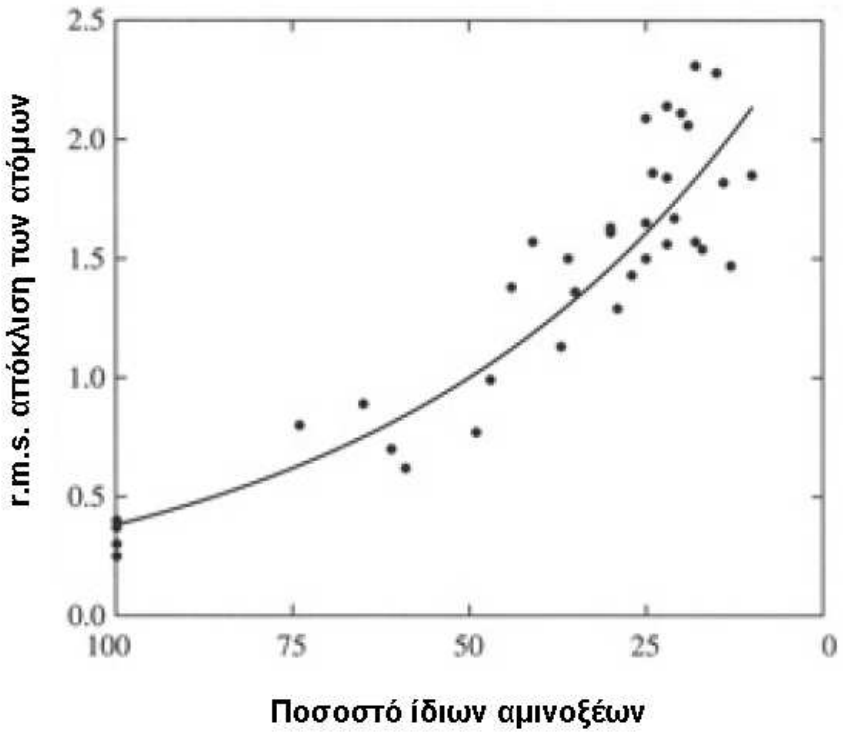
[9] Πηγή: C. Clothia and A.M. Lesk, Relationship between the divergence of sequence and structure in proteins, The EMBO Journal 5, 1986, 823-6

[10] Πηγή: http://www.pdg.cnb.uam.es/cursos/Oeiras2004/practicas/P_homology/





Σχήμα 5.15: Στοιχεία για την πρωτεΐνη hen egg white lysozyme.



Σχήμα 5.16: Στοιχεία για 32 ζευγάρια ομόλογων πρωτεϊνών.

5.3 Μοντέλα αναπαράστασης, γλώσσες προγραμματισμού

Ως τώρα στην περιγραφή κάθε λειτουργίας αναφέρεται τουλάχιστον ένας αλγόριθμος που ακολουθείται για την επιτυχημένη εφαρμογή της. Επιπλέον, έχει αναφερθεί και περίπτωση χρήσης γράφων (Παράγραφος 5.2.3), καθώς και κανονικών γραμματικών (Παράγραφος 5.2.4) για την αναπαράσταση και λύση ενός προβλήματος.

Στην παρούσα ενότητα αναφέρονται παραδείγματα δύο μοντέλων που χρησιμοποιούνται αρκετά συχνά για την ανακάλυψη της λύσης των προβλημάτων, τα hidden Markov models και τα νευρωνικά δίκτυα. Αυτά έχουν εφαρμογή σε περισσότερες από μία λειτουργίες. Επιπλέον, αναφέρεται ο ρόλος που έχουν οι script γλώσσες προγραμματισμού στον τομέα της βιοπληροφορικής.

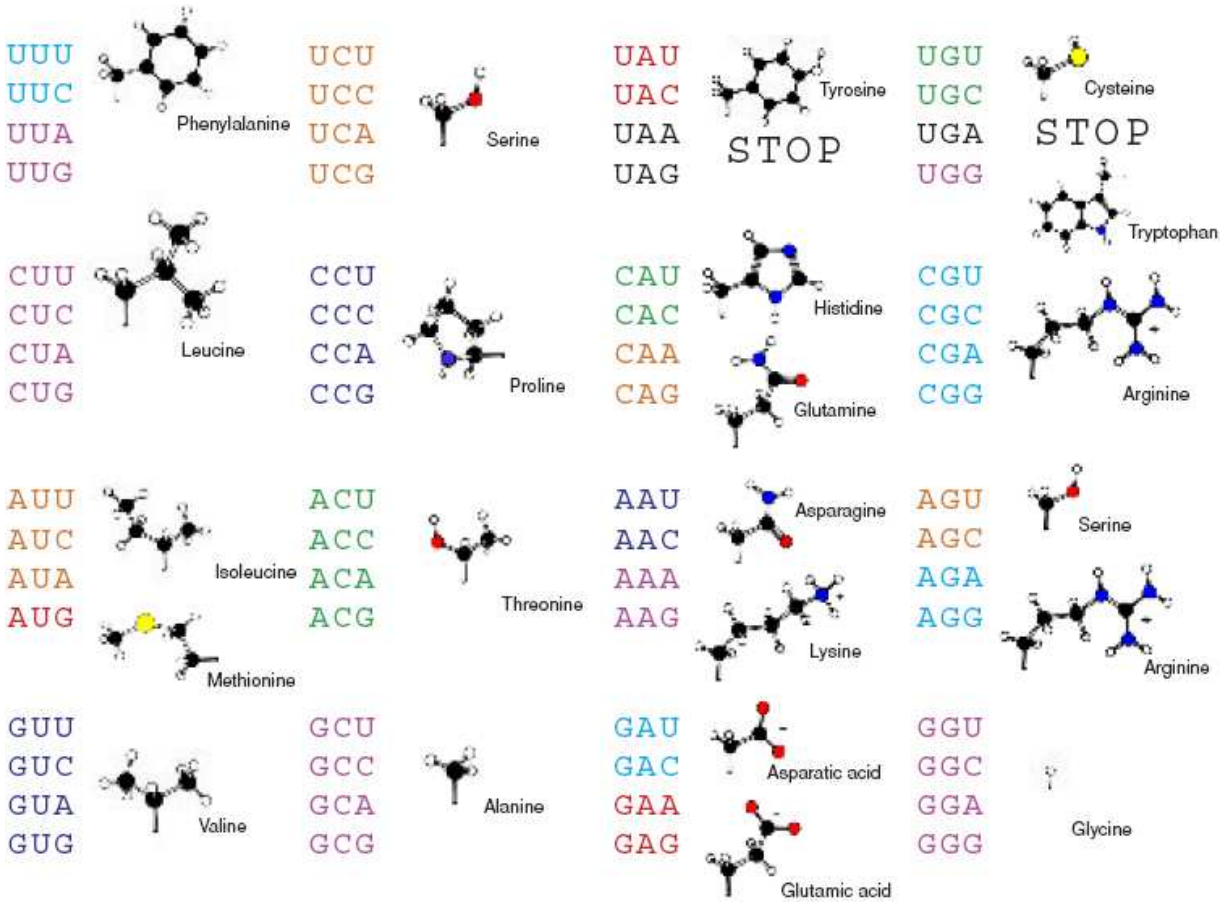
Επειδή στις ενότητες 5.3.2 και 5.3.3 παρουσιάζονται παραδείγματα που χρησιμοποιούν το γενετικό κώδικα, για λόγους ευκολίας το Σχήμα 5.18 [11] τον υπενθυμίζει. Στο Κεφάλαιο 3 έχει δοθεί πιο αναλυτική περιγραφή του.

5.3.1 Hidden Markov Models (HMM)

Ένα hidden Markov model (HMM) είναι ένα στοχαστικό αυτόματο πεπερασμένων καταστάσεων (stochastic FST). Στη συγκεκριμένη ενότητα δε γίνεται εισαγωγή στη θεωρία αυτών,

[11] Το Σχήμα αυτό όπως και το 5.20 προέρχονται από το [3]

Σχήμα 5.17: Μοντέλο της homology modelling (κόκκινο) και πραγματική δομή (γκρι).



Σχήμα 5.18: Ο γενετικός κώδικας.

αλλά παρουσιάζεται η χρησιμότητά τους για τη λύση προβλημάτων των βιοεπιστημών μέσω ενός παραδείγματος. Πιο αναλυτικό παράδειγμα μπορεί να βρεθεί στο κεφάλαιο του A. Krogh, An introduction to hidden Markov models for biological sequences που υπάρχει στο [52].

Έστω ότι ζητούμενη είναι η περιγραφή του DNA structural motif που ακολουθεί:

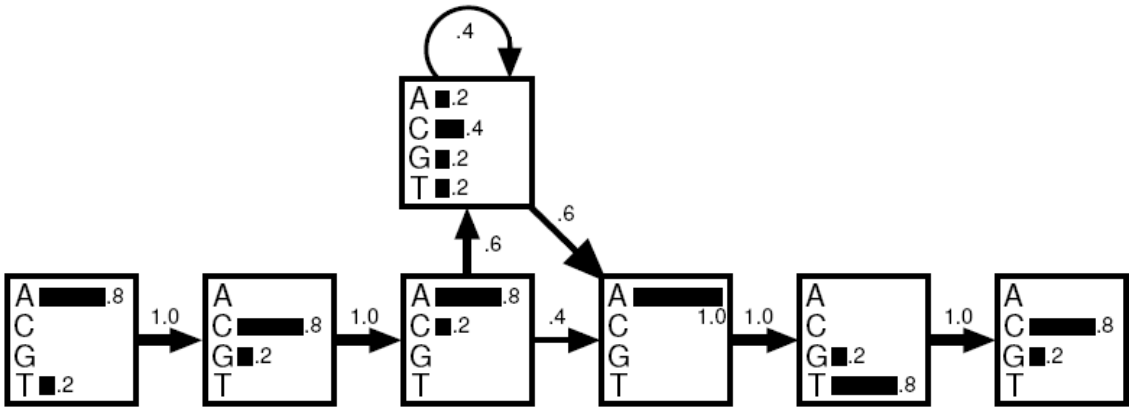
```
1.        A C A - - - A T G

2.        T C A A C T A T C

3.        A C A C - - A G C

4.        A G A - - - A T C

5.        A C C G - - A T C
```



Σχήμα 5.19: Το HMM που αντιστοιχεί στο παράδειγμα.

Σε αυτή την ομάδα αντιστοιχεί το HMM του Σχήματος 5.19. Η διαφορά από ένα κλασικό αυτόματο είναι η ύπαρξη των πιθανοτήτων. Οι έξι καταστάσεις της κάτω γραμμής του μοντέλου αντιστοιχίζονται στα έξι πρώτα και έξι τελευταία γράμματα των ακολουθιών. Η επιπλέον κατάσταση είναι εκείνη που περιγράφει την ύπαρξη των πιθανών ενδιάμεσων κενών μιας ακολουθίας. Για κάθε γράμμα η πιθανότητα προκύπτει ως το κλάσμα των εμφανίσεών του σε μια θέση, δια του 5. Για την επιπλέον κατάσταση η πιθανότητα να εμφανιστεί A ή T ή G είναι $\frac{1}{5}$, ενώ για το C βγαίνει $\frac{2}{5}$.

Επιπλέον, οι 3 από τις 5 ακολουθίες μπαίνουν στην επιπλέον κατάσταση, αφού έχουν τουλάχιστον ένα γράμμα στις τρεις μεσαίες θέσεις. Για αυτό, η αντίστοιχη πιθανότητα είναι $\frac{3}{5}$. Η πιθανότητα εξόδου από την κατάσταση αυτή υπολογίζεται ως εξής: φεύγει η ακολουθία 3, όταν έχει τοποθετήσει το C, φεύγει η ακολουθία 5, όταν έχει τοποθετήσει το G και η ακολουθία 2, όταν έχει τοποθετήσει τα C, T, ενώ μένει η ακολουθία 2 για να τοποθετήσει το C και μένει η ίδια ακολουθία για να τοποθετήσει το T. Έτσι, στις 3 από τις 5 περιπτώσεις

γίνεται έξοδος, ενώ στις 2 από τις 5 παραμένουν στην ίδια κατάσταση, δηλαδή η πιθανότητα εξόδου είναι 0.6 και η πιθανότητα παραμονής 0.4 .

Το πλεονέκτημα του HMM σε σχέση με μια κανονική γραμματική είναι εμφανές. Δε δηλώνει απλά όλες τις εναλλακτικές διατάξεις των γραμμάτων της ακολουθίας αλλά και την πιθανότητα να βρεθεί το καθένα στην κάθε θέση. Μια κανονική γραμματική για το παράδειγμα που δόθηκε είναι:

```
[AT] [CG] [AC] [ACGT]* A [TG] [GC]
```

Τα HMM βρίσκουν εφαρμογή σε αρκετές περιπτώσεις που ενδιαφέρουν τις βιοεπιστήμες. Η πιο συνηθισμένη είναι η εύρεση υπακολουθιών των νουκλεϊκών οξέων ή των πρωτεϊνών που έχουν ιδιαίτερη βιολογική αξία, όπως είναι τα motifs και τα patterns.

5.3.2 Νευρωνικά δίκτυα

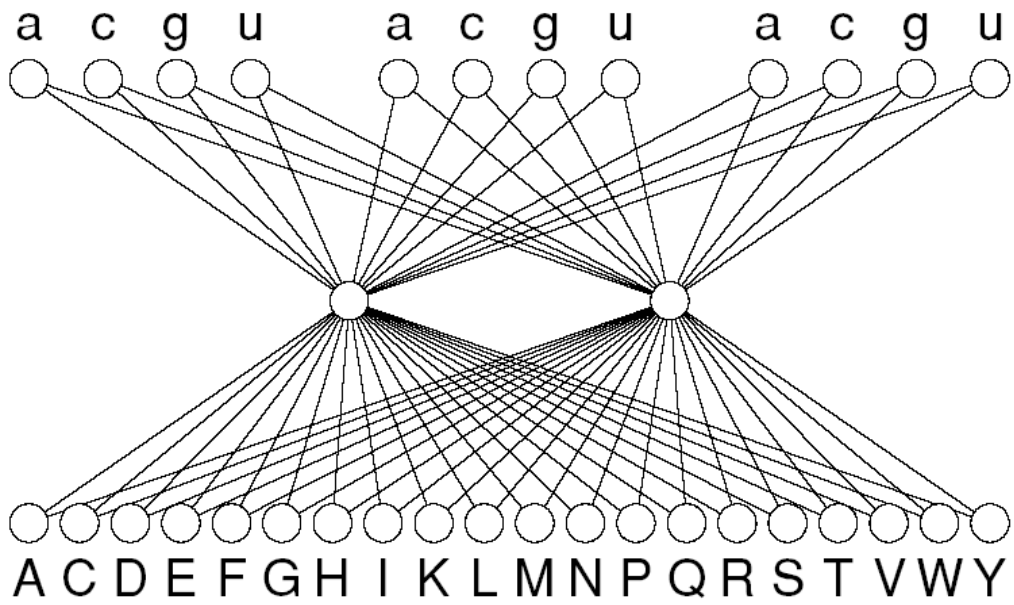
Τα νευρωνικά δίκτυα είναι μοντέλα που στηρίζονται στη λογική της μίμησης της κατανεμημένης λειτουργίας του ανθρώπινου εγκεφάλου. Το κύριο χαρακτηριστικό τους είναι η δυνατότητά τους να μαθαίνουν. Όπως και στην προηγούμενη παράγραφο, έτσι και εδώ στόχος δεν είναι μια εισαγωγή στον τομέα αυτό, αλλά η επίδειξη με ένα παράδειγμα του τρόπου με τον οποίο μπορούν να χρησιμεύσουν αυτά τα μοντέλα για την επίλυση θεμάτων των βιοεπιστημών.

Στο Σχήμα 5.20 φαίνεται η αρχιτεκτονική ενός νευρωνικού δικτύου. Σκοπός του είναι η αντιστοίχιση του κατάλληλου κωδικονίου νουκλεϊκού οξέος σε αμινοξύ πρωτεΐνης (Κεφάλαιο 3). Ο πίνακας του γενετικού κώδικα επαναλαμβάνεται και στο Σχήμα 5.18 για ευκολία. Υπάρχουν 12 είσοδοι χωρισμένες σε τρεις ομάδες, ώστε από καθεμία να είναι μία ακριβώς είσοδος ενεργή πάντα. Έτσι, σχηματίζεται το κωδικόνιο. Οι έξοδοι είναι 20, δηλαδή όσα και τα δυνατά αμινοξέα.

Όπως δείχνει και το Σχήμα 5.20, απαιτούνται $12 \cdot 2 + 2 \cdot 20 = 64$ βάρη και $2 + 20 = 22$ κατώφλια, ενώ η υλοποίηση χρειάζεται τουλάχιστον δύο κόμβους στο κρυμμένο επίπεδο. Ο προσδιορισμός των συντελεστών γίνεται με τον αλγόριθμο backpropagation, έτσι ώστε ένας μόνο κόμβος εξόδου να εμφανίζεται σε κάθε περίπτωση νικητής. Αυτό σημαίνει πως η έξοδός του πρέπει να είναι στο 1 και όχι στο 0, όπως των υπολοίπων κόμβων εξόδου, όταν στην είσοδο εμφανίζεται το κατάλληλο κωδικόνιο, δηλαδή είναι στο 1 οι αντίστοιχες τρεις είσοδοι.

Για παράδειγμα, έστω ότι το γράμμα A συμβολίζει την αλανίνη. Τότε, η αλανίνη είναι το αμινοξύ-απάντηση του νευρωνικού δικτύου, όταν αυτό εμφανίζει έξοδο 10000000000000000000. Αν έχει γίνει σωστά η εκπαίδευση, αυτή η απάντηση θα πρέπει να εμφανίζεται σε τέσσερις περιπτώσεις εισόδου. Αυτές είναι οι 0010 0100 0001 (GCU), 0010 0100 0100 (GCC), 0010 0100 1000 (GCA), 0010 0100 0010 (GCG).

Για την εν λόγω εκπαίδευση ο κύκλος για κάθε κωδικόνιο επαναλαμβάνεται αντιστρόφως ανάλογα προς το συνολικό αριθμό των κωδικονίων που αντιστοιχούν σε κάθε αμινοξύ. Έτσι, καθένα από τα έξι κωδικόνια της λευκίνης εκπαιδεύτηκε έξι φορές λιγότερο από το ένα και μοναδικό της μεθιονίνης.



Σχήμα 5.20: Νευρωνικό δίκτυο για τη μάθηση του γενετικού κώδικα.

5.3.3 Perl

Για τις περισσότερες από τις λειτουργίες που περιγράφηκαν στην ενότητα 5.2 υπάρχουν διαθέσιμα προγράμματα και λογισμικά πακέτα που τις εκτελούν. Συχνά, όμως, είναι απαραίτητο να χρησιμοποιήσει κανείς περισσότερα από ένα, λ.χ. να πραγματοποιήσει alignments μεταξύ διαφόρων ακολουθιών (με το CLUSTAL) και μετά μια φυλογενετική ανάλυση για κάποιες από αυτές σύμφωνα με τα αποτελέσματα από το προηγούμενο στάδιο (με το PHYLIP) [6].

Σε τέτοιες περιπτώσεις είναι χρήσιμο να γράψει ένα πρόγραμμα που να αυτοματοποιεί κάποιες ενέργειες, όπως την κλήση άλλων εργαλείων και τη μετατροπή ανάμεσα στα διαφορετικά format που χρησιμοποιούν προγράμματα ή βάσεις. Για το λόγο αυτό είναι αρκετά διαδεδομένες οι script γλώσσες, περισσότερο μάλιστα οι Perl και Python.

Στη συνέχεια παρουσιάζεται ένα παράδειγμα με τη γλώσσα Perl[12] (Practical Extraction and Report Language), που είναι και η παλαιότερη και έτσι πιο καθιερωμένη από τις δύο προγραμματιστικές γλώσσες που αναφέρθηκαν. Το πρόγραμμα δέχεται σαν είσοδο ακολουθία νουκλεοτιδίων και παράγει στην έξοδο την αντίστοιχη ακολουθία αμινοξέων. Για να γίνει, βέβαια, αυτή η μετατροπή, δηλαδή η μετάφραση, χρησιμοποιείται ο γενετικός κώδικας (Κεφάλαιο 3 και Σχήμα 5.18).

[12] Πηγή: [45]

```
#!/usr/bin/perl

#translate.pl

# translate nucleic acid sequence to protein sequence

# according to standard genetic code

# set up table of standard genetic code

%standardgeneticcode = (

"ttt"=> "Phe", "tct"=> "Ser", "tat"=> "Tyr", "tgt"=> "Cys",

"ttc"=> "Phe", "tcc"=> "Ser", "tac"=> "Tyr", "tgc"=> "Cys",

"tta"=> "Leu", "tca"=> "Ser", "taa"=> "TER", "tga"=> "TER",

"ttg"=> "Leu", "tcg"=> "Ser", "tag"=> "TER", "tgg"=> "Trp",

"ctt"=> "Leu", "cct"=> "Pro", "cat"=> "His", "cgt"=> "Arg",

"ctc"=> "Leu", "ccc"=> "Pro", "cac"=> "His", "cgc"=> "Arg",

"cta"=> "Leu", "cca"=> "Pro", "caa"=> "Gln", "cga"=> "Arg",

"ctg"=> "Leu", "ccg"=> "Pro", "cag"=> "Gln", "cgg"=> "Arg",

"att"=> "Ile", "act"=> "Thr", "aat"=> "Asn", "agt"=> "Ser",

"atc"=> "Ile", "acc"=> "Thr", "aac"=> "Asn", "agc"=> "Ser",

"ata"=> "Ile", "aca"=> "Thr", "aaa"=> "Lys", "aga"=> "Arg",

"atg"=> "Met", "acg"=> "Thr", "aag"=> "Lys", "agg"=> "Arg",

"gtt"=> "Val", "gct"=> "Ala", "gat"=> "Asp", "ggt"=> "Gly",

"gtc"=> "Val", "gcc"=> "Ala", "gac"=> "Asp", "ggc"=> "Gly",
```

```
"gta"=> "Val", "gca"=> "Ala", "gaa"=> "Glu", "gga"=> "Gly",

"gtg"=> "Val", "gcg"=> "Ala", "gag"=> "Glu", "ggg"=> "Gly" );


# process input data

while ($line = <DATA>) { # read in line of input

    print "$line"; # transcribe to output
    chop(); # remove end-of-line character

    @triplets = unpack("a3" x (length($line)/3), $line);
    # pull out successive triplets

    foreach $codon (@triplets) { # loop over triplets

        print "$standardgeneticcode{$codon}";
        # print out translation of each

    } # end loop on triplets

    print "\n\n"; # skip line on output

} # end loop on input lines

__END__
```

ΕΙΣΟΔΟΣ

```
atgcatccctttaat
tctgtctga
```

ΕΞΟΔΟΣ

```
atgcatccctttaat
MetHisProPheAsn

tctgtctga
SerValTER
```

Κεφάλαιο 6

Προβλήματα και λύσεις

Το παρόν κεφάλαιο ασχολείται με προβλήματα που εμφανίζονται σχετικά με όσα έχουν καταγραφεί στα προηγούμενα κεφάλαια, καθώς και λύσεις που έχουν προταθεί και είναι υπό συζήτηση. Φυσικό είναι να αναφέρονται προβλήματα που καλείται να λύσει ο ερευνητής των βάσεων δεδομένων και όχι εκείνα που έχουν να κάνουν γενικά με την αντιμετώπιση των δεδομένων των βιοεπιστημών.

Τα θέματα προσεγγίζονται όπου είναι δυνατόν με παραδείγματα, όμως αρκετά συχνά και περιγραφικά. Στόχος είναι η απόκτηση της εικόνας του συνόλου των δυσκολιών που υπάρχουν και όχι η λεπτομερής ανάλυση κάποιων από αυτές. Έτσι, θεωρείται ότι εισάγονται τα θεμέλια, ώστε ο αναγνώστης να εντρυφήσει από εκεί και πέρα σε όποιο θέμα κρίνει καταλληλότερο.

6.1 Προβλήματα

Στην ενότητα αυτή καταγράφονται σημαντικά προβλήματα που έχει να αντιμετωπίσει ο ερευνητής των βάσεων δεδομένων. Εξηγούνται οι ιδιαιτερότητες των περιορισμών, όπως και των patterns για τις βιοεπιστήμες και τα συστήματα βάσεων δεδομένων. Περιγράφονται, επίσης, τα προβλήματα της προέλευσης και της ενοποίησης των δεδομένων (data provenance, data integration). Τέλος, αναφέρονται οι δυσκολίες που πρέπει να ξεπεραστούν, για να ευδοκιμήσει η διεπιστημονική έρευνα.

6.1.1 Πολίτες α΄ κατηγορίας

Στα συστήματα βάσεων δεδομένων κάποιοι τύποι δεδομένων θεωρούνται πιο σημαντικοί από κάποιους άλλους. Αυτό μεταξύ των άλλων σημαίνει ότι συνήθως ερωτήματα για αυτούς μπορούν να γραφούν πιο εύκολα σε σχέση με ερωτήματα για τους άλλους. Οι συγκεκριμένοι τύποι δεδομένων συχνά αναφέρεται ότι είναι πολίτες πρώτης κατηγορίας σε ένα σύστημα.

Χαρακτηριστικό παράδειγμα είναι ο τύπος δεδομένων των αντικειμένων. Στα Αντικειμενοσχεσιακά μοντέλα η διαχείριση των αντικειμένων γίνεται μέσω επεκτάσεων στα κλασικά σχεσιακά. Αντίθετα, το Αντικειμενοστρεφές μοντέλο έχει σχεδιαστεί για να εξυπηρετήσει ακριβώς τις ανάγκες επεξεργασίας των αντικειμένων, επομένως αυτά αποτελούν πολίτες πρώτης κατηγορίας [60].

Στην κοινότητα των βάσεων δεδομένων έχει γίνει τα τελευταία δύο χρόνια η παρατήρηση ότι τα βιοδεδομένα έχουν δύο είδη στα οποία αξίζει να δοθεί περισσότερη προσοχή από ότι γίνεται συνήθως. Θεωρείται πως οι *περιορισμοί* και τα *patterns* (Κεφάλαιο 4) αξίζουν να αντιμετωπίζονται ως πολίτες πρώτης κατηγορίας από ένα μοντέλο που δίνει έμφαση στις ανάγκες των βιοεπιστημών ([42]).

Τα παραδοσιακά συστήματα έχουν ενσωματώσει τρόπους αντιμετώπισης ορισμένων τύπων περιορισμών. Αρκετές φορές αναφέρονται σε περιορισμούς πεδίων τιμών, όπως με χρήση της λέξης κλειδί check στον ορισμό πίνακα στην SQL. Άλλες πάλι, αυτοί οι τύποι αφορούν λογικούς κανόνες. Οι περιορισμοί ακεραιότητας αναφοράς (referential integrity constraints) ανήκουν σε αυτή την ομάδα και το ξένο κλειδί των σχεσιακών συστημάτων είναι ένα παράδειγμα αυτών.

Στον ακόλουθο ορισμό του πίνακα account το πεδίο του branch-name είναι κλειδί ενός άλλου πίνακα, του branch. Ορίζεται, επιπλέον, η τιμή του πεδίου balance να είναι αυστηρά θετική. Οι δύο τελευταίες γραμμές του ορισμού δηλώνουν ακριβώς τους παραπάνω περιορισμούς [54].

```
create table account (

    account-number  char(10),
    branch-name     char(15),
    balance         integer,

    primary key (account-number),
    foreign key (branch-name) references branch,
    check (balance >= 0)
)
```

Ωστόσο, χρειάζεται να παρατηρηθούν δύο αδυναμίες. Η πρώτη είναι το γεγονός ότι τα παραδοσιακά συστήματα βάσεων δεδομένων δεν έχουν καθιερώσει κάποιο μηχανισμό ελέγχου μη τοπικών περιορισμών. Η δεύτερη έχει ήδη σκιαγραφηθεί και αφορά την αναβάθμιση μιας μεγάλης κατηγορίας λογικών και μαθηματικών περιορισμών, με τη δυνατότητα όχι μόνο αποθήκευσης αλλά και επεξεργασίας αυτών μέσω ερωτημάτων.

Τα παραπάνω δύο ζητήματα χρειάζεται να μελετηθούν στο πλαίσιο των περιορισμών που συναντώνται στις βιοεπιστήμες. Αυτές χρειάζονται εκτός από τους λογικούς κανόνες, μαθηματικές ισότητες/ανισότητες και μάλιστα όλα αυτά να αντιμετωπίζονται ως πολίτες πρώτης κατηγορίας. Επιπλέον, συναντώνται αρκετές περιπτώσεις στις οποίες είναι απαραίτητη η χρήση μη τοπικών περιορισμών.

Για την καλύτερη κατανόηση της έννοιας του τοπικού περιορισμού αναφέρεται ένα παράδειγμα από το Κεφάλαιο 4. Στις βιοεπιστήμες έχουν σημαντική εφαρμογή οι αρχές διατήρησης της ενέργειας, της ορμής και της μάζας, οι οποίες βέβαια εκφράζονται με μαθηματικές σχέσεις. Τοπικός είναι ο περιορισμός της διατήρησης της μάζας αντιδρώντων και προϊόντων σε μία χημική αντίδραση. Μη τοπικός είναι η διατήρηση της ενέργειας σε έναν κύκλο αντιδράσεων της

θερμοδυναμικής.

Τέλος, όπως ήδη επιπώθηκε, οι ερευνητές των βιοεπιστημών μελετούν ιδιαίτερα και τα patterns. Ενδιαφέρονται να είναι σε θέση να τα αποθηκεύουν, να τα ομαδοποιούν και να τα επεξεργάζονται. Είναι, λοιπόν, επόμενο να θεωρείται λογική η αντιμετώπιση και αυτών ως τύπων δεδομένων πρώτης κατηγορίας.

6.1.2 Προέλευση (data provenance)

Η προέλευση και η ιστορική πορεία των δεδομένων μιας βάσης είναι θέματα που δεν έχουν απασχολήσει σε μεγάλο βαθμό ως τώρα την ερευνητική κοινότητα. Κάποιες μελέτες έχουν γίνει κυρίως από επιστήμονες που ασχολούνται με αποθήκες δεδομένων [56]. Ωστόσο, το ζήτημα αυτό αξίζει να τεθεί υπό στενότερη εξέταση υπό το πρίσμα της διαχείρισης δεδομένων των βιοεπιστημών.

Η προέλευση των δεδομένων ενδιαφέρει κυρίως σε περιπτώσεις, όπως στις βιολογικές και γενικότερα επιστημονικές βάσεις, όπου οι αποθηκευμένες πληροφορίες ανανεώνονται αρκετά συχνά. Με άλλα λόγια, όταν οι εγγραφές των δεδομένων είναι αποτέλεσμα πολλών σταδίων επεξεργασίας. Αυτή είναι η κατάσταση που ισχύει και στις βιοεπιστήμες. Υπάρχουν περίπου 500 βάσεις δεδομένων εκ των οποίων, όμως, περίπου 10 είναι εκείνες από τις οποίες αντλούν όλες οι υπόλοιπες το περιεχόμενό τους (Κεφάλαιο 4). Το γεγονός αυτό σε συνδυασμό με τη συχνή παραγωγή στοιχείων για υπάρχουσες ή νέες εγγραφές από την έρευνα δείχνει το δυναμικό χαρακτήρα των δεδομένων των βάσεων.

Υπάρχουν τουλάχιστον τέσσερις λόγοι που φανερώνουν την αξία που έχει η γνώση της προέλευσης των δεδομένων. Ταυτόχρονα υποδηλώνουν ότι η ενημέρωση σχετικά με την επεξεργασία στην οποία υποβάλλονται χρειάζεται να γίνεται με αυτόματο τρόπο από το σύστημα διαχείρισης της βάσης και όχι χειροκίνητα. Σημαντικό, επίσης, είναι να μπορούν να γίνονται και ερωτήματα σχετικά με την προέλευση.

Ένας σπουδαίος λόγος σχετίζεται με την αξιοπιστία των δεδομένων. Έχει σημασία για εκείνον που χρησιμοποιεί τα δεδομένα μιας βάσης να γνωρίζει ποιος τα έχει εναποθέσει. Συχνό είναι να εμπιστεύεται κανείς κάποιες πηγές γνώσης περισσότερο από κάποιες άλλες. Ειδικά στην περίπτωση των δευτερογενών βιολογικών βάσεων, δηλαδή εκείνων που τροφοδοτούνται και από άλλες, μια εγγραφή μπορεί να έχει εισαχθεί όντως από κάποια πρωτογενή ή από εκείνους που διατηρούν τη δευτερογενή βάση. Επιπλέον, αν βρεθεί λάθος σε μια εγγραφή, θα είναι εύκολο να ανατρέξει κανείς στην αρχική πηγή και να διορθώσει αυτό στη ρίζα του.

Ο δεύτερος λόγος για την υποστήριξη της τήρησης και επεξεργασίας της πληροφορίας της προέλευσης και της πορείας των δεδομένων είναι το γεγονός ότι αποτελεί κίνητρο για τη διακίνηση της γνώσης. Όπως η αναφορά σε δημοσιεύσεις κάποιου ερευνητή ή ομάδας (citations) αυξάνει το κύρος και τη φήμη του ή της, έτσι μπορεί να λειτουργήσει και η αναφορά σε δεδομένα, των οποίων η ύπαρξη σε κάποια βάση οφείλεται στην εργασία συγκεκριμένων προσώπων.

Εξάλλου, το καθεστώς στη δημοσίευση πορισμάτων βιολογικών μελετών έχει αλλάξει σημαντικά τα τελευταία χρόνια. Λόγω του μεγάλου όγκου πολλών πειραματικών αποτελε-

σμάτων αλλά και της ευκολίας χρήσης του διαδικτύου, οι ερευνητές στηρίζουν τις απόψεις που διατυπώνουν σε άρθρα περιοδικών, σε δεδομένα που διαθέτουν σε ιστοτόπους του παγκόσμιου ιστού. Έτσι, εμφανίζεται μια επιπλέον ανάγκη διατήρησης πληροφοριών σχετικών με την προέλευση. Μάλιστα είναι απαραίτητη η δυνατότητα επανάκτησης κάθε προηγούμενης έκδοσης των δεδομένων, γιατί διαφορετικά, παλαιότερα άρθρα που αναφέρονταν σε κάποια από αυτές καθίστανται άχρηστα ή μη υποκείμενα πλέον σε έλεγχο.

Ένας τέταρτος λόγος αφορά ξανά τη διακίνηση της γνώσης, αλλά από μια άλλη οπτική γωνία. Είναι γνωστό ότι υπάρχουν αρκετές mailing lists ή newsgroups, στα οποία συμμετέχουν άτομα που ενδιαφέρονται για κάποιο συγκεκριμένο επιστημονικό θέμα, λ.χ. για μια βάση βιοδεδομένων. Όταν κάποιος από αυτούς ανακαλύπτει ένα λάθος ή μια επιπλέον πληροφορία για κάποια εγγραφή της βάσης, συχνά τη μοιράζεται με τους τρόπους που αναφέρθηκαν. Η διακίνηση της γνώσης αυτής σε όλους απαιτεί από τους τελευταίους να ενημερώνονται συνεχώς. Πιο εξυπηρετική και σίγουρη μέθοδος θα ήταν να μπορεί να προσθέτει ο ίδιος το σχόλιο στη βάση μέσω κάποιου εργαλείου. Μάλιστα, το σχόλιο δεν επηρεάζεται από το σχήμα της βάσης.

6.1.3 Ενοποίηση (data integration)

Η έννοια της ενοποίησης αναφέρεται περισσότερο στην ανάγκη ύπαρξης εύκολου τρόπου μετάβασης από το format των δεδομένων σε ένα σύστημα στο format ενός άλλου, ώστε να είναι δυνατή η επεξεργασία τους, παρά στην καθιέρωση ενιαίου μοντέλου για όλα τα δυνατά είδη δεδομένων, παρότι αυτό θα ήταν το ιδανικό. Το πρόβλημα της ενοποίησης, επίσης, περικλείει και άλλα επιμέρους προβλήματα, για τα οποία γίνεται λόγος στη συνέχεια, καθώς και στο [42].

Τις περισσότερες φορές, για μια αλυσίδα εφαρμογών οι ερευνητές των βιοεπιστημών χρησιμοποιούν γκάμα πηγών δεδομένων και ευρύ φάσμα προγραμμάτων για την επεξεργασία τους, όπως αναφέρεται και στο [6]. Είναι δυνατόν να αντλούν τα πρωτογενή δεδομένα τους από πολλές διαφορετικές βάσεις, να εκτελούν κάποιες λειτουργίες σε αυτά μέσω εργαλείων και στη συνέχεια να διαλέγουν κάποια, για να συνεχίσουν να τα επεξεργάζονται.

Τα κυριότερα προβλήματα που εμφανίζονται σε αυτή τη διαδικασία είναι δύο. Το πρώτο είναι το γεγονός ότι δεν υπάρχει κοινό format αποθήκευσης των δεδομένων, αλλά ούτε και πρότυπο μετατροπής των διαφορετικών μορφών. Το δεύτερο αφορά την πρακτική με την οποία έχει κανείς πρόσβαση στα προγράμματα τις περισσότερες φορές. Η υποβολή των queries και η παραλαβή των αποτελεσμάτων γίνεται με φόρμες του διαδικτύου (web-based forms). Έτσι, αν ο χρήστης επιθυμεί την παραπέρα επεξεργασία τους, καταφεύγει στη λύση –μοναδική συνήθως– του copy-paste.

Ακόμη και όταν τα αποτελέσματα είναι διαθέσιμα σαν αρχείο κειμένου, είναι εμφανές ότι δεν εξυπηρετούν για την άμεση χρήση τους ως ορίσματα σε πρόγραμμα. Με άλλα λόγια, η αντιμετώπιση των δεδομένων από τις υπάρχουσες μεθόδους αποθήκευσης γίνεται σε πολύ χαμηλό επίπεδο και στερείται σημασιολογίας. Βοηθητικό είναι το παράδειγμα της ενότητας 6.2.2, για να κατανοήσει κανείς καλύτερα τη διαφορά ανάμεσα στην προσέγγιση που αντιμε-

τωπίζει τα δεδομένα απλά σαν αριθμούς ή λέξεις και σε εκείνην που είναι υψηλότερη και τους προσδίδει σημασιολογικό περιεχόμενο.

Η λύση να καθιερωθεί ένα κοινό πρότυπο αποθήκευσης των βιοδεδομένων δε φαίνεται να υποστηρίζεται ιδιαίτερα. Καταρχήν, είναι τόσα τα διαφορετικά είδη αυτών, που είναι δύσκολο να βρεθεί κάποιο format που να τα καλύπτει, αλλά και να ανταποκρίνεται στις ανάγκες όλων. Επίσης, οι ίδιοι οι ερευνητές δείχνουν προτίμηση στη χρήση των μικρών και εξειδικευμένων βάσεων δεδομένων, επειδή κυρίως είναι περισσότερο εύκολο να διαπιστώνεται η αξιοπιστία τους. Είναι, έτσι, περισσότερο βολικό κάθε είδος δεδομένου, δηλαδή κάθε βάση, να υιοθετεί το πρότυπο που ταιριάζει περισσότερο στις ανάγκες των δεδομένων της.

Ένα ακόμη επιμέρους πρόβλημα που υφίσταται είναι η αμφισημία των όρων της βιολογίας. Αρκετές –κυρίως εξειδικευμένεσ– βάσεις δεδομένων τείνουν να εισάγουν σαφώς καθορισμένη ορολογία για τις λέξεις-κλειδιά των βιοεπιστημών που χρησιμοποιούν (EcoCyc, RiboWeb, Gene Ontology, Ontology for Molecular Biology (OMB), RiboWeb, TAMBIS Ontology (TaO)). Ωστόσο, δεν υπάρχει πάντοτε σαφής αντιστοιχία από βάση σε βάση. Ακόμη, δεν είναι απίθανο ο ίδιος όρος να έχει διαφορετική σημασία από το ένα σύστημα σε ένα άλλο. Το γεγονός αυτό είναι ενδεικτικό της αβεβαιότητας που εμφανίζεται, σε σχέση με την ορθή χρήση των δεδομένων από βάση σε βάση.

Η κατάσταση δυσκολεύει ακόμη περισσότερο τη σύγκριση με στόχο την εύρεση σφαλμάτων στις βάσεις δεδομένων. Το λεγόμενο data curation προβληματίζει και σε αυτό το χώρο εφαρμογής των βάσεων δεδομένων. Μάλιστα εμφανίζει τα δικά του χαρακτηριστικά. Δεν περιλαμβάνει μόνο τον καθαρισμό των δεδομένων, αλλά και την ορθή τοποθέτηση ή επανατοποθέτηση των σχολίων και γενικότερα των δευτερευόντων στοιχείων που συνοδεύουν το καθαρό περιεχόμενο (π.χ. πειραματικό αποτέλεσμα) μιας εγγραφής (annotation) . Ο έλεγχος των βιολογικών δεδομένων περιλαμβάνει κυρίως τον έλεγχο των πειραματικών συνθηκών και παραδοχών και προς το παρόν, μπορεί να γίνει μόνο από τους αντίστοιχους ερευνητές. Στόχος είναι να μπορεί να γίνεται αυτοματοποιημένα, λόγω του μεγάλου όγκου και ρυθμού αύξησής τους.

6.1.4 Διεπιστημονική έρευνα

Για τη λύση των μυστηρίων της δομής και λειτουργίας των έμβιων όντων η έρευνα έχει ανάγκη τουλάχιστον τρεις επιστημονικούς χώρους. Αυτοί είναι βέβαια η κοινότητα των βιοεπιστημών, εκείνη της διαχείρισης βάσεων δεδομένων και γνώσεων, καθώς επίσης η νέα και γοργά αναπτυσσόμενη ομάδα της βιοπληροφορικής (bioinformatics).

Είναι σαφές ότι είναι απαραίτητη η εναρμόνιση των προσπαθειών αυτών των τριών κλάδων. Τα συστήματα βάσεων δεδομένων έχουν αντιμετωπίσει με επιτυχία πολλές προκλήσεις σε άλλα πεδία εφαρμογών οι οποίες, όμως, είναι πολύ πιθανό να παρουσιάζουν κοινά χαρακτηριστικά με προβλήματα των επιστημών του ευρύτερου χώρου της βιολογίας. Δεν υπάρχει, επομένως, κανένας λόγος η βιοπληροφορική να ανακαλύψει εκ νέου τεχνικές ήδη γνωστές στους ερευνητές των βάσεων. Απο την άλλη, είναι αναγκαίο οι τελευταίοι να ξέρουν τις μεθόδους που χρησιμοποιούν οι πρώτοι για τη λύση προβλημάτων, ώστε οι προσπάθειές τους να

τους βοηθήσουν να μην αποβούν άκαρπες. Φυσικά, η γνώση του αντικειμένου της βιολογίας για την αποτελεσματική συμβολή όλων των παραπάνω επιστημόνων στο χώρο κρίνεται εκ των ων ουκ άνευ.

Οι δυσκολίες που εμφανίζονται στο εγχείρημα της αλληλεπίδρασης των γνώσεων και ικανοτήτων των διαφόρων ερευνητικών χώρων είναι αρκετές. Η πιο σημαντική αλλά και ουσιώδης από αυτές είναι η εύρεση μιας κοινής γλώσσας συνεννόησης μεταξύ τους. Ιδιαίτερα ανάμεσα στους πληροφορικούς και στους βιολόγους, το χάσμα είναι περισσότερο αισθητό από ότι ανάμεσα σε πληροφορικούς των βάσεων και βιοπληροφορικούς, όπου και εκεί βέβαια δεν είναι αμελητέο. Τα αίτια που δημιουργούν αυτό το κενό στην επικοινωνία είναι κυρίως ο διαφορετικός τρόπος σκέψης και προσέγγισης των προβλημάτων, στον οποίο έχουν εκπαιδευτεί οι δύο επιστημονικές ομάδες, όπως και η άγνοια και από τις δύο πλευρές του αντικειμένου της άλλης.

Ένας ακόμη ανασταλτικός παράγοντας είναι περισσότερο τεχνικού χαρακτήρα, αλλά τοποθετεί ένα αξιοσημείωτο εμπόδιο στη γόνιμη συνεργασία. Ο παράγοντας αυτός δεν είναι άλλος από τον χρόνο. Υπάρχει μια διαφορά φάσης ανάμεσα στη χρονική στιγμή που οι βιοεπιστήμες χρειάζονται μια λύση και σε εκείνην που οι βάσεις δεδομένων μπορούν να τους τη δώσουν. Ερευνητικά αποτελέσματα παράγονται με ταχύτατους ρυθμούς, ενώ και η χρηματοδότηση των επιστημονικών προγραμμάτων αυξάνει τη χρονική πίεση. Συνήθως, το διάστημα που είναι απαραίτητο για την ανάπτυξη νέων τεχνολογιών βάσεων δεδομένων τις περισσότερες φορές είναι αρκετά μεγαλύτερο από εκείνο που είναι διατεθειμένοι οι βιολόγοι και βιοπληροφορικοί να περιμένουν.

Είναι γεγονός, ωστόσο, ότι συνεργασία μεταξύ διαφορετικού αντικειμένου ερευνητών έχει πολλές φορές επιτευχθεί στο παρελθόν με πολύτιμα οφέλη για όλες τις εμπλεκόμενες πλευρές. Όσον αφορά ιδιαίτερα την κοινότητα των βάσεων δεδομένων, αν προσπαθήσει έγκαιρα, θα έχει την ευκαιρία να προσφέρει λύσεις σε έναν χώρο που πραγματικά αξίζει τον κόπο, και να στεφθεί τις ανάλογες δάφνες.

6.2 Προτεινόμενες λύσεις

Στην ενότητα αυτή αναφέρονται διάφορες λύσεις που έχουν προταθεί για να ανταποκριθούν στις ανάγκες των βιοεπιστημών. Η ανάλυση δεν καλύπτει σε όλο το δυνατό βάθος τα θέματα που πραγματεύεται, γιατί στόχος της είναι να φωτίσει μεθόδους αντιμετώπισης και να φέρει ενδεικτικά παραδείγματα, παρά να εξηγήσει ολοκληρωμένους τρόπους λύσης. Πιστεύεται ότι έτσι ο αναγνώστης θα σχηματίσει στο νου του μια άποψη για το πώς θα πρέπει να εργαστεί, αν θέλει να συμβάλει στον τομέα υπό συζήτηση, και θα κατευθύνει κατάλληλα την εκτενέστερη μελέτη του.

6.2.1 Επεκτάσεις στην SQL

Μία προσέγγιση για τη λύση των προβλημάτων της βιολογίας είναι η κατάλληλη επέκταση της SQL, έτσι ώστε να μπορούν να υποστηριχθούν εξειδικευμένες εργασίες. Στα επόμενα

περιγράφεται συνοπτικά ένα τέτοιο συγκεκριμένο παράδειγμα, για να γίνει κατανοητή αυτή η προσέγγιση.

Το σύστημα MoBIoS (Molecular Biological Information System) είναι ένα σύστημα διαχείρισης βάσεων δεδομένων ειδικά σχεδιασμένο για τις απαιτήσεις των βιοεπιστημών. Αναπτύσσεται από το τμήμα πληροφορικής του University of Texas at Austin[1]. Περιλαμβάνει τύπους δεδομένων της βιολογίας και δομές ευρετηρίων που τις εξυπηρετούν, ενώ προσθέτει επιπλέον λογικούς τελεστές.

Πιο συγκεκριμένα, η mSQL, που είναι η εν λόγω επέκταση της SQL, έχει και τύπους για ακολουθίες, όπως και για τη δευτεροταγή και τεταρτοταγή διάταξη των πρωτεϊνών. Ειδικά για τις ακολουθίες, οι πρωτογενείς τύποι ονομάζονται DNA, RNA, Peptide και είναι υποτύποι του Sequence. Ένα Sequence μπορεί να θεωρηθεί ότι είναι ένα string, με τη διαφορά ότι το λεξιλόγιό του είναι πιο περιορισμένο (π.χ. ACTG για το DNA) και ότι υπάρχει ο τελεστής revcomp() που παράγει το συμπληρωματικό μιας ακολουθίας νουκλεοτιδίων, ενώ στη γενική περίπτωση των strings ο τελεστής αυτός δεν έχει νόημα.

Για παράδειγμα, το σχήμα για τις ακολουθίες DNA και πρωτεϊνών που έχουν πρωτεύον κλειδί το SID είναι:

```
DNA_Sequence(SID, Organism, seq)

Protein_Sequence(SID, Organism, seq)
```

Δύο λογικοί τελεστές πάνω στις ακολουθίες αυτές είναι οι Createfragments(), merge(). Όπως δηλώνουν και τα ονόματά τους, ο πρώτος σχηματίζει υπακολουθίες, ενώ ο δεύτερος τις ενώνει. Για να φανεί πιο συγκεκριμένα παράδειγμα για τον Createfragments(), έστω ότι έχουμε τις ακολουθίες DNA:

```
(R1, Rice, ACAA)        (R2, Rice, ACTCA)
```

Τα αποτελέσματα του ακόλουθου ερωτήματος φαίνονται στον Πίνακα 6.1. Το πρώτο όρισμα της αγκύλης της τρίτης στήλης του αντιστοιχεί στην απόσταση του πρώτου γράμματος της υπακολουθίας, που είναι το δεύτερο όρισμα της αγκύλης, από το πρώτο γράμμα της αρχικής ακολουθίας (R1 ή R2). Με άλλα λόγια είναι το offset.

```
select SID, Organism, Createfragments(seq,3)
from DNA_Sequence
```

Το πρώτο όρισμα του τελεστή Createfragments() είναι η ακολουθία από την οποία είναι επιθυμητό να δημιουργηθούν υπακολουθίες και το δεύτερο είναι το μήκος αυτών των υπακολουθιών. Το αποτέλεσμα του τελεστή είναι όλες οι δυνατές συνεχόμενες υπακολουθίες του καθορισμένου μεγέθους.

Ο τελεστής merge() είναι ο αντίστροφος του Createfragments(). Χρησιμοποιεί μάλιστα τον κλασικό τελεστή group by της SQL, για να ομαδοποιήσει πρώτα τις υπακολουθίες που

[1] Περισσότερα στο [46]

Πίνακας 6.1: Αποτελέσματα του δοθέντος query

SID	Organism	Createfragments(seq,3)
R1	Rice	[0, ACA]
R1	Rice	[1, CAA]
R2	Rice	[0, ACT]
R2	Rice	[1, CTC]
R2	Rice	[2, TCA]

ανήκουν στην ίδια ομάδα. Θεωρείται ότι μια ομάδα αποτελείται από υπακολουθίες που η διαφορά τους στο offset είναι μικρότερη από το μήκος τους.

Το σύστημα MoBIoS υποστηρίζεται ότι εξυπηρετεί σε αρκετές περιπτώσεις εφαρμογών. Παραδείγματα αυτών είναι η homology search (Κεφάλαιο 5), όπως και το πρόβλημα που σχετίζεται με τον προσδιορισμό της λεγόμενης πρωτεΐνης Rosetta Stone. Αυτή η ονομασία χρησιμοποιείται για οποιαδήποτε πρωτεΐνη είναι ομόλογη, δηλαδή παρουσιάζει πολλές ομοιότητες, με δύο άλλες πρωτεΐνες που είναι επιθυμητό να διευκρινιστεί κατά πόσο μεταξύ τους αυτές οι δύο αλληλεπιδρούν.

6.2.2 Ανάπτυξη μοντέλου και γλώσσας

Όπως φανερώθηκε από την ενότητα 6.1.3, δύο είναι οι βασικές επιταγές που καλούνται οι ερευνητές της κοινότητας των βάσεων δεδομένων και γνώσεων να αντιμετωπίσουν. Αυτές είναι να αναπτύξουν εξειδικευμένες τεχνικές επεξεργασίας των τύπων των δεδομένων των βιοεπιστημών, καθώς επίσης περισσότερο επεκτάσιμα και ευέλικτα συστήματα βάσεων δεδομένων.

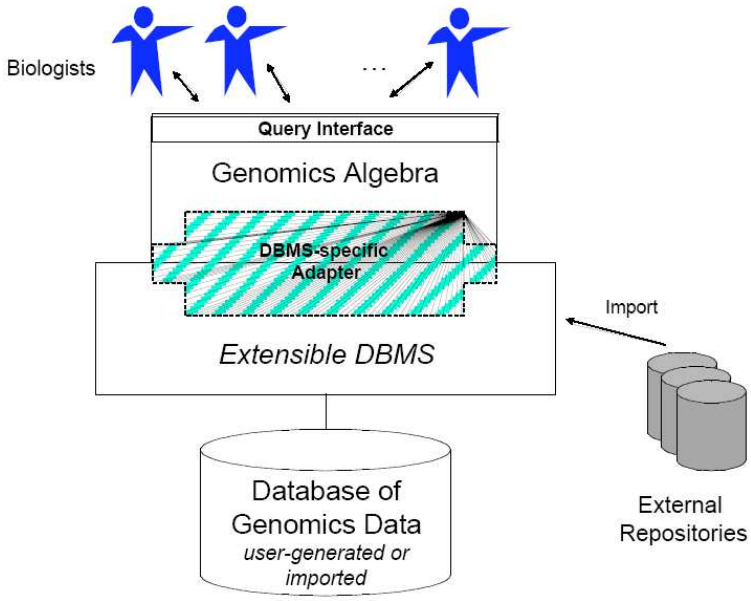
Η κύρια πρόταση που αυτή τη στιγμή φαίνεται να κυριαρχεί είναι η εισαγωγή τριών στοιχείων στο μέχρι τώρα τρόπο εργασίας. Χρειάζονται πρότυπα δεδομένων για την ανταλλαγή πληροφοριών ανεξαρτήτως μορφής, γλώσσα ερωτημάτων ικανή να εκφράσει αποδοτικά τα αναγκαία και ειδικά APIs για τη διαπροσωπεία. Μάλιστα, για τον έλεγχο της καταλληλότητας των διαφόρων προσεγγίσεων, υπάρχει η ιδέα της δημιουργίας συνόλων δεδομένων προς δοκιμή, όπως και benchmarks.

Προς την κατεύθυνση της ανάπτυξης ενός νέου μοντέλου, που να έχει τουλάχιστον κάποια από τα παραπάνω χαρακτηριστικά, γίνονται προσπάθειες, ίσως όμως σχετικά περιορισμένες. Ένα παράδειγμα που θα χρησιμεύσει στη συνέχεια της συζήτησης είναι το GenAlg project[2]. Για τη συγκεκριμένη προσπάθεια είναι υπεύθυνο το τμήμα Πληροφορικής του University of Florida, Gainesville.

Το κύριο χαρακτηριστικό του εν λόγω έργου είναι το γεγονός ότι εισάγει τα genomics ontology, genomics algebra. Μέσω αυτών λύνει κάποια από τα προβλήματα που αναφέρθηκαν στην ενότητα 6.1.3. Συγκεκριμένα, παρουσιάζει ένα πρότυπο δεδομένων και μια κατάλλη-

[2]Περισσότερα στο [13]

λα προσαρμοσμένη γλώσσα ερωτήσεων, ώστε να αντιμετωπίζει με περισσότερη ευελιξία και εκφραστικότητα τα δεδομένα. Η αρχιτεκτονική του συστήματος φαίνεται στο Σχήμα 6.1.



Σχήμα 6.1: Η αρχιτεκτονική του συστήματος GenAlg

Ένα από τα πιο αξιοσημείωτα χαρακτηριστικά του είναι η περιγραφή βιολογικών όρων με σκοπό την άρση της αμφισημίας. Στο Σχήμα 6.2 εικονίζεται το διάγραμμα που αντιστοιχεί σε αυτό που οι δημιουργοί του ονομάζουν gene onion, λόγω σχήματος και δομής. Φαίνονται, επίσης, οι σχέσεις εξάρτησης ανάμεσα στις διάφορες οντότητες.

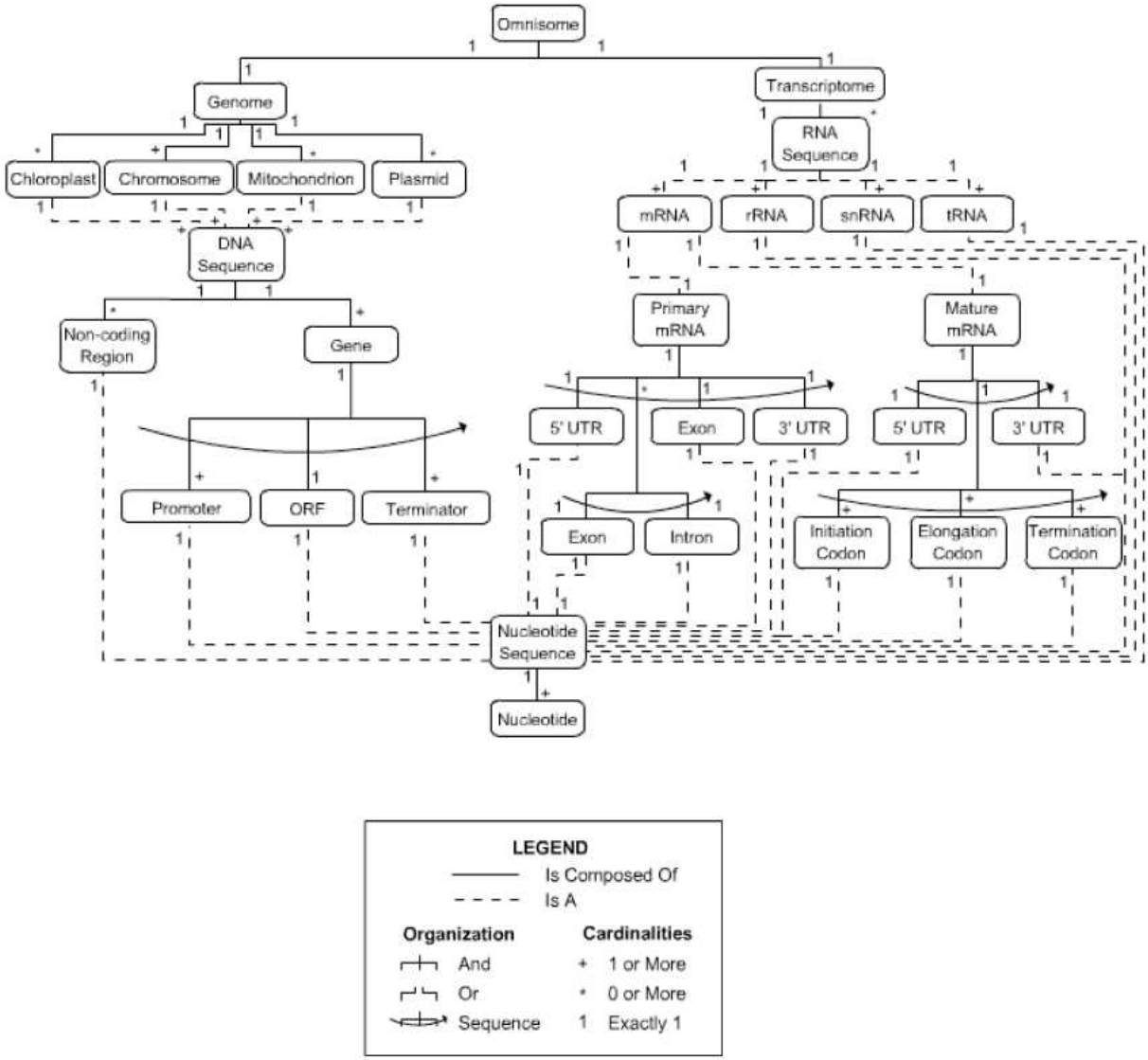
Από την άλλη, οι τελεστές της άλγεβρας που εισάγονται κάνουν το μοντέλο πιο εύχρηστο. Για παράδειγμα, υπάρχουν τελεστές που εκτελούν τη μεταγραφή, την ωρίμανση και τη μετάφραση (Κεφάλαιο 3) μεταξύ των οντοτήτων, που επίσης εισάγει το μοντέλο:

```
transcribe: Gene -> PrimarymRNA
splice: PrimarymRNA -> mRNA
translate: mRNA -> Protein
```

Η διαφορά που έχει στη χρήση το μοντέλο αυτό σε σχέση με το κλασικό με την SQL παρουσιάζεται με ένα παράδειγμα ερώτησης. Έστω ότι είναι επιθυμητή η πρόβλεψη της δευτεροταγούς δομής πρωτεϊνών με βάση τις αντίστοιχες ακολουθίες νουκλεοτιδίων, που βρίσκονται αποθηκευμένες σε μία βάση δεδομένων (Κεφάλαιο 4), όταν αυτές οι ακολουθίες μοιράζονται κάποιο κοινό χαρακτηριστικό (molecular function).

Η κλασική προσέγγιση απαιτεί τα ακόλουθα βήματα:

- Εκτέλεση ενός ερωτήματος για την εύρεση των κατάλληλων ακολουθιών που θα χρησιμοποιηθούν για την πρόβλεψη. Ο κωδικός GO:0003724 αντιστοιχεί στην RNA-helicase, που είναι το κοινό χαρακτηριστικό που αναφέρθηκε, ενώ sequence-ID είναι το πρωτεύον κλειδί.



Σχήμα 6.2: Gene Onion

```
SEQUENCE_FUNCTION (
    sequence-ID:integer,
    sequence:varchar(255),
    length:integer,
    GO_ID:varchar(16)
);

Select sequence_id, sequence

From SEQUENCE_FUNCTION

Where GO_ID=''GO:0003724'';
```

- Καθεμιά από τις παραπάνω ακολουθίες νουκλεοτιδίων μπαίνει ως είσοδος σε αλγόριθμο πρόβλεψης της πιο πιθανής περιοχής κωδικοποίησης. Με βάση αυτήν γίνεται η μετάφραση και προκύπτει η ακολουθία των αμινοξέων.
- Το αποτέλεσμα του προηγούμενου βήματος χρησιμοποιείται σε αλγόριθμο πρόβλεψης της δευτεροταγούς δομής.

Την ίδια λειτουργία μέσω του GenAlg μπορεί κανείς να την εκτελέσει με το ακόλουθο μόνο query. Μάλιστα, η υλοποίησή του παραμένει κρυφή από τον χρήστη και η διατύπωσή του είναι ανεξάρτητη από το σύστημα. Πάντως, στη συγκεκριμένη περίπτωση θεωρήθηκε ότι η υλοποίηση χρησιμοποιεί αντικειμενοσχεσιακό σύστημα βάσης δεδομένων.

```
Select predict2DStructure(translate(predictCDS(n))

From nucleotide_sequence n

Where n.getMolecularFunction() = ''RNA-helicase'';
```

Το GenAlg, αν και δεν παρέχει λύση σε όλα τα προβλήματα, είναι ένα παράδειγμα του πόσο πιο εύκολα μπορούν να γίνουν οι διάφορες λειτουργίες των βιοεπιστημών με τη συμβολή των ερευνητών των βάσεων δεδομένων.

6.2.3 Βελτίωση και εφαρμογή νέων αλγορίθμων και προγραμμάτων

Στο Κεφάλαιο 5 περιγράφηκαν αρκετές λειτουργίες που επιθυμούν οι βιοεπιστήμονες να κάνουν, αλλά και αλγόριθμοι για αυτές. Έχει αναφερθεί ξανά ότι υπάρχουν προγράμματα και πακέτα που υλοποιούν τους κατάλληλους αλγορίθμους, για να δώσουν λύση στα προβλήματα. Είναι γεγονός, όμως, ότι υπάρχουν σημαντικά περιθώρια βελτίωσης της πολυπλοκότητας και της απόδοσης. Από την άλλη, υπάρχουν ακόμη ζητήματα στα οποία δεν έχει δοθεί ικανοποιητική λύση, όπως λ.χ. ο προσδιορισμός της τριτοταγούς δομής μιας πρωτεΐνης.

Είναι, λοιπόν, φανερό ότι γίνεται αρκετή δουλειά και προς αυτήν την κατεύθυνση. Παράδειγμα είναι η προσπάθεια εύρεσης καλύτερων τρόπων εκτέλεσης του alignment. Ωστόσο, ο τομέας αυτός είναι αντικείμενο της βιοπληροφορικής (bioinformatics) και όχι άμεσα των βάσεων δεδομένων. Οι αλγόριθμοι και τα προγράμματα επηρεάζουν τον τρόπο που αποθηκεύονται και τίθενται υπό επεξεργασία τα δεδομένα, αλλά η ανακάλυψη νέων δεν είναι το πεδίο έρευνας των βάσεων δεδομένων.

Για το λόγο αυτό, στην παράγραφο αυτή επιλέχθηκε να γίνει απλή αναφορά στο θέμα και όχι ανάλυση. Η αναφορά κρίθηκε αναγκαία, επειδή ο εν λόγω είναι ένας χώρος με έντονη ερευνητική δραστηριότητα και οι επιπτώσεις του γίνονται αισθητές και στους ερευνητές των βάσεων που ασχολούνται με τις βιοεπιστήμες.

6.2.4 Συνεργασία

Στην ενότητα 6.1.4 περιγράφηκαν οι λόγοι για τους οποίους είναι αναγκαία η συνεργασία μεταξύ των τριών άμεσα εμπλεκόμενων επιστημονικών κοινοτήτων, όπως και τα προβλήματα που εμφανίζονται σε αυτήν. Θυμίζεται ότι η αναφορά γίνεται για τους βιοεπιστήμονες, τους ερευνητές των συστημάτων βάσεων γνώσεων και δεδομένων και τους βιοπληροφορικούς. Στην παρούσα ενότητα δίνονται ορισμένα στοιχεία που πιθανώς ενισχύσουν την προσπάθεια αυτή.

Είναι φανερό ότι απαραίτητη προϋπόθεση μιας δημιουργικής, ισότιμης συνεργασίας είναι η διάθεση και η καταβολή προσπάθειας από όλες τις πλευρές. Ένα σημαντικό βήμα προς αυτήν την κατεύθυνση είναι η εξοικείωση με τα βασικά στοιχεία που συνθέτουν το περιεχόμενο κάθε αντικειμένου. Άλλωστε, και αυτή η διπλωματική εργασία έχει γίνει στα πλαίσια της προσέγγισης που μόλις αναφέρθηκε.

Ιδιαίτερη αξία έχει, όμως, η διαπροσωπική επαφή με ερευνητές των άλλων επιστημών. Συγκεκριμένα, για την κοινότητα των συστημάτων των βάσεων, στην οποία και απευθύνεται η παρούσα εργασία, είναι πολύ σπουδαία η ανταλλαγή γνώσεων με την ομάδα των βιολόγων, καθώς είναι δύσκολο να εντοπιστούν όλες οι ανάγκες τους μόνο από τη μελέτη σχετικών άρθρων. Εκτός αυτού, χρειάζεται η βοήθειά τους για την κατανόηση λεπτών σημείων της επιστήμης τους. Η οργάνωση συνεδρίων αλλά και συνεργασιών είναι σχεδόν σίγουρο ότι θα αποφέρει καρπούς για την επίτευξη αυτού του στόχου.

Η ιδιοσυγκρασία της βιολογίας

Ένας μηχανικός υπολογιστών είναι πιθανό να ξενιστεί από τον τρόπο σκέψης και προσέγγισης των θεμάτων από τους βιοεπιστήμονες. Λογικό είναι να ισχύει και το αντίστροφο, βέβαια, αφού οι δύο κοινότητες έχουν εκπαιδευτεί για να ανταποκρίνονται σε άλλου τύπου απαιτήσεις. Για το λόγο αυτό είναι χρήσιμο ο πληροφορικός να αντιληφθεί από νωρίς κάποια χαρακτηριστικά των επιστημών της βιολογίας, που δε συνηθίζει να απαντά στη δική του επιστήμη.

- Η βιολογία είναι μία *στατιστική* επιστήμη. Αυτό σημαίνει ότι τα πορίσματά της, τις περισσότερες φορές, δεν είναι αδιάσειστες αποδείξεις που πηγάζουν από αυστηρά ορισμένη θεωρία, όπως στα μαθηματικά. Αντίθετα, προέρχονται από πειράματα. Οι θεωρίες, δηλαδή, στηρίζονται σε σημαντικές ενδείξεις. Η ισχύς τους οφείλεται στο γεγονός ότι

δεν υπάρχουν μέχρι στιγμής στοιχεία που να τις καταρρίπτουν. Ο μηχανισμός αυτός μπορεί, φυσικά, να μην έχει την αυστηρότητα άλλων, αλλά δε στερείται λογικής. Όταν κάτι έχει πειραματικά διαπιστωθεί πως συμβαίνει στο μεγαλύτερο ποσοστό των περιπτώσεων, είναι επόμενο να θεωρείται ο κανόνας. Εξάλλου, έτσι δημιουργούνται και για αυτό υφίστανται οι εξαιρέσεις.

- Η βιολογία είναι μία σχετικά *νέα* επιστήμη. Αν και από την αρχαιότητα έχει προβληματίσει τον άνθρωπο (π.χ. Αριστοτέλης), όλες οι αρχές της, ακόμη και σήμερα, δεν πιστεύεται ότι είναι πλήρως γνωστές. Σε μεγάλο βαθμό η κατάσταση αυτή οφείλεται στη μεγάλη αδυναμία πειραματικών τεχνικών και μέσων, που υπήρχε μέχρι τον εικοστό αιώνα. Το πλαίσιο έχει φανερά αλλάξει πλέον με την υπολογιστική δύναμη της πληροφορικής αλλά και τις εξειδικευμένες μεθόδους υψηλής τεχνολογίας που έχουν αναπτυχθεί. Τα πειράματα και οι μετρήσεις των ημερών μας ίσως έχουν να προσφέρουν τα πιο πολύτιμα οφέλη για την κατανόηση των έμβιων όντων.

- Η βιολογία *δεν είναι φυσική* επιστήμη. Η προηγούμενη πρόταση δεν αμφισβητεί το γεγονός ότι μέσα στους ζωντανούς οργανισμούς λαμβάνουν χώρα φυσικά φαινόμενα ή ότι χρησιμεύουν στη βιολογία πειραματικές τεχνικές της φυσικής και της χημείας. Ουσιαστικά, υποδηλώνει δύο θεμελιώδεις διαφορές σε σχέση με αυτές τις επιστήμες, που φανερώνονται στη συνέχεια.

  Η βιολογία έχει τους δικούς της νόμους. Οι δομές και οι λειτουργίες μέσα σε έναν ζωντανό οργανισμό διαφέρουν αισθητά σε σχέση με όσα έχουν μελετήσει οι φυσικοί και οι χημικοί. Εκείνοι έχουν ως αντικείμενο απλές έως εξαιρετικά πολύπλοκες δομές με ένα, όμως, κοινό χαρακτηριστικό: είναι περιοδικές. Στη βιολογία κανείς έχει να αντιμετωπίσει μη περιοδικούς κρυστάλλους, οι οποίοι ακολουθούν διαφορετικούς νόμους. "Η διαφορά στη δομή είναι παρόμοια μ΄ αυτήν που υπάρχει ανάμεσα σε ένα κοινό χαρτί ταπετσαρίας, όπου επαναλαμβάνεται διαρκώς το ίδιο σχέδιο με τακτική περιοδικότητα, και σ΄ ένα αριστουργηματικό κέντημα, όπως ας πούμε του Ραφαέλο, που δεν παρουσιάζει ανιαρές επαναλήψεις, αλλά ένα περίτεχνο, συναφές, μεστό νοήματος σχέδιο δημιουργημένο από τον μεγάλο δάσκαλο." [53].

  Είναι σχετικά απίθανο οι νόμοι αυτοί να αποτελέσουν ένα ενιαίο σύστημα μαθηματικών εξισώσεων, σε αντίθεση με ό,τι συμβαίνει στις φυσικές επιστήμες. Η άποψη αυτή[3] υποστηρίζεται λόγω της πολυπλοκότητας των δεδομένων. Έχουν αναπτυχθεί υπολογιστικές τεχνικές για το χειρισμό ορισμένων ειδών ξεχωριστά. Ωστόσο, θεωρείται εξαιρετικά δύσκολο να συντεθεί με επαγωγική διαδικασία ένα σύνολο μαθηματικών σχέσεων ικανό να περιγράψει όλους τους τύπους δεδομένων. Άλλοι τρόποι περιγραφής τους, λ.χ. οι τυπικές γραμματικές και τα διαγράμματα, φαίνονται περισσότερο κατάλληλοι για την αναπαράσταση αυτής της γνώσης.

Τα παραπάνω τρία χαρακτηριστικά της επιστήμης της βιολογίας οδηγούν και σε ένα συμπέρασμα για αυτήν: *οι νόμοι της συνεχώς αλλάζουν*. Τα πειραματικά αποτελέσματα ολοένα

[3] François Rechenman, From Data to Knowledge, Bioinformatics, Oxford University Press 2000

και πληθαίνουν, όπως αναφέρθηκε. Μερικές φορές είναι αφορμή για νέες παρατηρήσεις ή αναθεωρήσεις. Με αυτόν τον τρόπο αλλάζουν τα στατιστικά στοιχεία και έτσι αναγκάζονται σε αλλαγή και οι κανόνες που ως τότε ίσχυαν ή προκύπτει μία νέα εξαίρεση.

Το πόρισμα αυτό είναι καλό να υπάρχει πάντα στο πίσω μέρος του μυαλού του μηχανικού υπολογιστών ο οποίος ασχολείται με προβλήματα των βιοεπιστημών. Είναι ένας παράγοντας που μεταξύ των άλλων του υπενθυμίζει την ανάγκη της ευελιξίας για τα μοντέλα που υιοθετεί. Από την άλλη, τον προστατεύει από πιθανές δυσάρεστες εκπλήξεις, όταν κάτι που θεωρούσε ως τότε σίγουρο παύει να θεωρείται σωστό.

Υπάρχουν αρκετά παραδείγματα τα οποία σκιαγραφούν το χαρακτήρα της βιολογίας, έτσι όπως περιγράφηκε ως τώρα. Στη συνέχεια αναφέρονται ενδεικτικά έξι για την πληρότητα της ανάλυσης.

Το 1944 ήταν μεταξύ των άλλων το έτος που άλλαξε στους επιστήμονες την αντίληψη που είχαν για το βιολογικό ρόλο του DNA και των πρωτεϊνών. Τα πειράματα που έγιναν (Avery, Mac-Cleod, McCarthy) ήταν ισχυρές ενδείξεις ότι δεν φέρουν τη γενετική πληροφορία οι πρωτεΐνες, αλλά το DNA. Η οριστική πειραματική απόδειξη ήρθε το 1952 (Hershey, Chase).

Μόλις το 1961 άρχισαν οι μελέτες σχετικά με τη ρύθμιση των γονιδίων (Κεφάλαιο 3). Πρόκειται για θεμελιώδη ερωτήματα, όπως για ποιο λόγο ενεργοποιείται ένα γονίδιο σε έναν οργανισμό και όχι σε έναν άλλο, πώς καθορίζεται ο ρυθμός με τον οποίο παράγονται οι πρωτεΐνες και πότε. Για την ιστορία, τα πρώτα πειράματα έγιναν από τους Jacob, Monod.

Μέχρι το 1977, οι επιστήμονες πίστευαν ότι τα γονίδια των ευκαρυωτικών οργανισμών είναι συνεχείς περιοχές στο γονιδίωμά τους. Οι ομάδες των Sharp, Roberts ανακάλυψαν, όμως, τότε την ύπαρξη των εσωνίων και αποκάλυψαν ότι τα περισσότερα γονίδια είναι διακεκομμένα.[4] Τα εσώνια είναι περιοχές που παρεμβάλλονται ανάμεσα σε εκείνες που μεταφράζονται (Κεφάλαια 2, 3).

Ως το 1982, οι πρωτεΐνες θεωρούνταν οι μοναδικοί καταλύτες των ζωντανών οργανισμών. Πίστευαν, δηλαδή, ότι τα ένζυμα είναι οι μοναδικοί καταλύτες (όμως, όλες οι πρωτεΐνες δεν είναι ένζυμα, βλ. Κεφάλαιο 2). Ωστόσο, εκείνη τη χρονιά βρέθηκε σε πρωτόζωο ένα εσώνιο το οποίο είναι τμήμα του RNA του και μπορεί να αυτοκαταλύεται, δηλαδή να κόβει τον εαυτό του χωρίς τη βοήθεια πρωτεϊνικών ενζύμων. Αυτό το RNA ονομάστηκε μάλιστα ριβόζυμο.

Τέλος, δύο πρόσφατες ανακαλύψεις ανατρέπουν ξανά τις καθιερωμένες πεποιθήσεις[5]. Η μία αφορά το γεγονός ότι το ίδιο γονίδιο είναι δυνατόν να είναι υπεύθυνο για την παραγωγή παραπάνω από μίας πρωτεΐνης. Αυτό εξηγεί εν μέρει το λόγο για τον οποίο αναμενόταν μεγαλύτερος ο αριθμός των γονιδίων από αυτόν που βρέθηκε να είναι με την αποκωδικοποίηση διαφόρων γονιδιωμάτων. Η δεύτερη έδειξε πως ο αριθμός των αμινοξέων που συνθέτουν τις πρωτεΐνες των έμβιων όντων μπορεί να υπερβαίνει κατά περιπτώσεις το 20.

[4] Για την εργασία τους αυτή τιμήθηκαν το 1993 με το βραβείο Nobel.

[5] D.B. Searls, Grand Challenges in Computational Biology. In Computational Methods in Molecular Biology, S.L. Salzberg, D.B. Searls and S. Kasif, Eds. Elsevier Amsterdam, The Netherlands, 1998

Όλα τα παραπάνω συνθέτουν ένα εξαιρετικά ενδιαφέρον τοπίο για έρευνα. Οι προκλήσεις είναι πολλές, αλλά και οι ανακαλύψεις που φαίνεται ότι περιμένουν για να γίνουν ακόμη περισσότερες.

Κεφάλαιο 7

Μελέτη του προγράμματος BLAST και του συστήματος PathCase

Αποτέλεσμα της προσπάθειας να δοθούν αποδοτικές λύσεις σε πρακτικά προβλήματα και ερωτήματα ερευνητών, που μελετούν συμπεριφορές οργανισμών σε μοριακό κατά κύριο λόγο επίπεδο, είναι η ανάπτυξη προγραμμάτων και συστημάτων την τελευταία δεκαπενταετία, όπως το BLAST και το PathCase. Το Basic Local Alignment Search Tool (BLAST) ανακαλύπτει δοθείσα ακολουθία χημικών μορίων, κυρίως νουκλεοτιδίων ή αμινοξέων, σε μεγάλες ίδιου τύπου ακολουθίες. Το Pathways Database System (PathCase) μπορεί να αποθηκεύσει, επεξεργαστεί και παρουσιάσει αντιδράσεις που λαμβάνουν χώρα μέσα σε οργανισμούς σε διάφορα επίπεδα λεπτομέρειας - γενετικό, μοριακό, βιοχημικό, ολόκληρου οργανισμού. Το κεφάλαιο αυτό στοχεύει στην κατανόηση του θεωρητικού υπόβαθρου αυτών των εργαλείων καθώς και στην πειραματική διερεύνηση των λειτουργιών και δυνατοτήτων τους.

Αν και όπως και στην υπόλοιπη εργασία η μελέτη είναι κυρίως βιβλιογραφική, δε θα μπορούσαν να λείπουν παραδείγματα χρήσης των δύο συστημάτων. Στις περισσότερες περιπτώσεις αναφέρονται παραδείγματα που πρώτα βρέθηκαν στη σχετική βιβλιογραφία και μετά χρησιμοποιήθηκαν στην πράξη. Ο λόγος είναι ότι το αντικείμενο που πραγματεύονται τα δύο εργαλεία είναι αρκετά εξειδικευμένο και μπορεί εύκολα να πειραματιστεί κανείς με περιπτώσεις χρήσης υπαρκτές αλλά χωρίς νόημα. Αναφέρονται, επίσης, σε κάθε περίπτωση οι δυνατές επιλογές που έχει ο χρήστης του BLAST ή του PathCase για να πραγματοποιήσει την εργασία του.

Για το BLAST οι πηγές είναι άφθονες στο Internet λόγω της ευρείας χρήσης του, αλλά και του γεγονότος ότι χρησιμοποιείται ήδη δεκαπέντε χρόνια. Αντίθετα, για το PathCase, που είναι σημαντικά νεότερο, τα στοιχεία για μελέτη είναι μεν αρκετά, αλλά προέρχονται σχεδόν όλα από την ίδια πηγή, δηλαδή τους δημιουργούς του. Τα σχήματα που εμφανίζονται στην εργασία αυτή για το BLAST έχουν ως κύρια πηγή τις πληροφορίες που δίνονται το site του NCBI[1]. Τα σχήματα για το PathCase προέρχονται κατά κύριο λόγο από τα εγχειρίδια χρήσης που δίνονται στο αντίστοιχο site [21].

[1] http://www.ncbi.nlm.nih.gov/blast/ http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html

7.1 BLAST

Η ενότητα αυτή είναι αφιερωμένη στο ένα από τα δύο εργαλεία στα οποία επικεντρώνεται η παρούσα εργασία. Εξηγεί το πρόβλημα που λύνει το Basic Local Alignment Search Tool (BLAST) καθώς και τον αλγόριθμο με τον οποίο καταφέρνει να το κάνει. Επιπλέον, αναφέρονται οι δυνατοί τρόποι χρήσης του BLAST, όπως και η μορφή που δίνει το αποτέλεσμα (output) σαν παράδειγμα χρήσης του. Τέλος, παρουσιάζονται οι βάσεις δεδομένων πρωτεϊνών και νουκλεϊκών οξέων που μπορούν να χρησιμοποιηθούν με αυτό, ενώ από την εισαγωγή αναφέρονται τα συγκριτικά πλεονεκτήματα του εργαλείου σε σχέση με παρόμοιά του.

7.1.1 Εισαγωγή

Το BLAST δημιουργήθηκε με τη συνεργασία επιστημόνων βιολογίας και πληροφορικής υπό την αιγίδα αρκετών φορέων με κυριότερο το National Center for Biotechnology Information (NCBI), National Library of Medicine. Η πρώτη δημοσίευση για αυτό έγινε το 1990 [1]. Από τότε, όμως, έγιναν βελτιώσεις ή επεκτάσεις λόγω της ευρείας αποδοχής και χρήσης του και ακολούθησαν και άλλες δημοσιεύσεις, όπως η [7].

Εκτός από τις τροποποιήσεις που έγιναν στο BLAST από το NCBI, το Washington University ανέπτυξε και εκείνο μια ανεξάρτητη ομάδα προγραμμάτων. Αν και βασίζονται στις ίδιες γενικές αρχές, χρησιμοποιούν διαφορετικές μεθόδους στατιστικής ανάλυσης –για την αποτίμηση σκορ κατά τη σύγκριση ακολουθιών– και τα αποτελέσματα που βγάζουν συχνά είναι διαφορετικά. Στην εργασία αυτή εξετάζεται το NCBI-BLAST και όχι το WU-BLAST. Περισσότερες πληροφορίες για το δεύτερο μπορούν να αναζητηθούν στο αντίστοιχο url [14].

Έχουν δημιουργηθεί πολλά εργαλεία που έχουν ίδιο στόχο με το BLAST. Αρκετά από αυτά χρησιμοποιούν μεθόδους δυναμικού προγραμματισμού για την εύρεση της βέλτιστης ακολουθίας, κάτι το οποίο τα καθιστά εξαιρετικά αργά. Άλλα, όπως το FASTP, χρησιμοποιούν ευριστικούς αλγορίθμους και έχουν ικανοποιητική συμπεριφορά. Ωστόσο, το BLAST είναι ίσως το πιο συχνά χρησιμοποιούμενο πρόγραμμα για τη σύγκριση ακολουθιών για βιολογικούς σκοπούς και έχει σχεδόν καθιερωθεί.

7.1.2 Χαρακτηριστικά του προβλήματος

Το BLAST παρέχει τη δυνατότητα σύγκρισης ακολουθιών αμινοξέων ή νουκλεοτιδίων. Το πρόβλημα αυτό έχει ιδιαίτερη σημασία στην μοριακή βιολογία, ενώ όπως προδηλώνει και το όνομα του BLAST γίνεται local και όχι global alignment. Περισσότερα στοιχεία για το alignment έχουν δοθεί στο Κεφάλαιο 5.

Χρειάζεται να τονισθεί ότι η σύγκριση που ενδιαφέρει αφορά μεγάλου μήκους ακολουθίες. Η συνηθέστερη περίπτωση είναι να συγκρίνεται μια ακολουθία των δομικών στοιχείων που αναφέρθηκαν με ακολουθίες αποθηκευμένες σε κάποια βάση δεδομένων (πρωτεϊνών ή νουκλεϊκών οξέων). Για παράδειγμα, η σύγκριση μπορεί να αφορά δύο ολόκληρα γονιδιώματα, το οποίο σημαίνει ότι μπορεί να αναφερόμαστε σε ακολουθίες δισεκατομμυρίων βάσεων (Κεφάλαιο 2).

Αν και άλλοι αλγόριθμοι χρησιμοποιούνται από παρεμφερή προγράμματα, το BLAST χρησιμοποιεί με κάποιες τροποποιήσεις αλγόριθμο που βασίζεται στην κατασκευή του dot matrix (Κεφάλαιο 5). Τα μειονεκτήματα της μεθόδου αυτής παρακάμπτονται, αν δεν χτίζεται όλος ο πίνακας και χρησιμοποιείται μια τεχνική αξιολόγησης των επιμέρους διαγωνίων.

7.1.3 Κύρια σημεία του αλγορίθμου

Στόχος είναι να συγκριθεί μια ακολουθία, έστω a, αμινοξέων ή νουκλεοτιδίων με τις ανάλογες ακολουθίες που βρίσκονται αποθηκευμένες σε μία βάση δεδομένων. Θα περιγραφούν τα βήματα που ακολουθεί ο αλγόριθμος του BLAST, για να βγάλει το αποτέλεσμα.

Αρχικά η a φιλτράρεται ώστε να αποκλειστούν, αν υπάρχουν, περιοχές για τις οποίες είναι εκ των προτέρων γνωστό ότι δεν έχει νόημα να συμμετέχουν στο alignment. Αυτές είναι συνήθως περιοχές χαμηλής πολυπλοκότητας.

Το υπόλοιπο της a χωρίζεται σε υπακολουθίες, δηλαδή σε λέξεις. Αν πρόκειται για βάση πρωτεϊνών, οι λέξεις έχουν μήκος 3, ενώ για βάση που περιέχει DNA το μήκος τους είναι 11. Έστω ότι το μήκος των λέξεων είναι στη γενική περίπτωση w.

Στη συνέχεια υπολογίζεται το σκορ καθεμιάς λέξης όταν γίνεται aligned με λέξη ίσου μήκους της βάσης δεδομένων. Για τον υπολογισμό αυτού του σκορ χρησιμοποιείται ο BLOSUM στην περίπτωση των πρωτεϊνών, που αναφέρθηκε στο Κεφάλαιο 5. Στην περίπτωση των DNA, RNA δίνονται τιμές +5 για τα όμοια ζευγαρώματα και -4 για τα υπόλοιπα. Αυτό είναι το λεγόμενο MSP (Maximal Segment Pair) σκορ.

Από το σύνολο των λέξεων κρατώνται μόνο εκείνες που έχουν σκορ μεγαλύτερο από ένα κατώφλι, έστω T. Αυτές οργανώνονται αποδοτικά σε ένα δέντρο αναζήτησης, ώστε να συγκριθούν γρήγορα με τις ακολουθίες της βάσης δεδομένων για την εύρεση πιθανού ταιριάσματος (match).

Μετά από τη σύγκριση με τη βάση οι λέξεις που έχουν match επεκτείνονται τόσο από την μια πλευρά όσο και από την άλλη, έτσι ώστε να μεγαλώσει το σκορ που είχε η αρχική λέξη. Το επόμενο βήμα είναι να εκτιμηθεί αν το νέο σκορ κάθε αρχικής λέξης - που ονομάζεται HSP, High Scoring Segment Pair - είναι μεγαλύτερο από ένα κατώφλι, έστω S. Η τιμή αυτού είναι τις περισσότερες φορές εμπειρική.

Οι τιμές αυτών των σκορ υπόκεινται σε στατιστική ανάλυση. Από εκείνα τα matches που περνούν και αυτό το στάδιο, τελικά εμφανίζονται στον χρήστη του BLAST μόνο όσα περνούν το κατώφλι που ο ίδιος έθεσε όταν έδωσε το query.

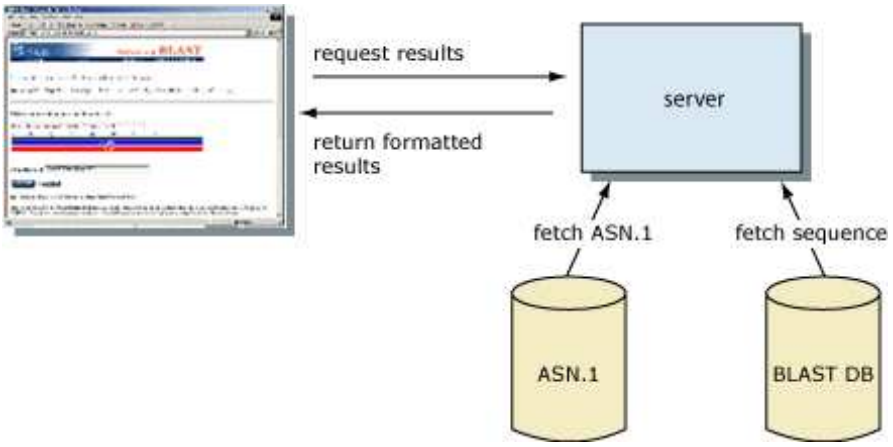
Αυτήν την προσέγγιση ακολουθεί ο αρχικός αλγόριθμος του BLAST. Επόμενες εκδόσεις του, όπως το [7] ακολουθούν ελαφρώς διαφορετικά βήματα, για παράδειγμα στην επιλογή των κατωφλίων ή στην επέκταση των λέξεων προς όλες τις κατευθύνσεις.

Συνοπτικά, ο αλγόριθμος του BLAST βρίσκει τμήματα των διαγωνίων του dot matrix και στη συνέχεια προσπαθεί να βρει εκείνα που μπορούν να επεκταθούν κατά το περισσότερο δυνατό.

7.1.4 Τρόποι χρήσης

Υπάρχουν δύο τρόποι με τους οποίους μπορεί κανείς να χρησιμοποιήσει το BLAST. Ο πρώτος είναι να θέσει το query του στην αντίστοιχη ιστοσελίδα (η οποία δίνεται στις αναφορές της ενότητας) και μέσω Internet να λάβει την απάντηση. Ο δεύτερος είναι να εγκαταστήσει στο δικό του υπολογιστικό σύστημα το BLAST ("stand-alone" BLAST).

Στην πρώτη περίπτωση ο χρήστης εισάγει την ακολουθία που θέλει και πιθανώς μερικά ακόμη στοιχεία –όπως το μέγεθος λέξης και την αναμενόμενη τιμή– και επιλέγει με ποια βάση δεδομένων επιθυμεί να γίνει η σύγκριση. Στη συνέχεια ο server τού επιστρέφει το αποτέλεσμα, αφού έχει συγκρίνει με τις ανάλογες ακολουθίες της βάσης, έχει δηλαδή εκτελέσει τον αλγόριθμο του BLAST και έχει εισαγάγει το αποτέλεσμα σε μια δομή δεδομένων, την SeqAlign σε ASN.1. Η δομή αυτή περιέχει μια αναφορά στις ακολουθίες της βάσης και όχι τις ίδιες τις ακολουθίες. Η πορεία αυτή φαίνεται στο Σχήμα 7.1.



Σχήμα 7.1: Η πορεία που ακολουθείται στο σύστημα για την απάντηση ενός ερωτήματος.

Η δεύτερη περίπτωση εξυπηρετεί περισσότερο χρήστες που θέλουν να συγκρίνουν μια ακολουθία με τη δική τους τοπική βάση δεδομένων ή θέλουν να ρυθμίσουν το BLAST να ταιριάζει καλύτερα στις ανάγκες τους και κατεβάζουν κάποιες βάσεις δεδομένων από το NCBI στο σύστημά τους. Υπάρχουν δύο μορφές του "stand-alone" BLAST. Η μία είναι εκτελέσιμα προγράμματα των οποίων ο χειρισμός μπορεί να γίνει από τη γραμμή εντολών. Η δεύτερη είναι το στήσιμο στον υπολογιστή μιας έκδοσης των BLAST Web pages.

7.1.5 Οι χρησιμοποιούμενες βάσεις δεδομένων

Στο site του NCBI-BLAST υπάρχει ο πλήρης κατάλογος των πρωτεϊνικών βάσεων, των βάσεων DNA, όπως και ολόκληρων γονιδιωμάτων, στις οποίες μπορεί ο χρήστης να αναζητήσει ομοιότητα με δική του ακολουθία ή υπάρχουσα ακολουθία σε κάποια από αυτές τις βάσεις. Αυτές είναι οι βάσεις τις οποίες χρησιμοποιεί το NCBI. Περισσότερες πληροφορίες μπορούν να αναζητηθούν στο Κεφάλαιο 4.

Επιπλέον, στο εν λόγω site υπάρχουν οδηγίες που βοηθούν το χρήστη να επιλέξει το κατάλληλο πρόγραμμα, δηλαδή την κατάλληλη έκδοση του BLAST, για κάθε μία βάση. Ο κατάλογος είναι μεγάλος και για αυτό ενδεικτικά παρατίθεται το Σχήμα 7.2. Εκεί φαίνεται ένας βοηθητικός πίνακας επιλογής βάσης και προγράμματος.

Table 3.2 Program Selection for Protein Queries

Length [1]	Database	Purpose	Program	Explanation
15 residues or longer	Peptide	Identify the query sequence or find protein sequences similar to the query	Standard Protein BLAST (blastp)	Learn more ...
		Find members of a protein family or build a custom position-specific score matrix	PSI-BLAST	Learn more ...
		Find proteins similar to the query around a given pattern	PHI-BLAST	Learn more ...
		Find conserved domains in the query	CD-search (RPS-BLAST)	Learn more ...
		Find conserved domains in the query and identify other proteins with similar domain architectures	Conserved Domain Architecture Retrieval Tool (CDART)	Learn more ...
	Nucleotide	Find similar proteins in a translated nucleotide database	Translated BLAST (tblastn)	Learn more ...
5-15 residues	Peptide	Search for peptide motifs	Search for short, nearly exact matches	Learn more ...

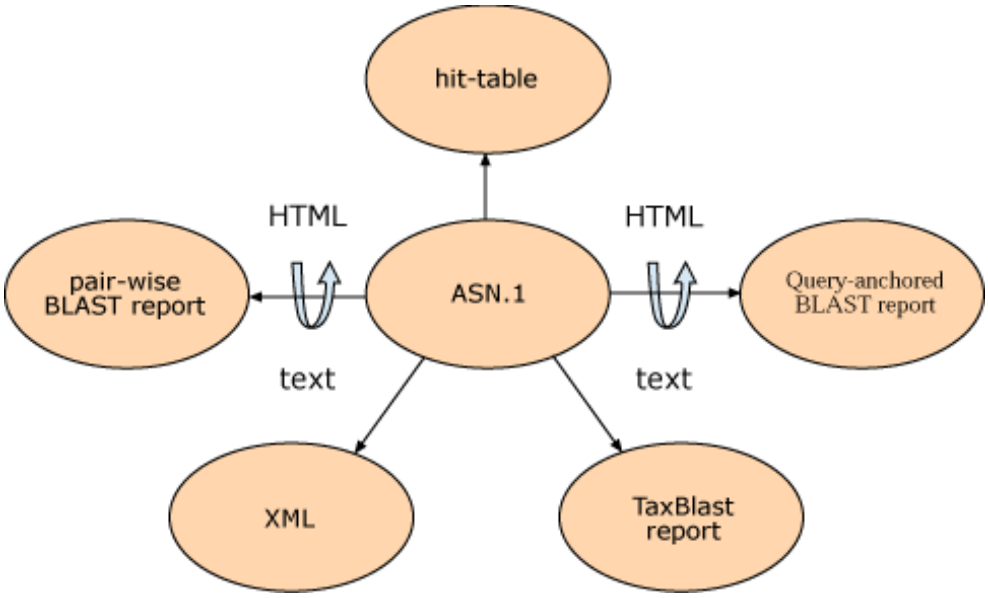
Note:
[1] The cut-off is only a recommendation. For short queries, one is more likely to get matches if the "Search for short, nearly exact matches" page is used. Detailed discussion is in Section 4 below.

Σχήμα 7.2: Ενδεικτικός πίνακας των δυνατών επιλογών.

7.1.6 Μορφή των αποτελεσμάτων

Το Σχήμα 7.3 ουσιαστικά αποτελεί συνέχεια εκείνου που δόθηκε στην ενότητα 7.1.4 και δείχνει όλες τις δυνατές μορφές του output ενός προγράμματος BLAST. Τα αποτελέσματα μπορούν να δοθούν ως plain text, XML, ASN.1, ως hit table και στην παραδοσιακή HTML μορφή. Το ακόμη σημαντικότερο είναι ότι δε χρειάζεται να επαναληφθεί το τρέξιμο του προγράμματος για να μετατραπεί το αποτέλεσμα από τη μια μορφή στην άλλη.

Η παραδοσιακή μορφή των αποτελεσμάτων σε HTML αποτελείται από τρία στοιχεία. Δίνεται η επικεφαλίδα, που περιέχει πληροφορίες για την ακολουθία που ήταν στο query αλλά



Σχήμα 7.3: Δυνατές μορφές των αποτελεσμάτων.

και για τη βάση δεδομένων, περιγράφεται κάθε ακολουθία της βάσης που βρέθηκε να ταιριάζει με αυτήν του query και παρέχεται το alignment για κάθε τέτοιο ζευγάρι. Τα Σχήματα 7.4-7.7 αντιστοιχούν ακριβώς σε αυτά: στην επικεφαλίδα (Σχήματα 7.4, 7.5), στις ακολουθίες της βάσης (Σχήμα 7.6) και στο αντίστοιχο alignment (Σχήμα 7.7) για ένα συγκεκριμένο παράδειγμα αναζήτησης.

Η τελευταία μορφή με την οποία μπορούν να παρασταθούν τα αποτελέσματα είναι αυτή του hit table. Η μορφή αυτή εξυπηρετεί κυρίως επιστήμονες που τρέχουν ένα μεγάλο αριθμό συγκρίσεων ακολουθιών για συγκεκριμένο σκοπό και ενδιαφέρονται για ένα υποσύνολο των πληροφοριών που δίνονται στην παραδοσιακή HTML μορφή των αποτελεσμάτων. Η μορφή αυτή ακολουθεί συγκεκριμένο format, όπως φαίνεται στο Σχήμα 7.8.

NCBI results of BLAST

BLASTP 2.2.1 [Apr-13-2001]

Reference:
Altschul,Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

RID: 1009580302-26840-4362

Query- RAB protein
(656 letters)

Database: Non-redundant SwissProt sequences
102,387 sequences; 37,391,913 total letters

If you have any problems or questions with the results of this search please refer to the **BLAST FAQs**

Taxonomy reports

Σχήμα 7.4: Επικεφαλίδα των αποτελεσμάτων.

Distribution of 41 Blast Hits on the Query Sequence

Mouse-over to show defline and scores. Click to show alignments

Color Key for Alignment Scores

<40 | 40-50 | 50-80 | 80-200 | >=200

1

0 100 200 300 400 500 600

Σχήμα 7.5: Γραφική μορφή της επικεφαλίδας των αποτελεσμάτων.

Sequences producing significant alignments:		Score (bits)	E Value
(a)	(b)	(c)	(d)
gi|116365|sp|P26374|RAE2_HUMAN	Rab proteins geranylgeranylt...	1216	0.0
gi|21431807|sp|P24386|RAE1_HUMAN	Rab proteins geranylgerany...	879	0.0
gi|585775|sp|P37727|RAE1_RAT	Rab proteins geranylgeranyltra...	846	0.0
gi|13626886|sp|Q61598|GDIC_MOUSE	RAB GDP dissociation inhib...	127	5e-29
gi|729566|sp|P39958|GDI1_YEAST	SECRETORY PATHWAY GDP DISSOC...	127	5e-29
gi|13626813|sp|O97556|GDIB_CANFA	Rab GDP dissociation inhib...	126	1e-28
gi|13638229|sp|P50397|GDIB_MOUSE	RAB GDP dissociation inhib...	125	3e-28
gi|1707888|sp|P50398|GDIA_RAT	RAB GDP dissociation inhibito...	124	7e-28
gi|121108|sp|P21856|GDIA_BOVIN	Rab GDP dissociation inhibit...	124	7e-28
gi|21903424|sp|P50396|GDIA_MOUSE	Rab GDP dissociation inhib...	124	7e-28
gi|13626812|sp|O97555|GDIA_CANFA	RAB GDP dissociation inhib...	124	8e-28
gi|1707886|sp|P31150|GDIA_HUMAN	Rab GDP dissociation inhibi...	123	9e-28
gi|13638228|sp|P50395|GDIB_HUMAN	Rab GDP dissociation inhib...	122	2e-27
gi|1707891|sp|P50399|GDIB_RAT	RAB GDP DISSOCIATION INHIBITO...	121	5e-27
gi|1723467|sp|Q10305|YD4C_SCHPO	Putative secretory pathway ...	120	8e-27
gi|585776|sp|P32864|RAEP_YEAST	RAB proteins geranylgeranylt...	97	7e-20
gi|10720243|sp|O93831|RAEP_CANAL	RAB proteins geranylgerany...	74	9e-13
gi|2498411|sp|Q49398|GLF_MYCGE	UDP-galactopyranose mutase	35	0.63
gi|11135401|sp|Q9XBQ9|STHA_AZOVI	Soluble pyridine nucleotid...	34	1.0
gi|11135075|sp|O05139|STHA_PSEFL	Soluble pyridine nucleotid...	33	1.3
gi|11135195|sp|P57112|STHA_PSEAE	Soluble pyridine nucleotid...	33	1.8
gi|22257022|sp|Q8TZJ8|RLA0_PYRFU	Acidic ribosomal protein P...	33	2.1
gi|3915516|sp|P94488|YNAJ_BACSU	Hypothetical symporter ynaJ	32	3.4
gi|231788|sp|P30599|CHS2_USTMA	CHITIN SYNTHASE 2 (CHITIN-UD...	32	3.7
gi|2498412|sp|P75499|GLF_MYCPN	UDP-galactopyranose mutase	32	4.2
gi|547891|sp|P36225|MAP4_BOVIN	Microtubule-associated prote...	32	4.2
gi|586602|sp|P37747|GLF_ECOLI	UDP-galactopyranose mutase	32	4.6

Σχήμα 7.6: Οι ακολουθίες της βάσης που ταιριάζουν στο ερώτημα.

```
>gi|126365|sp|P26374|RAB2_HUMAN   Rab proteins geranylgeranyltransferase component A 2 (Rab escort
          protein 2) (REP-2) (Choroideraemia-like protein)
          Length = 656

 Score =  846 bits (2186), Expect = 0.0
 Identities = 432/632 (68%), Positives = 489/632 (77%), Gaps = 13/632 (2%)

Query: 1   MADNLPTEFDVVIIGTGLPESILAAACSRSGQRVLHIDSRSYYGGNWASFSFSGLLSWLK 60
           MADNLP++FDV++IGTGLPESI+AAACSRSGQRVLH+DSRSYYGGNWASFSFSGLLSWLK
Sbjct: 1   MADNLPSDFDVIVIGTGLPESIIAAACSRSGQRVLHVDSRSYYGGNWASFSFSGLLSWLK 60

Query: 61  EYQQNNDIGEESTVVWQDLIHETEEAITLRKKDETIQHTEAFPYASQDMEDNVEEIGALQ 120
           EYQ+NND+  E++ +WQ+ I E EEAI L  KD+TIQH E F YASQD+  +VEE GALQ
Sbjct: 61  EYQENNDVVTEN5-MWQEQILENEEAIPLSSKDETIQHVEVFCYASQDLHKDVEEAGALQ 119

Query: 121 KNPSLGVS----NTFTEVLDSALPEESQLSYFNSDEMPAKHTQESDTEISLEVTDVEESV 176
           KN +   S              S LP   +      S E+PA+ +Q    E S EV D E +
Sbjct: 120 KNHASVTSAQSAEAAEAAETSCLPTAVEPLSMGSCEIPAEQSQCPGPESSPEVNDAEATG 179

Query: 177 EKEKYCGDKTCMHTVXXXXXXXXXXXXTVEDKADEPIRNRITYSQIVKEGRRFNIDLVSK 236
           +KE      +                V+D  + P +NRITYSQI+KEGRRFNIDLVS+
Sbjct: 180 KKENSDAKSS--------TEEPSENVFKVQDNTETFKKNRITYSQIIKEGRRFNIDLVSQ 231

Query: 237 LLYSQGLLIDLLIKSDVSRYVEFKNVTRILAFREGKVKQVPCSRADVFNSKELTMVEKRM 296
           LLYS+GLLIDLLIKS+VSRY EFKN+TRILAFREG VEQVPCSRADVFNSK+LTMVEKRM
Sbjct: 232 LLYSRGLLIDLLIKSNVSRYAEFKNITRILAFREGTVKQVPCSRADVFNSKQLTMVEKRM 291

Query: 297 LMKFLTFCLEYEQHPDEYQAFRQCSFSEYLKTKKLTPNLQHFVLHSIAMTSESSCTTIDG 356
           LMKFLTFC+EYE+HPDEY+A+   +FSEYLKT+KLTPNLQ+FVLHSIAMTSE++  T+DG
Sbjct: 292 LMKFLTFCVEYEEHPDEYRAYEGTTFSEYLKTQKLTPNLQYFVLHSIAMTSETTSCTVDG 351

Query: 357 LNATKNFLQCLGRFGNTPFLFPLYGQGEIPQGFCRMCAVFGGIYCLRHKVQCFVVDKESG 416
           L ATK FLQCLGR+GNTPFLFPLYGQGE+PQ FCRMCAVFGGIYCLRH VQC VVDKES
Sbjct: 352 LKATKKFLQCLGRYGNTPFLFPLYGQGELPQCFCRMCAVFGGIYCLRHSVQCLVVDKESR 411

Query: 417 RCKAIIDHFGQRINAKYFIVEDSYLSEETCSNVQYKQISRAVLITDQSILKTDLDQQTSI 476
           +CKA+ID FGQRI +K+FI+EDSYLSE TCS VQY+QISRAVLITD S+LKTD DQQ SI
Sbjct: 412 KCKAVIDQFGQRIISKHFIIEDSYLSENTCSRVQYRQISRAVLITDGSVLKTDADQQVSI 471

Query: 477 LIVPPAEPGACAVRVTELCSSTMTCMKDTYLVHLTCSSSKTAREDLESVVKKLFTPYTET 536
           L VP  EPG+  VRV ELCSSTMTCMK TYLVHLTC SSKTAREDLE VV+KLFTPYTE
Sbjct: 472 LAVPAEEPGSFGVRVIELCSSTMTCMKGTYLVHLTCMSSKTAREDLERVVQKLFTPYTEI 531

Query: 537 EINKEELTKPKLLWALYFNMRDSSGISRSSYNGLPSNVYVCSGPDCGLGNEHAVKQAETL 596
           E   E++ KPKLLWALYFNMRDSS ISR  YN LPSNVYVCSGPD GLGN++AVKQAETL
Sbjct: 532 EAENEQVEKPKLLWALYFNMRDSSDISRDCYNDLPSNVYVCSGPDSGLGNDNAVKQAETL 591

Query: 597 FQXXXXXXXXXXXXXXXXXXXXDGDDKQPEAP 628
           FQ                   DGD  Q E P
Sbjct: 592 FQQICPNEDFCPAPPNPEDIVLDGDSSQQEVP 623
```

Σχήμα 7.7: Το alignment του αποτελέσματος.

```
# BLASTN 2.2.1 [Aug-1-2001]
# Database: ecoli
# Query:gi|4730899|dbj|AP000130.1|Homo sapiens genomic DNA of 21q22.1, GART and AML, f43D11-119B8 region, segment 5/10.
# Fields: Query id, Subject id, % identity, alignment length, mismatches, gap openings, q. start, q. end, s. start, s. end, e-value, bit
gi|4730899|dbj|AP000130.1|AP000130     gi|2367099|gb|AE000133.1|AE000133     100.00   1198    0       0       52913   54110   10943
gi|4730899|dbj|AP000130.1|AP000130     gi|1789919|gb|AE000427.1|AE000427     100.00   1198    0       0       52913   54107   2347
gi|4730899|dbj|AP000130.1|AP000130     gi|1789607|gb|AE000401.1|AE000401     100.00   1198    0       0       52913   54107   6037
gi|4730899|dbj|AP000130.1|AP000130     gi|1788338|gb|AE000294.1|AE000294     100.00   1198    0       0       52913   54107   5700
gi|4730899|dbj|AP000130.1|AP000130     gi|1787588|gb|AE000231.1|AE000231     100.00   1198    0       0       52913   54107   4146
gi|4730899|dbj|AP000130.1|AP000130     gi|1786875|gb|AE000170.1|AE000170     100.00   1198    0       0       52913   54107   2321
gi|4730899|dbj|AP000130.1|AP000130     gi|1786751|gb|AE000160.1|AE000160     100.00   1198    0       0       52913   54107   9133
gi|4730899|dbj|AP000130.1|AP000130     gi|1788508|gb|AE000308.1|AE000308     99.92    1198    1       0       52913   54107   11740
gi|4730899|dbj|AP000130.1|AP000130     gi|2367181|gb|AE000381.1|AE000381     99.83    1198    2       0       52913   54107   2030
gi|4730899|dbj|AP000130.1|AP000130     gi|1788298|gb|AE000291.1|AE000291     99.58    1198    5       0       52910   54107   5290
gi|4730899|dbj|AP000130.1|AP000130     gi|1787633|gb|AE000234.1|AE000234     93.21    1105    72      3       52912   54014   1146
```

Σχήμα 7.8: Η μορφή του hit table.

7.2 PathCase

Στην ενότητα αυτή εξετάζεται το Pathways Database System (PathCase). Ερευνώνται οι τύποι δεδομένων που αποθηκεύονται σε αυτό και αναφέρεται το μοντέλο δεδομένων που χρησιμοποιείται, καθώς και η αρχιτεκτονική του συστήματος. Επιπλέον, διερευνώνται οι πιθανοί τρόποι χρήσης του και αναλύονται οι δυνατότητες που προσφέρουν καθένα από τα εργαλεία του. Τέλος, δίνονται κατάλληλα παραδείγματα χρήσης αυτών των εργαλείων.

7.2.1 Εισαγωγή

Το Pathways Database System αναπτύχθηκε στο Case Western Reserve University με τη συνεργασία επιστημόνων πληροφορικής και γενετικής. Η ανάπτυξή του ξεκίνησε στις αρχές του 2000 και αυτήν την εποχή τρέχουσα είναι η έκδοση 2.0.

Υπάρχουν αρκετά φημισμένα pathways systems, όπως αναφέρεται στο [49] και στο [6], με κυριότερο από αυτά το KEGG. Επίσης αξιόλογο θεωρείται το Ecocyc pathway DB [25], [44]. Ωστόσο, το PathCase ασχολείται τόσο με τα μεταβολικά μονοπάτια όσο και με αυτά των σημάτων - τα συστήματα που αναφέρθηκαν χειρίζονται μόνο το πρώτο είδος - ενώ υποστηρίζεται ότι προσφέρει μεγαλύτερη γκάμα δυνατοτήτων και εργαλείων για την αποθήκευση, παρουσίαση και επεξεργασία των σχετικών πληροφοριών.

7.2.2 Τύπος δεδομένων

Το είδος των πληροφοριών που διαχειρίζεται το PathCase είναι βιολογικά μονοπάτια (biological pathways) για τα οποία έχει γίνει λόγος στο Κεφάλαιο 4. Η χρησιμότητα της διαχείρισης τέτοιων δεδομένων τονίζεται στο [44]. Είναι αδύνατο να αναπτυχθεί μια ενιαία και πλήρης επιστημονική θεωρία για ένα πολύπλοκο σύστημα χωρίς τη βοήθεια μιας κατάλληλα σχεδιασμένης βάσης από γεγονότα και αλληλεπιδράσεις μεταξύ τους. Η πιο φιλόδοξη μάλιστα προσπάθεια είναι να κωδικοποιηθεί η επιστημονική γνώση σε βάσεις στους υπολογιστές. Η αναπαράστασή της γραφικά αλλά και οι δυνατότητες επεξεργασίας της ανοίγουν νέους ορίζοντες.

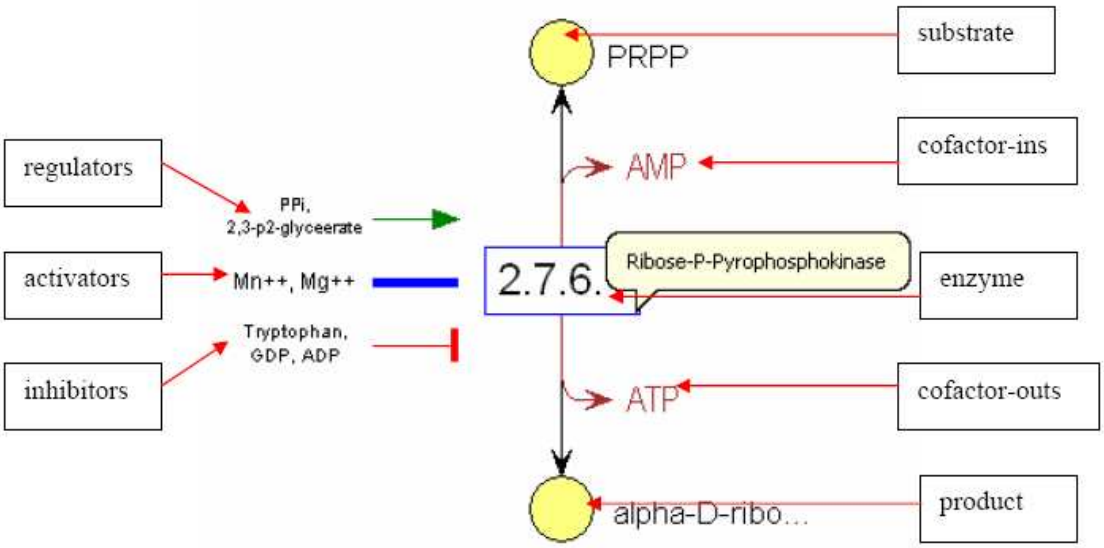
Παράδειγμα χρήσιμης αλλά παραδοσιακά σχεδόν αδύνατης ερώτησης είναι κατά πόσο υπάρχουν πολλά ένζυμα που μπορούν να καταλύσουν μια συγκεκριμένη αντίδραση. Σε περίπτωση που αυτά είναι πολλά, τότε η παραγωγή φαρμάκου που αναστέλλει τη δράση ενός από αυτά δε θα έχει νόημα. Το ερώτημα αυτό σήμερα μπορεί να απαντηθεί τουλάχιστον για το μεταβολικό δίκτυο της Escherichia Coli.

Στους τύπους δεδομένων χρειάζεται να συμπεριληφθούν και τα μόρια ή συμπλέγματα μορίων που ¨κυκλοφορούν στα μονοπάτια¨. Αυτά είναι κυρίως τα αντιδρώντα και προϊόντα χημικών αντιδράσεων (χημικές ενώσεις), οι καταλύτες (ένζυμα), ουσίες που ενεργοποιούν μια αντίδραση καθώς και ουσίες μέσω των οποίων μεταφέρονται σήματα.

7.2.3 Μοντέλο δεδομένων

Τα βασικά στοιχεία που συνθέτουν το μοντέλο δεδομένων περιλαμβάνουν τους τύπους δεδομένων που αναφέρθηκαν στην προηγούμενη παράγραφο. Είναι, δηλαδή, όλα τα στοιχεία που λαμβάνουν μέρος έχοντας οποιοδήποτε ρόλο σε μια αντίδραση, καθώς και δύο ακόμα.

Τα τελευταία είναι εκείνα που ονομάζονται μοριακές οντότητες (molecular entities) και διαδικασίες (processes). Ο όρος molecular entity αναφέρεται σε οποιοδήποτε στοιχείο παίρνει μέρος σε αντίδραση, ενώ ως process εννοείται οποιαδήποτε αντίδραση στην οποία συμμετέχουν μοριακές οντότητες. (Σχήμα 7.9).



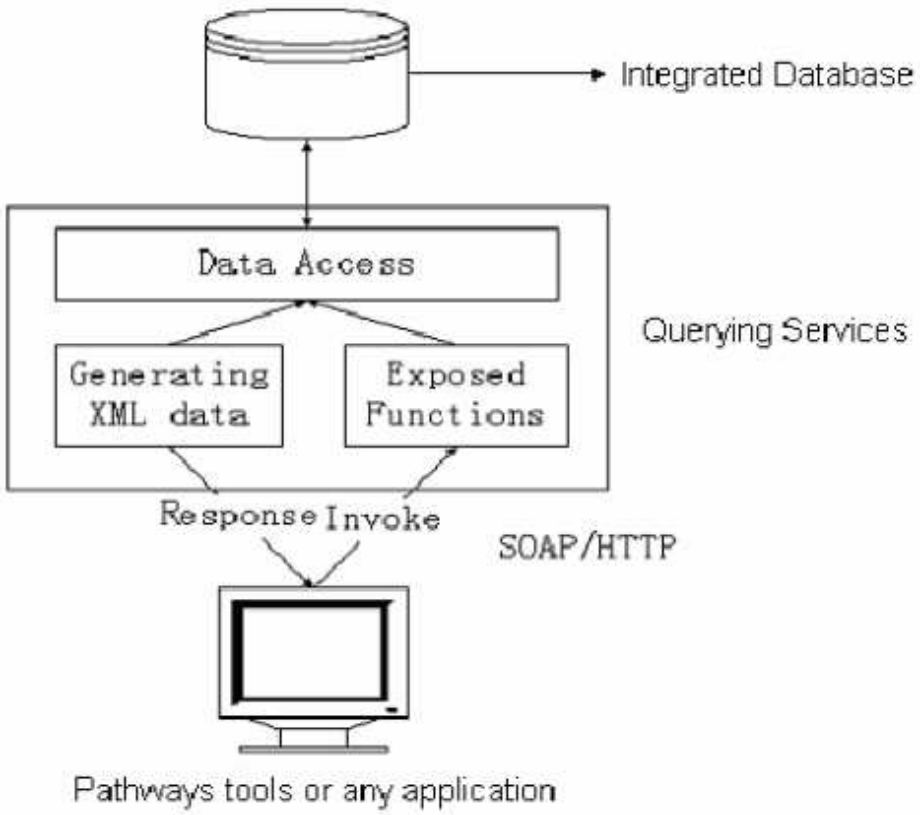
Σχήμα 7.9: Μία process και οι εμπλεκόμενες molecular entities.

Τα metabolic pathways μοντελοποιούνται ως σύνολα από processes. Πιο συγκεκριμένα έχουν τη μορφή γράφου με κορυφές τα processes και πιθανές ακμές ουσίες που μετακινούνται από κορυφή σε κορυφή, για παράδειγμα ως προϊόν μιας αντίδρασης και αντιδρών μιας άλλης. Ωστόσο, στη βάση δεδομένων αποθηκεύονται απλά κάποια χαρακτηριστικά του μονοπατιού (όνομα, τύπος, αναφορές, κ.ά.) και το σύνολο των processes που εμπεριέχονται σε αυτό.

Από την άλλη τα signaling pathways, που κωδικοποιούν την επικοινωνία μεταξύ κυττάρων, μοντελοποιούνται ως σύνολα από signaling steps (βήματα σημάτων). Ένα signaling step θεωρείται σαν μια κατευθυνόμενη ακμή από την οντότητα που στέλνει το σήμα προς εκείνη που το λαμβάνει. Ταυτόχρονα τηρούνται διάφορα στοιχεία για τις οντότητες που στέλνουν τα σήματα, όπως η τοποθεσία τους, ο ρόλος, η οικογένεια μορίων στην οποία ανήκουν. Χρειάζεται να σημειωθεί ότι δεν είναι απόλυτα σαφές ποια ακολουθία βημάτων συνθέτει ένα μονοπάτι σημάτων. Για το λόγο αυτό στη βάση του PathCase κρατώνται όλοι οι δυνατοί συνδυασμοί.

7.2.4 Αρχιτεκτονική συστήματος

Το σύστημα αποτελείται από τρία επίπεδα και έχει αναπτυχθεί στην πλατφόρμα .Net, ενώ είναι υπό εξέλιξη και μία Java έκδοσή του. Στο υψηλότερο επίπεδο βρίσκονται όλα τα εργαλεία που προσφέρουν τη διαπροσωπεία με το χρήστη. Αμέσως χαμηλότερα βρίσκονται το σύστημα που εξυπηρετεί τα queries (Pathway Querying Services) και εκείνο που ασχολείται με την εξαγωγή των δεδομένων (Data Extraction Services). Στο χαμηλότερο επίπεδο βρίσκεται μια σχεσιακή βάση δεδομένων. Στο Σχήμα 7.10 που δίνεται στη συνέχεια φαίνεται η αρχιτεκτονική στην οποία βασίζονται τα Querying Services.



Σχήμα 7.10: Η αρχιτεκτονική του συστήματος.

7.2.5 Τρόποι χρήσης

Δύο είναι οι μέθοδοι με τις οποίες μπορεί κανείς να χρησιμοποιήσει το PathCase. Η πρώτη είναι εγκαθιστώντας το στο δικό του υπολογιστικό σύστημα (PathCase Desktop Client) και η δεύτερη είναι απευθείας μέσω του αντίστοιχου δικτυακού τόπου που δίνεται στις αναφορές (PathCase Web Client).

Με τον πρώτο τρόπο παρέχονται όλα τα εργαλεία και οι δυνατότητες του συστήματος. Για να μπορέσει κανείς να το εγκαταστήσει στο σύστημά του, είναι απαραίτητο να λάβει σχετική άδεια στέλνοντας email στη διεύθυνση που αναφέρεται στο σχετικό web site. Προκειμένου

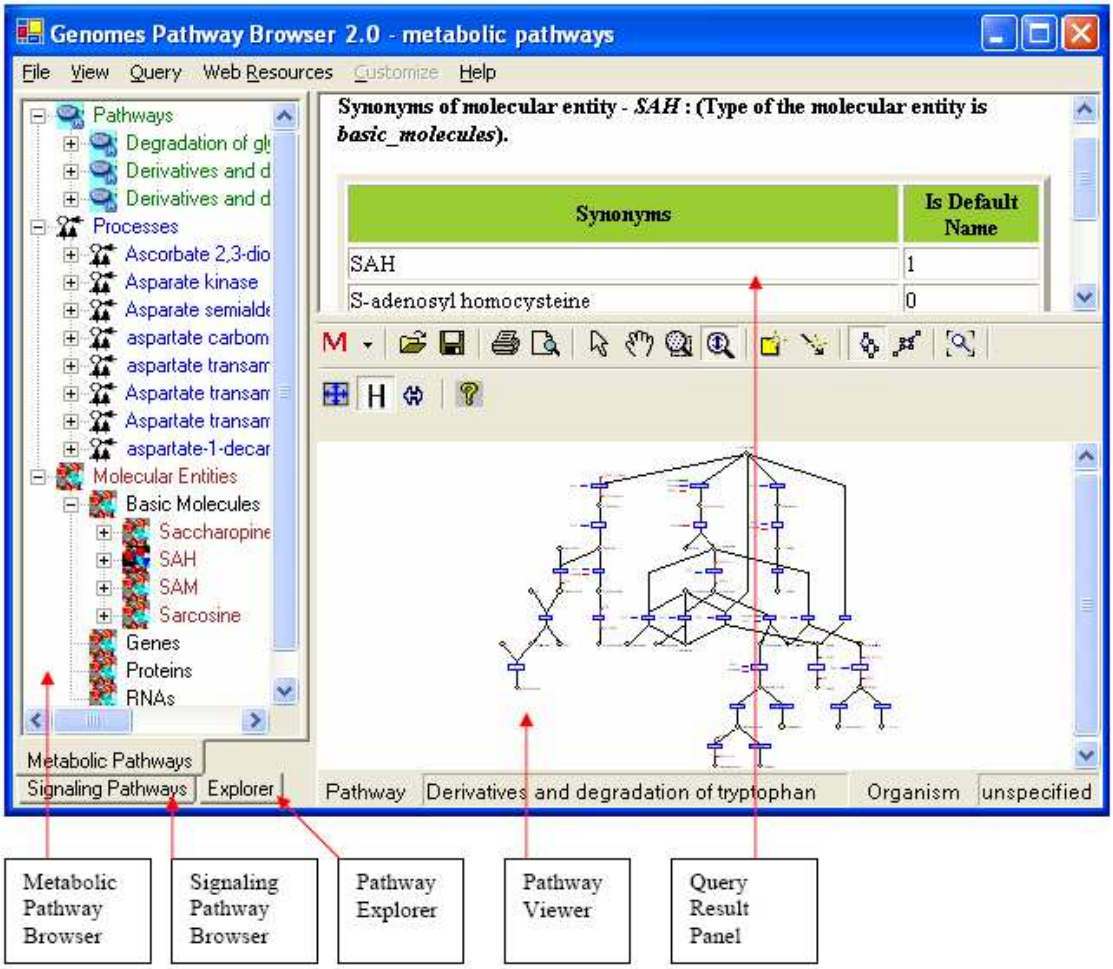
να μπορεί να τρέχει τις εφαρμογές από το δικό του σύστημα είναι απαραίτητη σύνδεση με το Internet τουλάχιστον 56k.

Ο τρόπος μέσω του Web Client, αν και απαιτεί προφανώς σύνδεση στο Internet, δεν έχει άλλους περιορισμούς. Προσφέρονται, όμως, μόνο οι Querying Services. Στην παρούσα εργασία, ωστόσο, το PathCase χρησιμοποιήθηκε με τον δεύτερο τρόπο, καθώς κρίθηκε ικανοποιητικός.

Τέλος, το σύνολο των εργαλείων που αποτελούν το PathCase τρέχει σε Windows και .Net.

7.2.6 Βασικά εργαλεία

Στο Σχήμα 7.11 φαίνονται τα κυριότερα στοιχεία που συνθέτουν την εικόνα του PathCase. Αυτά είναι οι δύο Pathway Browsers (Metabolic and Signaling) και οι Pathway Explorer και Viewer. Επιπλέον αυτών των εργαλείων δε φαίνονται στην εικόνα τα Pathway Editor και J-Viewer.

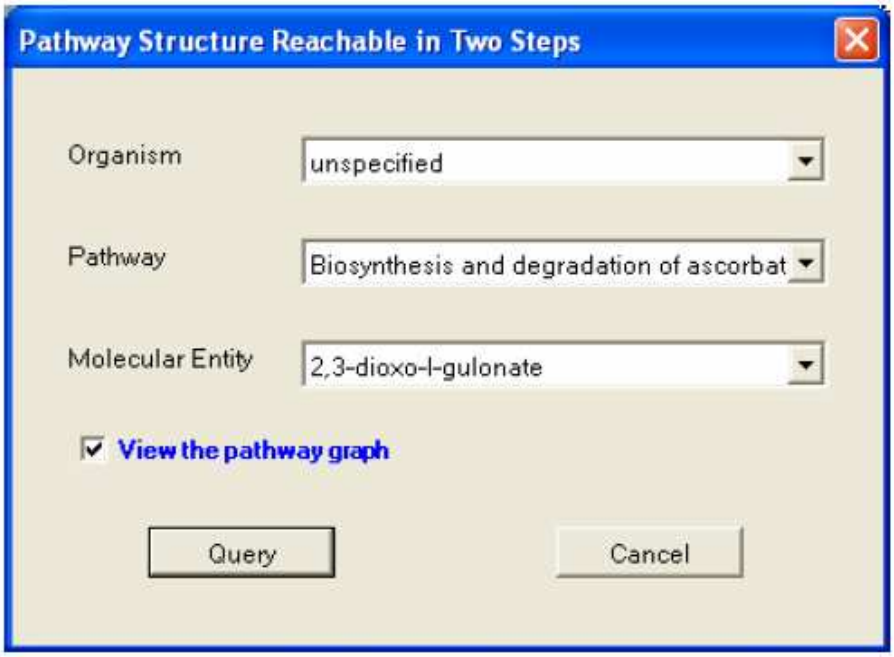


Σχήμα 7.11: Το γραφικό περιβάλλον του PathCase.

Στόχος της παρούσας ενότητας είναι να δώσει τη γενική εικόνα για τη χρήση του PathCase, δηλαδή τη χρήση αυτών των εργαλείων, χωρίς όμως να αναφερθεί σε λεπτομέρειες που εύκολα μπορούν να αναζητηθούν και καλύπτονται στα αντίστοιχα manuals.

Pathway Browser

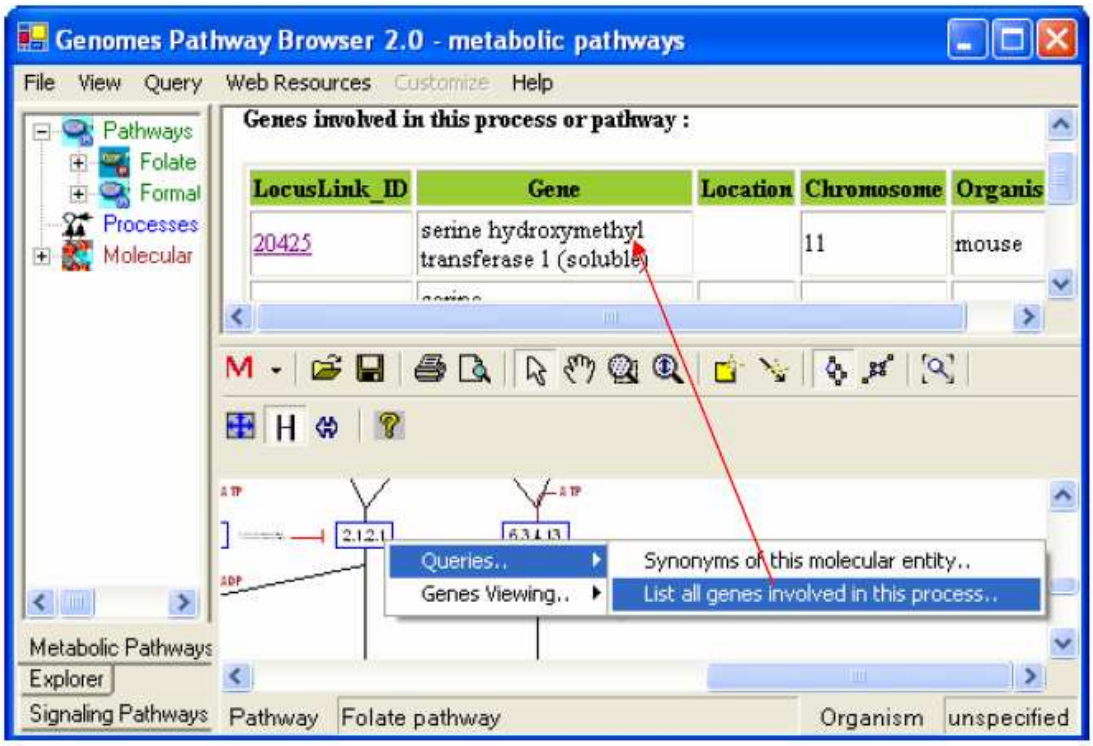
Παρέχει σε μορφή δέντρου καθένα από τα δύο είδη μονοπατιών (metabolic, signaling) που υποστηρίζει, καθώς και τα στοιχεία που τα αποτελούν, δηλαδή processes, molecular entities στην πρώτη περίπτωση και signaling steps, molecular entities στη δεύτερη, με βάση όσα αναφέρθηκαν στην ενότητα 7.2.3. Για τα μονοπάτια μεταβολισμού μάλιστα, μπορούν να τεθούν queries είτε προκαθορισμένα είτε παραμετροποιήσιμα από το χρήστη. Παράδειγμα δίνεται στο Σχήμα 7.12.



Σχήμα 7.12: Παράθυρο του Pathway Browser.

Pathway Viewer

Όπως προδηλώνει το όνομά του, το εργαλείο αυτό δίνει στον χρήστη εικόνα σε μορφή δέντρου ενός ή περισσοτέρων μονοπατιών. Επιτρέπει, επίσης, τη μεταπήδηση από signaling σε metabolic pathways και αντίστροφα, όπου αυτό είναι δυνατό. Στο Σχήμα 7.13 φαίνεται ένα query που μπορεί να τεθεί σαν αποτέλεσμα της αλληλεπίδρασης του Pathway Browser και του Pathway Viewer.



Σχήμα 7.13: Ενδεικτικό query.

Pathway Explorer

Μέσω αυτού μπορεί να τεθεί ένας μεγάλος αριθμός ερωτημάτων για τα μονοπάτια χρησιμοποιώντας φόρμες σε μορφή δέντρου. Χρειάζεται να προσδιοριστεί το μονοπάτι, η αντίδραση, το βιομόριο και ο οργανισμός που ενδιαφέρουν. Για καθένα από αυτά χρειάζεται να επιλεχθούν τα στοιχεία που είναι επιθυμητό να εμφανίζονται στο αποτέλεσμα. Αυτό το αποτέλεσμα δίνεται σε μορφή πίνακα και γράφου.

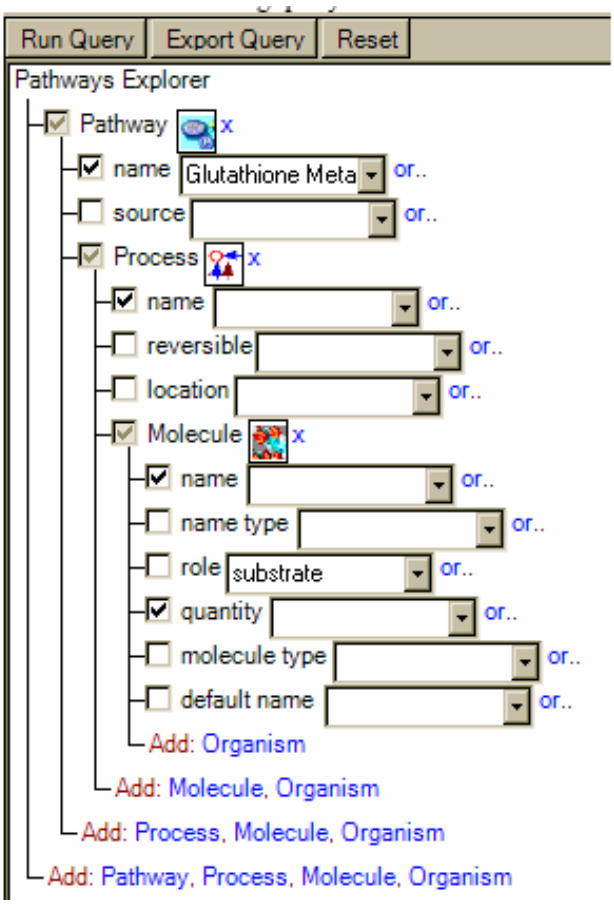
Στα Σχήματα 7.14-7.16 φαίνονται το ερώτημα στον pathway explorer και το αποτέλεσμα που προκύπτει στις δύο μορφές αναπαράστασης. Στο Σχήμα 7.17 φαίνεται μια ακόμη πιο ενδιαφέρουσα περίπτωση, όπου έχει τεθεί στο ερώτημα ένα join.

Pathway Editor

Με το εργαλείο αυτό ο χρήστης μπορεί να δημιουργήσει καινούρια pathways ή να τροποποιήσει ήδη υπάρχοντα αλλάζοντας το όνομα οντότητας, μετακινώντας processes ή να προσθέσει links από ένα pathway σε ένα άλλο. Το Σχήμα 7.18 δείχνει τη φόρμα στην οποία ζητώνται να συμπληρωθούν τα στοιχεία ενός νέου μονοπατιού.

J-Viewer

Πρόκειται για μια Java έκδοση ενός γραφικού εργαλείου για την αναπαράσταση των γράφων των μονοπατιών. Έχει αναπτυχθεί στα πλαίσια της δημιουργίας μιας Java έκδοσης του Path-

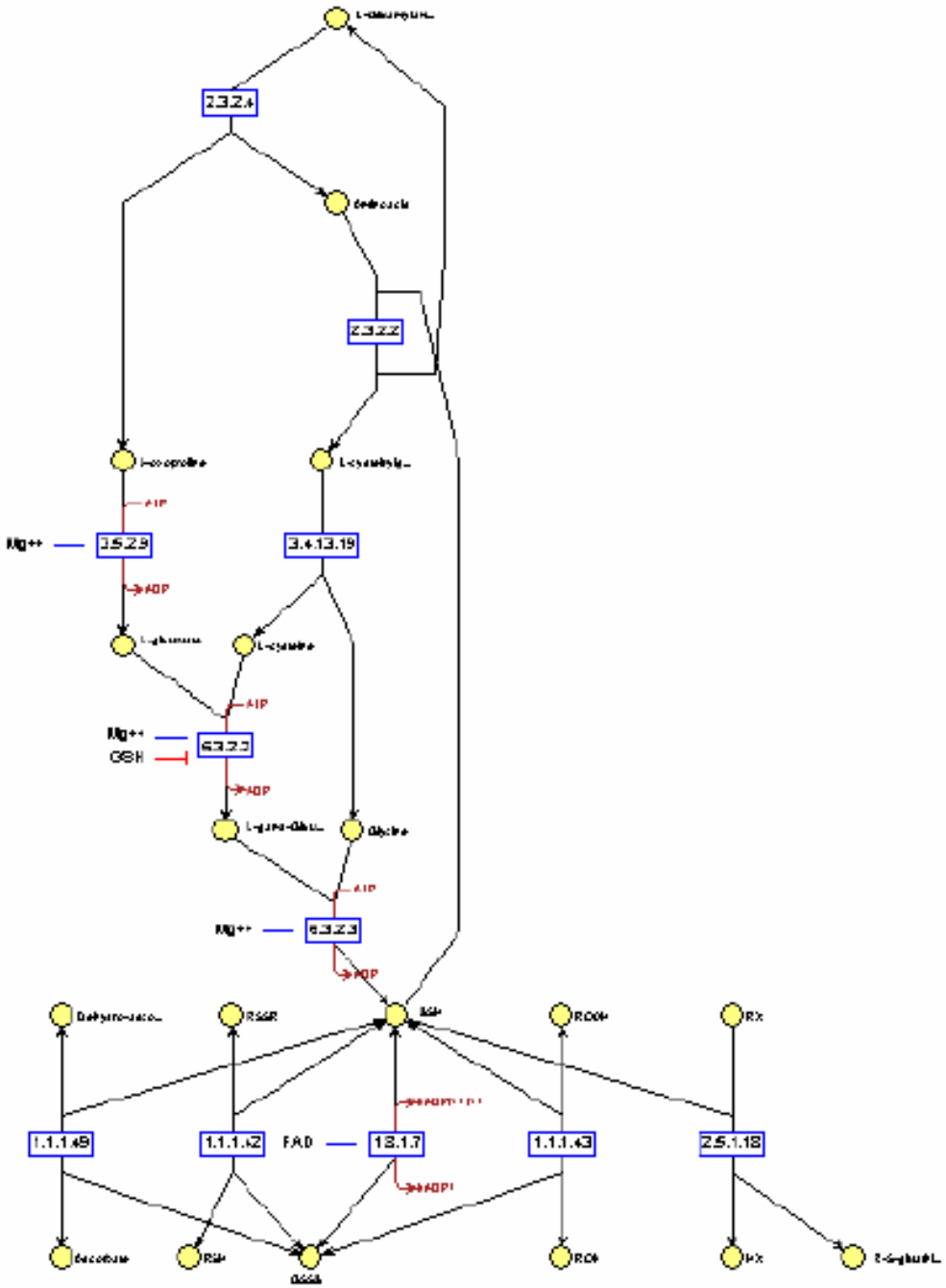


Σχήμα 7.14: Query στον Pathway Explorer.

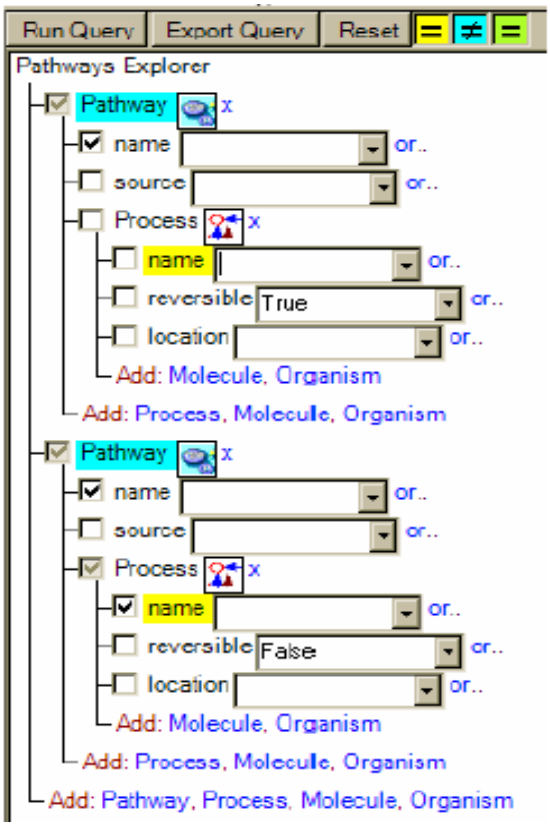
(46 results)

Pathway: name	Process: name	Molecule: name	Molecule: quantity
Glutathione Metabolism	5-Oxoprolinase(ATP hydolyzing)	5-oxoproline	1
Glutathione Metabolism	5-Oxoprolinase(ATP hydolyzing)	H2O	1
Glutathione Metabolism	5-Oxoprolinase(ATP hydolyzing)	su	1
Glutathione Metabolism	5-Oxoprolinase(ATP hydolyzing)	water	1
Glutathione Metabolism	Dipeptidase	H2O	1
Glutathione Metabolism	Dipeptidase	L-cysteinylglycine	1
Glutathione Metabolism	Dipeptidase	su	1
Glutathione Metabolism	Dipeptidase	water	1
Glutathione Metabolism	gamma-glutamylcyclotransferase	L-g-glutamylamino acid	1
Glutathione Metabolism	gamma-glutamylcyclotransferase	L-Glutamylamino acid	1
Glutathione Metabolism	gamma-glutamyltranspeptidase(gamma GT)	aa	1
Glutathione Metabolism	gamma-glutamyltranspeptidase(gamma GT)	Amino acid	1
Glutathione Metabolism	gamma-glutamyltranspeptidase(gamma GT)	Glutathione	1
Glutathione Metabolism	gamma-glutamyltranspeptidase(gamma GT)	GSH	1
Glutathione Metabolism	gamma-glutamyltranspeptidase(gamma GT)	Red. glutathione	1
Glutathione Metabolism	gamma-glutamyltranspeptidase(gamma GT)	Red. glutathione (GSH)	1
Glutathione Metabolism	Glutamate cysteine ligase	L-cysteine	1
Glutathione Metabolism	Glutamate cysteine ligase	L-glutamate	1
Glutathione Metabolism	Glutathione dehydrogenase(ascorbate)	Dehydro-ascorbate	1
Glutathione Metabolism	Glutathione dehydrogenase(ascorbate)	Glutathione	1
Glutathione Metabolism	Glutathione dehydrogenase(ascorbate)	GSH	1

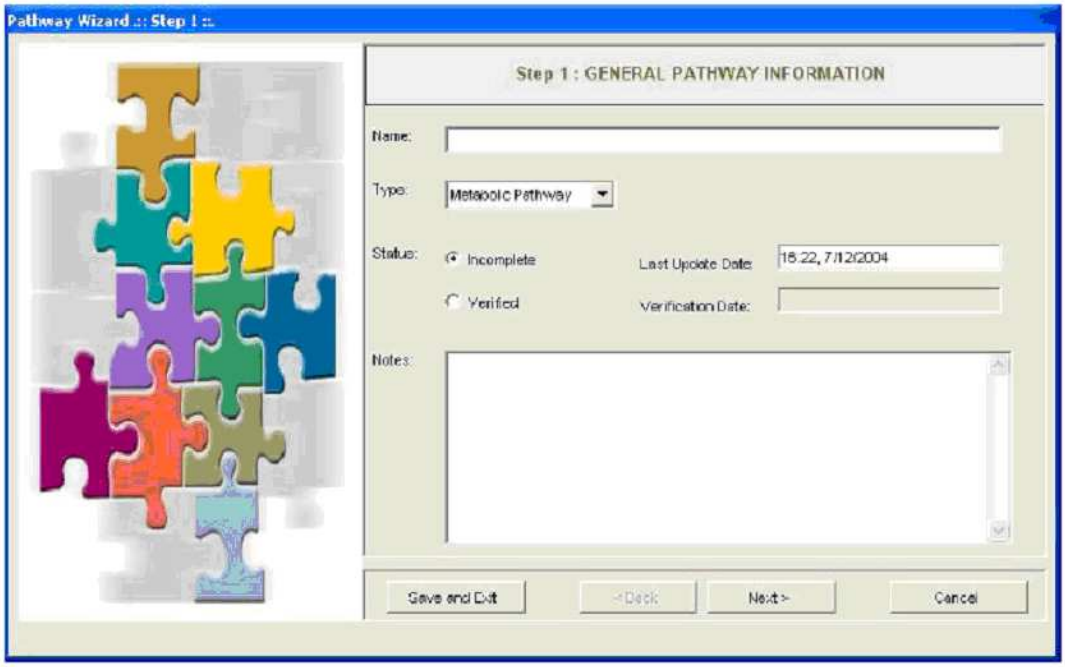
Σχήμα 7.15: Λίστα των απαντήσεων στο query του Σχήματος 7.14.



Σχήμα 7.16: Η απάντηση στο query του Σχήματος 7.14 σε μορφή γράφου.

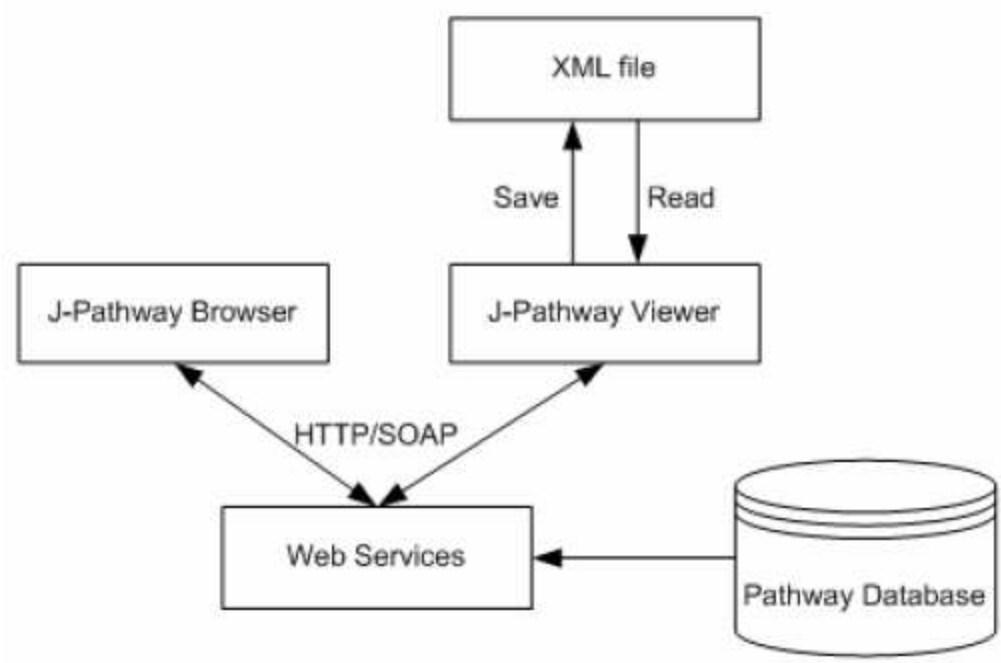


Σχήμα 7.17: Ένα ερώτημα join.



Σχήμα 7.18: Το παράθυρο του Pathway Editor.

Case. Η αρχιτεκτονική του νέου συστήματος, που περιλαμβάνει και τον J-Browser, φαίνεται στο Σχήμα 7.19.



Σχήμα 7.19: Η αρχιτεκτονική του συστήματος με τον J-Viewer.

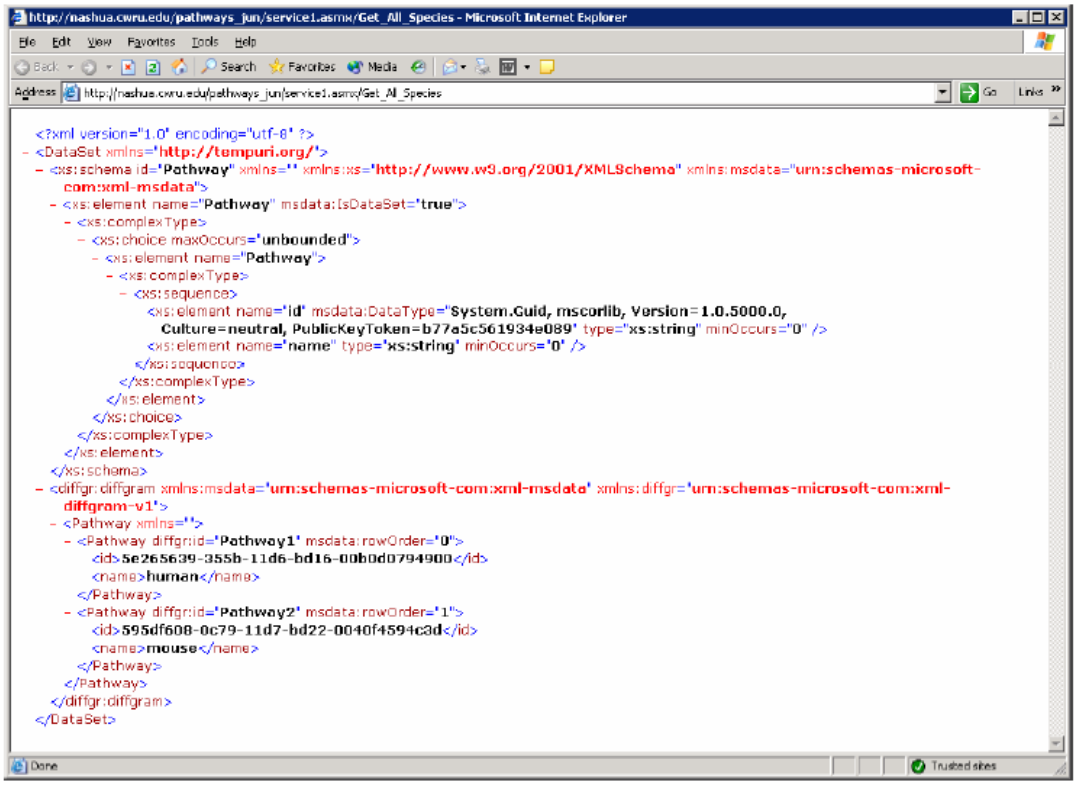
7.2.7 Web Services

Εκτός από τα βασικά εργαλεία ο χρήστης του PathCase έχει επιπλέον στη διάθεσή του περισσότερες από 40 έτοιμες συναρτήσεις ερωτημάτων (query functions) στις οποίες μπορεί να έχει πρόσβαση μέσω της αντίστοιχης ηλεκτρονικής σελίδας. Οι συναρτήσεις αυτές χωρίζονται σε 5 κατηγορίες ανάλογα με το ποιον αφορούν: μοριακές οντότητες (15), αντιδράσεις (10), μονοπάτια (9), οργανισμούς (3), άλλα (2).

Στη συνέχεια παρατίθεται από ένα παράδειγμα query function για κάθε κατηγορία με τη σειρά που αυτές αναφέρθηκαν.

- Εύρεση όλων των ρόλων που λαμβάνει μία μοριακή οντότητα σε όλα τα pathways ενός οργανισμού.
- Εύρεση των ids των αντιδράσεων που καταλύονται από συγκεκριμένο προϊόν ενός γονιδίου.
- Εύρεση όλων των μονοπατιών που το όνομά τους περιέχει δοσμένο string.
- Εύρεση των πληροφοριών για τα χρωμοσώματα ενός οργανισμού.
- Εύρεση όλων των ονομάτων και των ids όλων των ειδών.

Στην περίπτωση του τελευταίου ερωτήματος, η απάντηση που δίνεται φαίνεται στο Σχήμα 7.20.



Σχήμα 7.20: Η απάντηση στο ερώτημα για την εύρεση των ονομάτων και των ids.

7.3 Σύνοψη

Στο κεφάλαιο αυτό έγινε μια εισαγωγική παρουσίαση δύο σημαντικών εργαλείων. Μελετήθηκαν το πρόγραμμα BLAST και το σύστημα PathCase.

Το BLAST και το PathCase ασχολούνται με δύο καίριας σημασίας, αλλά διαφορετικές μεταξύ τους κατηγορίες προβλημάτων. Η πρώτη κατηγορία είναι η σύγκριση ακολουθιών των δομικών στοιχείων των πρωτεϊνών ή των νουκλεϊκών οξέων, ενώ η δεύτερη είναι η αποτύπωση και επεξεργασία των διαδρομών μορίων, συμπλεγμάτων μορίων ή ηλεκτρικών σημάτων μέσα στους οργανισμούς.

Η σύγκριση ακολουθιών των βιομορίων που αναφέρθηκαν, στον ίδιο ή σε διαφορετικούς οργανισμούς, έχει μεγάλη αξία στη μοριακή βιολογία. Η εύρεση όμοιων τέτοιων ακολουθιών μπορεί να οδηγήσει στην ανακάλυψη κοινού προγόνου βάσει της εξελικτικής θεωρίας, στην πρόβλεψη νέων μελών σε οικογένειες γονιδίων, στην εύρεση του γονιδίου που είναι υπεύθυνο για την κωδικοποίηση συγκεκριμένων πρωτεϊνών. (Κεφάλαιο 5)

Η γνώση και εύκολη επεξεργασία των μεταβολικών και βιοχημικών μονοπατιών, όπως και των διαδρομών που ακολουθούνται για τη σύνθεση των πρωτεϊνών και τη μετάδοση σημάτων

παρέχουν επίσης ένα πολύ ισχυρό εργαλείο στη μοριακή βιολογία. Αφενός συστηματοποιείται και αποθηκεύεται ένα πολύ μεγάλο κομμάτι επιστημονικής γνώσης, δηλαδή όχι μόνο δεδομένα που προέρχονται από μετρήσεις. Αφετέρου γίνεται δυνατή η απάντηση ερωτημάτων που διαφορετικά θα ήταν σχεδόν αδύνατη, όπως το αν υπάρχουν πολλά ένζυμα που μπορούν εναλλακτικά να καταλύσουν την ίδια χημική αντίδραση.

Ο λόγος για τον οποίο τα παραπάνω ενδιαφέρουν την τεχνολογία των βάσεων δεδομένων είναι το γεγονός ότι η τελευταία κατέχει σημαντική θέση στην επίλυση αυτών των προβλημάτων. Η σύγκριση των ακολουθιών σε όλες σχεδόν τις περιπτώσεις εμπλέκει μια βάση δεδομένων, καθώς η μία τουλάχιστον από τις δύο συγκρινόμενες ακολουθίες βρίσκεται εκεί. Ο τρόπος με τον οποίο είναι αποθηκευμένη, αλλά και αυτός με τον οποίο γίνεται η προσπέλασή της, επηρεάζει σημαντικά τη χρονική κυρίως πολυπλοκότητα της σύγκρισης. Ταυτόχρονα, η τεχνολογία των βάσεων δεδομένων έχει ακόμη πολλά να προσφέρει στην αποδοτική αποθήκευση νέας μορφής δεδομένων, όπως είναι τα τρισδιάστατα πρωτεϊνικά μόρια αλλά και στην αποθήκευση και επεξεργασία επιστημονικής γνώσης.

Πιστεύεται πως η έρευνα είναι αρκετή για να δώσει το στίγμα καθενός από αυτά, αλλά και να ενθαρρύνει τη μεγαλύτερη εξέτασή τους. Δυνατές επεκτάσεις πάνω στο θέμα συζητώνται στην Ενότητα 8.2.

Κεφάλαιο 8

Επίλογος

Οι τελευταίες αυτές σελίδες συνοψίζουν όσα αναφέρθηκαν στις προηγούμενες και προτείνουν κατευθύνσεις για περαιτέρω έρευνα. Αν και δεν είναι δυνατόν να περιληφθεί σε αυτές το περιεχόμενο ολόκληρης της εργασίας, ο ρόλος τους είναι να βοηθήσουν τον αναγνώστη να θυμηθεί τις διαδρομές από τις οποίες πέρασε και να δει τους καινούριους δρόμους που διαφαίνονται.

8.1 Σύνοψη και συμπεράσματα

Η παρούσα διπλωματική εργασία φιλοδοξεί να εξυπηρετήσει ως αναφορά για τα θέματα που αγγίζει. Ευελπιστεί να θέσει τα θεμέλια, ώστε να μπορούν να γίνουν μελέτες και έρευνες σε πιο εξειδικευμένες πτυχές τους.

Το περιεχόμενό της, επομένως, επιλέχθηκε να είναι τέτοιο που να καλύπτει ένα όσο το δυνατόν μεγαλύτερο εύρος θεμάτων. Ο αναγνώστης μπορεί να αντλήσει από αυτήν τις βασικές γνώσεις βιολογίας που είναι απαραίτητες στον μηχανικό υπολογιστών, για να αρχίσει την έρευνά του. Επίσης, βρίσκει τα πιο σημαντικά στοιχεία που συνθέτουν το τοπίο που η βιοπληροφορική έχει δημιουργήσει, στην προσπάθειά της να λύσει προβλήματα των βιοεπιστημών. Έχοντας αποσαφηνίσει το χαρακτήρα των δεδομένων και των ερωτημάτων στο χώρο, ο ερευνητής των βάσεων δεδομένων αναγνωρίζει πλέον το ρόλο του στα προβλήματα στα οποία καλείται εκείνος να δώσει λύση.

Πιστεύεται ότι ο μελετητής της εργασίας θα συμμεριστεί την άποψη πως υπάρχει μεγάλο ενδιαφέρον αλλά και σπουδαίες προκλήσεις στις εφαρμογές της τεχνολογίας των βάσεων δεδομένων για τις βιοεπιστήμες. Η συγγραφέας φέρει αποκλειστικά την ευθύνη για πιθανά λάθη ή ασάφειες που ίσως τον ταλαιπωρήσουν.

8.2 Μελλοντικές επεκτάσεις

Ο εισαγωγικός χαρακτήρας της εργασίας ανοίγει πολλά ζητήματα για επέκταση. Εξάλλου, είναι πολύ νέος ο χώρος των εφαρμογών των συστημάτων βάσεων δεδομένων στις βιοεπιστήμες, γεγονός που δείχνει ότι μάλλον έχει πολλά ακόμα να προσφέρει.

Με εξαίρεση τα Κεφάλαια 2 και 3, όλα τα υπόλοιπα επιδέχονται επεκτάσεων. Μπορεί κανείς να μελετήσει πιο συγκεκριμένα και περισσότερα πρότυπα και formats των βάσεων δεδομένων που υπάρχουν ήδη (Κεφάλαιο 4). Ίσως επιθυμεί να ψάξει παραπάνω στοιχεία για κάποια λειτουργία που γίνεται πάνω στα δεδομένα ή να ασχοληθεί με προγράμματα και συστήματα που τις εκτελούν (Κεφάλαιο 5). Ιδιαίτερο ενδιαφέρον πιθανώς να βρει στην ανακάλυψη των λεπτομερειών που συνθέτουν τα προβλήματα που υπάρχουν, ώστε να προσεγγίσει πιο εύκολα τη λύση τους (Κεφάλαιο 6). Είναι, επίσης, δυνατή μια εξαντλητική μελέτη των BLAST, PathCase με τη διερεύνηση όλων των δυνατών τρόπων χρήσης τους, καθώς και των δυνατοτήτων τους (Κεφάλαιο 7). Στα επόμενα, περιγράφονται πιο αναλυτικά οι προτάσεις αυτές.

Το κριτήριο, που φαίνεται να είναι το πιο λογικό να επιλεχθεί για να ακολουθηθούν οι πιο καίριες επεκτάσεις, είναι οι ίδιες οι ανάγκες των ερευνητών των βιοεπιστημών. Με άλλα λόγια, για να διαλέξει κανείς πώς θα βοηθήσει, χρειάζεται να ξέρει τι είναι ήδη διαθέσιμο και ποιο είναι το πραγματικό ζητούμενο. Ενδιαφέρει η ερώτηση κατά πόσον οι υπάρχοντες πόροι (εργαλεία και βάσεις δεδομένων) ανταποκρίνονται με τον καλύτερο δυνατό τρόπο (καταβολή ελάχιστης δυνατής προσπάθειας, βέλτιστη χρονική και χωρική πολυπλοκότητα) στις εργασίες που γίνονται από τους εν λόγω επιστήμονες. Η αναζήτηση της απάντησης γίνεται μέσω της έρευνας αρκετών πηγών, όπως εξηγείται στη συνέχεια.

Η εξέταση των βάσεων που εξυπηρετούν σήμερα την αποθήκευση των δεδομένων είναι μια καλή εκκίνηση. Η εξέταση αυτή μπορεί να αφορά σε αρχική φάση τις πρωταρχικές βάσεις, εκείνες που αποτελούν την πηγή όλων των υπολοίπων, και να συνεχίσει στις πιο εξειδικευμένες, οι οποίες δεν χρησιμοποιούνται λιγότερο. Η μελέτη είναι δυνατό να αφορά τόσο τα μοντέλα και format αποθήκευσης, όσο και τα διάφορα εργαλεία που συνήθως διαθέτουν οι βάσεις αυτές για την αναζήτηση ή άλλες εργασίες κατά περίπτωση.

Αρκετές από τις υποψήφιες βάσεις αναφέρονται στην ενότητα 4.4. Συνιστάται η εξερεύνηση του site του NCBI που προσφέρει πρόσβαση στην GenBank και στην PubMed, οι οποίες βασίζονται στο ASN.1, ενώ δίνει και τη δυνατότητα χρήσης του συστήματος Entrez. Οι βάσεις δεδομένων SWISS-PROT, PDB είναι συχνά χρησιμοποιούμενες και διαφέρουν ως προς τη μορφή από τη GenBank, αλλά και μεταξύ τους. Στον τομέα των βιομονοπατιών η KEGG είναι η πλέον αναγνωρίσιμη. Από εκεί και πέρα, πλήθος μικρότερων βάσεων είναι διαθέσιμο. Για παράδειγμα, η Gene Ontology, η οποία περιέχει λέξεις και όρους [2] ή η EcoCyc, η οποία έχει στοιχεία αποκλειστικά για το βακτήριο Escherichia coli [44].

Έχει νόημα, επίσης, να ψάξει κανείς σε μεγαλύτερο βάθος πρότυπα που δε χρησιμοποιούνται ευρέως ακόμη αλλά είναι αρκετά γνωστά. Στην ενότητα 4.3.2 αναφέρονται κάποια που σχετίζονται με την XML. Η ανακάλυψη λ.χ. των διαφορών τους προσφέρει αρκετή βοήθεια στην κατανόηση των ιδιαίτερων απαιτήσεων που έχει κάθε είδος δεδομένων, αλλά και στις διαφορετικές προσεγγίσεις που είναι δυνατές για την αντιμετώπιση ίδιου τύπου δεδομένων.

Από την άλλη πλευρά, η εκτενής μελέτη των λειτουργιών και προγραμμάτων μπορεί να αποφέρει και αυτή σημαντικούς καρπούς. Τουλάχιστον δύο τρόποι έρευνας είναι ορατοί. Ο πρώτος αφορά τη μελέτη μεμονωμένων λειτουργιών, ενώ ο δεύτερος μια πιο σύνθετη.

Η μία πρόταση είναι η κατάλληλη επιλογή κάποιας λειτουργίας και η μελέτη της σε λε-

πτομέρεια. Είναι εμφανές ότι η επέκταση μπορεί να σχετίζεται και με τους αλγορίθμους που υπάρχουν για αυτήν αλλά και με τα αντίστοιχα προγράμματα. Η πιο βασική, ίσως, λειτουργία από εκείνες που περιγράφονται στην ενότητα 5.2 είναι το alignment, ενώ ταυτόχρονα είναι και η περισσότερο μελετημένη. Αντίθετα, ο προσδιορισμός της τρισδιάστατης δομής των μακρομορίων είναι λιγότερο εύκολα αντιμετωπίσιμος, με ανοιχτό ακόμη αρκετά το ερευνητικό μέτωπο. Πολύτιμος οδηγός για τα διαθέσιμα προγράμματα που αντιστοιχούν στις διάφορες λειτουργίες είναι το [4].

Μία περισσότερο πολύπλοκη προσέγγιση είναι η μελέτη αλυσίδας λειτουργιών. Έχει αναφερθεί ξανά (Παράγραφοι 5.3.3 και 6.1.3) ότι οι ερευνητές συνήθως εφαρμόζουν σειρά λειτουργιών για την εκτέλεση κάποιας εργασίας. Πιθανώς, λοιπόν, να δώσει πιο ρεαλιστικά αποτελέσματα η μελέτη συνδυασμού λειτουργιών. Τα αρχικά δεδομένα να τροφοδοτούν πρόγραμμα της μιας λειτουργίας και μέρος των αποτελεσμάτων της να είναι είσοδος για το πρόγραμμα της άλλης κ.ο.κ. Πέρα από τη βιβλιογραφία, είναι χρήσιμη και η επαφή με τους ερευνητές των βιοεπιστημών, ώστε να εντοπίσουν συγκεκριμένους τέτοιους συνδυασμούς.

Ο τομέας των προβλημάτων και λύσεων του έκτου κεφαλαίου οπωσδήποτε δεν στερείται διαθέσιμων επεκτάσεων. Καθεμιά από τις δυσκολίες που αναφέρονται στις εν λόγω παραγράφους μπορεί από μόνη της να αποτελέσει αντικείμενο έρευνας. Ιδιαίτερα η προέλευση των δεδομένων (data provenance) είναι θέμα σχετικά λιγότερο εξιχνιασμένο [56]. Επιπλέον, η ενοποίηση των δεδομένων (data integration) εμπεριέχει αρκετά επιμέρους προβλήματα, όπως είναι ο καθαρισμός (curation). Σημαντικός οδηγός στην προσπάθεια αντιμετώπισης των προβλημάτων αυτών είναι το [42].

Τέλος, το Κεφάλαιο 7 προσφέρει και αυτό αρκετό υλικό πάνω στο οποίο μπορεί να γίνει μεγαλύτερη έρευνα. Τόσο το εργαλείο BLAST όσο και το σύστημα PathCase είναι δυνατό να μελετηθούν σε περισσότερη ανάλυση. Στην παρούσα εργασία έγινε η αρχική παρουσίασή τους.

Όπως αναφέρεται και στην Ενότητα 7.1, το BLAST έχει αρκετές διαφορετικές μορφές. Ανάλογα με το σκοπό και τη βάση στην οποία θέλει κανείς να εκτελέσει το alignment χρησιμοποιεί και διαφορετική μορφή του προγράμματος. Οι δυνατοί συνδυασμοί είναι αρκετοί (το Σχήμα 7.2 είναι ενδεικτικό) και διαφέρουν αρκετά μεταξύ τους ως προς το είδος δεδομένων (νουκλεοτίδια ή αμινοξέα) αλλά και το μέγεθος των υπό σύγκριση ακολουθιών. Αναλυτικά βοηθητικά κείμενα είναι διαθέσιμα στην [37].

Το σύστημα PathCase έχει και εκείνο περιθώρια εξερεύνησης. Οι δυνατοί τρόποι χρήσης του είναι δύο (Ενότητα 7.2.5). Στην εργασία αναφέρεται ότι ο ένας από τους δύο προτιμήθηκε. Οι δυνατότητες του συστήματος, ωστόσο, γίνονται καλύτερα αντιληπτές, αν εγκατασταθεί το σύστημα στην πλατφόρμα του υπολογιστή. Η χρησιμοποίησή του, πέρα από τις web φόρμες που προσφέρονται στο [21], απαιτεί για την εγκατάσταση ειδική άδεια και ικανοποιητική σύνδεση με το Internet, όπως διευκρινίζεται και στο έβδομο κεφάλαιο.

Είναι βέβαιο πως ο ενδιαφερόμενος θα βρει παραπάνω από ένα μονοπάτι, για να επεκτείνει θέματα αυτής της εργασίας. Είθε να ευοδωθούν οι προσπάθειές του με τον καλύτερο τρόπο.

Βιβλιογραφία

[1] Stephen F. Altschul, Warren Gish, Webb Miller add Eugene W.Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215, 1990.

[2] M. Bada, D. Turi, R. McEntire, and R. Stevens. Using reasoning to guide annotation with Gene Ontology terms in GOAT. *SIGMOD Record*, 33(2), 2004.

[3] Pierre Baldi and Sören Brunak. *Bioinformatics: The Machine Learning Approach.* Massachusetts Institute of Technology, 2001.

[4] Andreas D. Baxevanis and B. F. Francis Ouellette. *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins.* John Wiley and Sons, Inc., 2001.

[5] Akmal B. Chaudhri, Awais Rashid, and Roberto Zicari. *XML Data Management: Native XML and XML-Enabled Database Systems*, chapter 10. Addison Wesley, 2003.

[6] Jacques Cohen. Bioinformatics: An introduction for computer scientists. *ACM Computing Surveys*, 36(2), 2004.

[7] Stephen F.Altschul, Thomas L.Madden, Alejandro A.Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J.Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17), 1997.

[8] ftp://ftp.expasy.ch.

[9] ftp://ftp.research.microsoft.com/pub/debull/A04dec/issue1.htm. Special Issue on Database Support for the Sciences, IEEE Data Engineering Bulletin, 27 (4), 2004.

[10] ftp://ftp.research.microsoft.com/pub/debull/A04sept/issue1.htm. Special Issue on Querying Biological Sequences, IEEE Data Engineering Bulletin, 27 (3), 2004.

[11] Cynthia Gibas and Per Jambeck. *Developing Bioinformatics Computer Skills.* O"Reilly and Associates, Inc., 2001.

[12] Jim Gray and Alex Szalay. Where the rubber meets the sky: Bridging the gap between databases and science. *IEEE Data Engineering Bulletin*, 27(4), 2004.

[13] J. Hammer and M. Schneider. The GenAlg project. *SIGMOD Record*, 33(2), 2004.

[14] http://blast.wustl.edu/.

[15] http://ca.expasy.org/enzyme/.

[16] http://ca.expasy.org/sprot/.

[17] http://corba.ebi.ac.uk/Biocatalog/.

[18] http://genome www.stanford.edu/.

[19] http://mckoi.com/database/.

[20] http://molbio.info.nih.gov/molbio/db.html.

[21] http://nashua.cwru.edu/pathways.

[22] http://pir.georgetown.edu.

[23] http://www.aaai.org/Library/Classic/hunter.php.

[24] http://www.accessexcellence.org/AB/GG.

[25] http://www.biocyc.org.

[26] http://www.bioml.com.

[27] http://www.bsml.org.

[28] http://www.cbcb.umd.edu/ salzberg/.

[29] http://www.fruitfly.org/annot/gamexml.dtd.txt.

[30] http://www.genome.gov/.

[31] http://www.genome.jp/kegg/.

[32] http://www.informatik.uni-trier.de/~ley/db/journals/tkde/tkde17.html. Special Issue on Mining Biological Data, IEEE Transactions on Knowledge and Data Engineering, 17 (8), 2005.

[33] http://www.informatik.uni-trier.de/~ley/db/journals/vldb/vldb14.html. Special Issue on Data Management, Analysis, and Mining for the Life Sciences, The VLDB Journal, 14 (3), 2005.

[34] http://www.mcs.anl.gov/compbio.

[35] http://www.mged.org/Workgroups/MAGE/mage.html.

[36] http://www.ncbi.nlm.nih.gov.

[37] http://www.ncbi.nlm.nih.gov/BLAST.

[38] http://www.rcsb.org/pdb/.

[39] http://www.sigmod.org/sigmod/record/issues/0406/index.html. Special Section on Data Engineering for Life Sciences, SIGMOD Record, 33 (2), 2004.

[40] http://www.visualgenomics.ca/gordonp/xml.

[41] http://www.wikipedia.org/.

[42] H. V. Jagadish and F. Olken. Database management for life sciences research. *SIGMOD Record*, 33(2), 2004.

[43] Tamer Kahveci and Ambuj Singh. Progressive searching of biological sequences. *IEEE Data Engineering Bulletin*, 27(3), 2004.

[44] Peter D. Karp. Pathway databases: A case study in computational symbolic theories. *Science*, 293, 2001.

[45] Arthur M. Lesk. *Introduction to Bioinformatics*. Oxford University Press, 2002.

[46] Daniel P. Miranker, Willard J. Briggs, Rui Mao, Shulin Ni, and Weijia Xu. Biosequence use cases in mobios sql. *IEEE Data Engineering Bulletin*, 27(3), 2004.

[47] Sushmita Mitra and Tinku Acharya. *Data Mining: Multimedia, Soft Computing and Bioinformatics*. John Wiley and Sons, Inc., 2003.

[48] David W. Mount. *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor Laboratory University Press, 2001.

[49] Z. M. Ozsoyoglu, J. Nadeau, G. Ozsoyoglu, and M. Tasan. Towards an integrated software system for biological pathways. *IEEE Data Engineering Bulletin*, 27(4), 2004.

[50] Human Genome Program. Primer on Molecular Genetics. U.S. Department of Energy, June 1992.

[51] Francois Rechenmann. From data to knowledge. *Bioinformatics*, 16(2), 2000.

[52] S.L. Salzberg, D.B. Searls, and S. Kasif. *Computational Methods in Molecular Biology*. Elsevier, Amsterdam, The Netherlands, 1998.

[53] Erwin Schrödinger. *What is life*. Cambridge University Press, 1993.

[54] A. Silberschatz, H. F. Korth, and S. Sudarshan. *Database System Concepts*. The McGraw-Hill Companies Inc., 2002.

[55] A. K. Singh, B. S. Manjunath, and R. F. Murphy. A distributed database for biomolecular images. *SIGMOD Record*, 33(2), 2004.

[56] Wang-Chiew Tan. Research problems in data provenance. *IEEE Data Engineering Bulletin*, 27(4), 2004.

[57] Β. Αλεπόρου-Μαρίνου, Α. Αργυροκαστρίτης, Α. Κομητοπούλου, Π. Πιαλόγλου, Β. Σγουρίτσα. *Βιολογία Θετικής Κατεύθυνσης, Γ΄ Τάξης Ενιαίου Λυκείου.* Οργανισμός Εκδόσεων Διδακτικών Βιβλίων, Αθήνα, 2000.

[58] Α. Καψάλης, Ι. Ε. Μπουρμπουχάκης, Β. Περάκη, Σ. Σαλαμαστράκης. *Βιολογία Γενικής Παιδείας, Β΄ Τάξης Ενιαίου Λυκείου.* Οργανισμός Εκδόσεων Διδακτικών Βιβλίων, Αθήνα, 1999.

[59] Φ. Μπαρωνα-Μάμαλη, Ι. Μπότσαρης, Ι. Μπουρμπουχάκης, Β. Περάκη. *Βιολογία Γενικής Παιδείας, Γ΄ Τάξης Ενιαίου Λυκείου.* Οργανισμός Εκδόσεων Διδακτικών Βιβλίων, Αθήνα, 2000.

[60] Τίμος Σελλής. *Προχωρημένα Θέματα Βάσεων Δεδομένων, Συμπληρωματικές Σημειώσεις.* Ε. Μ. Π., 2005.