

**ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΓΕΩΠΟΝΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΓΕΩΠΟΝΙΑΣ ΙΧΘΥΟΛΟΓΙΑΣ
ΚΑΙ ΥΔΑΤΙΝΟΥ ΠΕΡΙΒΑΛΛΟΝΤΟΣ**

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Καταγραφή βάσεων δεδομένων κωδικοποιημένων και
μη κωδικοποιημένων
περιοχών των γονιδιωμάτων οστεϊχθύων και εφαρμογές τους

Νεκτάριος Δαβίδης

ΒΟΛΟΣ 2025

**UNIVERSITY OF THESSALY
SCHOOL OF AGRICULTURAL SCIENCES
DEPARTMENT OF ICHTHYOLOGY AND AQUATIC
ENVIRONMENT**

POSTGRADUATE MASTER'S THESIS

«Documentation of databases of coding and non coding regions of
osteichthyan genomes and their applications»

Nektarios Davidis

VOLOS 2025

Τριμελής Εξεταστική Επιτροπή:

1. Εξαδάκτυλος Αθανάσιος, Καθηγητής, Γενετική Υδροβίων Ζωϊκών Οργανισμών, Τμήμα Γεωπονίας Ιχθυολογίας και Υδάτινου Περιβάλλοντος, Σχολή Γεωπονικών Επιστημών, Πανεπιστήμιο Θεσσαλίας, Επιβλέπων.
1. Γεώργιος Γκάφας, Αναπληρωτής Καθηγητής, Μοριακή Βιολογία της Διατήρησης Θαλάσσιων Θηλαστικών και Ιχθυοαποθεμάτων, Τμήμα Γεωπονίας Ιχθυολογίας και Υδάτινου Περιβάλλοντος, Σχολή Γεωπονικών Επιστημών, Πανεπιστήμιο Θεσσαλίας, Μέλος.
3. Γρηγόριος Αμούτζιας, Καθηγητής Βιοπληροφορικής με έμφαση στην Μικροβιολογία, Τμήμα Βιοχημείας & Βιοτεχνολογίας, Σχολή Επιστημών Υγείας, Πανεπιστήμιο Θεσσαλίας, Μέλος.

ΕΥΧΑΡΙΣΤΙΕΣ

Κατά την εκπόνηση της παρούσας μεταπτυχιακής διπλωματικής εργασίας, θα ήθελα να εκφράσω την ειλικρινή μου ευγνωμοσύνη και τις θερμές μου ευχαριστίες σε όλους όσους συνέβαλαν με τον τρόπο τους στην επιτυχή ολοκλήρωσή της.

Πρώτα απ' όλα, οφείλω ιδιαίτερες ευχαριστίες στον επιβλέποντα καθηγητή μου, κ. Εξαδάκτυλο Αθανάσιο, για την εμπιστοσύνη που μου έδειξε εξ αρχής αναθέτοντάς μου το συγκεκριμένο θέμα, για την αδιάκοπη επιστημονική του καθοδήγηση, τις εύστοχες υποδείξεις του, καθώς και για το διαρκές ενδιαφέρον και τη στήριξή του καθ' όλη τη διάρκεια της εργασίας.

Θερμές ευχαριστίες επίσης απευθύνω στους καθηγητές Γεώργιο Γκάφα και Γρηγόριο Αμούτζια, μέλη της τριμελούς εξεταστικής επιτροπής, για τις επικοινωνητικές παρατηρήσεις τους και τη σημαντική τους συμβολή στην επιστημονική εμβάθυνση και ολοκλήρωση της παρούσας εργασίας.

Τέλος, θα ήθελα να ευχαριστήσω θερμά όλους τους συναδέλφους μου από την Περιφέρεια Θεσσαλίας με τους οποίους παρακολουθήσαμε από κοινού αυτό το Μεταπτυχιακό Πρόγραμμα. Η εμπειρία αυτή μας άνοιξε νέους ορίζοντες και μας έδωσε τη δυνατότητα να εμβαθύνουμε ερευνητικά σε ένα γνωστικό πεδίο εντελώς διαφορετικό από αυτό του βασικού μου πτυχίου στην Πληροφορική, συνδυάζοντας γνώσεις από διαφορετικές ειδικότητες

ΠΕΡΙΛΗΨΗ

Η παρούσα πτυχιακή εργασία επικεντρώνεται στη συστηματική καταγραφή, ανάλυση και αξιολόγηση των πιο σημαντικών βάσεων δεδομένων που αφορούν τόσο τις κωδικοποιημένες όσο και τις μη-κωδικοποιημένες περιοχές των γονιδιωμάτων των οστεϊχθύων – της πολυπληθέστερης ομάδας σπονδυλωτών στον πλανήτη. Εστιάζει ειδικότερα στη διερεύνηση των μεθόδων με τις οποίες ο μελετητής μπορεί να αναζητήσει πληροφορίες βιολογικού περιεχομένου και πώς μπορεί να αξιοποιήσει τα αποτελέσματα αυτά σε ερευνητικό ή εφαρμοσμένο πλαίσιο.

Η εργασία υλοποιήθηκε μέσα από δύο βασικούς άξονες: αφενός, την εις βάθος μελέτη εργαλείων βιοπληροφορικής όπως τα BLAST, FASTA, PSI-BLAST, HMMER, MAFFT, MUSCLE και Clustal Omega – και αφετέρου, τη δημιουργία μιας ενδεικτικής διαδικτυακής εφαρμογής (Flask App), η οποία λειτουργεί ως πρακτικό demo του τρόπου με τον οποίο μπορεί να υλοποιηθεί μια παραμετροποιήσιμη πλατφόρμα αναζήτησης βιολογικών δεδομένων. Η εφαρμογή αποδεικνύει τη δυνατότητα του χρήστη να επιλέγει συγκεκριμένες βάσεις δεδομένων (π.χ. Ensembl, UniProt, RefSeq, GenBank, RNACentral, ZFIN, InterPro, DDBJ, BOLD), να ορίζει κριτήρια αναζήτησης (gene name, organism, είδος εξόδου όπως JSON, XML, FASTA) και να λαμβάνει στοχευμένα αποτελέσματα. Μέσα από αυτή την υλοποίηση αναδεικνύεται πώς μπορεί να σχεδιαστεί και να υλοποιηθεί μια ευέλικτη, επεκτάσιμη και επιστημονικά χρήσιμη εφαρμογή, ανάλογα με τις ανάγκες κάθε ερευνητή ή ομάδας μελέτης.

Η μελέτη ανέδειξε τη συμπληρωματικότητα των εργαλείων και των βάσεων, ενώ η εφαρμογή που υλοποιήθηκε παρέχει πρακτική και λειτουργική αξία στην αναζήτηση, φιλτράρισμα και αξιολόγηση βιολογικών δεδομένων. Ένα χαρακτηριστικό παράδειγμα που παρουσιάζεται είναι το γονίδιο ROCK1 στο *Sparus aurata*, όπου αναλύεται η χρήση του NCBI API για την ανάκτηση δεδομένων, συγκρινόμενα με άλλες διεθνείς βάσεις.

Τα συμπεράσματα της εργασίας καταδεικνύουν ότι η ενσωμάτωση παραμετρικών, συνδυαστικών αναζητήσεων από πολλαπλές βάσεις, με δυνατότητα τοπικής αποθήκευσης και offline επεξεργασίας, μπορεί να μεταμορφώσει την προσβασιμότητα, την ταχύτητα και την ακρίβεια της γονιδιωματικής έρευνας. Επιπλέον, προτείνονται τεχνικές βελτιστοποίησης όπως caching, επέκταση υποστήριξης για μικρότερες ή θεματικά εξειδικευμένες βάσεις, και εμπλουτισμός με πρόσθετα φίλτρα για περαιτέρω εστιασμένη ανάλυση. Η διεπιστημονική προσέγγιση – συνδυάζοντας πληροφορική και ιχθυογενετική – αποδεικνύει έμπρακτα τη δυνατότητα ανάπτυξης ευέλικτων, επεκτάσιμων και εξειδικευμένων εργαλείων για την υποστήριξη της σύγχρονης έρευνας στις βιολογικές επιστήμες.

Λέξεις κλειδιά:

Γονιδιώματα οστεϊχθύων, Βιοπληροφορικά εργαλεία, Βιολογικές βάσεις δεδομένων, Παραμετροποιημένη αναζήτηση, Δεδομένα γονιδιακής αλληλουχίας, Γονιδιωματική ανάλυση, Ανάκτηση μέσω API (REST), Flask εφαρμογή, BLAST, Ensembl, UniProt, RefSeq, GenBank, Εφαρμογές Python σε βιοπληροφορική

ABSTRACT

This thesis focuses on the systematic recording, analysis, and evaluation of the most significant biological databases that include both coding and non-coding regions of the genomes of osteichthyans—the most populous group of vertebrates on the planet. Special emphasis is placed on exploring the various methods by which researchers can retrieve biologically relevant information and how such data can be effectively utilized in research or applied contexts.

The project was implemented along two main axes: first, an in-depth study of key bioinformatics tools such as BLAST, FASTA, PSI-BLAST, HMMER, MAFFT, MUSCLE, and Clustal Omega; and second, the development of a **proof-of-concept web application (Flask App)** that serves as a functional demo for how a customizable biological data search platform can be implemented. The application demonstrates how a user can select specific databases (e.g., Ensembl, UniProt, RefSeq, GenBank, RNACentral, ZFIN, InterPro, DDBJ, BOLD), define search parameters (e.g., gene name, organism), and specify output formats (e.g., JSON, XML, FASTA) to retrieve targeted results. This implementation highlights how flexible, scalable, and research-oriented applications can be designed and adapted to the needs of individual scientists or research teams.

The study revealed the complementary nature of the examined tools and databases, while the developed application provides tangible value in the querying, filtering, and evaluation of biological information. A representative case study is presented using the gene *ROCK1* in *Sparus aurata*, analyzing data retrieval via the NCBI API in comparison with other international databases.

The findings suggest that integrating parameterized, multi-source queries with local data storage and offline processing capabilities can significantly improve the accessibility, speed, and precision of genomic research. Further optimizations are proposed, such as caching, expanding support for smaller or more specialized databases, and adding advanced filtering options for targeted data analysis. The interdisciplinary approach—combining informatics and ichthyogenetics—demonstrates the practical potential for developing flexible, extensible, and domain-specific tools that support modern biological research.

Keywords:

Osteichthyan genomes, Bioinformatics tools, Biological databases, Parameterized search, Genomic sequence data, Genomic analysis, Data retrieval via RESTful API, Flask application, BLAST, Ensembl, UniProt, RefSeq, GenBank, Python applications in bioinformatics

Πίνακας Περιεχομένων

ΠΕΡΙΛΗΨΗ.....	5
ABSTRACT	6
1. Εισαγωγή.....	11
Κωδικοποιημένες περιοχές.....	11
Μη κωδικοποιημένες περιοχές	11
Στόχοι της πτυχιακής.....	12
2. Εργαλεία αναζήτησης	12
2.1 NCBI BLAST [Basic Local Alignment Search Tool]	12
Βάσεις Δεδομένων όπου χρησιμοποιείται το BLAST	12
Πως λειτουργεί το BLAST.....	13
Παραδείγματα χρήσης του BLAST για την αναγνώριση κωδικοποιημένων και μη-κωδικοποιημένων αλληλουχιών	13
Παραδείγματα εφαρμογών του BLAST	14
Παραδείγματα εφαρμογών του BLAST	14
Οι τρόποι αναζήτησης και πρόσβασης στα δεδομένα.....	14
2.2 FASTA - shortening of "FAST-All"	15
Πως λειτουργεί το FASTA.....	16
Παραδείγματα εφαρμογών του FASTA.....	17
Τρόποι αναζήτησης και πρόσβασης στα δεδομένα των βάσεων που χρησιμοποιεί το FASTA.....	17
API για διασύνδεση με βάσεις δεδομένων - Παράδειγμα	18
Αναλυτικά μέρη του κώδικα	18
2.3 PSI-BLAST (Position-Specific Iterated BLAST)	20
Πως λειτουργεί το PSI-BLAST	20
Παραδείγματα εφαρμογών του PSI-BLAST.....	21
Τρόποι αναζήτησης και πρόσβασης στα δεδομένα των βάσεων που χρησιμοποιεί το PSI-BLAST	21
2.4 HMMER.....	22
Πως λειτουργεί το HMMER.....	22
Παραδείγματα εφαρμογών του HMMER.....	23
Βάσεις δεδομένων που χρησιμοποιεί το HMMER	23
Παράδειγμα με χρήση API σε Python	23
Σημαντικά Pfam IDs που σχετίζονται με Hox γονίδια στους οστεϊχθύες.....	24
2.5 Clustal Omega	24
Πως λειτουργεί το CLUSTAL OMEGA	25

Παραδείγματα εφαρμογών του CLUSTAL OMEGA.....	25
Τρόποι αναζήτησης και πρόσβασης στα δεδομένα των βάσεων που χρησιμοποιεί το Clustal Omega	25
2.6 MAFFT (Multiple Alignment using Fast Fourier Transform).....	27
Πώς λειτουργεί το MAFFT.....	28
Παράδειγμα Χρήσης του MAFFT σε Οστεϊχθύες – Γονίδιο <i>myod1</i>	28
Εφαρμογές του MAFFT	28
Τρόποι Αναζήτησης και Πρόσβασης στα Δεδομένα των Βάσεων Δεδομένων.....	29
2.7 MUSCLE (Multiple Sequence Comparison by Log-Expectation).....	29
Πώς λειτουργεί.....	30
Παραδείγματα Εφαρμογών του MUSCLE.....	30
Τρόποι Αναζήτησης και Πρόσβασης στα Δεδομένα των Βάσεων Δεδομένων.....	31
Κώδικας API για Διασύνδεση με τις Βάσεις Δεδομένων UniProt και RefSeq	31
3. Συνδυασμοί εργαλείων και βάσεων δεδομένων.....	31
4. Οι σημαντικότερες βάσεις.....	32
5.0 Περιγραφή και ανάλυση των βάσεων, εστιάζοντας στη δυνατότητά τους να παρέχουν πληροφορίες για κωδικοποιημένες και μη-κωδικοποιημένες περιοχές των γονιδιωμάτων.....	33
5.1 Ensembl.....	33
Είδη Δεδομένων που Περιέχει και Τρόποι Πρόσβασης.....	33
Τρόποι Πρόσβασης στα Δεδομένα στην Βάση	34
Η Ensembl ως ολοκληρωμένη πλατφόρμα ενοποίησης, σχολιασμού και πρόσβασης	40
5.2 UniProt.....	41
Είδη Δεδομένων που Περιέχει η UniProt προς αναζήτηση:	41
Τρόποι Πρόσβασης στα Δεδομένα.....	41
Ρόλος της UniProt σε σχέση με τις υπόλοιπες Βάσεις ως "κόμβος διασύνδεσης"	42
5.3 RefSeq (NCBI)	42
Είδη δεδομένων που περιέχει η βάση RefSeq	42
Τρόποι αναζήτησης και άντλησης δεδομένων από την βάση.....	43
Συμπέρασμα	43
Σύγκριση με άλλες βάσεις δεδομένων.....	43
5.4 GenBank	44
Είδη Δεδομένων που Περιέχει η Βάση.....	44
Τρόποι Αναζήτησης και Άντλησης Δεδομένων	44

Παραδείγματα χρήσης για 3 από τα σημαντικότερα E-utilities για το πώς χρησιμοποιούνται	46
Περιορισμοί στην χρήση τους για να αποφεύγεται η υπερφόρτωση των servers:	49
Σύγκριση με άλλες βάσεις δεδομένων.....	50
5.5 RNACentral	51
Τρόποι Αναζήτησης στην Βάση RNACentral	51
Είδη Δεδομένων.....	51
Τρόποι Πρόσβασης στα Δεδομένα.....	52
Σύγκριση με άλλες βάσεις δεδομένων.....	52
5.6 ZFIN.....	53
Τρόποι Αναζήτησης στην Βάση ZFIN	53
Είδη Δεδομένων που Περιέχει η ZFIN.....	53
Τρόποι Πρόσβασης στα Δεδομένα.....	54
Σύγκριση με άλλες βάσεις δεδομένων.....	54
Συμπέρασμα	54
5.7 InterPro	55
Τρόποι Αναζήτησης στην Βάση InterPro.....	55
Είδη Δεδομένων που Περιέχει και Τρόποι Πρόσβασης σε Αυτά.....	55
Τρόποι Πρόσβασης στα Δεδομένα.....	56
Σύγκριση με άλλες βάσεις δεδομένων.....	56
Συνολική Χρήση της InterPro	57
5.8 DDBJ DNA Data Bank of Japan.....	57
Τρόποι Αναζήτησης στην Βάση DDBJ	57
Είδη Δεδομένων που Περιέχει η DDBJ.....	58
Τρόποι Πρόσβασης στα Δεδομένα.....	58
Συμπέρασμα	59
Σύγκριση με άλλες βάσεις δεδομένων.....	59
5.9 BOLD Systems, ή Barcode of Life Data Systems.....	59
Τρόποι Αναζήτησης στην Βάση BOLD Systems	60
Web Interface (Ιστοσελίδα)	60
Αναζήτηση με εργαλεία (Search Tools)	60
BOLD API (Application Programming Interface).....	60
Είδη Δεδομένων που Περιέχει	60
Τρόποι Πρόσβασης στα Δεδομένα.....	61
Σύγκριση με άλλες βάσεις δεδομένων.....	62

Συμπέρασμα.....	62
6.0 Το γονίδιο ROCK1 στο Sparus aurata ως Μοντέλο Αναφοράς για Αναζήτηση, Ανάλυση και Εξαγωγή Δεδομένων από το NCBI	64
Σύνδεση με το Θεωρητικό Υλικό της Εργασίας.....	64
7.0 Σχεδίαση και Περιγραφή Ιστοσελίδας	65
Αρχικός Σχεδιασμός και Μελλοντικό Πλάνο	65
Υποδομή και Προγραμματιστικά Εργαλεία	66
Flask Project – Δομή και Περιεχόμενο.....	67
Εκτέλεση του Flask App.....	68
Υλοποιημένος κώδικας.....	70
8.0 Συζήτηση – Συμπεράσματα	71
9.0 Τεχνολογικά Εργαλεία για την υλοποίηση της εργασίας.....	76
1. Εργαλεία Υποβοηθούμενης Συγγραφής και Αναδιατύπωσης	76
2. Εργαλεία Προγραμματισμού και Ανάλυσης API	76
3. Πηγές Δεδομένων και Βάσεις Βιοπληροφορικής.....	76
10.0 Βιβλιογραφία	77
ΠΑΡΑΡΤΗΜΑ	79

Πίνακας εικόνων

Εικόνα 1 Ανάκτηση αλληλουχιών πρωτεΐνης από την Uniport.....	19
Εικόνα 2 Αναζήτηση Hox γονιδίων με HMMER API.....	23
Εικόνα 3 Χρήση του Pfam API για επιπλέον πληροφορίες	24
Εικόνα 4 Παράδειγμα Python API για την UniProt βάση	26
Εικόνα 5 Αναζήτηση για το γονίδιο DMRT1 από τη βάση δεδομένων UniProt	26
Εικόνα 6 Αναζήτηση για το γονίδιο DMRT1 από τη βάση δεδομένων UniProt.....	27
Εικόνα 7 Ανάκτηση πρωτεϊνικών αλληλουχιών από το UniProt σε μορφή FASTA.....	31
Εικόνα 8 Ensembl Browser.....	35
Εικόνα 9 Αναζήτηση μέσω Ensembl Tools (VEP)	36
Εικόνα 10 Πρόσβαση μέσω FTP (FileTransferProtocol) στην βάση Ensembl.....	37
Εικόνα 11 Αναζήτηση με findstr	38
Εικόνα 12 Αναζήτηση για την catalase.....	45
Εικόνα 13 Παράδειγμα API σε Python για αναζήτηση στην βάση RefSeq.....	48
Εικόνα 14 Tunneling εργαλείο Ngrok	69
Εικόνα 15 GUI της εφαρμογής της πτυχιακής.....	79
Εικόνα 16 NCBI αποτέλεσμα αναζήτησης rock1, sparus.....	80
Εικόνα 17 NCBI αποτέλεσμα αναζήτησης στο Proteins	80
Εικόνα 18 NCBI αποτέλεσμα αναζήτησης στο Genoms.....	81
Εικόνα 19 NCBI αποτέλεσμα αναζήτησης στο Literature/PubMed	81

1. Εισαγωγή

Για να καταγράψουμε τις διαθέσιμες βάσεις δεδομένων σχετικές με τα γονιδιώματα, να αναλύσουμε τις τεχνικές αναζήτησης, τις εφαρμογές τους και να υλοποιήσουμε ένα ευέλικτο web-based εργαλείο για την εμφάνιση και φιλτράρισμα αποτελεσμάτων, είναι απαραίτητη η κατανόηση των κωδικοποιημένων και μη κωδικοποιημένων περιοχών του γονιδιώματος και της σημασίας τους στη μελέτη των οστεϊχθύων.

Η ακρίβεια των αποτελεσμάτων εξαρτάται από την ορθή παραμετροποίηση του κώδικα και την κατάλληλη στρατηγική αναζήτησης στις σχετικές βάσεις.

Οι οστεϊχθές (Osteichthyes) είναι η μεγαλύτερη ομάδα σπονδυλωτών, με πάνω από 28.000 είδη, και χαρακτηρίζονται από σκελετό με οστά. Έχουν σημαντικό ρόλο στα οικοσυστήματα και μεγάλη εμπορική αξία.

Τα γονιδιώματά τους περιλαμβάνουν περιοχές με διαφορετικές λειτουργίες, οι οποίες θα αναλυθούν στην πρώτη ενότητα, με στόχο την κατανόηση της βιολογικής τους σημασίας.

Κωδικοποιημένες περιοχές

Οι κωδικοποιημένες περιοχές του γονιδιώματος μεταγράφονται σε mRNA και μεταφράζονται σε πρωτεΐνες. Περιλαμβάνουν:

Λειτουργικά γονίδια: Συμμετέχουν σε ανάπτυξη, μεταβολισμό, ανοσία και κυτταρική σηματοδότηση.

Συντηρημένα γονίδια: Κοινά μεταξύ ειδών, δείγμα της εξελικτικής τους σημασίας.

Εξώνια: Τμήματα που καθορίζουν τη σύνθεση πρωτεϊνών.

Σημασία: Είναι κρίσιμες για τη βιολογική λειτουργία, μελέτες γονιδιακής θεραπείας και την κατανόηση ασθενειών μέσω μεταλλάξεων. Η ανάλυσή τους γίνεται με DNA αλληλούχιση και βιοπληροφορικά εργαλεία.

Μη κωδικοποιημένες περιοχές

Δεν παράγουν πρωτεΐνες αλλά ρυθμίζουν τη γονιδιακή έκφραση. Περιλαμβάνουν:

ncRNA [Μη μεταφραζόμενα RNA με ρυθμιστικό ρόλο], miRNA, lncRNA, siRNA [Ελέγχουν την έκφραση και σταθερότητα του mRNA], Ρυθμιστικά στοιχεία [Promoters, enhancers, silencers], Δομικές περιοχές [Τελομερή και κεντρομερή για τη σταθερότητα του γονιδιώματος].

Στόχοι της πτυχιακής

Καταγραφή των διαθέσιμων βάσεων δεδομένων που σχετίζονται με τα γονιδιώματα.
Ανάλυση των εργαλείων και των τεχνικών αναζήτησης βιολογικών δεδομένων.
Διερεύνηση των συσχετισμών μεταξύ διαφορετικών βάσεων και μεθόδων αναζήτησης.
Υλοποίηση ενός παραδειγματικού web-based περιβάλλοντος επίδειξης της χρήσης των εργαλείων, με δυνατότητα ταυτόχρονης αναζήτησης σε πολλαπλές βάσεις δεδομένων και χρήσης παραμετροποιημένων κριτηρίων, όπως:

Όρος αναζήτησης (π.χ. gene name, organism)

Επιλογή εξόδου (output format): JSON, XML

Επιλογή υπο-βάσης (π.χ. gene, nucleotide, protein, PubMed)

Στόχος: Η ανάπτυξη ενός ευέλικτου εργαλείου που θα διευκολύνει την ολοκληρωμένη και στοχευμένη πρόσβαση σε γονιδιωματικά δεδομένα.

2. Εργαλεία αναζήτησης

Εκτός από το **BLAST**, υπάρχουν και άλλα σημαντικά εργαλεία για την αναζήτηση ομοιοτήτων σε αλληλουχίες **DNA**, **RNA** ή **πρωτεΐνης**. Παρακάτω παρουσιάζονται ορισμένα από τα πιο διαδεδομένα, συνοδευόμενα από σύντομη περιγραφή, βασικές χρήσεις και τις κύριες **βάσεις δεδομένων** στις οποίες εφαρμόζονται.

2.1 NCBI BLAST [Basic Local Alignment Search Tool]

Το **BLAST** (*Basic Local Alignment Search Tool*) είναι ένα βασικό εργαλείο βιοπληροφορικής, αναπτυγμένο από το **NCBI**, που χρησιμοποιείται για την **αναγνώριση ομοιοτήτων σε αλληλουχίες DNA, RNA ή πρωτεϊνών**. Αποτελεί θεμέλιο για συγκριτική γονιδιωματική ανάλυση.

Βάσεις Δεδομένων όπου χρησιμοποιείται το BLAST

EBI (European Bioinformatics Institute) *Ensembl BLAST*: Αναζήτηση σε πολλά γονιδιώματα ειδών, *EMBL-EBI BLAST*: Πρόσβαση σε ποικίλες βάσεις νουκλεοτιδίων και πρωτεϊνών.

UCSC Genome Browser Προσφέρει το εργαλείο **BLAT** (BLAST-like Alignment Tool) για ευρύ φάσμα γονιδιωμάτων.

DDBJ (DNA Data Bank of Japan) Υποστηρίζει BLAST αναζητήσεις σε νουκλεοτιδικές και πρωτεϊνικές βάσεις.

ExPASy (Expert Protein Analysis System) Εστιάζει σε **πρωτεϊνικά δεδομένα** με πρόσβαση σε σχετικές βάσεις.

MG-RAST Προσανατολισμένο στη **μεταγονιδιατική ανάλυση** μέσω BLAST.

PDB (Protein Data Bank) BLAST αναζητήσεις σε **δομικά δεδομένα πρωτεϊνών** (3D).

TAIR (The Arabidopsis Information Resource) Ειδικό εργαλείο BLAST για το **γονιδίωμα του Arabidopsis**.

Πως λειτουργεί το BLAST

Εισαγωγή Αλληλουχίας: Ο χρήστης εισάγει μια αλληλουχία DNA, RNA ή πρωτεΐνης ως *query sequence*.

Επιλογή Βάσης Δεδομένων: Η αλληλουχία συγκρίνεται με αλληλουχίες από μια επιλεγμένη βάση δεδομένων (π.χ. GenBank, Ensembl).

Εύρεση Ομοιοτήτων: Το BLAST χρησιμοποιεί αλγόριθμους ευθυγράμμισης για να εντοπίσει περιοχές υψηλής ομοιότητας ανάμεσα στην εισαγόμενη και τις αποθηκευμένες αλληλουχίες.

Παρουσίαση Αποτελεσμάτων: Επιστρέφεται λίστα αποτελεσμάτων που περιλαμβάνει:

Βαθμολογίες ευθυγράμμισης (scores), Ποσοστά ταυτότητας (identity), Στατιστικά σημαντικότητας (e-values)

Παραδείγματα χρήσης του BLAST για την αναγνώριση κωδικοποιημένων και μη-κωδικοποιημένων αλληλουχιών

Για την ανάλυση κωδικοποιημένων αλληλουχιών, εισάγεται μια αλληλουχία DNA στο εργαλείο BLASTX. Το εργαλείο μεταφράζει την αλληλουχία σε πιθανές πρωτεΐνες και τις συγκρίνει με αλληλουχίες πρωτεϊνών που υπάρχουν στη βάση δεδομένων. Αν εντοπιστούν ομοιότητες με γνωστές πρωτεΐνες, αυτό υποδηλώνει ότι η συγκεκριμένη DNA αλληλουχία είναι πιθανόν κωδικοποιητική.

Στην περίπτωση μη-κωδικοποιημένων αλληλουχιών, εισάγεται μια RNA αλληλουχία στο εργαλείο BLASTN. Το εργαλείο συγκρίνει την αλληλουχία με άλλες νουκλεοτιδικές αλληλουχίες που περιέχονται στη βάση δεδομένων. Αν τα αποτελέσματα δείξουν ομοιότητες με γνωστές μη-κωδικοποιημένες αλληλουχίες, όπως είναι τα tRNA ή τα miRNA, αυτό μπορεί να υποδεικνύει ότι η συγκεκριμένη RNA αλληλουχία έχει ρυθμιστικό ή δομικό ρόλο.

Παραδείγματα εφαρμογών του BLAST

Το εργαλείο BLAST είναι ιδιαίτερα ευέλικτο και χρησιμοποιείται ευρέως στη μοριακή βιολογία και τη βιοπληροφορική. Επιτρέπει τη σύγκριση βιολογικών αλληλουχιών με στόχο την αναγνώριση ομοιοτήτων, γεγονός που το καθιστά πολύτιμο σε ποικίλα ερευνητικά και κλινικά πεδία. Η χρήση του εκτείνεται από τη βασική έρευνα μέχρι την εφαρμοσμένη βιοτεχνολογία και την ιατρική.

Παραδείγματα εφαρμογών του BLAST

[Αναγνώριση Γονιδίων], ένας ερευνητής που αλληλουχίζει το γονιδίωμα ενός νέου βακτηριακού είδους μπορεί να χρησιμοποιήσει το BLAST για να συγκρίνει τις νέες αλληλουχίες με γνωστά γονίδια, εντοπίζοντας εκείνα που κωδικοποιούν πρωτεΐνες.
[Εύρεση Ορθόλογων Γονιδίων], μια επιστημονική ομάδα μπορεί να μελετά τη λειτουργία ενός γονιδίου στο ποντίκι και να αναζητήσει το αντίστοιχό του στον άνθρωπο, εντοπίζοντας ομόλογα γονίδια μέσω BLASTP και πιο ειδικά, ορθόλογα γονίδια μέσω ανταποδοτικού BLASTP.

[Ανάλυση Μεταγραφικών Δεδομένων], σε μεταγραφικές μελέτες, το BLASTX χρησιμοποιείται για τη σύνδεση μέσω ομολογίας μεταγραφικών προϊόντων RNA-Seq με πιθανές πρωτεΐνες, βοηθώντας στην ερμηνεία των δεδομένων.

[Εύρεση Συντηρημένων Περιοχών], η σύγκριση ενζύμων από διαφορετικά είδη με BLASTP βοηθά στην εντόπιση συντηρημένων περιοχών που είναι κρίσιμες για τη λειτουργία τους.

[Εξέταση Γενετικής Ποικιλότητας], η γενετική ποικιλότητα μπορεί επίσης να μελετηθεί με τη χρήση του BLASTN, συγκρίνοντας αλληλουχίες ενός γονιδίου από διαφορετικούς πληθυσμούς για τον εντοπισμό παραλλαγών.

[Ανάπτυξη Νέων Φαρμάκων], στην ανάπτυξη νέων φαρμάκων, η σύγκριση ενός ενζύμου-στόχου με σχετικές πρωτεΐνες αποκαλύπτει κρίσιμες λειτουργικές ή δομικές περιοχές που μπορούν να αξιοποιηθούν θεραπευτικά.

[Ανίχνευση Παθογόνων], τέλος, σε κλινικά περιβάλλοντα, το BLASTN μπορεί να χρησιμοποιηθεί για την ανίχνευση παθογόνων οργανισμών μέσω της ταυτοποίησης αλληλουχιών από δείγματα ασθενών, προσφέροντας ένα γρήγορο και αξιόπιστο εργαλείο για διαγνωστική χρήση.

Οι τρόποι αναζήτησης και πρόσβασης στα δεδομένα

Το **API** (Application Programming Interface) του NCBI BLAST (Basic Local Alignment Search Tool) επιτρέπει στους χρήστες να αλληλεπιδρούν προγραμματιστικά με την υπηρεσία BLAST για τη σύγκριση αλληλουχιών. Το **API** του BLAST αποτελεί μέρος των Προγραμματιστικών Εργαλείων **Entrez (E-utilities)** που παρέχονται από το NCBI. Παρακάτω παρατίθεται μια επισκόπηση του τρόπου χρήσης του API του BLAST, συμπεριλαμβανομένων των βασικών βημάτων και παραδειγμάτων αιτημάτων που μπορούν να γίνουν.

Για να **αξιοποιηθεί το API**, αρχικά χρειάζεται να προετοιμαστεί η νουκλεοτιδική ή πρωτεϊνική αλληλουχία που πρόκειται να συγκριθεί. Στη συνέχεια, υποβάλλεται στο σύστημα του NCBI μέσω κατάλληλης εντολής, επιλέγοντας το κατάλληλο εργαλείο ανάλογα με τον τύπο της αλληλουχίας και τον στόχο της ανάλυσης. Οι επιλογές περιλαμβάνουν τα εργαλεία **blastn**, **blastp**, **blastx**, **tblastn** και **tblastx**.

Μετά την υποβολή, ελέγχεται η κατάσταση του αιτήματος ώστε να επιβεβαιωθεί αν έχει ολοκληρωθεί η επεξεργασία. Μόλις ολοκληρωθεί, γίνεται η ανάκτηση των αποτελεσμάτων, τα οποία μπορούν να ληφθούν σε μορφές όπως XML ή JSON, ανάλογα με τις ανάγκες της ανάλυσης ή της ενσωμάτωσης τους σε άλλο σύστημα.

Βασικά API Endpoints [διεύθυνση σε έναν εξυπηρετητή, που δέχονται αιτήματα και επιστρέφουν απαντήσεις]

Υποβολή μιας αναζήτησης BLAST: URL: https://blast.ncbi.nlm.nih.gov/Blast.cgi Μέθοδος GET ή POST Παράμετροι CMD: Put, PROGRAM: blastn, blastp, κλπ., DATABASE: nt, nr, κλπ., QUERY: Τα δεδομένα της αλληλουχίας μας, FORMAT_TYPE: XML, JSON, Text	<pre>import requests def submit_blast(query_seq, program="blastn", database="nt"): url = "https://blast.ncbi.nlm.nih.gov/Blast.cgi" params = { "CMD": "Put", "PROGRAM": program, "DATABASE": database, "QUERY": query_seq } response = requests.post(url, data=params) return response.text</pre>
Έλεγχος της κατάστασης μιας αναζήτησης BLAST: URL: https://blast.ncbi.nlm.nih.gov/Blast.cgi Μέθοδος GET Παράμετροι CMD: Get, RID: Το Request ID που επιστράφηκε από το βήμα της υποβολής, FORMAT_OBJECT: SearchInfo	<pre>import requests def check_blast_status(rid): url = "https://blast.ncbi.nlm.nih.gov/Blast.cgi" params = { "CMD": "Get", "RID": rid, "FORMAT_OBJECT": "SearchInfo" } response = requests.get(url, params=params) return response.text</pre>
Ανάκτηση αποτελεσμάτων μιας αναζήτησης BLAST: URL: https://blast.ncbi.nlm.nih.gov/Blast.cgi Μέθοδος: GET Παράμετροι: CMD: Get, RID: Request ID, FORMAT_TYPE: XML, JSON, Text	<pre>def get_blast_results(rid, format_type="JSON2"): url = "https://blast.ncbi.nlm.nih.gov/Blast.cgi" params = { "CMD": "Get", "RID": rid, "FORMAT_TYPE": format_type } response = requests.get(url, params=params) return response.text # ή response.json() αν format_type=JSON2</pre>

2.2 FASTA - shortening of "FAST-All"

Το **FASTA** είναι ένα από τα πρώτα εργαλεία που αναπτύχθηκαν για την αναζήτηση ομοιοτήτων σε αλληλουχίες πρωτεϊνών ή νουκλεϊκών οξέων (DNA, RNA). Χρησιμοποιεί

ευρετικούς αλγόριθμους για να εντοπίσει περιοχές ομοιότητας μεταξύ μιας αλληλουχίας αναζήτησης και μιας βάσης δεδομένων αλληλουχιών. Το εργαλείο λειτουργεί παρόμοια με το BLAST, αλλά εφαρμόζει διαφορετική μεθοδολογία στον τρόπο που βρίσκει και ευθυγραμμίζει τις ομοιότητες, δίνοντας έμφαση στην **ανάλυση τοπικών ευθυγραμμίσεων**.

Το **FASTA** ξεκινά με το σπάσιμο της αλληλουχίας σε μικρά τμήματα (όπως τα k-tuples) και, στη συνέχεια, εστιάζει στις περιοχές όπου υπάρχουν πιθανές ομοιότητες για να κάνει μια πιο λεπτομερή ευθυγράμμιση. Αυτό μειώνει την υπολογιστική πολυπλοκότητα σε σχέση με την πλήρη ευθυγράμμιση, ενώ διατηρεί υψηλή ακρίβεια στις αναζητήσεις.

Πως λειτουργεί το FASTA

Εισαγωγή Αλληλουχίας: Ο χρήστης παρέχει μια αλληλουχία πρωτεΐνης ή νουκλεϊκού οξέος στο εργαλείο FASTA.

Σπάσιμο της Αλληλουχίας: Το FASTA σπάει την αλληλουχία σε μικρά τμήματα (συνήθως k-tuples) και τα συγκρίνει με τις αλληλουχίες στη βάση δεδομένων.

Αρχική Ευθυγράμμιση: Ανιχνεύει τις περιοχές με το μεγαλύτερο ποσοστό ταυτότητας ή ομοιότητας, δημιουργώντας τοπικές ευθυγραμμίσεις.

Βαθμολογία και Στατιστική Ανάλυση: Δίνει μια βαθμολογία για τις ευθυγραμμίσεις και υπολογίζει στατιστικά για να εκτιμήσει τη σημασία των αποτελεσμάτων.

Αποτελέσματα: Το εργαλείο παράγει μια λίστα με τις αλληλουχίες που έχουν την υψηλότερη ομοιότητα με την αλληλουχία αναζήτησης, συνοδευόμενη από στατιστικά όπως τα ποσοστά ταυτότητας και ομοιότητας, και P-values για την αξιολόγηση της στατιστικής σημασίας.

Παράδειγμα Χρήσης του FASTA για την Αναγνώριση Κωδικοποιημένων και Μη Κωδικοποιημένων Αλληλουχιών

Κωδικοποιημένη Αλληλουχία:

Ερευνητής διαθέτει μία DNA αλληλουχία που θεωρείται πιθανό να κωδικοποιεί πρωτεΐνη.

Βήμα 1: Η αλληλουχία μεταφράζεται σε πιθανό πρωτεϊνικό προϊόν (ORF).

Βήμα 2: Το εργαλείο **FASTA-P** (protein mode) χρησιμοποιείται για σύγκριση με τη βάση **UniProt**, που περιέχει γνωστές πρωτεϊνικές αλληλουχίες.

Ερμηνεία: Αν προκύψουν υψηλής ποιότητας ευθυγραμμίσεις με γνωστές πρωτεΐνες:

- Υψηλό **ποσοστό ταυτότητας** και **χαμηλό E-value**
- Ενδείξεις για πιθανή **κωδικοποιητική λειτουργία** της εισόδου
- Δυνατότητα λειτουργικής και εξελικτικής ερμηνείας

Μη Κωδικοποιημένη Αλληλουχία:

Η αλληλουχία δεν δίνει ικανοποιητικό ORF ή αναμένεται να αποτελεί μη κωδικοποιημένο RNA (π.χ. tRNA, miRNA).

Βήμα : Το εργαλείο **FASTA-N** (nucleotide mode) χρησιμοποιείται για σύγκριση με βάσεις δεδομένων **μη κωδικοποιημένων RNA**, όπως:

- **RNAcentral**
- **Rfam**
- **miRBase**

Ερμηνεία: Αν βρεθεί σημαντική ομοιότητα με γνωστά ncRNAs:

- Δείχνει ότι η αλληλουχία έχει πιθανώς **δομική ή ρυθμιστική λειτουργία**
- Δεν κωδικοποιεί πρωτεΐνες αλλά συμμετέχει σε γονιδιακή ρύθμιση, ωρίμανση RNA, κ.λπ.

Παραδείγματα εφαρμογών του FASTA

Εύρεση ομοιοτήτων σε νέες γονιδιακές αλληλουχίες: Οι ερευνητές χρησιμοποιούν το FASTA για να συγκρίνουν νέες αλληλουχίες γονιδίων με γνωστές βάσεις δεδομένων για να βρουν ομοιότητες και να προσδιορίσουν πιθανές λειτουργίες.

Αναγνώριση συντηρημένων αλληλουχιών: Το FASTA μπορεί να χρησιμοποιηθεί για τον εντοπισμό συντηρημένων γονιδιακών ή πρωτεϊνικών περιοχών μεταξύ διαφορετικών οργανισμών.

Ανάλυση γονιδιωμάτων και μεταγραφωμάτων: Το FASTA χρησιμοποιείται ευρέως για τη σύγκριση ολόκληρων γονιδιωμάτων ή μεταγραφωμάτων, επιτρέποντας την ταυτοποίηση ομοιοτήτων που σχετίζονται με τη λειτουργία των γονιδίων.

Τρόποι αναζήτησης και πρόσβασης στα δεδομένα των βάσεων που χρησιμοποιεί το FASTA

Το FASTA μπορεί να διασυνδεθεί με διάφορες βάσεις δεδομένων, όπως η **UniProt**, **EMBL-EBI**, και **GenBank**, χρησιμοποιώντας τις διαδικτυακές διεπαφές και τα **API** που αυτές παρέχουν. Τα περισσότερα εργαλεία προσφέρουν **διαδικτυακές πύλες** όπου οι αλληλουχίες μπορούν να αναζητηθούν αυτόματα.

- **UniProt:** Η UniProt είναι μια εκτενής βάση δεδομένων πρωτεϊνών. Παρέχει ένα API που μπορεί να χρησιμοποιηθεί για την αναζήτηση πρωτεϊνικών αλληλουχιών και τις σχετικές πληροφορίες.
- **EMBL-EBI:** Το EMBL-EBI παρέχει μια σειρά από βάσεις δεδομένων και εργαλεία για την ανάλυση βιολογικών δεδομένων. Το API τους μπορεί να χρησιμοποιηθεί για την εκτέλεση αναζητήσεων αλληλουχιών.
- **GenBank:** Η GenBank είναι μια βάση δεδομένων νουκλεοτιδικών αλληλουχιών που διαχειρίζεται το NCBI. Προσφέρει επίσης API για την αναζήτηση και ανάκτηση δεδομένων αλληλουχιών.

API για διασύνδεση με βάσεις δεδομένων - Παράδειγμα

Παράδειγμα κώδικα API σε **Python** για τη Διασύνδεση της Βάσης **UniProt** με το **FASTA**

χρησιμοποιούμε το **UniProt API** για να ανακτήσουμε αλληλουχίες πρωτεϊνών και να τις χρησιμοποιήσουμε ως βάση δεδομένων για τις αναζητήσεις μας με το FASTA.

Παρακάτω ένα παράδειγμα που:

Ανακτά αλληλουχίες πρωτεϊνών από το UniProt με βάση ένα συγκεκριμένο ερώτημα μέσω HTTP αιτήματος (REST API), **αποθηκεύει τις αλληλουχίες** σε ένα αρχείο σε μορφή FASTA, **εκτελεί το εργαλείο FASTA**, χρησιμοποιώντας μια αλληλουχία αναζήτησης και τη βάση δεδομένων που κατεβάσαμε από το UniProt (uniprot_human.fasta), για να εντοπίσει ομοιότητες.

Προαπαιτούμενα

- **Python 3.x** εγκατεστημένο στο σύστημά μας.
- Το εργαλείο **FASTA** εγκατεστημένο και προσβάσιμο από τη γραμμή εντολών.
- Η βιβλιοθήκη **requests** της Python για την πραγματοποίηση HTTP αιτημάτων. Μπορεί να εγκατασταθεί με την εντολή `>> pip install requests`

Αναλυτικά μέρη του κώδικα

```
fetch_uniprot_sequences(query, output_file)
```

Δημιουργεί ένα HTTP αίτημα προς το **UniProt REST API** >> Στέλνει το ερώτημα (π.χ. 'organism:9606' για τον άνθρωπο)>> Ζητά απάντηση σε **μορφή FASTA**>> Αν η απόκριση είναι επιτυχής (200 OK), αποθηκεύει το περιεχόμενο στο αρχείο output_file>> Διαφορετικά, εμφανίζει σφάλμα και τερματίζει το πρόγραμμα.

```
run_fasta(query_sequence_file, database_file, output_file)
```

Εκτελεί το εργαλείο **fasta36** στη γραμμή εντολών>> Χρησιμοποιεί ως είσοδο την αλληλουχία αναζήτησης (query_sequence_file) και τη βάση δεδομένων (database_file)>> Αποθηκεύει τα αποτελέσματα στο αρχείο output_file>> Αν υπάρξει σφάλμα κατά την εκτέλεση του FASTA, εμφανίζει μήνυμα και τερματίζει.

```

1 import requests
2 import subprocess
3 import sys
4
5 // Δηλώνουμε τις 2 βασικές συναρτήσεις def
6
7 def fetch_uniprot_sequences(query, output_file): //Αυτή η συνάρτηση στέλνει ένα HTTP αίτημα στο UniProt API για να ανακτήσει
8 //αλληλουχίες πρωτεϊνών που ταιριάζουν με το ερώτημα
9 //Οι αλληλουχίες αποθηκεύονται σε ένα αρχείο σε μορφή FASTA
10
11 base_url = 'https://www.uniprot.org/uniprot/' //Η βασική διεύθυνση URL για τα αιτήματα στο UniProt
12 params = { //Παράμετροι του αιτήματος
13     'query': query, //Το ερώτημα αναζήτησης (π.χ., organism:9606 για τον άνθρωπο).
14     'format': 'fasta', //Η μορφή των δεδομένων που θα επιστραφούν (fasta)
15     'compress': 'no', //Ορίζει αν τα δεδομένα θα είναι συμπίεμένα (no για μη συμπίεμένα).
16 }
17 response = requests.get(base_url, params=params)
18 if response.status_code == 200:
19     with open(output_file, 'w') as f:
20         f.write(response.text)
21     print(f"Οι αλληλουχίες αποθηκεύτηκαν στο αρχείο {output_file}")
22 else:
23     print(f"Σφάλμα κατά την ανάκτηση δεδομένων: {response.status_code}")
24     sys.exit(1)
25
26 def run_fasta(query_sequence_file, database_file, output_file): //Αυτή η συνάρτηση εκτελεί το πρόγραμμα FASTA χρησιμοποιώντας
27 //την αλληλουχία αναζήτησης και τη βάση δεδομένων δημιουργήθηκε.
28
29 fasta_command = ['fasta36', query_sequence_file, database_file] //Η εντολή που θα εκτελεστεί στη γραμμή εντολών (fasta36 είναι μια κοινή έκδοση του FASTA).
30 try:
31     with open(output_file, 'w') as out:
32         subprocess.run(fasta_command, stdout=out, check=True) //Εκτελεί την εντολή και αποθηκεύει την έξοδο στο αρχείο output_file.
33
34     print(f"Τα αποτελέσματα του FASTA αποθηκεύτηκαν στο αρχείο {output_file}")
35 except subprocess.CalledProcessError as e:
36     print(f"Σφάλμα κατά την εκτέλεση του FASTA: {e}")
37     sys.exit(1)
38
39 if __name__ == "__main__": // το εκτελέσιμο
40
41     # Βήμα 1: Ανάκτηση αλληλουχιών από το UniProt
42     database_query = 'organism:9606' // Καλούμε εδώ την fetch_uniprot_sequences με παράμετρο organism:9606 για όλες οι ανθρώπινες πρωτεΐνες
43     // Στο database_query, μπορεί να ζητηθεί ανάκτηση για διαφορετικές αλληλουχίες. Συγκεκριμένο πρωτεϊνικό
44     // όνομα: 'protein:Hemoglobin' λέξη-κλειδί: 'keyword:"Kinase"' η και συγκεκριμένο UniProt ID: 'accession:P69905'
45     database_file = 'uniprot_human.fasta' // έξοδος στο αρχείο uniprot_human.fasta
46     fetch_uniprot_sequences(database_query, database_file)
47
48     # Βήμα 2: Προετοιμασία της αλληλουχίας αναζήτησης
49     query_sequence_file = 'query_sequence.fasta' //Δημιουργούμε ένα αρχείο query_sequence.fasta στο ίδιο φάκελο με τον κώδικα.
50     //Εισάγουμε την αλληλουχία σας σε μορφή FASTA. Για παράδειγμα θα μπορούσε να είναι με την παρακάτω μορφή
51     // >my_query_sequence
52     // MTEITAAWVKELRESTGAGMMDCKNALSETQHEK...
53
54     # Βήμα 3: Εκτέλεση του FASTA
55     output_file = 'fasta_results.txt'
56     run_fasta(query_sequence_file, database_file, output_file) // Καλούμε εδώ την run_fasta
57

```

Εικόνα 1 Ανάκτηση αλληλουχιών πρωτεΐνης από την Uniprot

Παρατηρήσεις

Το UniProt API έχει όρια στον αριθμό των αιτημάτων [UniProt Rate Limiting]. Για να ανακτήσουμε μεγάλο όγκο δεδομένων, πρέπει να τηρήσουμε τους όρους χρήσης.

Για μεγάλες βάσεις δεδομένων, η διαδικασία μπορεί να είναι χρονοβόρα. Για βελτιστοποίηση θα δούμε το ενδεχόμενο να περιορίσουμε το ερώτημα ή να χρησιμοποιήσουμε προϋπάρχουσες βάσεις δεδομένων.

2.3 PSI-BLAST (Position-Specific Iterated BLAST)

Είναι μια προσαρμοσμένη έκδοση του BLAST που έχει σχεδιαστεί για να ανιχνεύει απομακρυσμένες ομοιότητες μεταξύ πρωτεϊνικών αλληλουχιών. Σε αντίθεση με το παραδοσιακό BLAST, το PSI-BLAST δεν βασίζεται απλά στη γραμμική ευθυγράμμιση των αλληλουχιών, αλλά χρησιμοποιεί ένα **βαθμολογικό πίνακα θέσης (PSSM - Position-Specific Scoring Matrix)**. Αυτό το μοντέλο δημιουργείται από την αρχική αναζήτηση και χρησιμοποιείται σε επαναληπτικές αναζητήσεις, προσδιορίζοντας ομοιότητες που μπορεί να μην ήταν ορατές σε μία μόνο αναζήτηση.

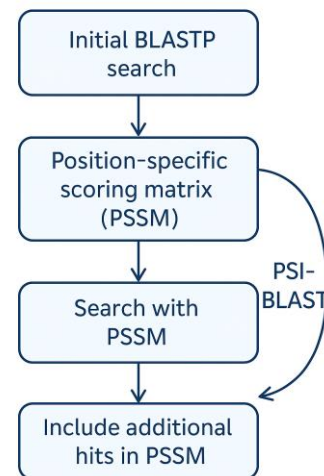
Το PSI-BLAST επιτρέπει την ανίχνευση απομακρυσμένων ομοιοτήτων και εξελικτικών σχέσεων που δεν μπορούν να εντοπιστούν με άλλα εργαλεία, καθιστώντας το ιδανικό για βαθύτερες αναλύσεις σε πρωτεϊνικές αλληλουχίες.

Πως λειτουργεί το PSI-BLAST

Το PSI-BLAST ξεκινά με μια σύγκριση της πρωτεϊνικής αλληλουχίας με μια βάση δεδομένων. Από αυτή την αναζήτηση δημιουργείται ένα μοντέλο (PSSM) που περιγράφει τη συχνότητα εμφάνισης αμινοξέων σε κάθε θέση. Το μοντέλο αυτό χρησιμοποιείται για επόμενες αναζητήσεις που στοχεύουν σε πιο δύσκολες ή μακρινές ομοιότητες.

Η διαδικασία επαναλαμβάνεται, κάθε φορά με βελτιωμένο μοντέλο, μέχρι να μη βρεθούν νέες σχετικές αλληλουχίες ή να ολοκληρωθούν οι προκαθορισμένοι κύκλοι.

Το PSI-BLAST χρησιμοποιείται για την ανίχνευση εξελικτικά απομακρυσμένων ομόλογων πρωτεϊνικών αλληλουχιών.



PSI-BLAST

Για τον εντοπισμό κωδικοποιημένων περιοχών, ξεκινά με μια γνωστή πρωτεϊνική αλληλουχία, την οποία συγκρίνει με βάσεις δεδομένων όπως το UniProt ή το NCBI Protein. Με βάση τα αποτελέσματα δημιουργείται ένα μοντέλο που βοηθά στον εντοπισμό συγγενικών ή εξελικτικά σχετιζόμενων πρωτεϊνών.

Για μη κωδικοποιημένες αλληλουχίες, το εργαλείο δεν εφαρμόζεται άμεσα, καθώς δεν αναλύει RNA ή DNA που δεν κωδικοποιούν πρωτεΐνες. Παρ' όλα αυτά, τα ευρήματα από τις πρωτεϊνικές συγκρίσεις μπορούν να συμβάλουν στην κατανόηση της λειτουργίας περιοχών του

γονιδιώματος και να εντοπίσουν περιοχές που πιθανώς σχετίζονται με τη ρύθμιση ή υποστήριξη πρωτεϊνικών γονιδίων.

Παραδείγματα εφαρμογών του PSI-BLAST

- **Ανίχνευση απομακρυσμένων ομοιοτήτων:** Εξαιρετικά χρήσιμο για την αναγνώριση πρωτεϊνών που μπορεί να έχουν εξελικτικές σχέσεις, αλλά δεν παρουσιάζουν άμεσες αλληλουχίες ομοιότητες.
- **Μελέτη εξελικτικών σχέσεων:** Μπορεί να χρησιμοποιηθεί για να εντοπιστούν πρωτεΐνες που έχουν κοινή εξελικτική προέλευση, παρόλο που έχουν υποστεί σημαντικές τροποποιήσεις.
- **Ανάλυση δομής και λειτουργίας πρωτεϊνών:** Οι απομακρυσμένες ομοιότητες που εντοπίζει το PSI-BLAST μπορούν να βοηθήσουν στην πρόβλεψη της λειτουργίας ή της δομής πρωτεϊνών που δεν έχουν μελετηθεί καλά.

Τρόποι αναζήτησης και πρόσβασης στα δεδομένα των βάσεων που χρησιμοποιεί το PSI-BLAST

Το PSI-BLAST μπορεί να χρησιμοποιηθεί με διάφορες βάσεις δεδομένων πρωτεϊνών όπως:

NCBI: Η κύρια βάση δεδομένων όπου το PSI-BLAST είναι διαθέσιμο είναι το NCBI (National Center for Biotechnology Information). Η βάση παρέχει ελεύθερη πρόσβαση σε αλληλουχίες DNA, RNA και πρωτεϊνών, επιτρέποντας στους χρήστες να εκτελούν αναζητήσεις PSI-BLAST.

Πρόσβαση: Μέσω της ιστοσελίδας του NCBI <https://blast.ncbi.nlm.nih.gov/Blast.cgi>.

UniProt: Χρησιμοποιείται επίσης για την αναζήτηση ομολογιών πρωτεϊνών. Προσφέρει πρόσβαση σε εκτενείς πρωτεϊνικές βάσεις δεδομένων.

Πρόσβαση: Μέσω της ιστοσελίδας του UniProt <https://www.uniprot.org/>.

Pfam: Χρησιμοποιείται για την αναγνώριση πρωτεϊνικών δομικών πεδίων και προσφέρει εργαλεία για αναλύσεις PSI-BLAST.

Πρόσβαση: Μέσω της ιστοσελίδας του Pfam.

Ensembl: Παρέχει δεδομένα από πολυάριθμα είδη, επιτρέποντας την αναζήτηση πρωτεϊνικών αλληλουχιών μέσω PSI-BLAST.

Πρόσβαση: Μέσω του Ensembl <https://www.ensembl.org/>.

2.4 HMMER

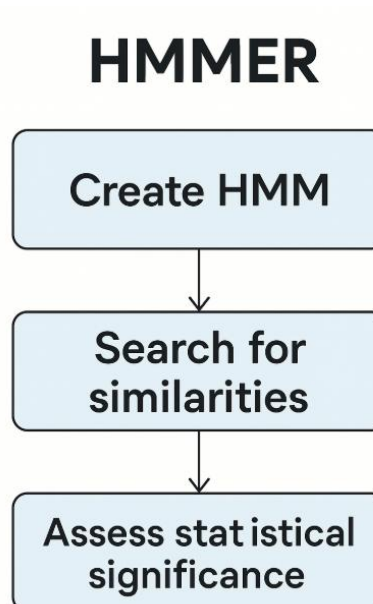
είναι ένα βιοπληροφορικό εργαλείο που χρησιμοποιεί **κρυφά μοντέλα Markov (Hidden Markov Models, HMMs)** για την αναζήτηση και την ευθυγράμμιση/στοίχιση αλληλουχιών πρωτεϊνών ή νουκλεϊκών οξέων. Το HMMER είναι ιδιαίτερα αποτελεσματικό στην ανίχνευση απομακρυσμένων ομοιοτήτων, καθώς χρησιμοποιεί στατιστικά μοντέλα για να αναγνωρίσει πρότυπα και συντηρημένες περιοχές στις αλληλουχίες.

Πως λειτουργεί το HMMER

Δημιουργία HMM: Το HMMER ξεκινάει με την κατασκευή ενός μοντέλου κρυφού Markov από μια ευθυγράμμιση αλληλουχιών. Το μοντέλο αυτό περιλαμβάνει στατιστικά στοιχεία για τη συχνότητα εμφάνισης κάθε αμινοξέος ή νουκλεοτιδίου σε κάθε θέση της αλληλουχίας.

Αναζήτηση Ομοιοτήτων: Το HMM που δημιουργήθηκε συγκρίνεται με άλλες αλληλουχίες πρωτεϊνών ή νουκλεϊκών οξέων, εντοπίζοντας περιοχές που παρουσιάζουν ομοιότητες με το μοντέλο. Λόγω της στατιστικής φύσης του HMM, το εργαλείο μπορεί να ανιχνεύσει απομακρυσμένες ομοιότητες που άλλες μέθοδοι δεν μπορούν να αναγνωρίσουν.

Εκτίμηση Στατιστικής Σημασίας: Το HMMER παρέχει στατιστικές βαθμολογίες για την εκτίμηση της σημασίας των αποτελεσμάτων, όπως το **E-value**, το οποίο δείχνει την πιθανότητα η ομοιότητα να είναι τυχαία.



Το HMMER μπορεί να χρησιμοποιηθεί για να διακρίνει αν μια αλληλουχία είναι κωδικοποιημένη ή όχι. **Στην περίπτωση των κωδικοποιημένων αλληλουχιών**, συγκρίνει την πρωτεϊνική αλληλουχία με πρότυπα από βάσεις όπως το Pfam ή το InterPro. Αν βρεθεί αντιστοιχία, είναι πιθανό η αλληλουχία να αφορά λειτουργική πρωτεΐνη.

Για μη κωδικοποιημένες αλληλουχίες, το HMMER συγκρίνει την είσοδο με μοντέλα RNA, όπως εκείνα για tRNA ή rRNA. Αν υπάρχει ομοιότητα, η αλληλουχία πιθανώς δεν μεταφράζεται σε πρωτεΐνη αλλά έχει άλλον ρόλο, όπως ρυθμιστικό ή δομικό.

Παραδείγματα εφαρμογών του HMMER

Αναγνώριση συντηρημένων περιοχών: Εντοπίζει περιοχές με εξελικτική σημασία σε πρωτεϊνικές αλληλουχίες.

Αναγνώριση δομικών περιοχών: Βοηθά στην ανίχνευση επαναλαμβανόμενων μοτίβων που εμφανίζονται σε διαφορετικές πρωτεΐνες.

Μελέτη απομακρυσμένων ομοιοτήτων: Αναγνωρίζει λειτουργικές ή δομικές ομοιότητες που δεν εντοπίζονται εύκολα με απλή αλληλουχιακή σύγκριση.

Βάσεις δεδομένων που χρησιμοποιεί το HMMER

- **Pfam**: Περιλαμβάνει συντηρημένα μοτίβα και οικογένειες πρωτεϊνών.
- **InterPro**: Συνδυάζει δεδομένα από πολλές βάσεις και παρέχει λειτουργικές πληροφορίες για πρωτεϊνικές περιοχές.
- **EMBL-EBI**: Προσφέρει εργαλεία και βάσεις για ανάλυση νουκλεοτιδικών και πρωτεϊνικών δεδομένων.

Παράδειγμα με χρήση API σε Python

Στάδιο 1: Αναζήτηση Hox γονιδίων με HMMER API

- Δίνεται η αλληλουχία ενός Hox γονιδίου.
- Το API εκτελεί αναζήτηση στη βάση **Pfam**.
- Τα αποτελέσματα αποθηκεύονται στη μεταβλητή `hmm_results` και περιλαμβάνουν HMM προφίλ που ταιριάζουν με την αλληλουχία.

```
import requests
import json

hmm_url = "https://www.ebi.ac.uk/Tools/hmmer/search/pfam"
// URL του HMMER API για αναζήτηση πρωτεϊνικής αλληλουχίας

sequence_data = {
    "seq": "MADEKPFQKQVTKLAGFTKAVHKKHTQHIYKLQKKSQ...",
    // Η πρωτεϊνική αλληλουχία του Hox γονιδίου
    "seqdb": "pfam"
    // Αναζήτηση στη βάση δεδομένων Pfam
}

response = requests.post(hmm_url, headers={"Content-Type": "application/json"},
                        data=json.dumps(sequence_data)) // Εκτέλεση αιτήματος POST για αναζήτηση

if response.status_code == 200:
    // Έλεγχος του κωδικού απάντησης
    hmm_results = response.json()
    // Αν η αναζήτηση ήταν επιτυχής, αποθήκευση των αποτελεσμάτων συμπεριλαμβανομένων
    // HMM [hidden Markov models] προφίλ που ταιριάζουν με την αλληλουχία.
    print(json.dumps(hmm_results, indent=4))
else:
    print(f"Error: {response.status_code}")
```

Εικόνα 2 Αναζήτηση Hox γονιδίων με HMMER API

Η μεταβλητή `hmm_results` περιέχει τα αποτελέσματα που προέρχονται από την αναζήτηση του HMMER στη βάση δεδομένων Pfam, συμπεριλαμβανομένων των HMM προφίλ που ταιριάζουν με την αλληλουχία που βάλαμε.

Στάδιο 2: Χρήση του Pfam API για επιπλέον πληροφορίες

- Από τα αποτελέσματα, εξάγονται τα **Pfam ID** (όπως το PF00046 για το homeobox).
- Το Pfam API επιστρέφει αναλυτικά δεδομένα για το κάθε προφίλ: βιολογικές λειτουργίες, εξελικτική σημασία και ρόλο των Hox γονιδίων.

```
import requests

pfam_id = "PF00046"
//Ορισμός του Pfam ID για το HMM Profile που θές να αναζητήσεις
//παράδειγμα εδώ για Hox)

pfam_url = f"https://pfam.xfam.org/family/{pfam_id}?output=json"
//URL του Pfam API για λήψη πληροφοριών σχετικά με κάθε ID

response = requests.get(pfam_url)
//Υποβολή αιτήματος στο Pfam API

if response.status_code == 200:
    //Ελεγχος αν το αίτημα ολοκληρώθηκε επιτυχώς

    pfam_data = response.json()
    //Ανάκτηση δεδομένων σε μορφή JSON

    print(f"Pfam ID: {pfam_data['pfam']['id']}")
    //Εμφάνιση αποτελεσμάτων

    print(f"Description: {pfam_data['pfam']['description']}")
    print(f"Number of Sequences: {pfam_data['pfam']['num_full']}")
    print(f"HMM Length: {pfam_data['pfam']['hmm_length']}")
else:
    print(f"Error: {response.status_code}")
```

Εικόνα 3 Χρήση του Pfam API για επιπλέον πληροφορίες

Έχουμε τώρα αντλήσει αναλυτικές πληροφορίες για τα **HMM profiles** που εντοπίζονται μέσω του HMMER API. Με βάση τα **Pfam IDs**, μπορούμε να ζητήσουμε περιγραφή, βιολογικό ρόλο και λειτουργίες κάθε τομέα.

Σημαντικά Pfam IDs που σχετίζονται με Hox γονίδια στους οστεϊθύες

- **PF00046 – Homeobox domain**
Περιέχει το βασικό μοτίβο 60 αμινοξέων που επιτρέπει στις Hox πρωτεΐνες να συνδέονται με DNA και να ρυθμίζουν την ανάπτυξη.
- **PF00126 – Hox9/10/11/13 family**
Καλύπτει τα Hox γονίδια Hox9 έως Hox13, τα οποία ελέγχουν τη διαμόρφωση περιοχών του σώματος, όπως τα πτερύγια και τα άκρα.

Pfam URL: Το URL για το αίτημα στο API περιλαμβάνει το Pfam ID και χρησιμοποιεί την παράμετρο output=json για να επιστρέψει τα δεδομένα σε μορφή JSON.

Αποτελέσματα: Επιστρέφονται Συμπέρασμα όπως η περιγραφή του domain, το μήκος του HMM profile, και ο αριθμός των αλληλουχιών που ανήκουν σε αυτό.

2.5 Clustal Omega

Είναι ένα εργαλείο πολλαπλής στοίχισης αλληλουχιών (multiple sequence alignment) που χρησιμοποιείται για την ανάλυση εξελικτικών σχέσεων μεταξύ αλληλουχιών DNA, RNA ή

πρωτεϊνών. Στηρίζεται σε προοδευτικές μεθόδους ευθυγράμμισης/στοίχισης, όπου οι αλληλουχίες ευθυγραμμίζονται διαδοχικά, χρησιμοποιώντας μια ιεραρχική προσέγγιση, και δημιουργεί μια ενιαία στοίχιση όλων των αλληλουχιών.

Πως λειτουργεί το CLUSTAL OMEGA

Το Clustal Omega υπολογίζει τις αποστάσεις μεταξύ αλληλουχιών με γρήγορο αλγόριθμο, δημιουργεί ένα φυλογενετικό δέντρο οδηγό βάσει αυτών των αποστάσεων και στη συνέχεια εκτελεί στοίχιση, ξεκινώντας από τις πιο όμοιες προς τις πιο διαφορετικές.

Χρησιμοποιείται για στοίχιση πολλαπλών αλληλουχιών DNA, RNA ή πρωτεϊνών, για την ανάλυση εξελικτικών σχέσεων και την κατασκευή φυλογενετικών δέντρων.

Στην περίπτωση κωδικοποιημένων αλληλουχιών, ευθυγραμμίζει πρωτεΐνες από διάφορα είδη, εντοπίζοντας κοινές λειτουργικές περιοχές. Για μη κωδικοποιημένες, όπως tRNA ή rRNA, συγκρίνει τις αλληλουχίες για την αναγνώριση επαναλαμβανόμενων λειτουργικών μοτίβων.

Παραδείγματα εφαρμογών του CLUSTAL OMEGA

Αναγνώριση συντηρημένων περιοχών σε γονίδια, δημιουργία φυλογενετικών δέντρων για την απεικόνιση εξελικτικών σχέσεων, πρόβλεψη λειτουργίας άγνωστων αλληλουχιών βάσει ομοιότητας με γνωστές.

Τρόποι αναζήτησης και πρόσβασης στα δεδομένα των βάσεων που χρησιμοποιεί το Clustal Omega

Το **Clustal Omega** μπορεί να χρησιμοποιηθεί μέσω της ιστοσελίδας του EMBL-EBI, όπου οι χρήστες μπορούν να εισάγουν τις αλληλουχίες τους ή να συνδέονται με τις παρακάτω βάσεις δεδομένων .

- **UniProt:** Παρέχει δεδομένα πρωτεϊνικών αλληλουχιών.
- **EMBL-EBI:** Παρέχει ευρεία γκάμα εργαλείων για ανάλυση βιολογικών αλληλουχιών.

Εναλλακτικά, μπορεί να χρησιμοποιηθεί μέσω προγραμματιστικών διεπαφών (API) για να γίνει αυτοματοποίηση αναζητήσεων και αναλύσεων αλληλουχιών.

Παράδειγμα Python **API** για την **UniProt** βάση:

αναζήτηση του γονιδίου **DMRT1** στη βάση δεδομένων UniProt χρησιμοποιώντας το UniProt API και κατόπιν χρήση του Clustal Omega για στοίχιση των αλληλουχιών.

```

import requests

def search_uniprot(query):
    url = 'https://rest.uniprot.org/uniprotkb/search'
    params = {
        'query': query,
        'format': 'json'
    }
    response = requests.get(url, params=params)
    if response.status_code == 200:
        return response.json()
    else:
        return None

results = search_uniprot('DMRT1')
if results:
    for entry in results['results']:
        print(f"Accession: {entry['primaryAccession']}, Organism: {entry['organism']['scientificName']},
              Function: {entry['proteinDescription']['recommendedName']['fullName']['value']}")

# Περαιτέρω επεξεργασία για χρήση στο Clustal Omega
# Θα μπορούσαμε να αποθηκεύσουμε τα αποτελέσματα για μελλοντική στοίχιση

```

Εικόνα 4 Παράδειγμα Python API για την UniProt βάση

Για κλήσεις στο **Clustal Omega** API, χρειάζεσαι ένα script που θα στείλει τις αλληλουχίες που θα στοιχίσουμε μέσω του EMBL-EBI's Clustal Omega web service. Αυτό το API δέχεται αλληλουχίες πρωτεϊνών ή DNA, τις στοιχίζει και επιστρέφει τα αποτελέσματα.

Βήματα:

Αναζητούμε τις αλληλουχίες για το γονίδιο **DMRT1** από τη βάση δεδομένων UniProt>>
 Στέλνουμε αυτές τις αλληλουχίες στο Clustal Omega API για να τις στοιχίσει>> Λαμβάνουμε τα αποτελέσματα στοίχισης.

Παράδειγμα για το πώς θα υλοποιηθεί σε Python χρησιμοποιώντας τη βιβλιοθήκη requests.

```

import requests
import time

def fetch_dmrt1_sequences():
    url = 'https://rest.uniprot.org/uniprotkb/search'
    params = {
        'query': 'DMRT1',
        'format': 'json',
        'fields': 'sequence'
    }
    response = requests.get(url, params=params)
    if response.status_code == 200:
        return [entry['sequence']['value'] for entry in response.json()['results']]
    else:
        return None

def submit_clustal_omega_job(sequences):
    url = 'https://www.ebi.ac.uk/Tools/services/rest/clustalo/run'
    fasta_data = ''.join([f">seq{i}\n{seq}\n" for i, seq in enumerate(sequences)])
    payload = {
        'sequence': fasta_data,
        'email': 'your_email@example.com',
        'outfmt': 'clustal'
    }
    response = requests.post(url, data=payload)
    if response.status_code == 200:
        return response.text
    else:
        return None

```

Εικόνα 5 Αναζήτηση για το γονίδιο DMRT1 από τη βάση δεδομένων UniProt

```

def get_clustal_omega_results(job_id):
    status_url = f'https://www.ebi.ac.uk/Tools/services/rest/clustalo/status/{job_id}'
    result_url = f'https://www.ebi.ac.uk/Tools/services/rest/clustalo/result/{job_id}/clustal'

    while True:
        status_response = requests.get(status_url)
        if status_response.text == 'FINISHED':
            # Αν έχει ολοκληρωθεί, λαμβάνουμε τα αποτελέσματα
            result_response = requests.get(result_url)
            if result_response.status_code == 200:
                return result_response.text
            else:
                return None
        elif status_response.text == 'ERROR':
            return None
        else:
            time.sleep(10)

    sequences = fetch_dmrt1_sequences()
    if sequences:
        job_id = submit_clustal_omega_job(sequences)
        if job_id:
            print(f"Job submitted. ID: {job_id}")
            alignment = get_clustal_omega_results(job_id)
            if alignment:
                print("Alignment results:")
                print(alignment)
            else:
                print("Failed to retrieve alignment results.")
        else:
            print("Failed to submit job to Clustal Omega.")
    else:
        print("Failed to fetch DMRT1 sequences from UniProt.")

```

Εικόνα 6 Αναζήτηση αγια το γονίδιο DMRT1 από τη βάση δεδομένων UniProt

Επεξήγηση του κώδικα:

fetch_dmrt1_sequences(): Κάνει κλήση στο UniProt API για να βρει τις αλληλουχίες του DMRT1 γονιδίου.

submit_clustal_omega_job(): Στέλνει τις αλληλουχίες στο Clustal Omega API για να εκτελέσει την στοίχιση.

get_clustal_omega_results(): Ελέγχει την κατάσταση του job μέχρι να ολοκληρωθεί και, όταν ολοκληρωθεί, επιστρέφει τα αποτελέσματα της στοίχισης.

Σημειώσεις: job_id [είναι ένας μοναδικός κωδικός που αποδίδεται στο κάθε request προς το Clustal Omega API], χρειάζεται ένα έγκυρο email στο payload καθώς αυτό απαιτείται από την υπηρεσία Clustal Omega, ο έλεγχος για την κατάσταση του job γίνεται ανά 10 δευτερόλεπτα.

2.6 MAFFT (Multiple Alignment using Fast Fourier Transform)

Εργαλείο για την ευθυγράμμιση πολλαπλών αλληλουχιών (Multiple Sequence Alignment - MSA), ταχύτατο και με ακρίβεια, ειδικά όταν χειρίζεται μεγάλους αριθμούς πρωτεϊνικών ή νουκλεοτιδικών αλληλουχιών. Χρησιμοποιείται ευρέως για την ανάλυση εξελικτικών σχέσεων και για τη δημιουργία φυλογενετικών δέντρων. Το MAFFT χρησιμοποιεί διάφορες προσεγγίσεις για να επιτύχει υψηλή απόδοση σε περιπτώσεις όπου πρέπει να γίνει ευθυγράμμιση σε χιλιάδες αλληλουχίες.

Πως λειτουργεί το MAFFT

Το **MAFFT** πραγματοποιεί ευθυγράμμιση αλληλουχιών σε δύο βασικά στάδια. Πρώτα εφαρμόζει μια ταχεία προσέγγιση με χρήση **μετασχηματισμού Fourier (FFT)**, μετατρέποντας τις αλληλουχίες σε αριθμητικά μοτίβα για να εντοπίσει περιοχές ομοιότητας. Στη συνέχεια, εκτελεί **προοδευτική ευθυγράμμιση**, ευθυγραμμίζοντας όλες τις αλληλουχίες με βάση τα αρχικά αποτελέσματα.

Το εργαλείο προσφέρει διάφορες μεθόδους, όπως οι **L-INS-i** και **G-INS-i**, ώστε ο χρήστης να επιλέγει ανάμεσα σε μεγαλύτερη ακρίβεια ή ταχύτερη εκτέλεση, ανάλογα με τις απαιτήσεις της ανάλυσης.

Παράδειγμα Χρήσης του MAFFT σε Οστεϊχθύες – Γονίδιο *myod1*

Το *myod1* είναι γονίδιο-κλειδί για τη ρύθμιση της μυϊκής ανάπτυξης. Στους οστεϊχθύες, εμφανίζει υψηλή συντήρηση, ιδιαίτερα στις κωδικοποιημένες περιοχές.

Συλλογή αλληλουχιών>> Ανακτώνται πρωτεϊνικές αλληλουχίες του *myod1* από διαφορετικά είδη οστεϊχθύων (όπως zebrafish, medaka, fugu, trout) μέσω UniProt ή Ensembl.

Ευθυγράμμιση στο MAFFT>> Οι αλληλουχίες αποθηκεύονται σε ένα αρχείο .fasta. Η ευθυγράμμιση γίνεται με: `mafft --auto myod1_fish.fasta > myod1_fish_aligned.fasta`

Ανάλυση αποτελεσμάτων>> Η έξοδος αποκαλύπτει συντηρημένες θέσεις αμινοξέων, οι οποίες πιθανώς σχετίζονται με τη ρύθμιση της μυογένεσης. Παράλληλα, επισημαίνονται διαφοροποιήσεις που μπορεί να συνδέονται με εξελικτική προσαρμογή.

Εφαρμογές του MAFFT

Εξελικτική βιολογία: Χρήση στην ευθυγράμμιση πρωτεϊνικών ή γονιδιακών αλληλουχιών για τη μελέτη εξελικτικών σχέσεων.

Δομική βιολογία: Ευθυγράμμιση πρωτεϊνών για την ανάλυση της ομολογίας και την πρόβλεψη δομής.

Βιολογία συστημάτων: Χρήση για ανάλυση γονιδιακών δικτύων και βιολογικών μονοπατιών.

Τρόποι Αναζήτησης και Πρόσβασης στα Δεδομένα των Βάσεων Δεδομένων

Μπορούμε να χρησιμοποιήσουμε APIs για τη σύνδεση με τις βάσεις δεδομένων **UniProt** και **Ensembl** για να κατεβάσουμε αλληλουχίες γονιδίων όπως το Sox9 για περαιτέρω ανάλυση μέσω του MAFFT:

Αναζήτηση μέσω API στην UniProt

```
response = requests.get("https://rest.uniprot.org/uniprotkb/search?query=gene:sox9+AND+organism_id:9606&format=fasta")
```

Αναζήτηση μέσω API στην Ensembl

```
response = requests.get("https://rest.ensembl.org/sequence/id/ENSG00000125398?content-type=text/x-fasta")
```

Το ENSG00000125398 είναι το Ensembl ID για το ανθρώπινο γονίδιο **SOX9**

Παράδειγμα χρησιμοποιώντας το **UniProt API**, για ανάκτηση πρωτεϊνικών αλληλουχιών για ανάλυση με το MAFFT σε μορφή FASTA:

```
response = requests.get("https://rest.uniprot.org/uniprotkb/search?query=gene:SOX9+AND+organism_id:9606&format=fasta")
```

```
with open("sox9_sequences.fasta", "w") as f:
```

```
    f.write(response.text) // αποθηκεύει τις FASTA αλληλουχίες στο αρχείο sox9_sequences.fasta
```

```
mafft --auto sox9_sequences.fasta > aligned_sox9.fasta // ανάλυση με FASTA
```

2.7 MUSCLE (Multiple Sequence Comparison by Log-Expectation)

Το MUSCLE είναι εργαλείο για πολλαπλή ευθυγράμμιση αλληλουχιών (Multiple Sequence Alignment), γνωστό για την υψηλή ακρίβεια και ταχύτητά του. Χρησιμοποιείται για την ευθυγράμμιση τόσο πρωτεϊνικών όσο και νουκλεοτιδικών αλληλουχιών.

Ξεκινά με τη δημιουργία μιας αρχικής ευθυγράμμισης βασισμένης σε δέντρο ομοιότητας και στη συνέχεια εφαρμόζει βελτιστοποίηση με τη μέθοδο προσδοκίας-μέγιστης ακρίβειας (log-expectation), ώστε να βελτιώσει τη στοίχιση σε κρίσιμες περιοχές.

Πώς λειτουργεί

Αρχικό στάδιο – Προοδευτική ευθυγράμμιση

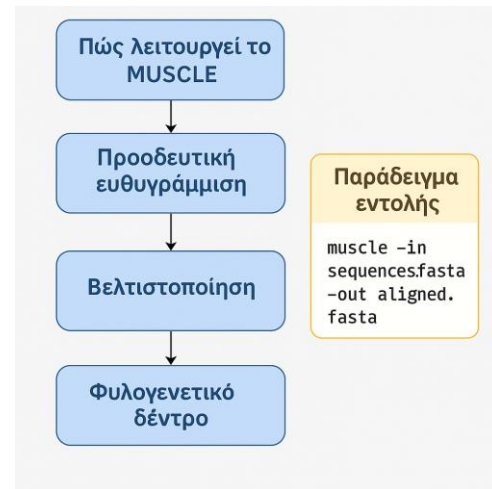
Δημιουργείται ένα γρήγορο δέντρο ομοιότητας, βασισμένο σε ευρετικές μεθόδους, για την πρώτη στοίχιση των αλληλουχιών.

Βελτιστοποίηση

Η στοίχιση επαναυπολογίζεται επανειλημμένα με στόχο τη βελτίωση της ακρίβειας, κυρίως σε συντηρημένες περιοχές.

Φυλογενετικό δέντρο

Μπορεί να παραχθεί δέντρο εξελικτικής σχέσης με βάση την τελική στοίχιση.



Το MUSCLE εφαρμόζεται τόσο σε κωδικοποιημένες όσο και μη κωδικοποιημένες αλληλουχίες, ανάλογα με τον σκοπό της ανάλυσης.

Για κωδικοποιημένες αλληλουχίες, όπως πρωτεΐνες από διαφορετικά είδη, το εργαλείο εντοπίζει συντηρημένες περιοχές που σχετίζονται με κρίσιμες λειτουργίες (π.χ. ενεργά κέντρα, περιοχές δέσμευσης).

Για μη κωδικοποιημένες αλληλουχίες, όπως tRNA ή rRNA, εντοπίζει δομικά μοτίβα που υποδεικνύουν σταθερότητα ή ρυθμιστικό ρόλο, παρά το γεγονός ότι δεν οδηγούν σε παραγωγή πρωτεΐνης.

Παραδείγματα Εφαρμογών του MUSCLE

Το MUSCLE χρησιμοποιείται ευρέως στην ανάλυση βιολογικών αλληλουχιών για:

Κατασκευή φυλογενετικών δέντρων, ώστε να απεικονιστούν οι εξελικτικές σχέσεις μεταξύ ειδών ή γονιδίων.

Εντοπισμό συντηρημένων περιοχών μεταξύ αλληλουχιών διαφορετικών οργανισμών, με στόχο την κατανόηση κοινής εξελικτικής προέλευσης.

Ανάλυση λειτουργικών και δομικών χαρακτηριστικών πρωτεϊνών, μέσω της ανίχνευσης περιοχών που πιθανώς σχετίζονται με συγκεκριμένες βιολογικές λειτουργίες.

Τρόποι Αναζήτησης και Πρόσβασης στα Δεδομένα των Βάσεων Δεδομένων

Το **MUSCLE** για την ανάκτηση αλληλουχιών συνδέεται με τις βάσεις δεδομένων:

UniProt: Παρέχει πρωτεϊνικές αλληλουχίες και δεδομένα για τη βιολογική λειτουργία πρωτεϊνών.

RefSeq: Μια εκτενής βάση δεδομένων που παρέχει νουκλεοτιδικές και πρωτεϊνικές αλληλουχίες.

Κώδικας API για Διασύνδεση με τις Βάσεις Δεδομένων UniProt και RefSeq

Ο παρακάτω κώδικας επιτρέπει την ανάκτηση πρωτεϊνικών αλληλουχιών από το **UniProt** σε μορφή FASTA:

```
import requests

def fetch_uniprot_sequences(query):
    """Ανακτά πρωτεϊνικές αλληλουχίες από το UniProt σε μορφή FASTA.
    //:param query: Το ερώτημα αναζήτησης (π.χ. όνομα πρωτεΐνης ή οργανισμός).
    //:return: Αλληλουχίες σε μορφή FASTA.

    url = 'https://www.uniprot.org/uniprot/'
    params = {
        'query': query,
        'format': 'fasta'
    }
    response = requests.get(url, params=params)
    if response.status_code == 200:
        return response.text
    else:
        print(f"Σφάλμα: {response.status_code}")
        return None

# Παράδειγμα χρήσης
query = "P53 human"
sequences = fetch_uniprot_sequences(query)
if sequences:
    print(sequences)
```

Εικόνα 7 Ανάκτηση πρωτεϊνικών αλληλουχιών από το UniProt σε μορφή FASTA

3. Συνδυασμοί εργαλείων και βάσεων δεδομένων.

BLAST (Basic Local Alignment Search Tool)>> Χρησιμοποιείται για αναζήτηση ομόλογων ακολουθιών και εντοπισμού ομοιοτήτων σε DNA, RNA ή πρωτεΐνες. **Συμβατές βάσεις:** Ensembl, UniProt, RefSeq, GenBank, Ensembl Genomes, RNACentral, ZFIN

FASTA >> Παρόμοια χρήση με το BLAST, κατάλληλο για αναλύσεις ομοιοτήτων. **Συμβατές βάσεις:** Ensembl, UniProt, RefSeq, GenBank, RNACentral, ZFIN

PSI-BLAST>> Εξειδικεύεται στην ανίχνευση απομακρυσμένων ομόλογων ακολουθιών.
Συμβατές βάσεις: UniProt, RefSeq, GenBank

HMMER>> Ιδανικό για αναγνώριση δομικών μοτίβων και απομακρυσμένων ομόλογων ακολουθιών μέσω HMMs. **Συμβατές βάσεις:** Pfam, InterPro, EMBL-EBI, RNACentral, Ensembl

MAFFT>> Για πολλαπλή ευθυγράμμιση αλληλουχιών, υψηλής ακρίβειας.
Συμβατές βάσεις: UniProt, Ensembl, RefSeq

MUSCLE>> Χρήσιμο για γρήγορη και ακριβή στοίχιση πρωτεϊνών ή νουκλεοτιδικών αλληλουχιών. **Συμβατές βάσεις:** UniProt, RefSeq, GenBank

Clustal Omega>> Κατάλληλο για εξελικτική ανάλυση και ευθυγράμμιση πολλών αλληλουχιών.
Συμβατές βάσεις: UniProt, Ensembl, EMBL-EBI

4. Οι σημαντικότερες βάσεις

Ensembl Εστιάζει σε πλήρη γονιδιώματα πολλών οργανισμών, με δεδομένα για κωδικοποιημένα και μη κωδικοποιημένα γονίδια. Παρέχει εξελικτικές αναλύσεις, συγκριτική γονιδιωματική και ανατομική συσχέτιση.

UniProt Πρωτεϊνική βάση δεδομένων με πληροφορίες για τις αλληλουχίες, τη λειτουργία και τη δομή πρωτεϊνών. Εστιάζει στα κωδικοποιημένα γονίδια, με περιορισμένη αναφορά σε ncRNAs.

RefSeq (NCBI) Περιλαμβάνει επιμελημένα δεδομένα για γονιδιώματα, RNA και πρωτεΐνες πολλών οργανισμών. Υψηλής ποιότητας και κατάλληλη για επιστημονικές αναφορές.

GenBank Παγκόσμια πρωτογενής αποθήκη γενετικών αλληλουχιών. Περιέχει ακατέργαστα δεδομένα DNA και RNA, τόσο κωδικοποιημένα όσο και μη κωδικοποιημένα.

DDBJ (DNA Data Bank of Japan) Παρόμοια με το GenBank. Περιλαμβάνει DNA και RNA αλληλουχίες και συνεργάζεται διεθνώς με GenBank και EMBL για παγκόσμια διάδοση δεδομένων.

RNACentral Αποκλειστικά αφιερωμένη στα μη κωδικοποιημένα RNAs (π.χ. lncRNAs, miRNAs, snoRNAs). Ενοποιεί δεδομένα από πολλές βάσεις.

ZFIN Εξειδικευμένη βάση για το zebrafish. Παρέχει γονιδιωματικά και φαινοτυπικά δεδομένα, με κάλυψη τόσο για κωδικοποιημένες όσο και για μη κωδικοποιημένες περιοχές.

InterPro Κατηγοριοποιεί πρωτεΐνες βάσει δομής και λειτουργίας, και εντοπίζει πρωτεϊνικές επικράτειες (domains). Παρέχει εξελικτική πληροφόρηση για πρωτεϊνικά μοτίβα, χωρίς έμφαση σε ncRNAs.

BOLD System Εστιάζει στην ταυτοποίηση ειδών μέσω DNA barcoding. Δεν παρέχει πληροφορίες για κωδικοποιημένες ή μη κωδικοποιημένες περιοχές γονιδιώματος.

Συγκριτικά:

	Βάση Δεδομένων	Καλύπτει Κωδικοποιημένα	Καλύπτει Μη Κωδικοποιημένα	Εστίαση
1	Ensembl	Ναι	Ναι	Πλήρη γονιδιώματα, εξελικτικά δεδομένα
2	UniProt RefSeq	Ναι	Περιορισμένα	Πρωτεΐνες & λειτουργία
3	(NCBI)	Ναι	Ναι	Επιμελημένα δεδομένα για γονίδια και RNA
4	GenBank	Ναι	Ναι	Πρωτογενής αποθήκη αλληλουχιών
5	RNAcentral	Όχι	Ναι	ncRNAs (μόνο μη κωδικοποιημένα)
6	ZFIN	Ναι	Ναι	Zebrafish (αναπτυξιακή γενετική)
7	InterPro	Ναι	Όχι	Δομή και λειτουργία πρωτεϊνών
8	DDBJ	Ναι	Ναι	Παγκόσμια ανταλλαγή DNA/RNA
9	BOLD System	Όχι	Όχι	Ταυτοποίηση ειδών (DNA barcoding)

5.0 Περιγραφή και ανάλυση των βάσεων, εστιάζοντας στη δυνατότητά τους να παρέχουν πληροφορίες για κωδικοποιημένες και μη-κωδικοποιημένες περιοχές των γονιδιωμάτων

5.1 Ensembl

Η **Ensembl** είναι μια ολοκληρωμένη βάση δεδομένων και πλατφόρμα βιοπληροφορικής που προσφέρει ανοιχτή πρόσβαση σε γονιδιώματα διαφόρων ειδών, συμπεριλαμβανομένων των ανθρώπων, των ζώων και άλλων οργανισμών. Αναπτύχθηκε και διαχειρίζεται από το **European Bioinformatics Institute (EMBL-EBI)** και το **Wellcome Trust Sanger Institute**. Σκοπός της Ensembl είναι η παροχή αναλυτικών δεδομένων για γονιδιώματα, που περιλαμβάνουν κωδικοποιημένες περιοχές (πρωτεϊνικά γονίδια) και μη κωδικοποιημένες περιοχές (ncRNA, ρυθμιστικά στοιχεία).

Η Ensembl βοηθά στη σύγκριση γονιδιωμάτων μεταξύ διαφορετικών ειδών, στη μελέτη των εξελικτικών σχέσεων και στη βιολογική έρευνα που σχετίζεται με τα γονίδια, τις πρωτεΐνες και τις μεταλλάξεις. Παρέχει επίσης πλήρως σχολιασμένα γονιδιώματα και λειτουργικά στοιχεία, όπως SNPs και δομές γονιδίων.

Είδη Δεδομένων που Περιέχει και Τρόποι Πρόσβασης

Η Ensembl περιλαμβάνει:

Γονιδιακά δεδομένα: Γονίδια, exons, πρωτεϊνικές αλληλουχίες, μετάφραση και τοπολογίες.

Γονιδιώματα: Πλήρη σχολιασμένα γονιδιώματα για πολλά είδη, συμπεριλαμβανομένων των κωδικοποιημένων και μη κωδικοποιημένων περιοχών.

Μεταλλάξεις (SNPs, INDELS): Δεδομένα σχετικά με τις γενετικές παραλλαγές που εμφανίζονται στα γονιδιώματα των οργανισμών.

Μη κωδικοποιημένα RNA (ncRNA): Αλληλουχίες και λειτουργικές περιοχές που δεν κωδικοποιούν για πρωτεΐνες.

Συγκριτική γονιδιωματική: Εξελικτικά δεδομένα, συγκρίσεις γονιδιωμάτων μεταξύ ειδών, όπως ομόλογα και παραλογικά γονίδια.

Ρυθμιστικά στοιχεία: Δεδομένα για ενισχυτές, promoters και άλλες περιοχές που ρυθμίζουν την έκφραση των γονιδίων.

Τρόποι Πρόσβασης στα Δεδομένα στην Βάση

Πρόσβαση μέσω Ιστού (https): Οι χρήστες μπορούν να κάνουν αναζητήσεις και να πλοηγούνται στα γονιδιώματα και τα γονιδιακά δεδομένα μέσω της ιστοσελίδας της Ensembl με τον Ensembl Browser και με το BioMart δίνει τη δυνατότητα εξαγωγής δεδομένων βάσει συγκεκριμένων φίλτρων

στην πλατφόρμα BioMart από την ιστοσελίδα της Ensembl. Στην επιλογή "Choose Database", επιλέγουμε Ensembl Genes 112, που είναι η πιο πρόσφατη διαθέσιμη έκδοση. Στη συνέχεια, στο "Choose Dataset", επιλέγουμε Danio rerio genes (GRCz11) για να περιορίσουμε την αναζήτησή μας στο zebrafish.

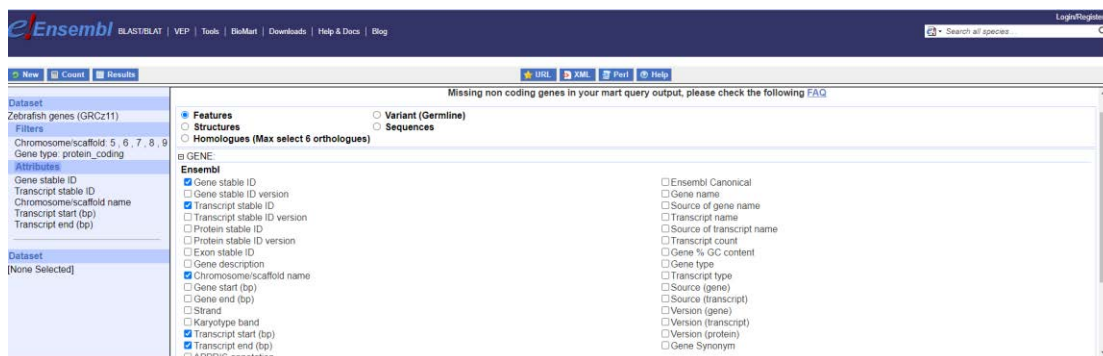
Από την αριστερή πλευρά, μεταβαίνουμε στην ενότητα **Filters**. Στην κατηγορία **REGION** μπορούμε να φιλτράρουμε ανά χρωμόσωμα, για παράδειγμα επιλέγοντας τα χρωμοσώματα 5, 6, 7, 8 ή 9, ώστε να περιορίσουμε τη γεωγραφική περιοχή αναζήτησης μέσα στο γονιδίωμα.

Για να αναζητήσουμε **κωδικοποιημένα γονίδια**, επιλέγουμε το φίλτρο **Gene Type** και ορίζουμε την τιμή **protein_coding**. Για **μη κωδικοποιημένα γονίδια**, μπορούμε να επιλέξουμε τιμές όπως **non_coding**, **lincRNA**, **miRNA**, ή **snRNA**, ανάλογα με το είδος του RNA που μας ενδιαφέρει.

Στην καρτέλα **Attributes**, επιλέγουμε ποια δεδομένα θέλουμε να εξάγουμε. Τυπικά περιλαμβάνουν:

Gene stable ID, Transcript stable ID, Chromosome/scaffold name, Transcript start, Transcript end

Αφού ορίσουμε τα φίλτρα και τα δεδομένα, πατάμε **Results**.



Εικόνα 8 Ensembl Browser

Τα αποτελέσματα εμφανίζονται άμεσα και μπορούμε να τα κατεβάσουμε σε μορφή **CSV**, **TSV** ή **XLS**, έτοιμα για ανάλυση ή εισαγωγή σε άλλο εργαλείο.

Πρόσβαση μέσω BLAST/BLAT Search: Επιτρέπει την αναζήτηση ομοιότητας αλληλουχιών DNA ή πρωτεϊνών με την εισαγωγή αλληλουχίας και σύγκρισή της με τα δεδομένα της Ensembl.

Πρόσβαση μέσω Ensembl Tools (VEP): Ένα εργαλείο που επιτρέπει την αναζήτηση και ανάλυση των επιπτώσεων γενετικών παραλλαγών στο γονιδίωμα.

Παράδειγμα για την αναζήτηση για το Ψάρι ζέβρα (Zebrafish) με χρήση του VEP επιλέγουμε το σωστό είδος, δηλαδή το **Zebrafish (Danio rerio)**

Οι παραλλαγές πρέπει να δίνονται σε μορφή CHROMOSOME:POSITION:REF, [πχ 1:15000:A:T]

Species: Zebratfish Assembly: GRCh38 [Change species](#)

Name for this job (optional):

Input data:

Either paste data:

Examples: [Ensembl default VCF](#), [Variant identifiers](#), [H2V3 notations](#), [EPQ](#)

Or upload file: Δεν επιλέχθηκε κανένα αρχείο.

Or provide file URL:

Transcript database to use:

☒ Ensembl transcripts

☐ RefSeq transcripts

☐ Ensembl and RefSeq transcripts

Additional configurations:

[Identifiers](#) Additional identifiers for genes, transcripts and variants

[Variants and frequency data](#) Co-located variants and frequency data

[Additional annotations](#) Additional transcript, protein and regulatory annotations

[Predictions](#) Variant deleteriousness predictions, e.g. SIFT, PolyPhen

[Filtering options](#) Pre-filter results by frequency or consequence type

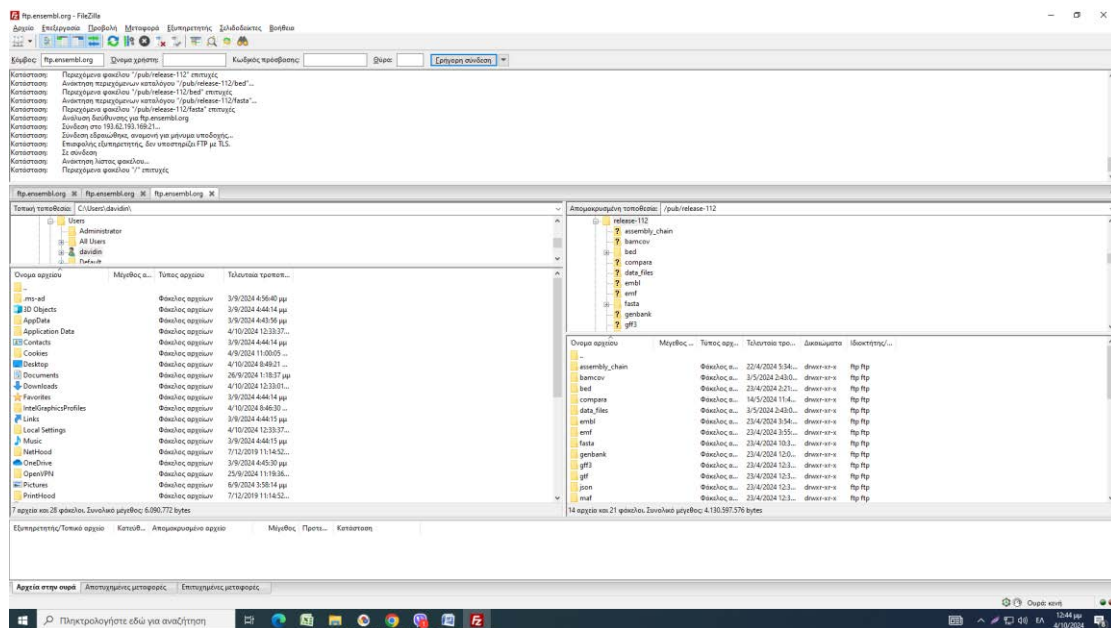
[Advanced options](#) Additional enhancements

Εικόνα 9 Αναζήτηση μέσω Ensembl Tools (VEP)

Πρόσβαση μέσω FTP (FileTransferProtocol): Παράδειγμα αναζήτησης στην βάση Ensembl για Sox9 γονίδια στους οστεϊχθες

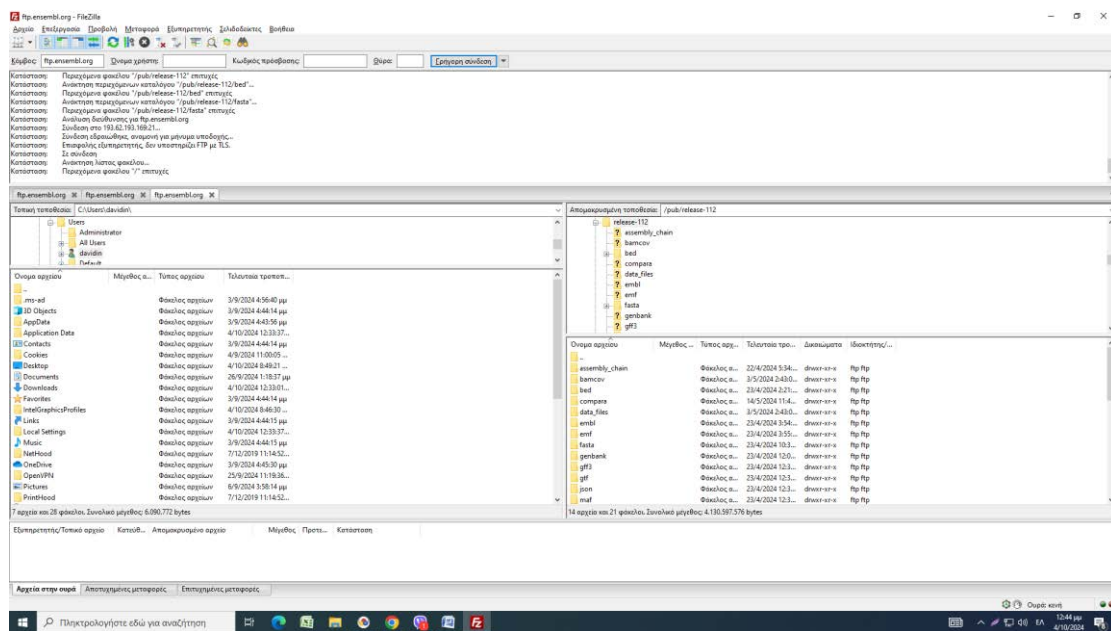
Σύνδεση με την χρήση του FileZilla στην FTP διεύθυνση <ftp://ftp.ensembl.org/pub/>

Στην κεντρική δομή του FTP, επιλέγουμε τον φάκελο release-113 (για το τελευταίο release), fasta, gtf, κ.ά. και αναζητούμε τα γονίδια Sox9 για τους οστεϊχθές, <ftp://ftp.ensembl.org/pub/release-109/>



Εικόνα 10 Πρόσβαση μέσω FTP (FileTransferProtocol) στην βάση Ensembl

Πρέπει να κατεβάσουμε τα αρχεία FASTA ή GTF που περιέχουν τις αλληλουχίες των γονιδίων <ftp://ftp.ensembl.org/pub/release-109/fasta/>



Μέσα στον κατάλογο fasta, υπάρχουν φάκελοι με τα ονόματα διάφορων ειδών οστεϊχθύων. Επιλέγουμε για παράδειγμα *Danio rerio* για ζέβραψαράκι. Μέσα στον φάκελο του είδους θα βρούμε αρχεία που περιέχουν τις γονιδιακές αλληλουχίες, π.χ.:

ftp://ftp.ensembl.org/pub/release-109/fasta/danio_rerio/cdna/Danio_rerio.GRCz11.cdna.all.fa.gz

Αποσυμπίεση του αρχείου και με εργαλεία όπως το `grep` ή `less` αναζητούμε το γονίδιο `Sox9`.

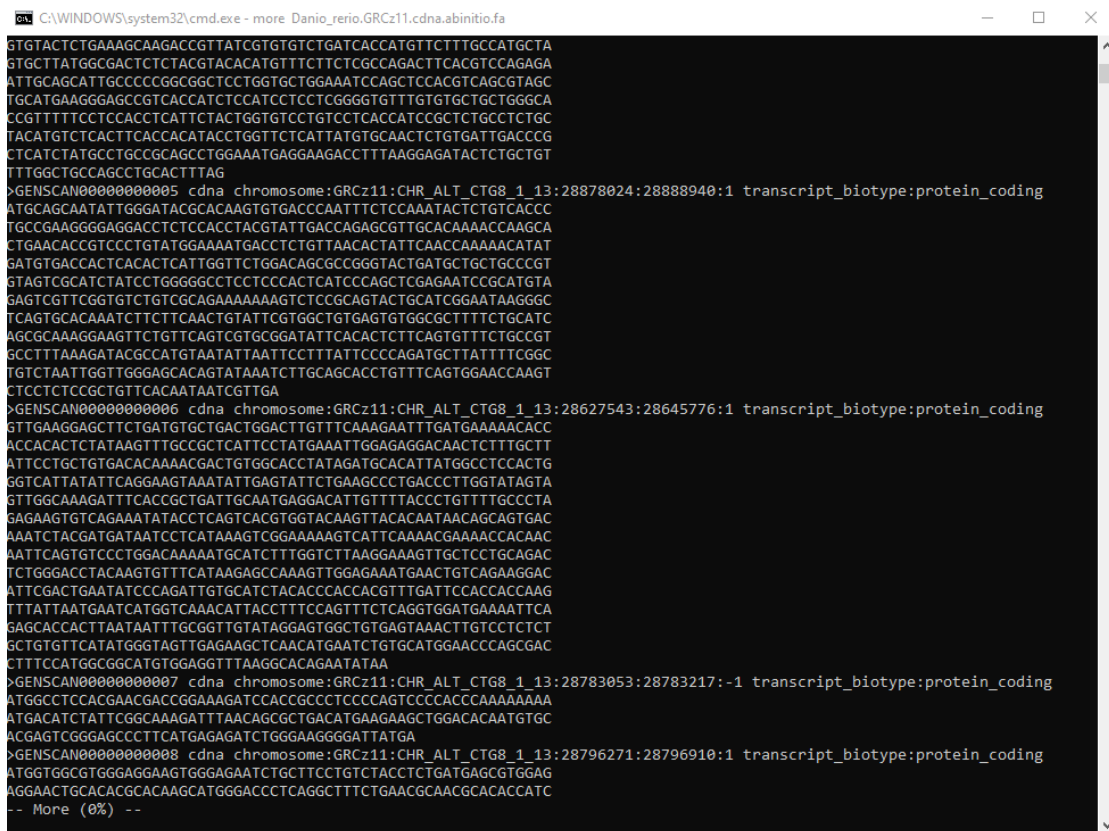
Για παράδειγμα:

Με το εργαλείο `findstr` θα αναζητήσουμε στο αρχείο κειμένου

`findstr /i "Sox9" Danio_rerio.GRCz11.cdna.all.fa`

Με το εργαλείο `more` μπορούμε να δούμε το περιεχόμενο του αρχείου

`more Danio_rerio.GRCz11.cdna.all.fa`



```
C:\WINDOWS\system32\cmd.exe - more Danio_rerio.GRCz11.cdna.abinitio.fa
GTGACTCTGAAAGCAAGACCGTTATCGTGTGCTGATCACCATGTTCTTTGCCATGCTA
GTGCTTATGGCGACTCTACGTACACATGTTTCTTCTGCCAGACTTCACGTCCAGAGA
ATTGCAGCATTTGCCCGCGCGGCTCTGGTGTGGAAATCCAGCTCCACGTCAGCGTAGC
TGCAATGAAGGAGCGCTCACCATCTCCATCTCTCGGGTGTGTTGTGTGCTGGGCA
CCGTTTTTCTCCACCTATTCTACTGGTGTCTCTCCATCCGCTCTGCTCTGCG
TACATGTCTCACTTCAACCATACCTGGTCTCATTATGTGCAACTCTGTGATTGACCG
CTCATCTATGCTGCCGCGACCTGGAAATGAGGAAGACCTTTAAGGAGATACTGCTGT
TTTGGTGCCAGCGCTGCACCTTAG
>GENSCAN0000000005 cdna chromosome:GRCz11:CHR_ALT_CTG8_1_13:28878024:28888940:1 transcript_biotype:protein_coding
ATGCAGCAATATTGGGATACGCACAAGTGTGACCAATTTCTCCAAATACTCTGCACCC
TGCCGAAGGGGAGGACCTCTCCACCTACGTATTGACAGAGCGTTGCACAAAACCAAGCA
CTGAACACCGTCCCTGTATGGAAATGACCTCTGTTAACACTATTCAACCAAAACATAT
GATGTGACCACTCACACTATTGGTTCTGGACAGCGCGGTAAGTGTGCTGCTGCCGT
GTAGTCGCTATCTCTGGGGGCTCTCTCCACTCATCCAGCTCGAAGATCCGATGTA
GAGTCGTTCCGTTGCTCTGCGCAGAAAAAAGTCTCCGCACTACTGCATCGGAATAAGGGC
TCAGTGCAAAATCTTCTCAACTGATTCTGGTGTGAGTGTGGCGCTTTTCTGCATC
AGCGCAAGGAAGTTCTGTTCACTGCTGCGGATATTCACTCTTCACTGTTCTGCCGT
GCCTTTAAAGATACGCCATGTAATTAATTCCTTTATCCCGAGATGCTTATTTCCGC
TGCTAATTGGTTGGGAGCACAGTATAAATCTTGACGACCTGTTTCACTGGAACCAAGT
CTCCTCTCCGCTGTTCAATAATCGTTGA
>GENSCAN0000000006 cdna chromosome:GRCz11:CHR_ALT_CTG8_1_13:28627543:28645776:1 transcript_biotype:protein_coding
GTTGAAGGAGCTTCTGATGTGCTGACTGGACTTGTTCAAAGAAATTTGATGAAAAACCC
ACCACACTCTATAAGTTTGGCGCTCATTCCTATGAAATTTGGAGAGGACAACCTTTGCTT
ATTCCTGCTGTGACACAAAACGACTGTGGACCTATAGATGCACATTATGGCTCCACTG
GGTCATTATATTCAAGGAAGTAAATATTGAGTATTCTGAAGCCCTGACCTTGGTATAGTA
GTTGGCAAGATTTTCCCGCTGATTGCAATGAGGACATTTTACCCTGTTTGGCCCTA
GAGAAGTGTGAGAAATATACCTCAGTCACGTGGTACAAAGTTACACAATAACAGCAGTGAC
AAATCTACGATGATAATCTCATAAAGTCGGAAGAAAGTCAATTCAAAACGAAACCAAC
AATTCACTGTCCTGGACAAAAATGCATCTTTGGTCTTAAGGAAGTTGCTCCTGCAGAC
TCTGGGACCTACAAGTGTTCATAAGAGCCAAAGTTGGAGAAATGAAGTGTGAGAGGAC
ATTCGACTGAATATCCAGATTGTGCATCTACACCCACACGTTTGATTCCACACCAAG
TTTATTAAATGAATCATGGTCAACATTACCTTTCCAGTTTCTCAGGTGGATGAAAAATCA
GAGCACCCTTAATAATTTGCGGTTGTATAGGAGTGGCTGTGAGTAACTTGTCTCTCT
GCTGTGTTTATATGGGTAGTTGAGAAGCTCAACATGAATCTGTGCATGGAACCCAGCGAC
CTTCCATGGCGGCTGTGGAGGTTTAAGGCACAGAAATATAA
>GENSCAN0000000007 cdna chromosome:GRCz11:CHR_ALT_CTG8_1_13:28783053:28783217:-1 transcript_biotype:protein_coding
ATGGCCTCCAGCAACGACCGGAAAGATCCACCGCCTCCCGAGTCCCGACCCAAAAAAA
ATGACATCTATTCGGCAAGATTTAAGCGCTGACATGAAGAGCTGGACACAATGTGC
ACGAGTCGGGAGCCCTTCATGAGAGATCTGGGAAGGGGATTATGA
>GENSCAN0000000008 cdna chromosome:GRCz11:CHR_ALT_CTG8_1_13:28796271:28796910:1 transcript_biotype:protein_coding
ATGTTGGCTGGGAGGAAGTGGGAGAAATCTGTTCTGTCTACCTCTGATGAGCGTGGAG
AGGAAGTGCACACGCACAAGCATGGGACCTCAGGCTTTCTGAACGCAACGCACACCATC
-- More (0%) --
```

Εικόνα 11 Αναζήτηση με `findstr`

Δυνατότητα αναζήτησης με εργαλεία Linux όπως το `grep` και `toless` με εγκατάσταση του Git Bash για Windows περιβάλλον

```
MINGW64; c:/Users/davidin/Desktop/Danio_rerio.GRCz11.cdna.abinitio.fa.gz
davidin@8157-deg MINGW64 ~/Desktop/Danio_rerio.GRCz11.cdna.abinitio.fa.gz
$ ls
Danio_rerio.GRCz11.cdna.abinitio.fa
davidin@8157-deg MINGW64 ~/Desktop/Danio_rerio.GRCz11.cdna.abinitio.fa.gz
$ zgrep -i 'Sox9' rerio.GRCz11.cdna.all.fa.gz
```

Το **Cygwin** είναι ένα πιο εκτεταμένο περιβάλλον που παρέχει πλήρη λειτουργικότητα τύπου Unix στα Windows, επιτρέποντάς σου να εγκαταστήσεις μια ευρεία γκάμα εργαλείων, όπως **grep**

```
MINGW64; c:/Users/davidin/Desktop/Danio_rerio.GRCz11.cdna.abinitio.fa.gz
davidin@8157-deg MINGW64 ~/Desktop/Danio_rerio.GRCz11.cdna.abinitio.fa.gz
$ ls
Danio_rerio.GRCz11.cdna.abinitio.fa
davidin@8157-deg MINGW64 ~/Desktop/Danio_rerio.GRCz11.cdna.abinitio.fa.gz
$ zgrep -i 'Sox9' rerio.GRCz11.cdna.all.fa.gz
```

Πρόσβαση μέσω API στον FTP server :

Δημιουργία API με Python κώδικα (χρήση βιβλιοθήκης **ftplib**) για να συνδεθούμε στον FTP και να κατεβάσουμε τα αρχεία αυτόματα.

```
from ftplib import FTP                                # Σύνδεση με τον FTP server του Ensembl

ftp = FTP('ftp.ensembl.org')

ftp.login()                                           # Πλοήγηση στον κατάλογο για το είδος Danio rerio

ftp.cwd('/pub/release-109/fasta/danio_rerio/cdna/')

files = ftp.nlst()                                   # Λίστα των αρχείων στον κατάλογο
```

```
print(files)

with open('Danio_rerio.GRCz11.cdna.all.fa.gz', 'wb') as fp:      # Κατέβασμα του αρχείου cdna

ftp.retrbinary('RETR Danio_rerio.GRCz11.cdna.all.fa.gz', fp.write)

ftp.quit()
```

Η Ensembl ως ολοκληρωμένη πλατφόρμα ενοποίησης, σχολιασμού και πρόσβασης

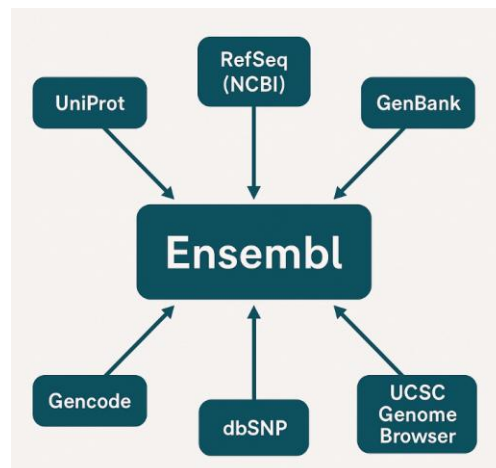
Οι περισσότερες βιολογικές βάσεις δεδομένων, όπως οι **UniProt**, **RefSeq**, **dbSNP**, είναι **εξειδικευμένες**. Η κάθε μία εστιάζει σε ένα συγκεκριμένο είδος πληροφορίας: για παράδειγμα, οι πρωτεΐνες, οι γενετικές παραλλαγές ή τα μη κωδικοποιημένα RNA.

Η σύνδεση μεταξύ τους γίνεται **μέσω παραπομπών (cross-references)**, χωρίς πλήρη ενοποίηση των δεδομένων. Οι μεταξύ τους δεσμοί είναι **χαλαροί** και δεν προσφέρουν ενιαία εμπειρία πλοήγησης ή συνδυασμένης ανάλυσης.

Σε αυτό το κατακερματισμένο περιβάλλον, η **Ensembl** διαδραματίζει **μοναδικό και ενοποιητικό ρόλο**. Συγκεκριμένα:

- **Ενσωματώνει δεδομένα** από πολλές ανεξάρτητες βάσεις (όπως UniProt, RefSeq, GenBank, dbSNP, Gencode) σε **μία ολοκληρωμένη πλατφόρμα**.
- **Συνδυάζει** διαφορετικούς τύπους πληροφορίας: γονιδιακή, πρωτεϊνική, ρυθμιστική και εξελικτική.
- Προσφέρει **ισχυρά εργαλεία** για αναζήτηση, ανάλυση και απεικόνιση, όπως ο **Ensembl Browser**, το **BioMart**, το **VEP**, REST API και FTP πρόσβαση.
- Βασίζεται σε **επιμελημένα δεδομένα υψηλής ακρίβειας**, μέσω συνεργασιών με εξειδικευμένες πηγές όπως το **Gencode** για RNA και το **RefSeq** για πρωτεΐνες.

Η Ensembl δεν είναι απλώς άλλη μία βάση· είναι **πυρήνας ενοποίησης**, κάνοντας προσβάσιμα και συγκρίσιμα τα δεδομένα της βιολογικής πληροφορίας σε ένα κοινό περιβάλλον.



5.2 UniProt

Η UniProt (Universal Protein Resource) αποτελεί μία από τις πληρέστερες και πιο αναγνωρισμένες βάσεις δεδομένων για πρωτεΐνες. Παρέχει αναλυτικές πληροφορίες σχετικά με τις αλληλουχίες και τις λειτουργίες πρωτεϊνών από ποικίλους οργανισμούς. Δημιουργήθηκε μέσα από συνεργασία τριών μεγάλων οργανισμών: του European Bioinformatics Institute (EMBL-EBI), του Swiss Institute of Bioinformatics (SIB) και του Protein Information Resource (PIR).

Ο βασικός της στόχος είναι η συγκέντρωση, επιμέλεια και διάθεση υψηλής ποιότητας δεδομένων για τις πρωτεΐνες, όπως η αλληλουχία, η λειτουργία, η δομή και οι μετα-μεταφραστικές τροποποιήσεις τους. Η UniProt είναι ιδιαίτερα σημαντική για τη μελέτη των πρωτεϊνών, τη βιοϊατρική έρευνα και τη βιοτεχνολογία, διευκολύνοντας την κατανόηση των βιολογικών μηχανισμών και την ερμηνεία παθολογικών καταστάσεων.

Τρόποι Αναζήτησης στην Βάση UniProt

Η UniProt προσφέρει διάφορες δυνατότητες αναζήτησης, καλύπτοντας τις ανάγκες τόσο απλών όσο και προχωρημένων χρηστών. Η πιο συνηθισμένη προσέγγιση είναι η αναζήτηση **μέσω της ιστοσελίδας της**, όπου μπορεί κανείς να χρησιμοποιήσει το όνομα πρωτεΐνης, το όνομα γονιδίου, τον κωδικό ταυτοποίησης (ID), λέξεις-κλειδιά ή το όνομα του οργανισμού. Επιπλέον, διατίθεται η **λειτουργία BLAST** για την εύρεση παρόμοιων αλληλουχιών με βάση μία δοθείσα πρωτεϊνική αλληλουχία. Τέλος, είναι δυνατή και η άμεση αναζήτηση **μέσω Accession Number**, προσφέροντας ταχύτατη πρόσβαση σε λεπτομερή δεδομένα για μια συγκεκριμένη πρωτεΐνη.

Είδη Δεδομένων που Περιέχει η UniProt προς αναζήτηση:

Αλληλουχίες Πρωτεϊνών, Λειτουργικές Πληροφορίες, Δομικές Πληροφορίες, Εξέλιξη και Σύγκριση Αλληλουχιών, Κλινικές και Παθολογικές Πληροφορίες, Σχολιασμένα Δεδομένα

Τρόποι Πρόσβασης στα Δεδομένα

Web Access (https) μέσω της ιστοσελίδας της UniProt, **FTP Access** πρόσβαση σε δεδομένα μαζικής λήψης μέσω FTP, **API (Application Programming Interface)**, **SPARQL Endpoint** πρόσβαση μέσω SPARQL για αναζητήσεις σε δομημένα δεδομένα και για τη σύνδεση με άλλες βάσεις δεδομένων που υποστηρίζουν το ίδιο πρωτόκολλο.

Ρόλος της UniProt σε σχέση με τις υπόλοιπες Βάσεις ως "κόμβος διασύνδεσης"

Η UniProt συγκεντρώνει, ενοποιεί και σχολιάζει πρωτεϊνικές πληροφορίες από άλλες βάσεις όπως:

Ensembl (γονιδιακές πληροφορίες), **RefSeq / GenBank** (νουκλεοτιδικές αλληλουχίες), **InterPro / Pfam** (δομικά/λειτουργικά μοτίβα), **PDB** (δομές πρωτεϊνών), **GO** (Gene Ontology), **KEGG / Reactome** (μεταβολικά μονοπάτια), **RNAcentral** (μη κωδικοποιημένα RNA)

Αντιστοίχιση ID και cross-references για κάθε καταχώρηση πρωτεΐνης, η UniProt παρέχει **cross-references** προς τα αντίστοιχα IDs στις άλλες βάσεις (π.χ. Ensembl gene ID, RefSeq mRNA ID, PDB ID).

API υποστήριξη για διασύνδεση: Παρέχει **RESTful API** που μπορεί να χρησιμοποιηθεί για την **εύρεση συνδέσμων και mapping μεταξύ ID** διαφόρων βάσεων.

Ποιοτικός έλεγχος & ενοποίηση: Ειδικά η **Swiss-Prot** (το σχολιασμένο υποσύνολο της UniProt) ελέγχει και ενοποιεί δεδομένα από πολλές πηγές, αυξάνοντας την αξιοπιστία.

5.3 RefSeq (NCBI)

Η RefSeq αποτελεί μέρος του οικοσυστήματος του NCBI και συνδέεται με πολλές άλλες βάσεις δεδομένων, διευκολύνοντας τη συνολική ανάλυση των βιολογικών δεδομένων. Παρέχει σύνδεση με τη UniProt για πληροφορίες σχετικά με τις πρωτεΐνες, με τη Gene για δεδομένα γονιδίων, και με τη GenBank για τις αρχικές καταθέσεις αλληλουχιών DNA και RNA. Επιπλέον, οι αλληλουχίες RefSeq εμφανίζονται δίπλα στις προβλέψεις του Ensembl, συμβάλλοντας σε συγκριτικές και φυλογενετικές μελέτες. Αυτή η δικτύωση υποστηρίζει μια ενοποιημένη, διεπιστημονική προσέγγιση, επιτρέποντας στους ερευνητές να συνδυάζουν και να συσχετίζουν πληροφορίες από διαφορετικές πηγές με μεγαλύτερη ακρίβεια και βάθος.

Είδη δεδομένων που περιέχει η βάση RefSeq

Η RefSeq περιέχει δεδομένα όπως Αλληλουχίες DNA (π.χ. γονίδια, cDNA, RNA), Πρωτεϊνικές αλληλουχίες, Σχολιασμένα γονιδιώματα από μια ευρεία ποικιλία οργανισμών, Πληροφορίες για γονιδιακές αλληλουχίες, θέσεις γονιδίων και γενετικές παραλλαγές.

Ένα από τα χαρακτηριστικά που διακρίνουν τη RefSeq από άλλες βάσεις δεδομένων είναι η επιμέλεια των δεδομένων από ειδικούς, καθώς και η παροχή μη πλεονάζουσων αλληλουχιών αναφοράς, κάτι που βοηθά στην αποφυγή συγχύσεων και διπλοκαταγραφών που μπορεί να εμφανιστούν σε άλλες βάσεις δεδομένων.

Τρόποι αναζήτησης και άντλησης δεδομένων από την βάση

Αναζήτηση μέσω της πλατφόρμας NCBI: Μπορείς να χρησιμοποιήσεις την κεντρική μηχανή αναζήτησης του NCBI (Entrez). Μπορείς να εισάγεις κλειδιά όπως όνομα γονιδίου ή πρωτεΐνης, αριθμούς πρόσβασης, ή το είδος του οργανισμού.

Παράδειγμα: Αναζήτηση για το ανθρώπινο γονίδιο **BRCA1** μέσω της Entrez.

FTP download: Η RefSeq παρέχει τη δυνατότητα μαζικής λήψης δεδομένων μέσω του FTP server της. Αυτός ο τρόπος είναι χρήσιμος όταν θέλεις να κατεβάσεις ολόκληρες αλληλουχίες ή μεγάλους όγκους δεδομένων.

Παράδειγμα: Λήψη δεδομένων για όλα τα γονιδιώματα ανθρώπων μέσω του FTP server.

API E-utilities: Μέσω του API της NCBI (E-utilities), οι χρήστες μπορούν να αντλήσουν δεδομένα από τη RefSeq με προγραμματιστικό τρόπο, κάτι που διευκολύνει την αυτοματοποίηση της αναζήτησης και λήψης δεδομένων.

Παράδειγμα: Χρήση του esearch για αναζήτηση του **TP53** γονιδίου και του efetch για να λάβεις την αλληλουχία του.

BLAST (Basic Local Alignment Search Tool): Μπορείτε να χρησιμοποιήσετε τη RefSeq για να βρείτε ομοιότητες αλληλουχιών με άλλες γνωστές αλληλουχίες μέσω του εργαλείου BLAST.

Χρήση: Αναζήτηση ομοιότητας αλληλουχιών, ώστε να βρείτε παρόμοιες αλληλουχίες στη βάση δεδομένων RefSeq.

Συμπέρασμα

Οι αλληλουχίες της RefSeq είναι επιμελημένες και σχολιασμένες από ειδικούς επιστήμονες, κάτι που προσδίδει υψηλή ακρίβεια και αξιοπιστία στη βάση δεδομένων.

Η βάση ενημερώνεται τακτικά με νέα δεδομένα και βελτιώσεις στις υπάρχουσες αλληλουχίες.

Η RefSeq είναι στενά συνδεδεμένη με πολλές άλλες βάσεις δεδομένων του NCBI, όπως το Gene, το Genome, και το Protein.

Σύγκριση με άλλες βάσεις δεδομένων

Η RefSeq υπερέρχει σε Αξιοπιστία και εγκυρότητα, Ποιότητα και επιμέλεια δεδομένων, Κλινική και ιατρική εφαρμογή, Διασύνδεση με άλλες βάσεις με API υποστήριξη

Υστερεί (σχετικά) σε Πληρότητα δεδομένων (έχει φιλτραρισμένες, μη πλεονάζουσες εγγραφές), Πρωτεϊνική λειτουργία (σε σχέση με UniProt), Φυλογενετική πληροφορία (σε σχέση με Ensembl)

Συνολικά, η RefSeq αποτελεί τη βάση επιλογής όταν προέχει η ποιότητα, η εγκυρότητα και η σταθερή αναφορά, ιδιαίτερα σε ιατρική και διαγνωστική έρευνα.

5.4 GenBank

GenBank είναι μια από τις μεγαλύτερες δημόσιες βάσεις δεδομένων για την αποθήκευση αλληλουχιών νουκλεϊνικών οξέων, και είναι διαχειριζόμενη από το Εθνικό Κέντρο Βιοτεχνολογικών Πληροφοριών (NCBI) στις ΗΠΑ. Ιδρύθηκε το 1982 και είναι μέρος της διεθνούς συνεργασίας με τις βάσεις EMBL-EBI (Ευρώπη) και DDBJ (Ιαπωνία). Σκοπός της GenBank είναι η συλλογή, καταχώρηση και παροχή ελεύθερης πρόσβασης σε αλληλουχίες DNA και RNA από όλα τα είδη οργανισμών.

Είδη Δεδομένων που Περιέχει η Βάση

Η GenBank περιέχει δεδομένα από διάφορους τύπους οργανισμών, όπως ιούς, βακτήρια, φυτά, ζώα, και ανθρώπους. Οι τύποι δεδομένων περιλαμβάνουν:

Αλληλουχίες DNA/RNA (γονιδιωματικές, cDNA, και RNA αλληλουχίες), **Γονίδια και μεταγραφές** (exons, introns, UTRs), **Πληροφορίες σχετικά με πρωτεΐνες** που έχουν αναγνωριστεί από τις αλληλουχίες (μέσω μετάφρασης των νουκλεϊνικών οξέων), **Βιβλιογραφία και αναφορές** από επιστημονικά άρθρα που συνδέονται με τις αλληλουχίες, **Επισημάνσεις γονιδίων** και άλλες πληροφορίες που αφορούν λειτουργίες γονιδίων ή μοριακά χαρακτηριστικά.

Εξειδικευμένα δεδομένα της GenBank

Η GenBank είναι εξειδικευμένη στην καταχώριση αλληλουχιών νουκλεϊνικών οξέων, και συχνά περιλαμβάνει πρωτογενή δεδομένα που κατατίθενται απευθείας από ερευνητές πριν αυτά διαμοιραστούν ή επεξεργαστούν από άλλες βάσεις. Ειδικά γονιδιωματικά δεδομένα ή πρωτογενή δεδομένα που δεν υπάρχουν αλλού μπορούν να βρεθούν πρώτα στη GenBank.

Τρόποι Αναζήτησης και Άντλησης Δεδομένων

Entrez Web Interface

Το NCBI (National Center for Biotechnology Information) έχει αναπτύξει το Entrez, ένα σύστημα ολοκληρωμένης αναζήτησης και ανάκτησης πληροφοριών, το οποίο παρέχει πρόσβαση σε δεκάδες βάσεις δεδομένων που καλύπτουν ευρύ φάσμα βιοϊατρικών και μοριακών βιολογικών πληροφοριών. Μέσω του Entrez, οι χρήστες μπορούν να αναζητήσουν και να συνδυάσουν πληροφορίες από διαφορετικές βάσεις, όπως η GenBank, το PubMed, το

RefSeq, το Gene, το Protein και πολλές άλλες. Ο στόχος του συστήματος είναι να διευκολύνει την ενιαία πρόσβαση και χρήση ετερογενών δεδομένων, επιτρέποντας την ολοκληρωμένη και πολυεπίπεδη ανάλυση της βιολογικής πληροφορίας.

Ο βασικός τρόπος πρόσβασης στο Entrez είναι μέσω της διαδικτυακής του διεπαφής (Web-based Interface) από την επίσημη ιστοσελίδα του NCBI. Μέσα από αυτή τη διεπαφή, οι χρήστες μπορούν να αναζητήσουν αλληλουχίες και σχετικές πληροφορίες χρησιμοποιώντας απλές ή σύνθετες λέξεις-κλειδιά, το όνομα του γονιδίου, το όνομα του οργανισμού, αναγνωριστικά εγγραφών (όπως accession numbers), ακόμη και τμήματα αλληλουχίας. Το Entrez υποστηρίζει την προβολή αποτελεσμάτων σε πολλαπλές μορφές, όπως FASTA ή GenBank, και προσφέρει δυνατότητα φιλτραρίσματος, διασύνδεσης με συναφείς εγγραφές άλλων βάσεων και άμεση πρόσβαση σε εργαλεία ανάλυσης όπως το BLAST και το Genome Viewer.

Επιπλέον, το Entrez είναι σχεδιασμένο να είναι προσβάσιμο και από χρήστες χωρίς προηγμένη τεχνική κατάρτιση, προσφέροντας απλότητα στη χρήση αλλά και πρόσβαση σε πολύπλοκες και διασυνδεδεμένες βιολογικές πληροφορίες. Η δομή του το καθιστά βασικό εργαλείο για ερευνητές, φοιτητές και επαγγελματίες που επιθυμούν να εξερευνήσουν, να εξάγουν και να αναλύσουν δεδομένα από ένα ευρύ φάσμα βιοϊατρικών πεδίων.

Παράδειγμα: Αναζήτηση για την catalase

The screenshot shows the NCBI Entrez search results for the query 'catalase'. The interface includes a search bar at the top with the query 'catalase' and a 'Search' button. Below the search bar, there are filters for 'Species' and 'Molecule types'. The search results are displayed in a list format, with the first result highlighted in a red box. The highlighted result is 'Rattus norvegicus strain Sprague-Dawley catalase (catalase) gene complete cds', which is a 10,544 bp linear DNA sequence. The result is linked to various databases including GenBank, FASTA, and Graphics. To the right of the search results, there are sections for 'Results by taxon', 'Find related data', and 'Search details'. The 'Results by taxon' section lists top organisms such as Salmonella enterica, Escherichia coli, Klebsiella pneumoniae, Staphylococcus aureus, and Pseudomonas aeruginosa. The 'Find related data' section has a dropdown menu for 'Database' and a 'Find items' button. The 'Search details' section shows the search query 'catalase[All Fields]' and a 'Search' button. The 'Recent activity' section is also visible at the bottom right.

Εικόνα 12 Αναζήτηση για την catalase

Entrez Programming Utilities (E-utilities)

Είναι ένα σύνολο από διαδικτυακά εργαλεία (APIs) που παρέχονται από το NCBI, τα οποία επιτρέπουν στους χρήστες να αλληλεπιδρούν προγραμματιστικά με τις βάσεις δεδομένων του NCBI, όπως το **PubMed**, το **Gene**, το **Nucleotide**, το **Protein** κ.ά. Οι E-utilities δίνουν τη δυνατότητα να εκτελούνται αναζητήσεις, να αντλούνται δεδομένα και να διαχειρίζονται σύνολα δεδομένων χωρίς να απαιτείται χρήση του γραφικού περιβάλλοντος της ιστοσελίδας.

Οι **E-utilities** βασίζονται σε HTTP αιτήσεις και επιστρέφουν δεδομένα σε μορφή XML, JSON ή απλού κειμένου, ανάλογα με τη συγκεκριμένη μέθοδο. Οι κύριες λειτουργίες που παρέχουν είναι οι εξής:

- **Esearch**: Χρησιμοποιείται για την αναζήτηση εγγραφών σε μια βάση δεδομένων του NCBI με βάση έναν όρο αναζήτησης.
- **Efetch**: Επιστρέφει τα πλήρη δεδομένα για μία ή περισσότερες εγγραφές που ανακτήθηκαν από τη βάση δεδομένων χρησιμοποιώντας τα μοναδικά αναγνωριστικά τους (UIDs).
- **Esummary**: Παρέχει μια σύντομη περίληψη για μία ή περισσότερες εγγραφές που βρέθηκαν στη βάση δεδομένων.
- **Elink**: Συνδέει εγγραφές από διαφορετικές βάσεις δεδομένων του NCBI.
- **Epost**: Επιτρέπει στους χρήστες να αποθηκεύσουν μοναδικά αναγνωριστικά εγγραφών (UIDs) για μελλοντική χρήση.
- **Egquery**: Επιστρέφει τον αριθμό των αποτελεσμάτων για έναν όρο αναζήτησης σε όλες τις βάσεις δεδομένων του NCBI.
- **Espell**: Παρέχει ορθογραφικούς ελέγχους για τους όρους αναζήτησης.

Παραδείγματα χρήσης για 3 από τα σημαντικότερα E-utilities για το πώς χρησιμοποιούνται

Esearch: Αναζήτηση για το γονίδιο *BRCA1* στον άνθρωπο στη βάση Nucleotide

<https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=nucleotide&term=BRCA1+human>

Αυτός ο σύνδεσμος επιστρέφει έναν κατάλογο με τα UIDs των σχετικών αλληλουχιών που περιέχουν το γονίδιο *BRCA1* στον άνθρωπο στην nucleotide database σε XML μορφή όπως παρακάτω.

```
<eSearchResult>
<Count>6754</Count>    //συνολικά records
<RetMax>20</RetMax>
<RetStart>0</RetStart>
<IdList>
<Id>2288627445</Id>
<Id>1519244207</Id>
<Id>2797126124</Id>
.....
</TranslationStack>
```

```
<QueryTranslation>BRCA1[All Fields] AND ("Homo sapiens"[Organism] OR human[All Fields])</QueryTranslation>
</eSearchResult>
```

Efetch: Ανάκτηση της αλληλουχίας με βάση το UID

Αν έχουμε το UID 12345 από το προηγούμενο βήμα, μπορούμε να ανακτήσουμε τα πλήρη δεδομένα της αλληλουχίας χρησιμοποιώντας την εντολή efetch:

```
https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=nucleotide&id=12345&rettype=fasta&retmode=text
```

Αυτό επιστρέφει την αλληλουχία σε μορφή FASTA. Η efetch μπορεί να επιστρέψει διάφορες επιπλέον μορφές αρχείων (XML, BinaryASN.1, Fullrecord κλπ)

Esummary: Λήψη σύντομης περίληψης για το UID

```
https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi?db=nucleotide&id=12345
```

Αυτό επιστρέφει μια περίληψη των δεδομένων που σχετίζονται με το συγκεκριμένο UID, όπως περιγραφή, οργανισμός και μέγεθος αλληλουχίας.

Elink: Σύνδεση μιας εγγραφής από το Gene με το PubMed

Αν θέλουμε να βρούμε άρθρα που σχετίζονται με ένα συγκεκριμένο γονίδιο στο PubMed, μπορούμε να χρησιμοποιήσουμε την elink:

```
https://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=protein&db=nucleotide&id=148762980
```

Αυτό επιστρέφει 3 ιδιότητες από ένα proteinrecord και προσδιορίζουμε και ποιο θα είναι το codon, τα άλλα 2 είναι Genomic DNA

```
<eLinkResult>
<LinkSet>
<DbFrom>protein</DbFrom>
<IdList>
<Id>148762980</Id>
</IdList>
<LinkSetDb>
<DbTo>nucleotide</DbTo>
<LinkName>protein_nucleotide</LinkName>
<Link>
<Id>2194972797</Id>
</Link>
```

```

<Link>
<Id>2025113600</Id>
</Link>
<Link>
<Id>568815579</Id>
</Link>
</LinkSetDb>
<LinkSetDb>
<DbTo>nuccore</DbTo>
<LinkName>protein_nuccore_mrna</LinkName>
<Link>
<Id>2025113600</Id>
</Link>
</LinkSetDb>
</LinkSet>
</eLinkResult>

```

Υπάρχει επίσης η δυνατότητα χρήσης **API (Application Programming Interface)** προγραμματιστικών αιτημάτων ώστε να αυτοματοποιηθούν αυτές οι κλήσεις που είδαμε παραπάνω προς τα εργαλεία **NCBI E-utilities**

Ολοκληρωμένο Παράδειγμα API σε Python για αναζήτηση στην βάση RefSeq για κλήσεις με τα NCBI E-utilities για το γονίδιο GHR (Growth Hormone Receptor) στους **οστεϊχθύες**

```

1 import requests
2
3
4 efetch_url = "https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi" # 1. Αναζήτηση GHR γονιδίου στους οστεϊχθύες (Osteichthyes) με το efetch
5 params_search = { # Αναζήτηση στη βάση δεδομένων nucleotide (DNA/RNA)
6     "db": "nucleotide", # Παράμετρος term περιλαμβάνει τον όρο αναζήτησης GHR AND Osteichthyes[organism]
7     "term": "GHR AND Osteichthyes[organism]", # Μορφή αποτελεσμάτων σε JSON αρχείο
8     "retmode": "json", # Επιστρέφουμε τα πρώτα 5 αποτελέσματα (μέσω της παραμέτρου retmax)
9     "retmax": "5"
10 }
11
12
13 response_search = requests.get(efetch_url, params=params_search) # Στέλνουμε το αίτημα αναζήτησης
14 search_data = response_search.json()
15
16
17 id_list = search_data['searchresult']['idlist'] # Παίρνουμε τα ID των πρώτων αποτελεσμάτων με την efetch σε λίστα
18
19 if len(id_list) == 0: # Αρχίζουμε να διαβάζουμε την λίστα
20     print("Δεν βρέθηκαν αποτελέσματα για το γονίδιο GHR στους οστεϊχθύες.")
21 else: # Παίρνουμε το πρώτο ID εφόσον υπάρχει valid επιστροφή στην λίστα
22     first_id = id_list[0]
23     print(f"Περίελα το ID: {first_id} για το γονίδιο GHR στους οστεϊχθύες.")
24
25
26 efetch_url = "https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi" # 2. Χρησιμοποιούμε το efetch για να κατεβάσουμε την αλληλουχία
27 # σε απλό κείμενο (μορφή FASTA για το πρώτο αποτέλεσμα)
28 params_fetch = {
29     "db": "nucleotide", # Βάση δεδομένων nucleotide (για DNA/RNA αλληλουχίες)
30     "id": first_id, # ID από την αναζήτηση
31     "rettype": "fasta", # Τύπος δεδομένων: FASTA
32     "retmode": "text" # Μορφή δεδομένων: απλό κείμενο
33 }
34
35
36 response_fetch = requests.get(efetch_url, params=params_fetch) # Στέλνουμε το αίτημα για τη λήψη της αλληλουχίας
37 fasta_sequence = response_fetch.text
38
39
40 print(f"Αλληλουχία FASTA για το γονίδιο GHR στους οστεϊχθύες:\n{fasta_sequence}") # Εκτύπωση της αλληλουχίας σε μορφή FASTA
41

```

Εικόνα 13 Παράδειγμα API σε Python για αναζήτηση στην βάση RefSeq

Όταν εκτελείς το παραπάνω πρόγραμμα, θα επιστρέψει την αλληλουχία FASTA για το γονίδιο GHR στους οστεϊχθύες από το RefSeq.

Τα **E-utilities** είναι χρήσιμα για:

Αυτοματοποιημένη ανάκτηση δεδομένων από βάσεις του NCBI.

Ανάλυση μεγάλων συνόλων δεδομένων χωρίς την ανάγκη χειροκίνητης πρόσβασης μέσω του web interface.

Συνδυαστική χρήση δεδομένων από διαφορετικές βάσεις δεδομένων του NCBI (π.χ. αναζήτηση άρθρων που σχετίζονται με μια αλληλουχία από το PubMed).

Τα E-utilities χρησιμοποιούν διάφορες παραμέτρους που καθορίζουν τον τύπο αναζήτησης, τη βάση δεδομένων και το μοντέλο επιστροφής δεδομένων:

db Η βάση δεδομένων στην οποία θέλεις να κάνεις αναζήτηση (π.χ. nuccore για αλληλουχίες DNA), **term** Ο όρος αναζήτησης, **id** Το μοναδικό αναγνωριστικό (UID) μιας εγγραφής, **rettype** Ο τύπος δεδομένων που θέλεις να ανακτήσεις (π.χ. fasta, xml), **retmode** Η μορφή δεδομένων που θέλεις να επιστραφεί (π.χ. text, json, xml).

Advanced Search Options: Το Entrez υποστηρίζει σύνθετες αναζητήσεις με τη χρήση λογικών τελεστών (AND, OR, NOT), φίλτρων και συνδυασμών κριτηρίων.

Παράδειγμα: Σύνθετη αναζήτηση στο Entrez για γονιδιώματα οστεϊχθύων

`Clupea harengus[Organism] AND genome[Title] AND Osteichthyes[Taxonomy]`

`Clupea harengus[Organism]`: Περιορίζει την αναζήτηση στη Ρέγγα του Ατλαντικού.

`genome[Title]`: Αναζητά καταχωρήσεις που περιλαμβάνουν τον όρο "genome" στον τίτλο, περιορίζοντας την αναζήτηση σε γονιδιωματικές αλληλουχίες.

`Osteichthyes[Taxonomy]`: Προσθέτει ένα φίλτρο για την ταξινομική κατηγορία των Οστεϊχθύων, αν θέλεις να εστιάσεις στα ψάρια με οστέινο σκελετό.

Περιορισμοί στην χρήση τους για να αποφεύγεται η υπερφόρτωση των servers:

Μέγιστος αριθμός αιτήσεων ανά δευτερόλεπτο: 3 , Μέγιστος αριθμός UUIDs ανά αίτηση: 10.000, Οι χρήστες πρέπει να περιλαμβάνουν το δικό τους email στις αιτήσεις για λόγους ταυτοποίησης και υποστήριξης.

Τα δεδομένα που επιστρέφονται από τις E-utilities είναι συνήθως σε μορφή **XML** ή **FASTA** και μπορούν να διαμορφωθούν ανάλογα με τις ανάγκες του χρήστη.

Η χρήση των E-utilities είναι δωρεάν και προσφέρει **API πρόσβαση** σε όλα τα δεδομένα που είναι διαθέσιμα μέσω του NCBI.

FTP από το NCBI

Το **NCBI** προσφέρει πρόσβαση σε **μεγάλα σύνολα δεδομένων** μέσω του **FTP server** του, επιτρέποντας τη **μαζική λήψη αρχείων** όπως GenBank αλληλουχίες, RefSeq δεδομένα, γονιδιωματικές καταχωρήσεις κ.ά.

Αυτό είναι ιδιαίτερα χρήσιμο για ερευνητές που χρειάζονται **bulk downloads** για offline ανάλυση ή ανάπτυξη bioinformatics pipelines.

Σύνδεση στον FTP Server του NCBI, με ένα FTP client όπως το γραφικό πρόγραμμα FileZilla, ή με command line

Σύνδεση με command line: ftp <ftp.ncbi.nlm.nih.gov>

User : anonymous

Password: ένα **email** (δεν απαιτείται πιστοποίηση — είναι απλώς πρακτική καταγραφής).

Συμπέρασμα

- Η GenBank ενημερώνεται σε τακτά χρονικά διαστήματα και περιλαμβάνει δεδομένα από διεθνείς πηγές.
- Όλες οι αλληλουχίες στη βάση δεδομένων είναι ελεύθερα διαθέσιμες για δημόσια χρήση, και οι χρήστες καλούνται να καταθέτουν τις δικές τους ανακαλύψεις.
-

Σύγκριση με άλλες βάσεις δεδομένων

Η **GenBank** είναι μια από τις μεγαλύτερες βάσεις δεδομένων γενετικών αλληλουχιών, προσφέροντας ευρεία κάλυψη DNA και RNA από όλους τους οργανισμούς. Πλεονεκτεί στην ταχύτητα καταγραφής νέων δεδομένων και στη διεθνή διασύνδεση μέσω του INSDC. Ωστόσο, υστερεί σε επιμέλεια και συνέπεια, καθώς οι εγγραφές δεν επανεξετάζονται συστηματικά, σε αντίθεση με τη RefSeq. Δεν προσφέρει εξειδικευμένα φίλτρα και λειτουργική ανάλυση ncRNA, όπως η RNACentral, ούτε παρέχει βιολογικές πληροφορίες για πρωτεΐνες, όπως η UniProt. Συνεπώς, είναι χρήσιμη για πρωτογενή αναζήτηση και αρχειοθέτηση, αλλά όχι για στοχευμένη ανάλυση ή επιμελημένα δεδομένα.

5.5 RNACentral

Η RNACentral είναι μια δημόσια βάση δεδομένων που συγκεντρώνει και ενοποιεί πληροφορίες για μη κωδικοποιημένα RNA (ncRNA) από πολλαπλές αξιόπιστες πηγές. Στόχος της είναι να προσφέρει μια ολοκληρωμένη και ενοποιημένη αναπαράσταση των δεδομένων ncRNA που είναι διασκορπισμένα σε διαφορετικές βιολογικές βάσεις.

Πρόκειται για μια πλατφόρμα που παρέχει πρόσβαση σε διάφορους τύπους μη κωδικοποιημένων RNA από όλα τα είδη οργανισμών, όπως:

lncRNA (long non-coding RNA), miRNA (microRNA), tRNA (transfer RNA), rRNA (ribosomal RNA), snoRNA (small nucleolar RNA), και πολλούς άλλους τύπους ncRNA

Ο βασικός σκοπός της RNACentral είναι να αποτελέσει μια εύχρηστη, κεντρική πηγή για την αναζήτηση και ανάλυση των μη κωδικοποιημένων RNA. Αξιοποιείται ευρέως από τη βιολογική και ιατρική ερευνητική κοινότητα για:

- την κατανόηση των λειτουργιών των ncRNA σε θεμελιώδεις βιολογικές διεργασίες,
- τη μελέτη της γονιδιακής ρύθμισης,
- και την έρευνα σε σχέση με ανθρώπινες ασθένειες και βιοδείκτες.

Τρόποι Αναζήτησης στην Βάση RNACentral

Η αναζήτηση στην RNACentral μπορεί να γίνει με τους παρακάτω τρόπους:

Αναζήτηση με κείμενο: Οι χρήστες μπορούν να εισάγουν λέξεις-κλειδιά, ονόματα RNA ή γονιδίων, οργανισμούς και άλλα κριτήρια για να εντοπίσουν τις σχετικές εγγραφές ncRNA.

Αναζήτηση αλληλουχίας (Sequence Search): Υπάρχει δυνατότητα αναζήτησης βασισμένη σε αλληλουχίες. Ο χρήστης μπορεί να εισάγει μία αλληλουχία RNA για να βρει παρόμοιες αλληλουχίες ή να ταυτοποιήσει άγνωστες αλληλουχίες.

Φίλτρα: Υπάρχουν διάφορα φίλτρα για να περιοριστεί η αναζήτηση ανάλογα με τον τύπο RNA, τον οργανισμό ή την πηγή δεδομένων.

Είδη Δεδομένων

Η RNACentral συγκεντρώνει δεδομένα για πολλούς τύπους ncRNA, όπως:

- **miRNA** (microRNAs): Μικρά RNA που ρυθμίζουν την έκφραση γονιδίων.
- **lncRNA** (long non-coding RNA): Μακρά ncRNAs που συμμετέχουν σε ρυθμιστικές διεργασίες.
- **tRNA** (transfer RNA): Μεταφορικά RNA που συμμετέχουν στη σύνθεση πρωτεϊνών.
- **rRNA** (ribosomal RNA): Συγκροτούν τα ριβοσώματα, τα οποία είναι υπεύθυνα για τη μετάφραση του mRNA σε πρωτεΐνες.
- **snoRNA** (small nucleolar RNA): Εμπλέκονται στη βιογένεση και τροποποίηση άλλων RNA.

Τα δεδομένα περιλαμβάνουν:

- **Αλληλουχίες RNA**: Αλληλουχίες από διαφορετικά είδη ncRNA.
- **Δομικά δεδομένα**: Πληροφορίες για τη δευτεροταγή δομή των RNA.
- **Λειτουργικά δεδομένα**: Πληροφορίες για τη βιολογική λειτουργία των ncRNA.
- **Πηγές και συνδέσεις**: Κάθε εγγραφή περιλαμβάνει συνδέσμους σε άλλες βάσεις δεδομένων που παρέχουν περισσότερες πληροφορίες για τα RNA.

Τρόποι Πρόσβασης στα Δεδομένα

Οι χρήστες μπορούν να έχουν πρόσβαση στα δεδομένα της RNAcentral μέσω:

1. **Web Interface (https)**: Η κύρια πλατφόρμα αναζήτησης και προβολής δεδομένων ncRNA μέσω της ιστοσελίδας της RNAcentral. Παρέχει πρόσβαση σε αλληλουχίες και συνδεδεμένες πληροφορίες με φιλικά εργαλεία.
2. **FTP Access**: Για τη μαζική λήψη δεδομένων, η RNAcentral προσφέρει πρόσβαση μέσω FTP, επιτρέποντας στους χρήστες να κατεβάσουν ολόκληρα datasets ή μεγάλα αρχεία δεδομένων.
3. **API (Application Programming Interface)**: Παρέχεται πρόσβαση μέσω API για προγραμματισμένη και αυτοματοποιημένη λήψη δεδομένων. Οι χρήστες μπορούν να αναζητήσουν ή να εξαγάγουν δεδομένα μέσω κώδικα.
4. **Cross-references**: Συνδέσεις με άλλες βάσεις δεδομένων για πιο ολοκληρωμένη προβολή των δεδομένων.

Σύγκριση με άλλες βάσεις δεδομένων

Η **RNAcentral** είναι μια ενοποιητική βάση δεδομένων για μη κωδικοποιημένα RNA (ncRNA), συγκεντρώνοντας δεδομένα από περισσότερες από 40 εξειδικευμένες πηγές, όπως η miRBase, Ensembl και RefSeq. Πλεονεκτεί στην παροχή ολοκληρωμένης πληροφορίας για όλους τους τύπους ncRNA, με δυνατότητα φιλτραρίσματος, αναζήτησης με αλληλουχία και πλούσιες διασυνδέσεις με άλλες βάσεις. Ωστόσο, περιλαμβάνει και μη επιμελημένα δεδομένα, καθώς η ποιότητα εξαρτάται από τις πηγές. Σε αντίθεση με τη RefSeq ή τη UniProt, δεν πραγματοποιεί δική της επιμέλεια ή λειτουργική ανάλυση, αλλά λειτουργεί ως κόμβος ενοποίησης και πρόσβασης στην υπάρχουσα πληροφορία.

5.6 ZFIN

Η **ZFIN (Zebrafish Information Network)** είναι η επίσημη βάση δεδομένων που διατηρείται για το zebrafish (**Danio rerio**), ένα ευρέως χρησιμοποιούμενο μοντέλο οργανισμού στη βιοϊατρική έρευνα. Ο κύριος σκοπός της ZFIN είναι να υποστηρίξει την ερευνητική κοινότητα παρέχοντας ολοκληρωμένα δεδομένα για το zebrafish, που περιλαμβάνουν γενετικές, γονιδιωματικές, φαινοτυπικές και αναπτυξιακές πληροφορίες. Η ZFIN λειτουργεί ως κεντρικός κόμβος για την αποθήκευση και τη διανομή δεδομένων, προωθώντας τη συνεργασία και την προώθηση της επιστημονικής γνώσης για τη βιολογία του zebrafish και τη χρήση του στη μελέτη των ασθενειών.

Τρόποι Αναζήτησης στην Βάση ZFIN

Η ZFIN παρέχει διάφορες επιλογές αναζήτησης, επιτρέποντας στους χρήστες να ανακτήσουν δεδομένα με πολλούς τρόπους:

1. **Αναζήτηση Γενετικής Πληροφορίας:**
 - ο **Γονίδια και γονιδιώματα:** Οι χρήστες μπορούν να αναζητήσουν συγκεκριμένα γονίδια, μεταγραφές, πρωτεΐνες ή περιοχές του γονιδιώματος.
 - ο **Μουταγόνες παραλλαγές:** Αναζήτηση μεταλλάξεων και παραλλαγών στο zebrafish.
2. **Αναζήτηση Φαινοτύπων:**
 - ο Οι χρήστες μπορούν να αναζητήσουν φαινοτυπικές πληροφορίες που σχετίζονται με γονίδια ή μεταλλάξεις στο zebrafish.
3. **Αναζήτηση Εργασιών και Πηγών Δεδομένων:**
 - ο Αναζήτηση επιστημονικών δημοσιεύσεων και εργασιών που σχετίζονται με το zebrafish.
 - ο Αναζήτηση σε συνεργαζόμενες βάσεις δεδομένων, όπως το PubMed.
4. **Αναζήτηση με βάση τους αναπτυξιακούς δείκτες:**
 - ο Πρόσβαση σε δεδομένα σχετικά με την ανάπτυξη του zebrafish, συγκεκριμένες φάσεις ανάπτυξης και τα γονίδια που εκφράζονται σε αυτές.

Είδη Δεδομένων που Περιέχει η ZFIN

Η ZFIN περιέχει ένα μεγάλο εύρος δεδομένων που καλύπτουν διάφορες πτυχές της βιολογίας του zebrafish:

1. **Γενετικά δεδομένα:**
 - ο Γονίδια, αλληλούχες DNA/RNA, γονιδιακές μεταλλάξεις και παραλλαγές.
2. **Φαινοτυπικά δεδομένα:**
 - ο Περιγραφή φαινοτύπων που σχετίζονται με συγκεκριμένα γονίδια, μεταλλάξεις και πειράματα στο zebrafish.
3. **Δεδομένα ανάπτυξης:**

- Αναλυτικές πληροφορίες για τα αναπτυξιακά στάδια του zebrafish, τα γονίδια που εκφράζονται σε αυτά και οι μοριακές διεργασίες.
- 4. **Δεδομένα απεικόνισης:**
 - Εικόνες και διαγράμματα που παρουσιάζουν την ανάπτυξη του zebrafish και τους φαινοτύπους.
- 5. **Συνδέσεις με επιστημονικές δημοσιεύσεις:**
 - Αναφορές και πληροφορίες από σχετικές ερευνητικές εργασίες.

Τρόποι Πρόσβασης στα Δεδομένα

1. **Πρόσβαση μέσω Ιστού:**
 - Οι χρήστες μπορούν να αναζητήσουν και να αποκτήσουν δεδομένα από τον ιστότοπο της ZFIN (<https://zfin.org/>). Παρέχονται εργαλεία για την αναζήτηση γονιδίων, φαινοτύπων και άλλων σχετικών δεδομένων μέσω του ενσωματωμένου συστήματος αναζήτησης.
2. **Αναλυτικά εργαλεία:**
 - Η ZFIN προσφέρει μια σειρά εργαλείων για την ανάλυση των δεδομένων, συμπεριλαμβανομένων εργαλείων για την κατηγοριοποίηση και τη μελέτη γονιδιακών μεταλλάξεων και φαινοτύπων.
3. **API και FTP:**
 - Προσφέρει πρόσβαση σε δεδομένα μέσω FTP για μαζική λήψη δεδομένων, όπως σύνολα γονιδίων και αλληλουχιών.
 - Διαθέσιμα API για προγραμματιστές, επιτρέποντας την αυτοματοποιημένη εξαγωγή δεδομένων για βιοπληροφορικές εφαρμογές.

Σύγκριση με άλλες βάσεις δεδομένων

Η ZFIN είναι η κεντρική βιολογική βάση δεδομένων για το είδος *Danio rerio* (ζέβραφαράκι), και παρέχει επιμελημένες πληροφορίες για γονίδια, RNA, φαινότυπα, μεταλλάξεις και έκφραση. Πλεονεκτεί στην ακρίβεια και την εξειδίκευση, αποτελώντας βασικό εργαλείο για ερευνητές που μελετούν το zebrafish ως πρότυπο οργανισμό. Περιλαμβάνει λειτουργικά και αναπτυξιακά δεδομένα, με διασύνδεση με βάσεις όπως Ensembl, NCBI και RNACentral. Ωστόσο, περιορίζεται αποκλειστικά στο zebrafish και δεν υποστηρίζει μαζική αναζήτηση ή επεξεργασία δεδομένων πολλών ειδών, όπως άλλες γενικευμένες βάσεις (π.χ. RNACentral, RefSeq, GenBank).

Συμπέρασμα

- **Πολύτιμος Πόρος για τη Βιοϊατρική Έρευνα:** Το zebrafish είναι ένα σημαντικό μοντέλο οργανισμού για τη μελέτη της ανθρώπινης βιολογίας και των ασθενειών, καθιστώντας τη ZFIN πολύτιμο εργαλείο για τη βιοϊατρική έρευνα.
- **Πρόσβαση σε Πειραματικά Δεδομένα:** Η βάση παρέχει πρόσβαση σε δεδομένα που προκύπτουν από πειράματα στο zebrafish, όπως οι γονιδιακές μεταλλάξεις και οι επιπτώσεις τους στη βιολογία του οργανισμού.

- **Συνεχής Ενημέρωση:** Η ZFIN ενημερώνεται συνεχώς με νέα δεδομένα από δημοσιευμένες εργασίες και πειράματα, εξασφαλίζοντας ότι οι ερευνητές έχουν πρόσβαση στα πιο πρόσφατα αποτελέσματα.

5.7 InterPro

Η **InterPro** είναι μια δημόσια διαθέσιμη βάση δεδομένων που παρέχει πληροφορίες για πρωτεϊνικές οικογένειες, πρωτεϊνικές επικράτειες (domains) και λειτουργικά μοτίβα (signatures). Η InterPro συνδυάζει δεδομένα από πολλές διαφορετικές βάσεις δεδομένων πρωτεϊνικών αναλύσεων για να προσφέρει μια ολοκληρωμένη αναπαράσταση της λειτουργίας και της δομής των πρωτεϊνών.

Ο κύριος σκοπός της InterPro είναι να ενοποιήσει πληροφορίες από διάφορες βάσεις δεδομένων που περιγράφουν χαρακτηριστικά πρωτεϊνών, όπως μοτίβα, τομείς και περιοχές. Η βάση επιτρέπει στους χρήστες να προβλέπουν τις λειτουργίες των πρωτεϊνών, να ταξινομούν πρωτεΐνες σε οικογένειες και να κατανοούν τη δομή και τη λειτουργία των πρωτεϊνών.

Τρόποι Αναζήτησης στην Βάση InterPro

Η αναζήτηση στην InterPro μπορεί να γίνει με διάφορους τρόπους:

1. **Αναζήτηση βάσει πρωτεΐνης:**
 - Μπορείτε να εισαγάγετε την αλληλουχία της πρωτεΐνης για να λάβετε πληροφορίες σχετικά με τα λειτουργικά μοτίβα, τους τομείς και τις οικογένειες πρωτεϊνών που ανήκει.
 2. **Αναζήτηση με όνομα ή κωδικό πρωτεΐνης:**
 - Η InterPro επιτρέπει την αναζήτηση με βάση το όνομα ή τον κωδικό της πρωτεΐνης, συμπεριλαμβανομένων των accession numbers από άλλες βάσεις δεδομένων.
 3. **BLAST και InterProScan:**
 - Οι χρήστες μπορούν να χρησιμοποιήσουν το **InterProScan**, ένα εργαλείο που επιτρέπει την αναζήτηση μοτίβων και τομέων πρωτεϊνών με βάση την αλληλουχία τους.
 - Το **BLAST** μπορεί επίσης να χρησιμοποιηθεί για να βρεθούν πρωτεϊνικές αλληλουχίες που μοιάζουν με τη δεδομένη αλληλουχία.
-

Είδη Δεδομένων που Περιέχει και Τρόποι Πρόσβασης σε Αυτά

Η InterPro περιέχει πληροφορίες που ενοποιούνται από διάφορες βάσεις δεδομένων όπως:

- **Πρωτεϊνικές Οικογένειες:** Δεδομένα που ταξινομούν τις πρωτεΐνες σε οικογένειες βασισμένες σε ομοιότητες αλληλουχίας.
- **Λειτουργικά Μοτίβα και Τομείς:** Πληροφορίες για συγκεκριμένα δομικά και λειτουργικά χαρακτηριστικά των πρωτεϊνών, όπως τομείς που σχετίζονται με ενζυματική δραστηριότητα ή σύνδεση με μόρια.
- **Εξελικτικές Σχέσεις:** Δεδομένα που δείχνουν πώς οι πρωτεΐνες σχετίζονται εξελικτικά, κάτι που βοηθά στη μελέτη της ομολογίας και των εξελικτικών οικογενειών πρωτεϊνών.

Τρόποι Πρόσβασης:

1. **Web Interface:** Η πρόσβαση μπορεί να γίνει μέσω της ιστοσελίδας της InterPro, όπου οι χρήστες μπορούν να πραγματοποιήσουν αναζητήσεις και να κατεβάσουν δεδομένα.
2. **InterProScan:** Εργαλείο ανάλυσης αλληλουχιών που μπορεί να χρησιμοποιηθεί για την πρόβλεψη πρωτεϊνικών λειτουργιών με βάση τα χαρακτηριστικά της πρωτεΐνης.
3. **REST API:** Η InterPro παρέχει **REST API**, που επιτρέπει την προγραμματισμένη πρόσβαση στα δεδομένα της.

Τρόποι Πρόσβασης στα Δεδομένα

1. **Web Interface:** Ο πιο κοινός τρόπος πρόσβασης στη βάση είναι μέσω της ιστοσελίδας της InterPro. Οι χρήστες μπορούν να αναζητήσουν πρωτεΐνες, να κατεβάσουν αλληλουχίες, και να εξερευνήσουν δεδομένα για πρωτεϊνικές οικογένειες και τομείς.
2. **InterProScan:** Προσφέρει την ανάλυση αλληλουχιών για την πρόβλεψη της λειτουργίας πρωτεϊνών. Οι χρήστες μπορούν να ανεβάσουν πρωτεϊνικές αλληλουχίες και να λάβουν πληροφορίες για τομείς και λειτουργίες.
3. **FTP Access:** Η InterPro προσφέρει πρόσβαση στα δεδομένα της μέσω του FTP server της, που επιτρέπει την μαζική λήψη δεδομένων.
4. **REST API:** Οι χρήστες που επιθυμούν να έχουν πρόσβαση στα δεδομένα της InterPro μέσω προγραμματισμού μπορούν να χρησιμοποιήσουν το REST API της, που επιτρέπει την αυτόματη εξαγωγή και ανάλυση δεδομένων.

Σύγκριση με άλλες βάσεις δεδομένων

Η **InterPro** είναι μια ενοποιημένη βάση δεδομένων που παρέχει πληροφορίες για **πρωτεϊνικές οικογένειες, περιοχές, τομείς και λειτουργικά μοτίβα**, ενοποιώντας δεδομένα από πολλαπλές εξειδικευμένες βάσεις όπως Pfam, SMART, PROSITE και TIGRFAMs. Πλεονεκτητά στην **ανάλυση και κατηγοριοποίηση πρωτεϊνών**, προσφέροντας λειτουργικές περιγραφές και προβλέψεις τομέων για αλληλουχίες από οποιονδήποτε οργανισμό. Υποστηρίζει αναζητήσεις με αλληλουχία, προσφέρει API και διαλειτουργεί με UniProt, Ensembl και άλλα συστήματα. Ωστόσο, η InterPro **επικεντρώνεται αποκλειστικά σε πρωτεΐνες** και δεν καλύπτει ncRNA ή γονιδιακές πληροφορίες όπως η RNAcentral, η Ensembl ή η GenBank.

Συνολική Χρήση της InterPro

Η InterPro είναι μια σημαντική πηγή για τη μελέτη των πρωτεϊνών, ενοποιώντας πληροφορίες από πολλές διαφορετικές βάσεις δεδομένων. Είναι ιδιαίτερα χρήσιμη για την κατανόηση των λειτουργικών χαρακτηριστικών των πρωτεϊνών, την ανάλυση αλληλουχιών και την κατηγοριοποίηση των πρωτεϊνών σε οικογένειες. Προσφέρει πληθώρα εργαλείων για αναζήτηση και ανάλυση δεδομένων, που καλύπτουν τις ανάγκες της βιολογικής έρευνας.

5.8 DDBJ DNA Data Bank of Japan

Η **DDBJ (DNA Data Bank of Japan)** ιδρύθηκε το 1986 στο National Institute of Genetics (NIG) στην Ιαπωνία και αποτελεί μια από τις τρεις κύριες βάσεις δεδομένων για την αποθήκευση και διάθεση γενετικών αλληλουχιών, μαζί με την **GenBank (NCBI)** και την **EMBL-EBI**. Αυτές οι τρεις βάσεις συνεργάζονται ως μέρος της **International Nucleotide Sequence Database Collaboration (INSDC)** για την ανταλλαγή δεδομένων γενετικών αλληλουχιών.

Ο βασικός σκοπός της DDBJ είναι η συλλογή, αποθήκευση και παγκόσμια διάθεση γενετικών αλληλουχιών DNA και RNA που υποβάλλονται από ερευνητές και επιστημονικά ιδρύματα. Η DDBJ εξυπηρετεί τις ανάγκες της βιολογικής έρευνας, προσφέροντας ελεύθερη πρόσβαση σε βιολογικά δεδομένα, με στόχο την υποστήριξη της επιστημονικής κοινότητας στην έρευνα της γενετικής, της βιοτεχνολογίας και της εξελικτικής βιολογίας.

Τρόποι Αναζήτησης στην Βάση DDBJ

DDBJ Search System

Η βασική μηχανή αναζήτησης της DDBJ επιτρέπει στους χρήστες να αναζητούν δεδομένα αλληλουχίας βάσει λέξεων-κλειδιών, accession numbers, οργανισμών και περιγραφών.

Οι χρήστες μπορούν να εισάγουν λέξεις-κλειδιά ή ονόματα οργανισμών για να αναζητήσουν συγκεκριμένες αλληλουχίες DNA ή RNA.

BLAST (Basic Local Alignment Search Tool)

Αναζήτηση ομοιότητας αλληλουχίας μέσω του BLAST. Χρησιμοποιείται για τη σύγκριση νέων αλληλουχιών με τις καταχωρήσεις της DDBJ και άλλων βάσεων δεδομένων.

Βοηθά στην εύρεση παρόμοιων αλληλουχιών και στην ταυτοποίηση άγνωστων αλληλουχιών DNA, RNA ή πρωτεϊνών.

ARSA (All-round Retrieval of Sequence and Annotation)

Ένα προηγμένο εργαλείο αναζήτησης που επιτρέπει την αναζήτηση και την ανάκτηση αλληλουχιών DNA και RNA με βάση χαρακτηριστικά όπως η αλληλουχία και οι σχολιασμοί.

Είδη Δεδομένων που Περιέχει η DDBJ

- **Αλληλουχίες Νουκλεϊκών Οξέων (DNA/RNA):** Η DDBJ αποθηκεύει αλληλουχίες DNA και RNA που υποβάλλονται από ερευνητές παγκοσμίως. Αυτές οι αλληλουχίες μπορεί να είναι κωδικοποιημένες ή μη-κωδικοποιημένες, περιλαμβάνοντας γονίδια, ρυθμιστικές περιοχές και μεταλλάξεις.
- **Πρωτεϊνικές Αλληλουχίες:** Περιλαμβάνει πρωτεϊνικές αλληλουχίες που προκύπτουν από τις αλληλουχίες DNA και RNA που καταχωρούνται στη βάση.
- **Βιολογικές Σημειώσεις (Annotations):** Περιέχει σχόλια και πληροφορίες σχετικά με τις αλληλουχίες, όπως γονιδιακές περιοχές, θέσεις ρύθμισης, και γενετικές μεταλλάξεις.
- **Πληροφορίες Οργανισμών:** Δεδομένα ταξινόμησης που περιγράφουν τους οργανισμούς από τους οποίους προέρχονται οι αλληλουχίες.

Τρόποι Πρόσβασης στα Δεδομένα

Πρόσβαση μέσω Ιστού (https):

Η πρόσβαση στα δεδομένα της DDBJ μπορεί να γίνει μέσω της επίσημης ιστοσελίδας, όπου οι χρήστες μπορούν να αναζητούν αλληλουχίες και να προβάλλουν δεδομένα χρησιμοποιώντας εργαλεία αναζήτησης.

Εργαλεία αναζήτησης: Οι χρήστες μπορούν να χρησιμοποιήσουν τον DDBJ Search System ή το ARSA για λεπτομερή αναζήτηση αλληλουχιών.

FTP (File Transfer Protocol):

Η DDBJ προσφέρει επίσης δυνατότητα μαζικής λήψης δεδομένων μέσω του **FTP server**. Χρήσιμο για ερευνητές που χρειάζονται μεγάλα σύνολα δεδομένων για ανάλυση ή αρχειοθέτηση.

Οι χρήστες μπορούν να κατεβάσουν ολόκληρα αρχεία δεδομένων αλληλουχιών για ανάλυση εκτός σύνδεσης.

API (Application Programming Interface):

Η DDBJ προσφέρει προγραμματισμένη πρόσβαση μέσω API, επιτρέποντας την αυτοματοποίηση της ανάκτησης δεδομένων. Εργαλεία όπως το efetch και το esearch μπορούν να χρησιμοποιηθούν για πρόσβαση σε συγκεκριμένες αλληλουχίες ή σύνολα δεδομένων.

BLAST Search:

Χρησιμοποιείται για τη σύγκριση των αλληλουχιών DNA ή RNA με άλλες καταχωρήσεις της DDBJ και διεθνών βάσεων δεδομένων, επιτρέποντας την ανακάλυψη ομοιοτήτων.

Συμπέρασμα

Ανοικτή πρόσβαση: Όλα τα δεδομένα στη DDBJ είναι ελεύθερα προσβάσιμα από το παγκόσμιο ερευνητικό κοινό, χωρίς περιορισμούς.

Τακτική Ενημέρωση: Τα δεδομένα της DDBJ ανανεώνονται τακτικά με νέες αλληλουχίες και δεδομένα που υποβάλλονται από ερευνητές.

Εκπαίδευση: Η DDBJ παρέχει τεχνική τεκμηρίωση και οδηγίες για τη χρήση των εργαλείων αναζήτησης και τη συλλογή δεδομένων.

Σύγκριση με άλλες βάσεις δεδομένων

Η DDBJ είναι η εθνική ιαπωνική βάση γενετικών δεδομένων και αποτελεί ένα από τα τρία μέλη του INSDC (μαζί με τις GenBank και EMBL-EBI), με στόχο τη συλλογή και παγκόσμια διάθεση νουκλεοτιδικών αλληλουχιών. Πλεονεκτεί στην ανοικτή υποβολή δεδομένων, στη διαρκή συγχρονισμένη ανταλλαγή με GenBank και EMBL, και στην υποστήριξη όλων των τύπων γενετικής πληροφορίας (DNA, RNA, ncRNA, μεταγραφές). Υποστηρίζει πρόσβαση μέσω FTP και API, καθώς και εργαλεία όπως το DFAST για σχολιασμό γονιδιωμάτων. Ωστόσο, όπως και η GenBank, η DDBJ δεν παρέχει επιμέλεια ή λειτουργική ερμηνεία, ούτε προσφέρει εξειδικευμένα φίλτρα για ncRNA ή πρωτεϊνικά δεδομένα, σε αντίθεση με βάσεις όπως η RefSeq, RNACentral ή InterPro.

5.9 BOLD Systems, ή Barcode of Life Data Systems

Το **BOLD Systems** (Barcode of Life Data Systems) είναι μια διαδικτυακή πλατφόρμα που έχει σχεδιαστεί για να υποστηρίζει τη συλλογή, αποθήκευση και ανάλυση δεδομένων DNA barcoding. Το DNA barcoding είναι μια μέθοδος για την ταυτοποίηση ειδών με βάση σύντομες τυποποιημένες περιοχές γονιδιωμάτων, οι οποίες λειτουργούν ως "μοριακό barcode" για κάθε είδος.

Ο κύριος σκοπός του BOLD είναι να προάγει τη συστηματική χρήση του DNA barcoding για την ταυτοποίηση και την καταγραφή της βιοποικιλότητας. Συλλέγει γενετικά δεδομένα από διάφορα είδη παγκοσμίως, παρέχοντας ένα εργαλείο για τη διαχείριση της βιοποικιλότητας, την επιστημονική έρευνα και την υποστήριξη της βιοασφάλειας. Η βάση αυτή χρησιμοποιείται ευρέως από επιστήμονες για την ανίχνευση νέων ειδών, την ταυτοποίηση οργανισμών και την μελέτη των οικοσυστημάτων.

Τρόποι Αναζήτησης στην Βάση BOLD Systems

Web Interface (Ιστοσελίδα)

Η κύρια πρόσβαση γίνεται μέσω της επίσημης πλατφόρμας: <https://www.boldsystems.org>

Περιλαμβάνει εργαλεία όπως:

Public Data Portal: Αναζήτηση ειδών, project, αλληλουχιών.

Taxonomy Browser: Πλοήγηση ανά taxon (βασίλειο, φύλο, γένος κ.λπ.).

BIN Database: Ομαδοποιήσεις barcodes με βάση γενετική απόσταση.

Specimen Page: Αναλυτικές πληροφορίες για κάθε δείγμα (εικόνες, τοποθεσία, αλληλουχία, metadata)

Αναζήτηση με εργαλεία (Search Tools)

Η πλατφόρμα υποστηρίζει:

Basic Search: Αναζήτηση με είδος, project, ή locality.

Sequence ID ή Process ID Search: Αναζήτηση συγκεκριμένης εγγραφής.

BLAST ή BOLD Identification Engine (ID Engine): Υποβολή αλληλουχίας για εύρεση ταυτοποίησης είδους.

BOLD API (Application Programming Interface)

Διαθέσιμο για **προγραμματισμένη πρόσβαση** σε δεδομένα BOLD:

API τεκμηρίωση: <http://v4.boldsystems.org/index.php/resources/api>

Υποστηρίζει ερωτήματα για: BINs, Είδη, Αλληλουχίες, Project, Geolocation

Τα δεδομένα επιστρέφονται σε μορφή **JSON ή XML**, κατάλληλη για ανάλυση μέσω Python, R, κ.λπ.

Είδη Δεδομένων που Περιέχει

Το BOLD Systems περιέχει δεδομένα που σχετίζονται με τη γενετική αλληλουχία των DNA barcodes από οργανισμούς. Περιλαμβάνει τα εξής είδη δεδομένων:

DNA Barcodes: Σύντομες τυποποιημένες περιοχές γονιδιωμάτων που χρησιμοποιούνται για την ταυτοποίηση ειδών.

Taxonomic Information: Ταξινομικά δεδομένα για τα είδη που έχουν καταγραφεί, συμπεριλαμβανομένων των επιστημονικών ονομάτων και των γεωγραφικών κατανομών τους.

Εικονογραφημένα Δεδομένα: Πληροφορίες όπως εικόνες των οργανισμών και δεδομένα ανατομίας.

Γεωγραφικές Συντεταγμένες: Δεδομένα σχετικά με τις τοποθεσίες από όπου συλλέχθηκαν τα δείγματα DNA.

Συνδεδεμένες Πληροφορίες Δείγματος: Περιγραφή των δειγμάτων που υποβλήθηκαν για ανάλυση DNA.

Τρόποι Πρόσβασης στα Δεδομένα

Η BOLD Systems προσφέρει πολλαπλές δυνατότητες πρόσβασης στα δεδομένα της, τόσο για απλούς χρήστες όσο και για ερευνητές και προγραμματιστές.

1. Πρόσβαση μέσω του Web Interface (<https://www.boldsystems.org>)

Το κύριο περιβάλλον χρήσης της BOLD είναι η ιστοσελίδα της, όπου οι χρήστες μπορούν να αναζητούν και να προβάλλουν δεδομένα με βάση:

- είδη (species),
- barcodes (γονιδιακές αλληλουχίες),
- γεωγραφική περιοχή,
- αριθμούς BIN (Barcode Index Numbers).

Παρέχεται επίσης το **Public Data Portal**, μέσω του οποίου διατίθενται ελεύθερα προσβάσιμα δεδομένα για την επιστημονική κοινότητα.

Η **BIN Database** προσφέρει πληροφορίες για γενετικά clusters που ενδέχεται να αντιστοιχούν σε διαφορετικά είδη, ακόμη και αν αυτά δεν έχουν ακόμη επίσημη ονομασία.

2. Πρόσβαση μέσω API (Application Programming Interface)

Για προχωρημένους χρήστες και προγραμματιστές, η BOLD παρέχει REST API, το οποίο επιτρέπει:

- αυτοματοποιημένη αναζήτηση και λήψη δεδομένων,
- ενσωμάτωση με εργαλεία βιοπληροφορικής,
- δημιουργία δικών τους εφαρμογών ή μελετών μεγάλης κλίμακας.

Η API τεκμηρίωση είναι διαθέσιμη στη διεύθυνση:
<http://v4.boldsystems.org/index.php/resources/api>

3. Εξαγωγή Δεδομένων (Data Export)

Οι χρήστες μπορούν να εξαγάγουν δεδομένα για offline ανάλυση σε διάφορες μορφές:

- **CSV:** για μεταδεδομένα και πίνακες πληροφοριών
- **FASTA:** για γονιδιακές αλληλουχίες
Η εξαγωγή μπορεί να γίνει από συγκεκριμένα είδη, projects ή γεωγραφικές περιοχές μέσω της ιστοσελίδας.

Σύγκριση με άλλες βάσεις δεδομένων

Η **BOLD** είναι μια εξειδικευμένη βάση δεδομένων για την **ταυτοποίηση ειδών μέσω DNA barcoding**, επικεντρωμένη κυρίως σε **μικρά τμήματα γονιδίων**, όπως το **COI** στα ζώα. Πλεονεκτεί στην **ταξινομική ακρίβεια**, προσφέροντας γενετικές αλληλουχίες συνδεδεμένες με επαληθευμένες αναγνωρίσεις ειδών, γεωγραφική κατανομή, φωτογραφίες δειγμάτων και metadata. Χρησιμοποιείται ευρέως στη βιοποικιλότητα, την οικολογία και την περιβαλλοντική διαχείριση. Ωστόσο, η **BOLD δεν παρέχει λειτουργικά δεδομένα για γονίδια ή RNA**, ούτε καλύπτει πλήρως το γονιδίωμα όπως βάσεις όπως η GenBank ή η Ensembl. Εστιάζει κυρίως σε **ταξινομητική εφαρμογή** και **δεν υποστηρίζει ανάλυση ncRNA ή πρωτεϊνών** όπως η RNACentral ή η InterPro.

Συμπέρασμα

Ενημερώσεις Δεδομένων: Τα δεδομένα στο BOLD ενημερώνονται συνεχώς με νέες αλληλουχίες που υποβάλλονται από ερευνητές παγκοσμίως, γεγονός που καθιστά τη βάση ζωτικής σημασίας για την ταυτοποίηση ειδών.

Πρόσβαση και Εργαλεία Ανάλυσης: Το BOLD παρέχει επίσης ενσωματωμένα εργαλεία ανάλυσης, όπως τη δυνατότητα συγκρίσεων μεταξύ DNA barcodes και γεωγραφικής ανάλυσης της βιοποικιλότητας.

Συνεισφορά στην Επιστήμη: Η βάση δεδομένων BOLD χρησιμοποιείται ευρέως για τη μελέτη της βιοποικιλότητας, την προστασία των ειδών και την εφαρμογή τεχνικών βιοασφάλειας.

	RNAcentral	Ensembl	RefSeq (NCBI)	UniProt	GenBank	BOLD	Entrez (NCBI)	ZFIN	InterPro	DDBJ
Περιεχόμενο/Εξειδίκευση	ncRNA δεδομένα από πολλαπλές πηγές	Ολοκληρωμένα γονιδιώματα, annotations	Επιμελημένα γονίδια και RNA	Πρωτεϊνικές αλληλουχίες και domains	Πρωτογενείς νουκλεοτιδικές αλληλουχίες	DNA barcoding και ταυτοποίηση ειδών	Ενοποιημένη πλατφόρμα αναζήτησης NCBI	Zebrafish-specific (Danio rerio) βάση	Ανάλυση πρωτεϊνικών δομών και μοτίβων	Ιαπωνική βάση για DNA/RNA
Είδη Δεδομένων	miRNA, lncRNA, tRNA, snoRNA κ.ά.	Γονίδια, μεταγραφές, πρωτεΐνες	mRNA, ncRNA, γονίδια	Πρωτεΐνες, τομείς, annotations	DNA/RNA (raw/unverified)	COI, 16S, ITS	Όλα μέσω PubMed, Gene, κ.λπ.	Γονίδια, φαινότυπα, έκφραση	Domains, motifs, protein families	DNA, RNA (όπως GenBank)
Επιμέλεια Περιεχομένου	Όχι (εξαρτάται από πηγές)	Ναι (προβλέψεις/annotations)	Ναι (curated, σταθερά IDs)	Ναι (Swiss-Prot curated)	Όχι (υποβολές από χρήστες)	Μερική (ειδικές ταυτοποιήσεις)	Όχι (είναι εργαλείο πλοήγησης)	Ναι (επιστημονική επιμέλεια)	Ναι (συνεργαζόμενες βάσεις)	Όχι (ανοιχτή υποβολή)
Δυνατότητες Φιλτραρίσματος	Πλήρης	Περιορισμένη	Μέτρια	Ελάχιστη	Υπάρχει, χωρίς φιλτράρισμα	Όχι	Εξαρτάται από βάση	Περιορισμένη	Όχι	Υπάρχει, χωρίς ανάλυση
Παραμετροποιημένη Αναζήτηση	Όχι	Ναι	Ναι	Πλήρης	Όχι	Όχι	Όχι	Περιορισμένα	Πλήρης	Όχι
API Πρόσβαση	Ναι	Ναι	Ναι (Entrez/E-utilities)	Ναι	Ναι	Ναι	Ναι	Ναι	Ναι	Ναι
Output Formats	FASTA, JSON	FASTA, JSON, XML	FASTA, GenBank	FASTA, JSON, TSV	GenBank, FASTA	JSON, XML	Όλα	FASTA, JSON	XML, JSON, TSV	FASTA, GenBank
Είδη Οργανισμών	Όλα	Όλα	Όλα	Όλα	Όλα	Ζώα/Έντομα	Όλα	Μόνο zebrafish	Όλα	Όλα
Διασύνδεση με άλλες βάσεις	Εκτεταμένη (40+ βάσεις)	Με UniProt, RNAcentral, κ.ά.	Μέτρια	Υψηλή (cross-references)	Περιορισμένη	Χαμηλή	Υψηλή με όλο το NCBI	Με Ensembl, NCBI, RNAcentral	Με UniProt, Pfam, PROSITE	Συγχρονισμός με GenBank & EMBL

Συνολικός συγκριτικός πίνακας των επιλεγμένων βάσεων δεδομένων

6.0 Το γονίδιο ROCK1 στο *Sparus aurata* ως Μοντέλο Αναφοράς για Αναζήτηση, Ανάλυση και Εξαγωγή Δεδομένων από το NCBI

Το είδος ***Sparus aurata*** (τσιπούρα) αποτελεί σημαντικό μοντέλο στην υδατοκαλλιέργεια και τη βιοϊατρική έρευνα, ιδίως για μελέτες που αφορούν τη μυϊκή ανάπτυξη, την αναπαραγωγή και τις προσαρμογές στο θαλάσσιο περιβάλλον. Ένα από τα γονίδια που έχει προσελκύσει ιδιαίτερο ερευνητικό ενδιαφέρον είναι το **ROCK1 (Rho-associated coiled-coil containing protein kinase 1)**, το οποίο εμπλέκεται σε διαδικασίες όπως:

- Η κυτταρική κινητικότητα και μορφολογία,
- Η ρύθμιση του κυτταροσκελετού,
- Η κυτταρική διαφοροποίηση και απόπτωση,
- Καθώς και σε παθολογικούς μηχανισμούς, όπως η φλεγμονή και η καρκινογένεση.

Η αναζήτηση του γονιδίου **ROCK1 στο *Sparus aurata*** μέσω βιοπληροφορικών βάσεων δεδομένων, και ειδικά μέσω του **NCBI**, μπορεί να αποκαλύψει:

- Την **πλήρη ή μερική αλληλουχία του γονιδίου**,
- Πληροφορίες για το **mRNA, CDS (coding sequence) ή πρωτεϊνικά προϊόντα**,
- Ομόλογες αλληλουχίες σε άλλα είδη (**BLAST**),
- Τη **δομή του γονιδίου**, μέσω εγγραφών **Gene, Nucleotide** και **Protein**.

Η χρήση του ROCK1 και τα χαρακτηριστικά του χρησιμοποιούνται ως σημείο αναφοράς στην εργασία αυτή ώστε να αποφασίσουμε ποια εργαλεία (BLAST, Ensembl, NCBI API, FASTA, MAFFT κ.ά.) και API θα χρειαστούν, επίσης τα αποτελέσματα των αναζητήσεων στην NCBI θα χρησιμοποιηθούν για τον έλεγχο της ορθότητας των δικών μας αποτελεσμάτων.

Σύνδεση με το Θεωρητικό Υλικό της Εργασίας

Τα κεφάλαια της θεωρίας που υποστηρίζουν και ενισχύουν τη μελέτη του γονιδίων όπως το **ROCK1 στο *Sparus aurata*** είναι:

- **Κεφάλαιο 1** – Παρουσιάζει βασικές έννοιες κωδικοποιημένων και μη κωδικοποιημένων περιοχών και την αναγκαιότητα της στοχευμένης αναζήτησης ανά γονίδιο και οργανισμό.
- **Κεφάλαιο 2 και 3** – Περιγράφουν εργαλεία όπως BLAST, MAFFT και Clustal Omega, τα οποία εφαρμόζονται σε επίπεδο αλληλουχιών για την αναγνώριση, ευθυγράμμιση και συγκριτική ανάλυση του γονιδίου.
- **Κεφάλαιο 4 και 5** – Εξετάζουν τις δυνατότητες των βάσεων NCBI, Ensembl και RefSeq, οι οποίες χρησιμοποιούνται για την αναζήτηση και εξαγωγή των εγγραφών που σχετίζονται με γονίδια όπως το ROCK1.

- **Κεφάλαιο 6 και 7** – Εφαρμογή: η τελική web εφαρμογή επιτρέπει την εισαγωγή γονιδίων (π.χ. ROCK1) και του οργανισμού (*Sparus aurata*) για την άντληση όλων των σχετικών εγγραφών μέσω του NCBI API.

7.0 Σχεδίαση και Περιγραφή Ιστοσελίδας

Όπως αναδείχθηκε στην προηγούμενη ανάλυση των βάσεων δεδομένων και των εργαλείων που έχουν αναπτυχθεί γύρω από αυτές, το κεντρικό ζητούμενο στη σύγχρονη βιοπληροφορική και τεχνολογία της πληροφορίας είναι η **αποτελεσματική αναζήτηση, ανάκτηση και αξιοποίηση δεδομένων και γνώσης**.

Οι ίδιες οι βάσεις δεδομένων λειτουργούν ως **δομημένα αποθετήρια πληροφορίας**, όπου πρωτεύοντα ρόλο παίζουν η **ταχύτητα πρόσβασης**, η **επέκταση** και η **αποθηκευτική ικανότητα**. Ωστόσο, η ουσιαστική αξία τους αναδεικνύεται μέσω των **εργαλείων και εφαρμογών** που τις συνοδεύουν και επιτρέπουν στον τελικό χρήστη – ακόμα και χωρίς εξειδικευμένες γνώσεις – να αλληλεπιδρά ουσιαστικά με το περιεχόμενό τους.

Τα τελευταία χρόνια, η ανάπτυξη **νέων γλωσσών προγραμματισμού** και τεχνολογιών, σε συνδυασμό με την ενσωμάτωση της **Τεχνητής Νοημοσύνης**, έχει οδηγήσει στην εμφάνιση εφαρμογών που χαρακτηρίζονται από **ευφυΐα, προσαρμοστικότητα και διαλειτουργικότητα**, με ιδιαίτερη έμφαση στην **επεξεργασία πληροφορίας** και τη **διευκόλυνση της αναζήτησης σε μεγάλα σύνολα δεδομένων**.

Στο πλαίσιο της παρούσας ενότητας, αξιοποιώντας το θεωρητικό και τεχνικό υπόβαθρο που προηγήθηκε, θα προχωρήσουμε στο σχεδιασμό και την υλοποίηση **μιας διαδικτυακής εφαρμογής (web-based)**, η οποία θα επιτρέπει την **άντληση δεδομένων από πολλαπλές βάσεις του διαδικτύου**. Παράλληλα, θα ενσωματώνονται λειτουργίες που βασίζονται στα **διαθέσιμα εργαλεία και APIs** που προσφέρουν οι οργανισμοί που διαχειρίζονται τις βάσεις αυτές, αποδεικνύοντας τη δυνατότητα ενοποιημένης και αποτελεσματικής πρόσβασης στην επιστημονική πληροφορία.

Αρχικός Σχεδιασμός και Μελλοντικό Πλάνο

Αρχικά, η ιστοσελίδα σχεδιάστηκε ως ένα **παραδειγματικό περιβάλλον επίδειξης** της χρήσης των εργαλείων που παρουσιάστηκαν προηγουμένως. Ωστόσο, στην πορεία εξελίχθηκε σε ένα **λειτουργικό εργαλείο**, με δυνατότητα **άντλησης πραγματικών δεδομένων από τη βάση NCBI και όλες τις εσωτερικές βάσεις (πχ Nucleotide, Gene, Protein, RefSeq, Genome, BLAST, PubMed)**, μέσω προγραμματισμένης πρόσβασης (API).

Ο μελλοντικός στόχος είναι η **πλήρης ολοκλήρωση** του συστήματος με **υποστήριξη αναζητήσεων σε όλες τις κύριες βιολογικές βάσεις δεδομένων** (όπως Ensembl, UniProt, RNACentral, GenBank, BOLD, κ.ά.). Η αναζήτηση θα πραγματοποιείται μέσω **ενιαίας φόρμας εισαγωγής στοιχείων**, με την εφαρμογή να επιστρέφει αποτελέσματα από όλες τις σχετικές βάσεις που υποστηρίζονται, ενσωματώνοντας **λογική εναλλακτικής (fallback)** σε περίπτωση που κάποια πηγή δεν ανταποκρίνεται ή δεν παρέχει δεδομένα.

Απώτερος σκοπός είναι η δημιουργία ενός **κεντριοποιημένου εργαλείου ενοποιημένης αναζήτησης**, που να διευκολύνει την ανάκτηση επιστημονικών δεδομένων από ποικίλες πηγές, ανεξάρτητα από το τεχνικό υπόβαθρο του τελικού χρήστη.

Υποδομή και Προγραμματιστικά Εργαλεία

Η υλοποίηση της ιστοσελίδας βασίστηκε σε μια απλή αλλά λειτουργική υποδομή, αξιοποιώντας σύγχρονα και ελαφριά εργαλεία ανάπτυξης. Το **frontend** (παρουσίαση της διεπαφής χρήστη) υλοποιήθηκε με **απλή HTML**, ενώ το **backend** (διαχείριση λογικής εφαρμογής και διασύνδεση με εξωτερικές βάσεις) αναπτύχθηκε σε **Python**.

Για το backend επιλέχθηκε το **Flask**, ένα ευέλικτο και ελαφρύ web framework για Python, το οποίο λειτούργησε ως **το βασικό περιβάλλον εκτέλεσης της εφαρμογής**. Μέσω του Flask πραγματοποιείται η ενσωμάτωση όλων των διαθέσιμων **API από τη βάση δεδομένων NCBI**, ενώ παρέχεται και η δυνατότητα λειτουργίας της εφαρμογής ως **τοπικός web server**, προσβάσιμος ακόμη και εκτός τοπικού δικτύου, με χρήση κατάλληλων εργαλείων.

Αναλυτικά, τα εργαλεία που χρησιμοποιήθηκαν είναι τα εξής:

- **Notepad++**: Ελαφρύς HTML editor για την επεξεργασία των σελίδων του frontend.
- **Visual Studio**: Περιβάλλον ανάπτυξης (IDE) για τον προγραμματισμό σε Python και τη διαχείριση του έργου.
- **Flask (Python Web Framework)**: Το βασικό framework για την υλοποίηση του backend και της λειτουργικότητας του web server.
- **ngrok**: Εργαλείο που επιτρέπει την **προσωρινή δημοσιοποίηση τοπικής εφαρμογής (localhost)** στο διαδίκτυο μέσω ενός δημόσιου URL, χωρίς να απαιτείται στατική IP ή πολύπλοκες ρυθμίσεις router.

Η επιλογή των παραπάνω εργαλείων έγινε με γνώμονα την **απλότητα, τη φορητότητα και την ευκολία επέκτασης**, ώστε το σύστημα να μπορεί να εξελιχθεί μελλοντικά χωρίς πλήρη αναδόμηση.

Flask Project – Δομή και Περιεχόμενο

Η διαδικασία εγκατάστασης του **Flask** είναι σχετικά απλή και δημιουργεί στον τοπικό υπολογιστή ένα οργανωμένο **δέντρο φακέλων και αρχείων**, στο οποίο αποθηκεύονται ο πηγαίος κώδικας, τα templates, τα στατικά στοιχεία και οι σχετικές βιβλιοθήκες της εφαρμογής.

Το δικό μας **Flask project** υλοποιήθηκε με σκοπό τη δημιουργία μιας δυναμικής web εφαρμογής και περιλαμβάνει την παρακάτω βασική δομή:

Python Scripts

app1_All.py: Το βασικό αρχείο εκκίνησης της εφαρμογής.

Περιέχει όλο τον backend κώδικα σε Python, συμπεριλαμβανομένων των εξής:

- Των **routes** (διαδρομών), που καθορίζουν ποια λειτουργία εκτελείται και τι προβάλλεται σε κάθε σελίδα της εφαρμογής.
- Της σύνδεσης με εξωτερικά APIs (π.χ. NCBI, Ensembl, UniProt).
- Της εξαγωγής αποτελεσμάτων σε μορφή PDF, CSV ή ZIP.

HTML Templates

templates/index.html: Το κύριο αρχείο διεπαφής χρήστη (frontend).

Σχεδιάστηκε με καθαρή HTML και περιλαμβάνει:

- Φόρμα αναζήτησης για είσοδο του χρήστη.
- Checkboxes επιλογής βάσεων δεδομένων.
- Οπτικά μηνύματα (pop-ups) για την παρουσίαση αποτελεσμάτων ή σφαλμάτων.

Static Files

fonts/DejaVuSans.ttf: Γραμματοσειρά που χρησιμοποιείται για την **παραγωγή PDF αρχείων** με υποστήριξη Unicode (ελληνικοί χαρακτήρες κ.λπ.).

Ο φάκελος static/ μπορεί επίσης να επεκταθεί με:

- CSS αρχεία για μορφοποίηση,
- JavaScript scripts για δυναμικές λειτουργίες,
- εικόνες ή icons για το περιβάλλον χρήστη.

Αυτή η δομή επιτρέπει **σαφή διαχωρισμό frontend και backend**, εύκολη συντήρηση του κώδικα και επεκτασιμότητα σε περίπτωση μελλοντικών βελτιώσεων ή προσθήκης νέων λειτουργιών.

Βήματα Εκτέλεσης

Εγκατάσταση Python 3.13 (64-bit)

Έλεγχος έκδοσης Python → `python --version`

Εγκατάσταση Flask → `pip install flask`

Έλεγχος → `pip show flask`

Δημιουργία Project Directory

C:\Users\NekDav\flask_app\

| app1_All.py

→ το κύριο Python αρχείο με τον Flask κώδικα

|— templates\

→ φάκελος για HTML (Flask ψάχνει εκεί)

| index.html

Εκτέλεση του Flask App

`http://127.0.0.1:5000/`

Από Command Prompt `cd C:\Users\NekDav\flask_app`

τρέχουμε: `python app1.py`

Ανοίγω τον browser και πηγαίνω: `http://127.0.0.1:5000/`

Αν όλα είναι σωστά, θα δούμε την index.html.

Πρόσβαση την σελίδα

Για να **έχει κάποιος άλλος εντός η εκτός δικτύου** πρόσβαση στη σελίδα που τρέχει τοπικά στο `http://127.0.0.1`, πρέπει να αλλάξει το **host**, να έχει **δημόσια IP** ή να χρησιμοποιήσουμε κάποιο εργαλείο **port tunneling**. Εδώ θα χρησιμοποιήσουμε το Ngrok

Αλλαγή host στο Flask app

Από: `app.run(debug=True)` >>> `app.run(host="0.0.0.0", port=5000, debug=True)`

Αυτό κάνει το Flask να **ακούει σε όλες τις IP** του συστήματος όχι μόνο localhost).

Η σελίδα παίζει πλέον και στο <http://127.0.0.1:5000>

Για πρόσβαση στο ίδιο δίκτυο Wi-Fi

μπορεί να δούμε τη σελίδα με:

`http://[η-τοπική-σου-IP]:5000`

Την τοπική IP θα τη βρούμε με: Την εντολή `ipconfig`

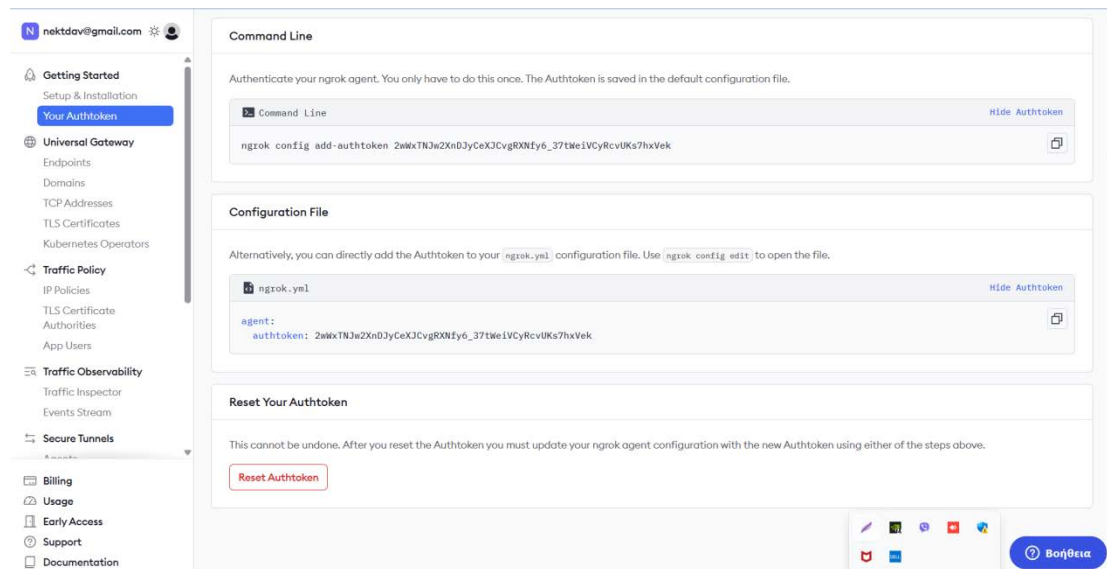
Για πρόσβαση εκτός δικτύου

Για πρόσβαση από εξωτερικά δίκτυα, χρειάζεσαι ένα tunneling εργαλείο όπως το Ngrok

Έγινε εγγραφή στο <https://dashboard.ngrok.com/get-started/setup/windows>

όπου και θα πάρουμε το authtoken και το οποίο θα το προσθέσουμε στο `ngrok.yml` [configuration file](#):

`ngrok config add-authtoken 2wWxTNJw2XnDjyCeXJCvgRXNfy6_37tWeiVCyRcvUKs7hxVek`



Εικόνα 14 Tunneling εργαλείο Ngrok

Εγκατάσταση Ngrok από το terminal: `pip install pyngrok`

Τρέχουμε τον Flask server : python app.py και ενεργοποιώ το <http://127.0.0.1:5000>

Ανοίγουμε άλλη κονσόλα και τρέχουμε το Ngrok : ngrok http http://127.0.0.1:5000/

Εχοντας το Authtoken και το ngrok.yml configuration file ήδη παραμετροποιημένα θα εμφανιστεί το **public link** που γινετε forward στο http://127.0.0.1:5000

Forwarding <https://955a-2a02-2149-88ab-2300-8d64-8c0a-9bf5-511b.ngrok-free.app> -> http://127.0.0.1:5000

Αντιγράφεις αυτό το **https** URL και το στέλνεις όπου θέλεις – ακόμα και κινητό ή εκτός δικτύου, παρακολουθούμε τις συνδεσεις.

Υλοποιημένος κώδικας

Κάνει αναζητήσεις σε βάσεις δεδομένων της NCBI και χρησιμοποιεί το **Entrez E-utilities API** για αναζητήσεις στις παρακατω βάσεις της NCBI

Gene (NCBI Gene), **Nucleotide** (NCBI Nucleotide database), **Protein** (Πρωτεϊνικές εγγραφές), **PubMed** (ιατρικά/βιολογικά άρθρα)

Κύρια Σημεία Κώδικα.

1. `get_all_gene_ids(query, organism, Αναζητά Gene IDs στη βάση gene του NCBI.`

2. `search_nucleotide_by_gene_id(gene_id, Επιστρέφει nucleotide IDs για ένα δοσμένο gene ID, με links προς το NCBI.`

3. `search_pubmed_literature(query, organism), Αναζητά PubMed άρθρα για το gene + organism (ή μόνο gene, με fallback). Χρησιμοποιεί esearch.fcgi και esummary.fcgi.`

4. `fetch_gene_details(gene_id, query, organism)`

Επιστρέφει:

- Gene summary (name, description, organism, chromosome)

- Nucleotide sequences (από `search_nucleotide_by_gene_id`)
- PubMed άρθρα μέσω `elink.fcgi`
- Ενισχυμένη αναζήτηση PubMed αν δεν υπάρχουν αρκετά άρθρα

5. `search_ncbi_protein(query, organism)`

Αναζητά **Protein IDs** από τη βάση `protein` του NCBI και επιστρέφει σχετικούς συνδέσμους.

6. Flask Route /

Ανταποκρίνεται σε GET και POST:

- Αν υποβληθεί φόρμα:

Ελέγχει αν έχει επιλεγθεί **τουλάχιστον ένα από τα checkboxes** (NCBI Gene, Protein), δεν προχωρά αν δεν επιλεγεί **ούτε NCBI ούτε Protein**

Καθαρίζει `gene_choices` και `selected_gene_id` πριν από κάθε νέα αναζήτηση

Αν NCBI επιλέχθηκε → αναζητά gene IDs → φέρνει λεπτομέρειες

Αν Protein επιλέχθηκε → αναζητά πρωτεΐνες

- Επιστρέφει το `index.html` με τα αποτελέσματα

8.0 Συζήτηση – Συμπεράσματα

Οι δυνατότητες που προσφέρουν οι σύγχρονες βιοπληροφορικές βάσεις δεδομένων είναι μεγάλες και εμπλουτίζονται καθημερινά με νέα γονίδια και δεδομένα. Η βασική πρόκληση δεν έγκειται μόνο στην πρόσβαση στην πληροφορία, αλλά κυρίως στη διαχείριση και αξιοποίησή της σε συνθήκες **υπερφόρτωσης δεδομένων (Big Data)**. Για το λόγο αυτό, απαιτούνται εξειδικευμένα εργαλεία και τεχνολογίες που υποστηρίζουν **αποτελεσματική ανάκτηση, φιλτράρισμα και οπτικοποίηση της πληροφορίας**, διασφαλίζοντας ταυτόχρονα την ακρίβεια, τη συνάφεια και την ορθότητα των αποτελεσμάτων.

Τα εργαλεία πρόσβασης που προσφέρονται από τις διεθνείς βάσεις δεδομένων, όπως το NCBI, Ensembl, UniProt κ.λπ., καλύπτουν ένα ευρύ φάσμα αναγκών. Περιλαμβάνουν:

- **Web-based διεπαφές χρήστη (web interfaces)** για απλούς χρήστες,
- **FTP servers** για μαζική μεταφόρτωση δεδομένων,
- **RESTful APIs** και γλώσσες προγραμματισμού όπως η **Python**, για ανάπτυξη προσαρμοσμένων και αυτοματοποιημένων pipelines ανάλυσης.

Στο πλαίσιο της παρούσας εργασίας, αναπτύχθηκε δοκιμαστική εφαρμογή Flask για **προγραμματισμένη πρόσβαση στη βάση δεδομένων του NCBI μέσω των E-utilities**.

Η βασική ροή της εφαρμογής βασίστηκε στην αρχική αναζήτηση του **Gene ID**, χρησιμοποιώντας το endpoint `esearch`. Αφού εντοπιστεί ο μοναδικός αναγνωριστικός αριθμός (UID), αξιοποιούνται οι υπηρεσίες `efetch`, `esummary` και `elink` για την εξαγωγή σχετικών πληροφοριών από τις βάσεις **gene**, **protein**, **nucleotide** και **biosample**, μέσω της παραμέτρου `db`.

Η κάθε βάση απαιτεί διαφορετικά φίλτρα και παραμέτρους, ανάλογα με τη φύση των δεδομένων. Για παράδειγμα, για την ανάκτηση μόνο RefSeq εγγραφών, εφαρμόζονται φίλτρα όπως `AND refseq[filter]`. Τα αποτελέσματα μπορούν στη συνέχεια να αναλυθούν περαιτέρω ή να αποθηκευτούν σε μορφές όπως **FASTA**, **XML** ή **JSON**.

Ιδιαίτερη έμφαση δόθηκε και στη δευτερογενή αναζήτηση επιστημονικής βιβλιογραφίας μέσω του **PubMed Central**, με εφαρμογή fallback μηχανισμού: σε περιπτώσεις όπου η αρχική αναζήτηση επιστρέφει λιγότερες από πέντε εγγραφές, ενεργοποιείται αυτόματα μια δεύτερη, ευρύτερη αναζήτηση μόνο βάσει του ονόματος του γονιδίου, χωρίς τον οργανισμό ως περιοριστικό όρο.

Η αναζήτηση για δεδομένα **genes**, **genomes** και **literature** υλοποιείται απευθείας μέσα από API endpoints. Αντίθετα, για τα **proteins** απαιτείται ανεξάρτητη αναζήτηση με χρήση της κατάλληλης βάσης (π.χ. `db=protein`), καθώς πρόκειται για ξεχωριστή πληροφοριακή κατηγορία εντός του NCBI.

Κατά την εκτέλεση της εφαρμογής, ο χρήστης αλληλεπιδρά με δυναμική φόρμα και επιλέγει τι τύπους δεδομένων επιθυμεί (π.χ. gene, protein), ενώ τα αποτελέσματα επιστρέφονται σε διακριτές καρτέλες (tabs). Η παρουσίαση των δεδομένων σε διασυνδεδεμένη μορφή επιτρέπει την καλύτερη κατανόηση της συσχέτισης μεταξύ **γονιδιακής πληροφορίας, σχετικής βιβλιογραφίας και βιολογικών εφαρμογών.**

Συνολικά, το προτεινόμενο εργαλείο αποδεικνύει ότι η συνδυαστική χρήση βιοπληροφορικών βάσεων δεδομένων, σε συνδυασμό με σύγχρονες τεχνικές ανάκτησης πληροφορίας, μπορεί να διευκολύνει ουσιαστικά τη βιολογική ερμηνεία γονιδιωματικών περιοχών και να συμβάλει σε εξελικτικές, λειτουργικές και εφαρμοσμένες μελέτες στους οστεϊχθύες.

Ένα τέτοιο εξατομικευμένο web interface μπορεί να αξιοποιηθεί από τελικούς χρήστες σε ακαδημαϊκά ιδρύματα, όπως φοιτητές, μεταπτυχιακούς ή διδακτορικούς ερευνητές, οι οποίοι επιδιώκουν να εκτελούν στοχευμένες αναζητήσεις στο πλαίσιο προσωπικής ή επιστημονικής έρευνας. Κάθε αναζήτηση βασίζεται σε συγκεκριμένες παραμέτρους που ορίζονται από τον ίδιο τον χρήστη, ανάλογα με τις απαιτήσεις της μελέτης του, και εκτελείται επιλεκτικά σε προκαθορισμένες βάσεις δεδομένων, με στόχο την ανάκτηση ακριβών και ουσιαστικών αποτελεσμάτων.

Ένα από τα βασικά πλεονεκτήματα της εφαρμογής είναι η δυνατότητα τοπικής αποθήκευσης των αποτελεσμάτων στον υπολογιστή του χρήστη, προσφέροντας πλήρη offline πρόσβαση και δυνατότητα επεξεργασίας οποιαδήποτε στιγμή. Παράλληλα, παρέχεται η δυνατότητα εξαγωγής των δεδομένων σε διάφορες μορφές αρχείων (όπως CSV, TXT, JSON ή PDF), ανάλογα με τη χρήση και τις ανάγκες του τελικού χρήστη.

Ίσως το σημαντικότερο πλεονέκτημα ενός τέτοιου custom web interface είναι ότι, αφού όλα τα δεδομένα έχουν ληφθεί σε ηλεκτρονική μορφή, μπορούν να προωθηθούν αυτόματα για περαιτέρω επεξεργασία μέσω διασύνδεσης με εξωτερικά εργαλεία και εφαρμογές, όπως πλατφόρμες στατιστικής ανάλυσης, βιοπληροφορικής ή διαχείρισης δεδομένων.

Για παράδειγμα, τα αποτελέσματα που ελήφθησαν για το γονίδιο ROCK1 στον οργανισμό *Sparus aurata* μπορούν να οδηγηθούν αυτόματα, μέσω προγραμματισμένης διαδικασίας (π.χ. POST σε εξωτερικό API), προς κατάλληλες εφαρμογές επεξεργασίας. Συγκεκριμένα:

- Τα Protein IDs μπορούν να προωθηθούν για ανάλυση σε εργαλεία όπως Clustal Omega ή MAFFT, για εκτέλεση πολλαπλής στοίχισης (MSA) με άλλα ομόλογα του ROCK1.
- Τα Nucleotide IDs μπορούν να χρησιμοποιηθούν ως είσοδος στο BLASTn για εντοπισμό ορθολογικών γονιδίων σε άλλα είδη.
- Τα άρθρα και οι τίτλοι μελετών μπορούν να εξαχθούν και να αναλυθούν με εργαλεία Text Mining, για εξαγωγή λέξεων-κλειδιών, σχετικών βιολογικών εννοιών και συσχετίσεων.

Η αρχιτεκτονική του συστήματος επιτρέπει στον χρήστη να καθορίζει τόσο τη μορφή παρουσίασης όσο και τον τρόπο αποθήκευσης των αποτελεσμάτων μέσω ενός φιλικού και διαμορφώσιμου περιβάλλοντος χρήστη (custom interface). Επιπλέον, η διαχείριση των παραμέτρων αναζήτησης πραγματοποιείται μέσα από μία ενιαία και ενοποιημένη διεπαφή, η οποία επιτρέπει την υποβολή κοινών αιτημάτων σε πολλαπλές βάσεις δεδομένων ταυτόχρονα, ανεξαρτήτως της διαφορετικής δομής και πολυπλοκότητας των επιμέρους API.

Η υλοποίηση βασίζεται στη χρήση εργαλείων ανοικτού κώδικα, όπως το Flask για την κατασκευή του backend, και την αξιοποίηση RESTful API endpoints για την ανάκτηση δεδομένων από εξωτερικές βάσεις. Ο web server μπορεί να φιλοξενείται τοπικά, ενώ μέσω υπηρεσιών όπως το Ngrok καθίσταται δυνατή η ασφαλής απομακρυσμένη πρόσβαση στην εφαρμογή, διευκολύνοντας τη συνεργασία μεταξύ χρηστών ή ερευνητικών ομάδων.

Με τον τρόπο αυτό, καθίσταται δυνατή η ανάπτυξη και χρήση ενός επεκτάσιμου, λειτουργικά ισχυρού και τεχνικά προσβάσιμου εργαλείου αναζήτησης, χωρίς την απαίτηση προηγμένων γνώσεων προγραμματισμού. Η εφαρμογή αξιοποιεί πλήρως τις

δυνατότητες που προσφέρουν τα διαθέσιμα API και οι ανοικτές τεχνολογίες, ενισχύοντας την ανεξαρτησία του ερευνητή και διευρύνοντας την πρόσβαση σε επιστημονική πληροφορία υψηλής αξίας.

9.0 Τεχνολογικά Εργαλεία για την υλοποίηση της εργασίας

Κατά την ανάπτυξη και τεκμηρίωση της παρούσας μεταπτυχιακής διπλωματικής εργασίας αξιοποιήθηκε ένας συνδυασμός τεχνολογικών εργαλείων, τα οποία υποστήριξαν τη διαδικασία συγγραφής, προγραμματισμού, ανάλυσης και ορθής επιστημονικής παρουσίασης των αποτελεσμάτων. Τα εργαλεία αυτά διακρίνονται σε τρεις βασικές κατηγορίες:

1. Εργαλεία Υποβοηθούμενης Συγγραφής και Αναδιατύπωσης

Grammarly / DeepL Write: Χρησιμοποιήθηκαν σε επιλεγμένα σημεία για έλεγχο γραμματικής, συντακτικής συνοχής και σαφήνειας στο επιστημονικό περιεχόμενο (όπου αυτό ήταν στα αγγλικά).

2. Εργαλεία Προγραμματισμού και Ανάλυσης API

Python 3.13: Χρησιμοποιήθηκε για την ανάπτυξη web εφαρμογής (Flask) και την εκτέλεση αιτημάτων μέσω REST APIs.

Jupyter Notebooks: Για τη σταδιακή καταγραφή και δοκιμή βιοπληροφορικών ερωτημάτων σε κώδικα Python. (εκτελείς κομμάτι-κομμάτι, και να βλέπεις άμεσα τα αποτελέσματα στην ίδια σελίδα)

Notepad++: Κύριο περιβάλλον ανάπτυξης της εφαρμογής και επεξεργασίας των Python scripts.

3. Πηγές Δεδομένων και Βάσεις Βιοπληροφορικής

NCBI E-Utilities: Χρήση των endpoints `esearch`, `esummary`, `efetch`, `elink` για αναζήτηση και ανάκτηση γονιδιακών και βιβλιογραφικών δεδομένων.

Ensembl REST API: Ανάκτηση annotations, mapping IDs και δεδομένων γονιδιώματος από οργανισμούς οστεϊχθύνων.

UniProt REST API: Απόκτηση λειτουργικών πληροφοριών πρωτεϊνών και μεταδεδομένων για RefSeq και Ensembl IDs.

RNAcentral API: Πρόσβαση σε δεδομένα μη-κωδικοποιημένων RNA (ncRNA), όπως lncRNA και miRNA.

10.0 Βιβλιογραφία

Εργαλεία ευθυγράμμισης/στοίχισης και αναζήτησης ομόλογων ακολουθιών

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), 403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)

Altschul, S. F., Madden, T. L., Schäffer, A. A., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17), 3389–3402. <https://doi.org/10.1093/nar/25.17.3389>

Sievers, F., & Higgins, D. G. (2014). Clustal Omega for making accurate alignments of many protein sequences. *Molecular Cell*, 56(1), 55–67. <https://doi.org/10.1016/j.molcel.2014.03.012>

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4), 772–780. <https://doi.org/10.1093/molbev/mst010>

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5), 1792–1797. <https://doi.org/10.1093/nar/gkh340>

Βάσεις Δεδομένων Βιοπληροφορικής

The UniProt Consortium. (2023). UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1), D523–D531. <https://doi.org/10.1093/nar/gkad195>

Cunningham, F., et al. (2022). Ensembl 2022. *Nucleic Acids Research*, 50(D1), D988–D995. <https://doi.org/10.1093/nar/gkab1049>

O'Leary, N. A., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status and future directions. *Nucleic Acids Research*, 44(D1), D733–D745. <https://doi.org/10.1093/nar/gkv1189>

Benson, D. A., et al. (2013). GenBank. *Nucleic Acids Research*, 41(D1), D36–D42. <https://doi.org/10.1093/nar/gks1195>

Mistry, J., et al. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1), D412–D419. <https://doi.org/10.1093/nar/gkaa913>

Blum, M., et al. (2021). The InterPro protein families and domains database: 20 years on. *Nucleic Acids Research*, 49(D1), D344–D354. <https://doi.org/10.1093/nar/gkaa977>

The RNAcentral Consortium. (2019). RNAcentral: a hub of information for non-coding RNA sequences. *Nucleic Acids Research*, 47(D1), D1250–D1251. <https://doi.org/10.1093/nar/gky1034>

Howe, D. G., et al. (2021). ZFIN, the Zebrafish Model Organism Database: Updates and new directions. *Nucleic Acids Research*, 49(D1), D822–D829. <https://doi.org/10.1093/nar/gkaa1035>

API και Υποστήριξη Ανάλυσης με Python

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace. <https://docs.python.org/3/>

Reitz, K. (2023). *Requests: HTTP for Humans*. <https://docs.python-requests.org/>

Richardson, L., & Ruby, S. (2007). *RESTful Web Services*. O'Reilly Media. <https://restfulapi.net/>

Cock, P. J. A., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11), 1422–1423. <https://doi.org/10.1093/bioinformatics/btp163>

WordPress Foundation. (2023). *WordPress Developer Resources*. <https://developer.wordpress.org/>

ΠΑΡΑΡΤΗΜΑ

Αναζήτηση τις Βάσεις : UniProt, NCBI/RefSeq/GenBank, Ensembl, RNAcentral, ZFIN, BOLD

Πλαίσιο 1: Φόρμα Εισαγωγής

Επιλέξτε Gene ID: Επιλέξτε ένα Καθαρισμός Gene ID

Όνομα γονιδίου:
ROCK1

Οργανισμός:
Sparus aurata

Accession number ή gene ID: "πχ Gene ID: 115572551 για ROCK1, Sparus aurata"

Επιλογή βάσεων αναζήτησης:

☐ Protein (NCBI)

☐ UniProt

☐ NCBI

☐ Ensembl

☐ RNAcentral

☐ ZFIN

☐ BOLD

☐ Εμφάνιση αλληλουχίας DNA/RNA/Protein

☐ Προβολή σε Genome Browser (Ensembl/UCSC)

☐ Προβολή λειτουργικών σχολίων (GO Terms)

☐ Σύνδεσμοι για PubMed / NCBI Gene

Αναζήτηση

Πλαίσιο 2: Αποτελέσματα

Αποτελέσματα:

Protein

- Protein ID: 1724710638
- Protein ID: 1724710636

NCBI

- Genomic ή Nucleotide Sequences από ESearch:
- Nucleotide 1714222218
- Nucleotide 1724710637
- Nucleotide 1724710635
- Gene ID: 115572551
- Όνομα: rock1
- Περιγραφή: Rho-associated, coiled-coil containing protein kinase 1
- Οργανισμός: Sparus aurata
- Χρωμόσωμα: 21
- Σχετικό Άρθρο:
- TreeGrafter: phylogenetic tree-based annotation of proteins with Gene Ontology terms and other annotations.
- Manual GO annotation of predictive protein signatures: the InterPro approach to GO curation.
- Ενίσχυση με failback αναζήτηση
- Extracellular matrix stiffness in endometrial cancer: driving progression and modulating treatment sensitivity via the ROCK1-YAP1 axis.
- Thymosin Beta-4 Modulates Cardiac Remodeling by Regulating ROCK1 Expression in Adult Mammals.
- Secretory Profile Analysis of Human Granulosa Cell Line Following Gonadotropin Stimulation.
- Sexual dimorphism of lung immune-regulatory units impairs pulmonary fibrosis.
- Multipotent Mesenchymal Stem Cell Therapy for Vascular Dementia.
- Design, Synthesis, and Evaluation of Novel ROCK Inhibitors for Glaucoma Treatment: Insights into In Vitro and In Vivo Efficacy and Safety.

Εικόνα 15 GUI της εφαρμογής της πτυχιακής

Αναζήτηση τις Βάσεις : UniProt, NCBI/RefSeq/GenBank, Ensembl, RNAcentral, ZFIN, BOLD

Πλαίσιο 1: Φόρμα Εισαγωγής

Επιλέξτε Gene ID: Επιλέξτε ένα Καθαρισμός Gene ID (ή αναζήτηση)

Όνομα γονιδίου:
rock1

Οργανισμός:
Sparus

Accession number ή gene ID: "πχ Gene ID: 115572551 για ROCK1, Sparus aurata"

Επιλογή βάσεων αναζήτησης: (Protein (NCBI) & NCBI μόνο ενεργό)

☒ Protein (NCBI)

☐ UniProt

☒ NCBI

☐ Ensembl

☐ RNAcentral

☐ ZFIN

☐ BOLD

☐ Εμφάνιση αλληλουχίας DNA/RNA/Protein

☐ Προβολή σε Genome Browser (Ensembl/UCSC)

☐ Προβολή λειτουργικών σχολίων (GO Terms)

☐ Σύνδεσμοι για PubMed / NCBI Gene

Αναζήτηση

Πλαίσιο 2: Αποτελέσματα

Αποτελέσματα:

Protein

- Protein ID: 295353498
- Protein ID: 3954981
- Protein ID: 24429815
- Protein ID: 24429813
- Protein ID: 24429811
- Protein ID: 62184959
- Protein ID: 62981792

NCBI

- Genomic ή Nucleotide Sequences από ESearch:
- Nucleotide 295353498
- Nucleotide 18959499
- Nucleotide 1196813841
- Nucleotide 52253119
- Gene ID: 60641
- Όνομα: scir6a
- Περιγραφή: SMY-like transcription factor 6a
- Οργανισμός: Drosophila
- Χρωμόσωμα: 12
- Σχετικό Άρθρο:
- Sirtuin 18 Defects in Human and Zebrafish Unravel Key Roles in Early Cartilage and Bone Development
- Juvenile ovarian to testis transition in zebrafish involves inhibition of oocytes.
- Genes that control of female sex determination in zebrafish
- Time point-based integrative analyses of deep-transcriptome identify four signal pathways in blastocyst reorganization of zebrafish lower jaw
- Regulation of zebrafish osteogenesis by esr2b and esr1b
- Fat-Duchenne syndrome coordinates cartilage differentiation and polarity during craniofacial development
- Endocrine and physiological effects of insulin and 5- α -dihydrotestosterone in zebrafish embryos
- Muscle contraction controls skeletal morphogenesis through regulation of chondrocyte convergent extension
- Delineated functional and temporal requirements for zebrafish Hnf1b during neural crest-derived craniofacial and peripheral neuron development
- Fgf1, a gene that is disrupted in cleft palate and has conserved function in zebrafish
- Dissection of the role of Pim in the development of zebrafish posterior rhombencephalic cartilages
- Notch signaling enhances bone regeneration in the zebrafish mandible
- A critical role of Foxo3a-estrogen regulatory T cells in maintaining immune homeostasis in zebrafish testis development
- Evolution of the osteoblast: osteoporosis in man and zebrafish
- Zebrafish sex determination and differentiation: involvement of F12-F1 genes
- FBLN2 is associated with Goldenhar syndrome and is essential for cranial neural crest cell development
- Evolution of vertebrate oil covers via shifts in an ancient Pou3f3 enhancer
- Baroreceptors control development using the same enhancer
- RTX transcription factors are essential for hearing in mice
- Visualisation of Gli activity in craniofacial tissues of hydroxyphenyl-ethoxy transporter transgenic zebrafish
- Control of cranial ectomesenchyme fate by Nr2f1 nuclear receptors

Εικόνα 16 GUI της εφαρμογής της πτυχιακής

79

Institutional Repository - Library & Information Centre - University of Thessaly
01/09/2026 12:30:58 EEST - 216.73.216.156



National Library of Medicine
National Center for Biotechnology Information

Log in

Search NCBI

rock1, sparus

×

Search

Results found in 6 databases

GENE

Was this helpful?

👍

👎

[ROCK1 orthologs from porgies and others](#)
Rho associated coiled-coil containing protein kinase 1
How are orthologs calculated?
[Genes with similar protein architectures](#)
How are similar genes calculated?

Literature

Bookshelf

0

MeSH

0

NLM Catalog

0

PubMed

0

PubMed Central

11

Genes

Gene

1

GEO DataSets

0

GEO Profiles

0

Proteins

Conserved Domains

0

Identical Protein Groups

0

Protein

2

Protein Family Models

0

Structure

0

Genomes

Assembly / Genome

NCBI Datasets

BioCollections

0

BioProject

0

BioSample

0

Nucleotide

3

Clinical

ClinicalTrials.gov

0

ClinVar

99

dbGaP

0

dbSNP

0

dbVar

0

PubChem

BioAssays

0

Compounds

0

Pathways

0

Substances

0

Εικόνα 17 NCBI αποτέλεσμα αναζήτησης rock1, sparus



National Library of Medicine
National Center for Biotechnology Information

Log in

Protein

Protein

rock1, sparus

×

Search

Create alert

Advanced

Help

Species

Animals (2)

Customize...

Source databases

RefSeq (2)

Customize...

Sequence length

Custom range...

Molecular weight

Custom range...

Release date

Custom range...

Revision date

Custom range...

[Clear all](#)
[Show additional filters](#)

Summary

Sort by Default order

Send to

Filters: [Manage Filters](#)

GENE

Was this helpful?

👍

👎

[ROCK1 orthologs from porgies and others](#)
Rho associated coiled-coil containing protein kinase 1
How are orthologs calculated?
[Genes with similar protein architectures](#)
How are similar genes calculated?

Items: 2

☐ **rho-associated protein kinase 1 isoform X2 [Sparus aurata]**
1. **1365 aa protein**
Accession: XP_030258595.1 GI: 1724710638
[BioProject](#) [Nucleotide](#) [Taxonomy](#)
[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

☐ **rho-associated protein kinase 1 isoform X1 [Sparus aurata]**
2. **1381 aa protein**
Accession: XP_030258594.1 GI: 1724710636
[BioProject](#) [Nucleotide](#) [Taxonomy](#)
[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

Analyze these sequences

Run BLAST

Align sequences with COBALT

Identify Conserved Domains with CD-Search

Find related data

Database: [Select](#)

Find items

Search details

rock1[All Fields] AND ("Sparus" [Organism] OR sparus[All Fields])

Search

See more...

Recent activity

Turn Off

Clear

rock1, sparus (2)

Protein

rho-associated protein kinase 1 isoform X2 [Sparus aurata]

Protein

Sparus aurata chromosome 21, ISPaAur1.1, whole genome shotgun sequence

Nucleotide

Mir21a, Mus musculus (2)

Nucleotide

Mus musculus strain C57BL/6J chromosome 11, GRChm39

Nucleotide

Εικόνα 18 NCBI αποτέλεσμα αναζήτησης στο Proteins

The screenshot shows the NCBI Genomes search results for the query 'rock1, sparus'. The interface includes a top navigation bar with the NIH logo and a search bar. The search results are displayed in a list format, showing three items:

- Item 1:** PREDICTED: *Sparus aurata* Rho associated coiled-coil containing protein kinase 1 (*rock1*). transcript variant X2. mRNA. 6,592 bp linear mRNA. Accession: XM_030402735.1. GI: 1724710637. Links: BioProject, Protein, Taxonomy, GenBank, FASTA, Graphics.
- Item 2:** PREDICTED: *Sparus aurata* Rho associated coiled-coil containing protein kinase 1 (*rock1*). transcript variant X1. mRNA. 5,340 bp linear mRNA. Accession: XM_030402734.1. GI: 1724710635. Links: BioProject, Protein, Taxonomy, GenBank, FASTA, Graphics.
- Item 3:** *Sparus aurata* chromosome 21. fSpaAur1.1. whole genome shotgun sequence. 35,160,130 bp linear DNA. Accession: NG_044297.1. GI: 1714222218. Links: Assembly, BioProject, BioSample, Protein, Taxonomy, GenBank, FASTA, Graphics.

The right sidebar contains filters and search options, including 'Find related data', 'Search details', and 'Recent activity'.

Εικόνα 19 NCBI αποτέλεσμα αναζήτησης στο Genoms

The screenshot shows the NCBI Literature/PubMed search results for the query 'rock1, sparus'. The interface includes a top navigation bar with the NIH logo and a search bar. The search results are displayed in a list format, showing five items:

- Item 1:** Exploring the Integrated Role of miRNAs and lncRNAs in Regulating the Transcriptional Response to Amino Acids and Insulin-like Growth Factor 1 in Gilthead Sea Bream (*Sparus aurata*) Myoblasts. Isabel Garcia-Perez, Bruno Oliveira Silva Duran, Mael Dal-Pai-Silva, Daniel Garcia de la serrana. Int J Mol Sci. 2024 Apr; 25(7): 3894. Published online 2024 Mar 31. doi: 10.3390/ijms25073894. PMID: PMC11011856. Article PDF-23M.
- Item 2:** An inside out journey: biogenesis, ultrastructure and proteomic characterisation of the ectoparasitic flatworm *Sparicotyle chrysophoni* extracellular vesicles. Enrique Riera-Ferrer, Hynok Mazanec, Ivona Mladineo, Peter Konik, M. Carla Piazon, Roman Kuchta, Oswaldo Palenzuela, Itziar Estensoro, Javier Sotillo, Ariadna Silja-Bobadilla. Parasit Vectors. 2024; 17: 175. Published online 2024 Apr 3. doi: 10.1186/s13071-024-06257-x. PMID: PMC10993521. Article PDF-3.7M.
- Item 3:** The spotted parrotfish genome provides insights into the evolution of a coral reef dietary specialist (Teleostei: Labridae: Scarini: *Cetoscarus ocellatus*). Yi-Kai Teo, Yulu Zhou, Kyle M. Ewart, Guo Cheng, Kazuhiko Kawasaki, Joseph D. DiBattista, Simon Y. W. Ho, Niallun Lo, Shaochua Fan. Ecol Evol. 2024 Mar; 14(3): e11140. Published online 2024 Mar 12. doi: 10.1002/ece3.11140. PMID: PMC10653694. Article PDF-3.3M.
- Item 4:** Using Vertebrate Stem and Progenitor Cells for Cellular Agriculture: State-of-the-Art Challenges and Future Perspectives. Teodora Knežić, Ljiljana Janjusević, Mila Džisalo, Supansa Yodmuang, Ivana Gadžanski. Biomolecules. 2022 May; 12(5): 699. Published online 2022 May 13. doi: 10.3390/biom12050699. PMID: PMC9138761. Article PDF-12M.
- Item 5:** RhoA improves cryopreservation of rooster sperm through the Rho/RhoA-associated kinase/cofilin pathway. Nuo Heng, Zhi-Xian Zhao, Yong Guo, Shan Gao, De-Lin Cai, Bo-Fan Fu, Xi-Hui Sheng, Xiang-Guo Wang, Kai Xing, Long-Fei Xiao, Cheng Long, He-Min Ni, Hua-Bin Zhu, Xiao-Long Qi.

The right sidebar contains filters and search options, including 'Find related data', 'Search details', and 'Recent activity'.

Εικόνα 20 NCBI αποτέλεσμα αναζήτησης στο Literature/PubMed