ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικόν και Καποδιστριακόν Πανεπιστήμιον Αθηνών
ΙΔΡΥΘΕΝ ΤΟ 1837

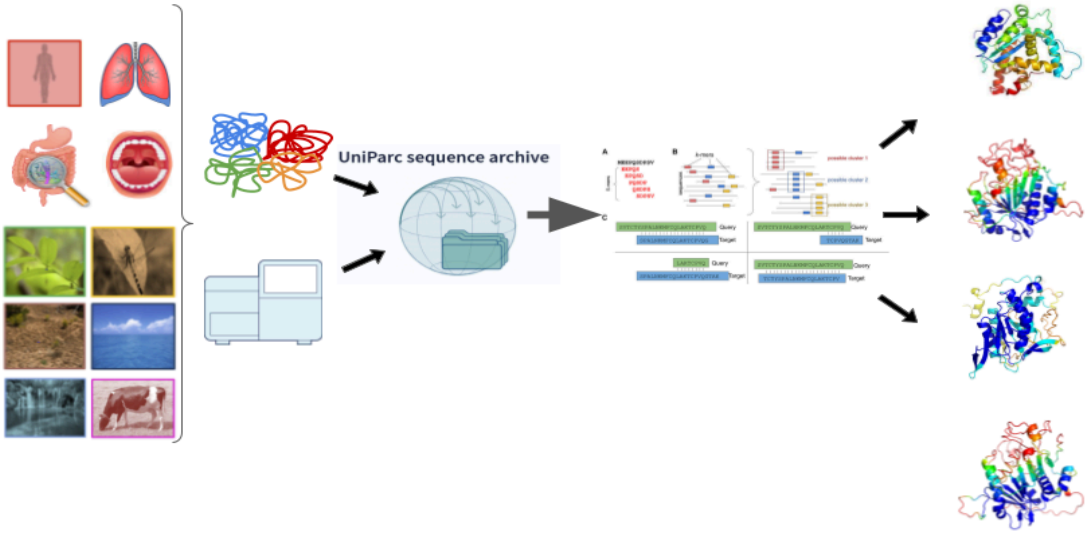
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ-ΥΠΟΛΟΓΙΣΤΙΚΗ ΒΙΟΛΟΓΙΑ»

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«Ανακάλυψη ενζύμων βιοτεχνολογικού ενδιαφέροντος μέσω μεταγονιδιωματικής ανάλυσης αλληλουχιών»



Νικόλαος Βεργουλίδης

Πτυχιούχος Ιατρικών Εργαστηρίων, Πανεπιστήμιο Θεσσαλίας

ΑΘΗΝΑ 2024

"ALEXANDER FLEMING"
Biomedical Sciences Research Center

ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικόν και Καποδιστριακόν Πανεπιστήμιον Αθηνών
ΙΔΡΥΘΕΝ ΤΟ 1837

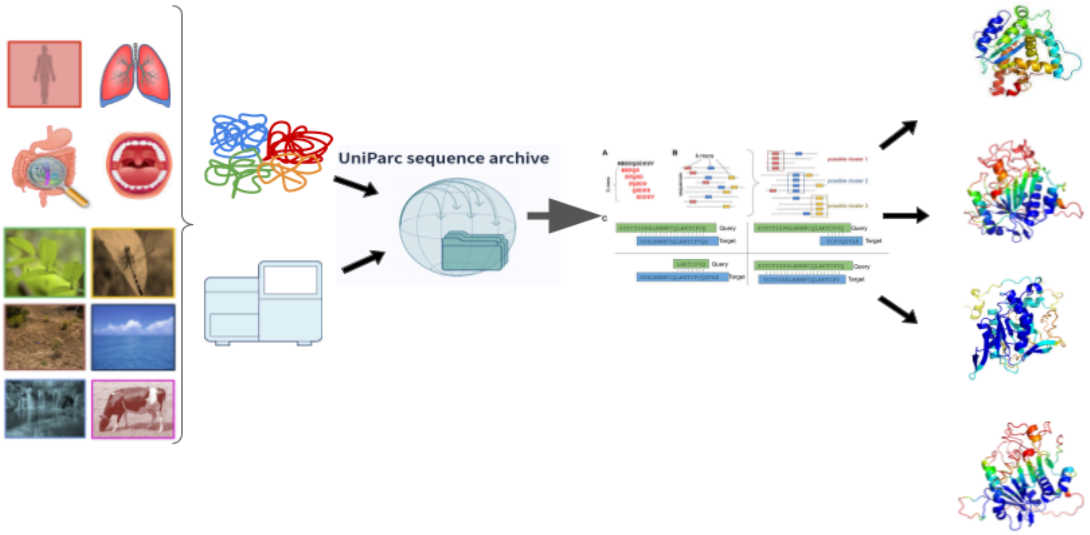
HELLENIC REPUBLIC
National and Kapodistrian University of Athens

SCHOOL OF SCIENCE
DEPARTMENT OF BIOLOGY

MASTER IN «BIOINFORMATICS-COMPUTATIONAL BIOLOGY»

Master Diploma Thesis

«Enzyme discovery through metagenomic sequence analysis»



Nikolaos Vergoulidis

Biomedical Scientist, University of Thessaly

ATHENS 2024

"ALEXANDER FLEMING"
Biomedical Sciences Research Center

ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικόν και Καποδιστριακόν
Πανεπιστήμιον Αθηνών
ΙΔΡΥΘΕΝ ΤΟ 1837

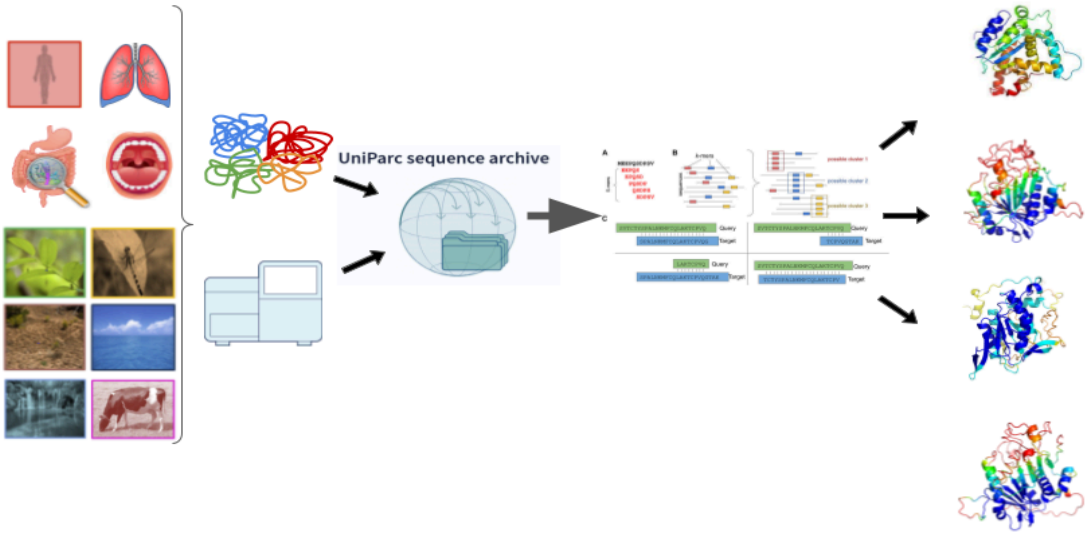
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ-ΥΠΟΛΟΓΙΣΤΙΚΗ ΒΙΟΛΟΓΙΑ»

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«Ανακάλυψη ενζύμων βιοτεχνολογικού ενδιαφέροντος μέσω μεταγονιδιωματικής ανάλυσης αλληλουχιών»



Τριμελής εξεταστική επιτροπή

Ερευνητής Δρ. Γεώργιος Α. Παυλόπουλος (Επιβλέπων ΠΜΣ)
Διευθυντής Ερευνών/Ερευνητής Α' Βιοπληροφορικής, Ε.ΚΕ.ΒΕ Αλέξανδρος Φλέμινγκ

Δρ. Ζωή Λίτου
Μέλος Ε.ΔΙ.Π., Τμήμα Βιολογίας, ΕΚΠΑ

Δρ. Νικόλαος Παπανδρέου
Μέλος Ε.ΔΙ.Π., Τμήμα Βιολογίας, ΕΚΠΑ

ΠΕΡΙΛΗΨΗ

Το μεταγονιδίωμα αναφέρεται στο σύνολο του γενετικού υλικού που περιέχεται σε ένα περιβαλλοντικό δείγμα. Η ανάλυση των μεταγονιδιωμάτων έχει ιδιαίτερη σημασία για την εύρεση και κατανόηση της αλληλουχίας, της δομής και λειτουργίας νέων πρωτεϊνών, καθώς και για τον εντοπισμό νέων ενζύμων που μπορούν να χρησιμοποιηθούν στη βιοτεχνολογία. Η αλληλούχιση μεταγονιδιωμάτων μέσω της μεθόδου shotgun είναι η πιο διαδεδομένη προσέγγιση για τη μελέτη και την ταξινόμηση μικροοργανισμών που προέρχονται από διαφορετικά βιοσυστήματα (biomes).

Στα πλαίσια αυτής της εργασίας, επιστρατεύτηκαν βιοπληροφορικές μέθοδοι με κεντρικό στόχο την ανακάλυψη ενζύμων βιοτεχνολογικού ενδιαφέροντος και τη λειτουργική ανάλυση πρωτεϊνικών αλληλουχιών από μεταγονιδιωματικά σύνολα δεδομένων. Η εργασία εστιάζει σε βακτηριακά ένζυμα με εφαρμογές στη βιοϊατρική και τη βιοτεχνολογία και χρησιμοποιεί προηγμένες μέθοδους αναζήτησης και ομαδοποίησης για δεδομένα που είναι καταχωρημένα σε εγκεκριμένες βάσεις μεταγονιδιωματικών δεδομένων, όπως οι IMG/M, MGnify και UniParc. Τα αποτελέσματα της ομαδοποίησης θα χρησιμοποιηθούν για τη δημιουργία τρισδιάστατων δομικών μοντέλων ενζύμων με τη χρήση της τεχνητής νοημοσύνης (A.I) ενώ τα μοντέλα αυτά θα αποτελέσουν τη βάση για την εξερεύνηση των μηχανισμών ενζυμικής δράσης, προκειμένου να επιλεχθούν τα κατάλληλα ένζυμα για την εκάστοτε εφαρμογή καθώς και να σχεδιαστούν στο μέλλον καινοτόμα ένζυμα.

Με το πέρας της εργασίας, έχει ολοκληρωθεί η ανάλυση της βάσης δεδομένων UniParc (~544 εκατομμύρια καταχωρήσεις) σε σχέση με συγκεκριμένες αυτοτελείς δομικές περιοχές (πρωτεϊνικά domains) από τέσσερις μεγάλες οικογένειες όπως οι: *Nattokinase, Feruloyl Esterases, Cocaine Esterases, Petases Pet Hydrolases*. Τα δεδομένα για τα πρωτεϊνικά προφίλ προήλθαν από τις βάσεις δεδομένων που είναι μέλη της InterPro: Pfam, PRINTS, PANTHER, PROSITE patterns, Tigrfams, SUPERFAMILY. Για την ανάλυση, χρησιμοποιήθηκαν διακριτές μέθοδοι βιοπληροφορικής ανάλυσης ενώ για το χειρισμό του μεγάλου όγκου δεδομένων αναπτύχθηκαν ροές σε διαφορες γλώσσες προγραμματισμού. Τα εργαλεία ανάλυσης ενσωματώθηκαν σε ολοκληρωμένες ροές εργασίας χρησιμοποιώντας τη γλώσσα nextflow και διατίθενται σε ένα από τα δημόσια αποθετήρια κώδικα όπως το GitHub. Τα αποτελέσματα της ανάλυσης, συμπεριλαμβανομένων των διακριτών στρατολογημένων πρωτεϊνών στις οποίες κατέληξε η εργασία, διατίθενται σε βάση δεδομένων με τη μορφή διαδικτυακής εφαρμογής (Meta-4) στην οποία ο χρήστης μπορεί να περιηγηθεί και να αναγνώσει τα χαρακτηριστικά των πρωτεϊνικών καταχωρήσεων, των τρισδιάστατων δομικών μοντέλων και τα μεταδεδομένα αυτών.

Επιστημονική Περιοχή: Βιοπληροφορική

Λέξεις κλειδιά: Μεταγονιδίωμα, Μεταγονιδιωματική Ανάλυση, Πρωτεϊνικά Domains, Ενζυμική Δραστηριότητα, Ομαδοποίηση Πρωτεϊνικών Αλληλουχιών

Abstract

A metagenome is the total amount of genetic material in an environmental sample. Metagenomic analysis is of paramount importance for identifying and understanding the complex mechanisms involved in sequence, structure and functions of de novo proteins as well as the recruitment of new enzymes for use in biotechnology. Metagenome shotgun sequencing has become the method of choice for studying and classifying microorganisms from various biomes.

The goal of this project is utilizing bioinformatics methods for enzyme discovery and engineering of protein sequences from metagenomic datasets. The project will focus on bacterial enzymes with biomedical and biotechnological applications and we will employ state-of-the-art search and clustering methods to recruit and cluster enzyme sequences from up-to-date curated data repositories such as IMG/M, MGnify and UniParc. The resulting clusters will be used to generate protein 3D models of the enzymes, through the use of Artificial Intelligence (A.I.) and the derived models will be used as the basis for studying the mechanisms of enzyme activity and designing novel drugs and inhibitors.

By the end of this project, UniParc database will have been analyzed (version by the time of initialization of the project included ~544 million entries) for the detection of protein domains of four superfamilies such as: *Nattokinase, Feruloyl Esterases, Cocaine Esterases, Petases Pet Hydrolases*. Protein profiles data will be derived from the database members of InterPro such as: Pfam, PRINTS, PANTHER, PROSITE patterns, Tigrfams, SUPERFAMILY. Distinct bioinformatic methods will be used for the analysis and scripts in a variety of programming languages will be used for handling the big-data. The overall workflow implementing the distinct tools used for the analysis will be developed in a global multi-layered pipeline using nextflow programming language and will be available in public repositories such as GitHub. The hits of the searching procedure, including the metadata and generated 3d models will be deposited in a web application database (Meta-4) with Graphical User Interface to be open and easily accessed for research/scientific purposes

Scientific Area: Bioinformatics

Keywords: Metagenomics, Biodiversity, Enzyme discovery, Metagenomic Sequence Analysis, Sequence Searching and Clustering, Protein Domains

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω για την παρούσα διπλωματική εργασία τον επιβλέποντα καθηγήτη Δρ. Γεώργιο Α. Παυλόπουλο, διευθυντή ερευνών/ερευνητή Α' βιοπληροφορικής στο Ε.ΚΕ.Β.Ε. "Αλέξανδρος Φλέμινγκ", για την πολύτιμη καθοδήγηση του και την ευκαιρία που μου έδωσε να ασχοληθώ με ένα τόσο ενδιαφέρον και πολυδιάστατο θέμα και κυρίως για την εμπιστοσύνη που μου έδειξε να δοκιμάσω τις δυνατότητες μου και να φέρω σε πέρας αυτό το σημαντικό επιστημονικό έργο.

Επιπλέον, θα ήθελα να ευχαριστήσω τα υπόλοιπα μέλη της τριμελούς επιτροπής και καθηγητές μου στο ΠΜΣ, Δρ. Ζωή Λίτου μέλος Ε.ΔΙ.Π., τμήματος Βιολογίας του ΕΚΠΑ και Δρ. Νικόλαο Παπανδρέου μέλος Ε.ΔΙ.Π., τμήματος Βιολογίας του ΕΚΠΑ, για την πολύτιμη διδασκαλία τους καθ' όλη τη διάρκεια των σπουδών, για τα εφόδια που μου έδωσαν και για την συνεχή στήριξη και καθοδήγηση που παρείχαν σε εμένα και όλους τους συμφοιτητές μου κατά την εκπαίδευσή μας στο πολύπλευρο και απαιτητικό επιστημονικό πεδίο της Βιοπληροφορικής.

Επίσης, θα ήθελα να ευχαριστήσω θερμά τους μεταδιδακτορικούς ερευνητές Δρ. Φώτη Μπαλτούμα του Ε.ΚΕ.Β.Ε. "Αλέξανδρος Φλέμινγκ" και Δρ. Ευάγγελο Καρατζά του European Bioinformatics Institute "EMBL - EBI", για την βοήθεια τους από την αρχή ως το τέλος αυτής της εργασίας, για την υπομονή τους, τις πολύτιμες συμβουλές τους και ενθάρρυνση που μου έδιναν, την μεταλαμπάδευση της εμπειρίας και της γνώσης τους τόσο σε τεχνικό όσο και στο θεωρητικό επίπεδο, χωρίς τους οποίους αυτή εργασία δεν θα μπορούσε να πραγματοποιηθεί. Μόνο τυχερός μπορεί να νιώθει κανείς αν έχει συνεργαστεί με τόσο εξαίρετους επιστήμονες και αξιοθαύμαστους ανθρώπους.

Τέλος, θα ήθελα να ευχαριστήσω τους δικούς μου ανθρώπους, την οικογένεια και τους φίλους μου, για την στήριξη τους από όταν πήρα την απόφαση να συμμετέχω στο ΠΜΣ και καθόλη την διάρκεια αυτού του δύσκολου ταξιδιού. Τα λόγια και οι πράξεις τους αποτέλεσαν την κινητήρια δύναμη για ότι έχω καταφέρει. Σας ευχαριστώ.

Νίκος Βεργουλίδης

Αθήνα, Σεπτέμβρης 2024

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

1. ΘΕΩΡΗΤΙΚΗ ΕΙΣΑΓΩΓΗ

1.1. Μεταγονιδίωμα

Ο συνολικός πληθυσμός των διαφόρων ειδών μικροοργανισμών του πλανήτη μας υπολογίζεται ότι ανέρχεται τουλάχιστον στο 1 τρις [1], ξεπερνώντας κατά πολύ σε αριθμό οποιαδήποτε άλλη μορφή ζωής. Το πλήθος των μικροβιακών κυττάρων επάνω στη Γη εκτιμάται ότι είναι 10^{30} , υπερέχοντας αριθμητικά των άστρων του Γαλαξία μας (~ 100 δισ. άστρα), ενώ αντίστοιχα αστρονομικά νούμερα φαίνεται να καταγράφονται αν απαριθμήσουμε το σύνολο των μικροβίων σε μια χούφτα χώμα [2], [3], [4]. Ευκαρυωτικοί μικροοργανισμοί όπως αρχαία, βακτήρια, πρωτόζωα και προκαρυωτικοί όπως οι ιοί, απαντώνται παντού στην βιόσφαιρα, σε ένα εύρος οικοσυστημάτων και μικρο-περιβάλλοντων από ποτάμια, λίμνες, έρημους έως το ανθρώπινο έντερο όπου οι μικροβιακοί πληθυσμοί που το παρασιτούν, εκτιμάται ότι απαρτίζονται από 10^{13} σε 10^{14} κύτταρα [2], [5].

Όλοι οι οργανισμοί οι οποίοι συνυπάρχουν σε ένα περιβάλλον αλληλεπιδρούν μεταξύ τους και έχει δειχθεί ότι οι μικροβιακές κοινότητες ή αλλιώς μικροβιώματα, κατέχουν σημαντικό ρόλο σε αυτή την αλληλεπίδραση είτε ως ρυθμιστές διαφόρων μεταβολικών ή λειτουργικών οδών είτε γενικότερα συμβάλλοντας στην ομοιόσταση, την υγεία, τη φυσιολογία, την συμπεριφορά και την οικολογική ισορροπία των ξενιστών τους. Συνεπώς, έχει ιδιαίτερη σημασία η μελέτη των μικροβιακών κοινοτήτων και των αλληλεπιδράσεων τους για την εξαγωγή συμπερασμάτων χρήσιμων σε ιατρικο-φαρμακευτικές, βιοτεχνολογικές και οικολογικές εφαρμογές, καθώς και για την γενικότερη ανάπτυξη και μελέτη του κλάδου της βιοποικιλότητας.

Παρόλη τη σημασία τους, λόγω δυσκολιών κατά την καλλιέργεια των προς μελέτη μικροοργανισμών, η πλειοψηφία τους και του γενετικού υλικού αυτών μένει ανεξερεύνητη, αντιπροσωπεύοντας αυτή τη στιγμή πολύ χαμηλά ποσοστά (μικρότερα του 1%) της παγκόσμιας μικροβιακής ταξινόμησης [3], [6].

Με την χρήση της μεταγονιδιωματικής μπορούν να ξεπεραστούν τέτοια εμπόδια. Αποτελεί σήμερα κρίσιμη τομή στην μελέτη του μικροβιώματος και ακαλλιέργητων μικροοργανισμών. Η μεταγονιδιωματική ανάλυση συνεπάγεται την αλληλούχιση και ανάλυση του συνόλου του γενετικού υλικού (DNA) το οποίο περιέχεται σε ένα περιβαλλοντικό δείγμα (περιβαλλοντικό DNA - eDNA), είτε αυτό πρόκειται για δείγμα χώματος από ένα δάσος, δείγμα από ύδατα μιας λίμνης ή του ωκεανού, μια απόχρεμψη του ανώτερου αναπνευστικού ή ακόμα και την επιφάνεια ενός τυριού. Το μεταγονιδίωμα του δείγματος εξάγεται απευθείας, χωρίς να χρειάζεται απομόνωση του από οργανισμούς που έχουν προηγουμένως καλλιεργηθεί στο εργαστήριο. Αυτή η προσέγγιση καθιστά

δυνατή την έρευνα της γενετικής ποικιλομορφίας, των λειτουργιών που επιτελούνται, καθώς και της δυναμικής σε ένα βιολογικό σύστημα [2], [7], [8].

Η πρόοδος στις σύγχρονες τεχνολογίες αλληλούχισης δίνει τη δυνατότητα για ακρίβεια στην εξαγωγή και τον χαρακτηρισμό των γονιδιωμάτων και η πλέον τυπική μέθοδος μεταγονιδιωματικής ανάλυσης συνοψίζεται στα εξης βηματα:

- **Αλληλούχιση**: Πραγματοποιείται αλληλούχιση του δείγματος σε κάποιον αναλυτή (sequencer), με αποτέλεσμα την εξαγωγή δεδομένων που περιέχουν θραύσματα DNA (DNA fragments) απο τους διάφορους οργανισμούς που βρίσκονται στο δείγμα. Το βάθος της αλληλούχισης καθώς και το μήκος των θραυσμάτων ποικίλουν αναλόγως το που στοχεύει η αλληλούχιση.
- **Ποιοτικός Έλεγχος:** Τα raw αρχεία ελέγχονται από υπολογιστικά εργαλεία για την εκτίμηση της ποιότητας αλληλούχισης καθώς και για τον καθαρισμό από περιττά δεδομένα όπως αυτά των εκκινητών ή των προσαρμογέων.
- **Assembly / Read Mapping:** Σε αυτό το στάδιο γίνεται συναρμολόγηση των μικρότερων θραυσμάτων DNA (reads) κατόπιν στοίχισης, σε μεγαλύτερες γονιδιωματικές αλληλουχίες, οι οποίες με τη σειρά τους συναρμολογούνται και διαμορφώνουν contigs (συνεχόμενες περιοχές στην DNA αλληλουχία) και ικριώματα (scaffolds) με τη χρήση είτε γονιδιώματος αναφοράς εφόσον είναι γνωστό (reference-based assembly) είτε de novo assembly εφόσον το γονιδίωμα αναφοράς δεν υπάρχει, ή υβριδική όπου υπάρχει γονιδίωμα αναφοράς για οδηγός και μερικώς γίνεται de novo assembly.
- **Ομαδοποίηση (Binning) και ανακατασκευή γονιδιώματος:** Οι συναρμολογημένες περιοχές του DNA (contigs) ομαδοποιούνται σε παρόμοιες ταξινομικές λειτουργικές μονάδες (OTUs) βάσει ομοιοτήτων στην νουκλεοτιδική αλληλουχία , αλληλοεπικάλυψη στην κατά ζεύγη στοίχιση κ.α. και τα γονιδιώματα που προκύπτουν από αυτή τη διαδικασία συνήθως ονομάζονται MAGs (Metagenome Assembled Genomes).
- **Χαρακτηρισμός (Annotation):** Τα MAGs υπόκεινται σε λειτουργικό και ταξινομικό χαρακτηρισμό.

Η ανάλυση του μεταγονιδώματος μέσω μεθόδων βιοπληροφορικής έχει σωρεία εφαρμογών, όπως στο πεδίο της περιβαλλοντικής μικροβιολογίας για τη διερεύνηση της μικροβιακής ποικιλομορφίας και την εξερεύνηση των διαφόρων λειτουργιών που επιτελούν μικροβιακοί πληθυσμοί σε ένα περιβάλλον [7], κάνοντας εφικτή την εις βάθος μελέτη της βιοποικιλότητας σε επίπεδο πρωτεϊνικών οικογενειών [3], στο πεδίο της αγροτικής παραγωγής ανοίγει νέους δρόμους για την μελέτη των μικροβιακών κοινοτήτων που βρίσκονται στο έδαφος και πώς αυτές επηρεάζουν την υγεία και απόδοση των καλλιεργειών, επιτρέποντας ταυτόχρονα την διερεύνηση νέων μεθόδων διατήρησης της ισορροπίας των τοπικών οικοσυστημάτων αξιοποιώντας μικροοργανισμούς όπως για παράδειγμα μύκητες, ως εναλλακτικές μεθόδους αντί για χημικά φυτοφάρμακα [9].

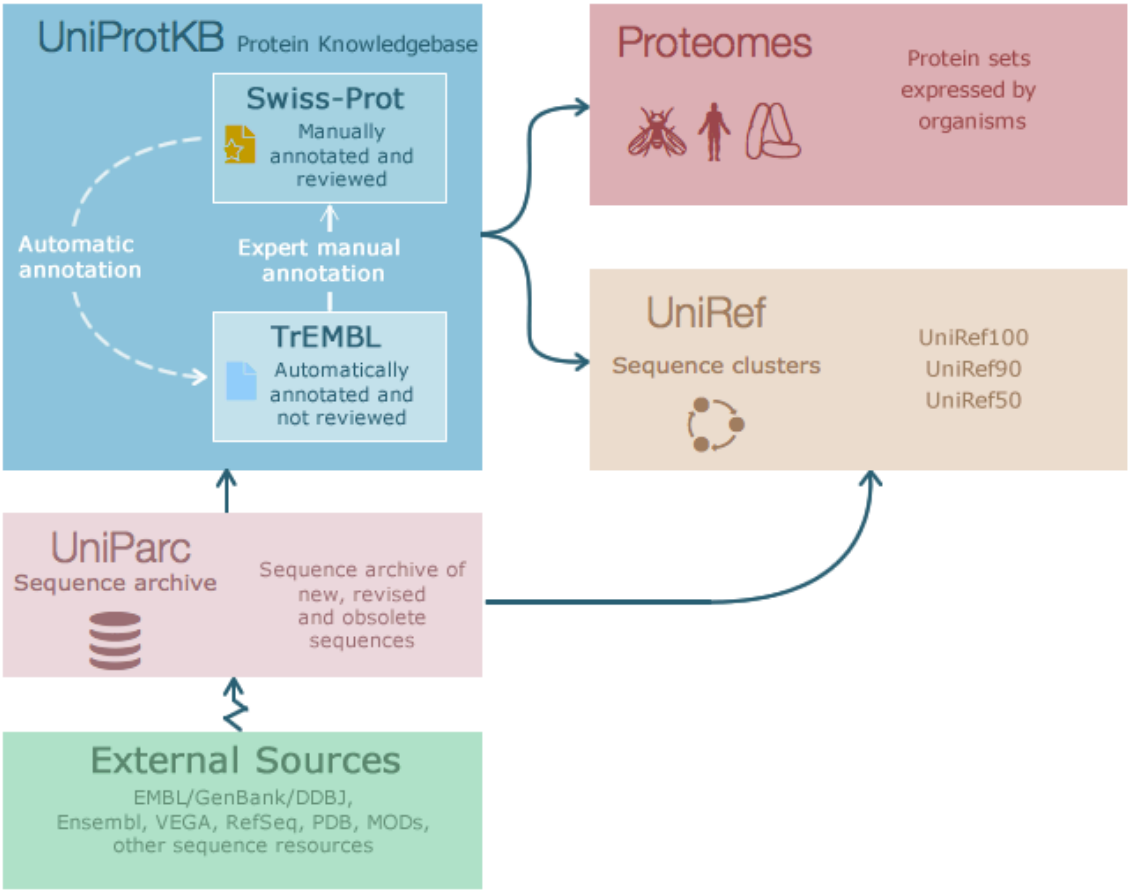
Επίσης τελευταία γίνονται όλο και περισσότερες προσπάθειες για την εγκαθίδρυση μεθόδων προσέγγισης με βάση το μεταγονιδίωμα και στην διαγνωστική ιατρική, ερμηνεύοντας το ανθρώπινο μικροβίωμα σε κλινικά δείγματα από διάφορες περιοχές του σώματος (έντερο, ανώτερο ή κατώτερο αναπνευστικό, δέρμα και άλλα όργανα), εξετάζοντας και αναλύοντας το πόσο και με ποιούς τρόπους το μικροβίωμα (βακτηρίωμα, ίωμα) και η διατάραξη αυτού επηρεάζουν την ομοιόσταση, τις ασθένεις, την φλεγμονώδη απόκριση, την φαρμακοαπόκριση, τα μεταβολικά μονοπάτια [5], [10], [11], [12]. Επιπλέον, τελευταίες σημαντικές έρευνες έχουν αξιοποιήσει την μεταγονιδιωματική για να ρίξουν φως σε παγκόσμιου επιπέδου πρωτεϊνικά δεδομένα με αχαρτογράφητες μέχρι τώρα λειτουργικές δυνατότητες [13]. Τέλος στο πεδίο της βιοτεχνολογίας, χρησιμοποιείται για την ανακάλυψη ενζύμων με επιθυμητές ιδιότητες μέσω διαλογής από βιβλιοθήκες και μεταγονιδιωματικά σύνολα δεδομένων από διάφορα περιβάλλοντα καθώς και για την ανάλυση λειτουργικών γονιδίων που εμπλέκονται σε μονοπάτια αποικοδόμησης, εφαρμόζοντας τέτοια ένζυμα στη βιομηχανία, την φαρμακευτική και την ιατρική και η παρούσα μελέτη επικεντρώθηκε σε αυτό το σκοπό [14], [15], [16].

1.2. Αποθετήρια δεδομένων

Τα δεδομένα και μεταδεδομένα που προκύπτουν από την παραπάνω μεθοδολογία, συμπεριλαμβανομένων DNA και πρωτεϊνικών αλληλουχιών, γονίδια που προβλέφθηκαν, σύνολα δεδομένων και χαρακτηριστικά, φιλοξενούνται σε δημόσιες, δωρεάν προσβάσιμες βάσεις δεδομένων και αποθετήρια. Επειδή η παρούσα εργασία αφορά την ενζυμική ανακάλυψη με στόχο βιοτεχνολογικές εφαρμογές και σχεδιασμό φαρμάκων, αξιοποιήθηκαν δεδομένα πρωτεϊνικών αλληλουχιών που προέκυψαν από ανάλυση μεταγονοδιωματικών πρωτεϊνικών αλληλουχιών με σκοπό την στρατολόγηση στόχων ενδιαφέροντος, βάσει των λειτουργικών δυνατοτήτων τους και της δομής, συγκεκριμένα τις αυτοτελείς δομικές περιοχές (domains) αυτών. Σε αυτό το παράρτημα, παρουσιάζονται οι βάσεις δεδομένων και τα αποθετήρια όπου πραγματοποιήθηκε η ανάλυση.

1.2.1. UniProt

Το σχήμα Universal Protein Resource ή αλλιώς UniProt [17], αποτελεί μια συνεργασία μεταξύ του EBI (European Bioinformatics Institute), του SIB (Swiss Institute of Bioinformatics) και του PIR (Protein Information Resource) και απαρτίζεται από τρεις βάσεις δεδομένων κάθε μια προορισμένη για συγκεκριμένη χρήση (Εικόνα **1.1**). Η UniProt Knowledgebase (UniProtKB) είναι η βάση δεδομένων με την καλύτερη και πιο εκτενώς επιμελημένη πληροφορία γύρω από τις πρωτεΐνες περικλείοντας γνώση γύρω από την λειτουργία, την ταξινόμηση καθώς και διασταυρούμενες αναφορές αυτών. Η UniRef (UniProt Reference Clusters) συνδυάζει στενά συσχετιζόμενες αλληλουχίες σε μια μοναδική καταγραφή ώστε να πετύχει μειωμένους χρόνους κατά την αναζήτηση ομοιοτήτων σε επίπεδο αλληλουχίας ενώ η UniParc (UniProt Archive) αποτελεί ένα συμπεριληπτικό αποθετήριο του συνόλου των πρωτεϊνικών αλληλουχιών, περιέχοντας μοναδικούς κωδικούς αναγνώρισης (identifiers) και αλληλουχιες. Είναι χαρακτηριστικό πως στην έκδοση 2024_4 η UniProt καταγράφει στο δυναμικό της 571,864 καλά χαρακτηρισμένες πρωτεϊνικές καταχωρήσεις (reviewed στην Swiss-Prot) και 245,324,902 μη αναθεωρημένες (unreviewed στην TrEMBL). Παρέχει εργαλεία αναζήτησης και στοίχισης, όπως το BLAST [18] και το ClustalW [19] καθώς επίσης και για αναζήτηση με βάση λίστα του χρήστη ή πεπτιδίου ενδιαφέροντος. Η UniProt δίνει τη δυνατότητα στον χρήστη για εκτενή αναζήτηση πληροφορίας σε βάθος, αφού κάθε καταχώρηση μπορεί να περιλαμβάνει από βιοφυσικά χαρακτηριστικά (αν πρόκειται πχ για μια πρωτεΐνη) μέχρι δημοσιεύσεις που αναφέρουν την συγκεκριμένη καταχώρηση



Εικόνα 1.1: *Γενική επισκόπηση της κοινοπραξίας UniProt σε σχηματική αναπαράσταση (πηγή: Από τον επίσημο ιστότοπο της UniProt στην καρτέλα "About" https://www.uniprot.org/help/about).*

1.2.2. UniParc

Η βάση δεδομένων UniProt Archive ή αλλιώς UniParc [17], είναι ένα αποθετήριο μη επαναλαμβανόμενων πρωτεϊνικών αλληλουχιών, το οποίο περιλαμβάνει όλες τις νέες και αναθεωρημένες πρωτεϊνικές αλληλουχίες από όλες τις δημόσια διαθέσιμες πηγές: EMBL-Bank / DDBJ / GenBank (βάσεις δεδομένων νουκλεοτιδικών ακολουθιών), Ensembl, EnsemblGenomes, EnsemblRapid, European Patent Office (EPO), FlyBase, H-Invitational Database (H-InvDB), International Protein Index (IPI), Japan Patent Office (JPO), Korean Intellectual Property Office (KIPO), Pathosystems Resource Integration Center (PATRIC), Protein Data Bank (PDB), Protein Research Foundation (PRF), RefSeq, Saccharomyces Genome database (SGD), TAIR Arabidopsis thaliana Information Resource, The Seed (SEED), TROME, USA Patent Office (USPTO), UniProtKB/Swiss-Prot, UniProtKB/Swiss-Prot protein isoforms, UniProtKB/TrEMBL, Vertebrate Genome Annotation database (VEGA), WormBase, WormBase ParaSite (WBParaSite). Επιπλέον συμπεριλαμβάνονται και θα συνεχίσουν να διατηρούνται διασταυρούμενες παραπομπές στις παρακάτω καταργημένες βάσεις: IPI, PIR, PIRARC, REMTREMBL, UniMES, TREMBLNEW, TrEMBL_varsplic. Ένας συγκεντρωτικός πίνακας με τις πρωτογενείς πηγές της UniParc φαίνεται στην Εικόνα **1.2**. Με αυτό τον τρόπο διασφαλίζεται πλήρης κάλυψη όλων των αλληλουχιών σε μια μοναδική ηλεκτρονικό ιστότοπο. Η μοναδικότητα των καταχωρήσεων εξασφαλίζεται με την συγχώνευση όλων εκείνων οι οποίες παρουσιάζουν 100% ομοιότητα σε ολόκληρο το μήκος της αλληλουχίας, ανεξαρτήτως οργανισμού προέλευσης. Πραγματοποιείται διασταυρούμενη αναφορά των νέων και ανανεωμένων καταχωρήσεων μέσω αριθμού πρόσβασης (accession number) της πηγαίας βάσης δεδομένων και αποδίδεται αριθμός έκδοσης για την αλληλουχία ο οποίος αυξάνεται με κάθε αλλαγή στην υποκείμενη αλληλουχία. Η βασική πληροφορία σε κάθε καταχώρηση της UniParc είναι:

- **Identifier -** Σταθερός μοναδικός κωδικός αποτελούμενος από το ακρωνύμιο “UPI” ακολουθούμενο από έναν συνδυασμό δέκα (10) δεκαεξαδικών αριθμών.
- **Sequence -** Η πρωτεϊνική αλληλουχία της εκάστοτε καταχώρησης.
- **Cyclic redundancy check number -** Μια αριθμητική τιμή της αλληλουχίας που υπολογίζεται μέσω ενός αλγορίθμου και αξιοποιείται ως μέτρο σύγκρισης της μοναδικότητας της αλληλουχίας.
- **Source database(s) with accession and version numbers -** Σύνδεσμοι προς την μητρική βάση δεδομένων στην οποία είναι καταχωρημένη η πρωτεϊνη, με πληροφορία έκδοσης και αριθμό πρόσβασης.

- **Time stamp -** Ημερομηνία πρώτης και τελευταίας εμφάνισης στα δημόσια αποθετήρια της συγκεκριμένης καταχώρησης.

Τέλος, κάθε σύνδεσμος με την βάση προέλευσης παρέχει πληροφορία για την κατάσταση της καταχώρησης στην βάση προέλευσης, αν η αλληλουχία υπάρχει ακόμα ή έχει διαγραφεί καθώς και αναφορές στο NCBI GI και TaxId όπου χρειάζεται.

UniParc sequence archive

The UniProt Archive (UniParc) is a comprehensive and non-redundant database that contains most of the publicly available protein sequences in the world. Proteins may exist in different source databases and in multiple copies in the same database. UniParc removes such redundancy by storing each unique sequence only once and giving it a stable and unique identifier (UPI) making it possible to identify the same protein from different source databases. A UPI is never removed, changed or reassigned. UniParc contains only protein sequences and cross-references. All other information about the protein must be retrieved from the source databases using the database cross-references. UniParc tracks sequence changes in the source databases and archives the history of all changes. UniParc has combined many databases into one at the sequence level and searching UniParc is equivalent to searching many databases simultaneously.

Start searching in UniParc »

Databases

Cross-reference	Number of UniParc entries
EMBLWGS	465,974,219
RefSeq	399,799,015
UniProtKB	290,335,821
EnsemblBacteria	108,708,156
EMBL CDS	88,909,771
PATRIC	86,394,939
EMBL_CON	60,850,596
EMBL_TSA	17,212,079
SEED	17,123,523
Ensembl	13,415,061
EnsemblFungi	12,316,705
EnsemblPlants	7,509,453
EnsemblMetazoa	6,694,327
WBParaSite	6,473,128

Cross-reference	Number of UniParc entries
UNIMES	6,028,179
USPTO	3,461,838
JPO	3,140,406
EnsemblRapid	2,922,635
EnsemblProtists	2,415,953
EPO	2,337,293
PRF	980,187
IPI	961,795
KIPO	923,231
TROME	734,989
VectorBase	684,777
TREMBLNEW	514,322
PIRARC	289,892
H-InvDB	268,343

Cross-reference	Number of UniParc entries
PIR	253,472
VEGA	205,220
PDB	174,701
REMTREMBL	126,336
UniProtKB/Swiss-Prot isoforms	79,498
FusionGDB	42,842
WormBase	42,612
TAIR	37,222
FlyBase	27,311
EMBL_TPA	23,203
SGD	6,269
TREMBL_VARSPLIC	866

Εικόνα 1.2: *Συγκεντρωτικοί πίνακες των βάσεων δεδομένων - μέλη της UniParc και των αριθμό καταχωρήσεων τους (πηγή: Στιγμιότυπο οθόνης από τον επίσημο ιστότοπο της UniProt https://www.uniprot.org/uniparc/).*

1.2.3. InterPro

Η βάση δεδομένων InterPro [20] παρέχει πληροφορίες για την λειτουργία των πρωτεϊνών χρησιμοποιώντας ένα σχήμα ταξινόμησης τους σε οικογένειες και προβλεπόμενα domains και χωροταξικά σημεία ενδιαφέροντος. Αυτό το πετυχαίνει με τη χρήση μοντέλων πρόβλεψης, τα λεγόμενα signatures, τα οποία παρέχονται από τις εκάστοτε βάσεις δεδομένων - μέλη της κοινοπραξίας InterPro (InterPro consortium member databases). Συνδυάζονται 13 βάσεις signature σε μία, παρέχοντας με αυτό τον τρόπο συμπληρωματικά επίπεδα πρωτεϊνικής πληροφορίας καθιστώντας έτσι την InterPro την

πιο ολοκληρωμένη πηγή σχετικά με πρωτεϊνικές οικογένειες, αυτοτελείς δομικές περιοχές (domains) και λειτουργικών sites. Όταν signatures από δύο ή περισσότερες διαφορετικές βάσεις-μέλη αναφέρονται στην ίδια βιολογική οντότητα, ενώνονται σε μια καταχώρηση προς αποφυγή επαναληψιμότητας. Πρόκειται για ένα ομολογουμένως πολύ δυνατό εργαλείο για τη διερεύνηση των δομικών χαρακτηριστικών των πρωτεϊνών και των λειτουργικών δραστηριοτήτων που συνδέονται με αυτές, με οργανωμένο και φιλικό προς τον χρήστη λειτουργικό περιβάλλον (Εικόνα **1.3**) και με ενσωματωμένες λειτουργίες πρόβλεψης πρωτεϊνικών αλληλουχιών και δομικών μοντέλων (Εικόνα **1.4**).

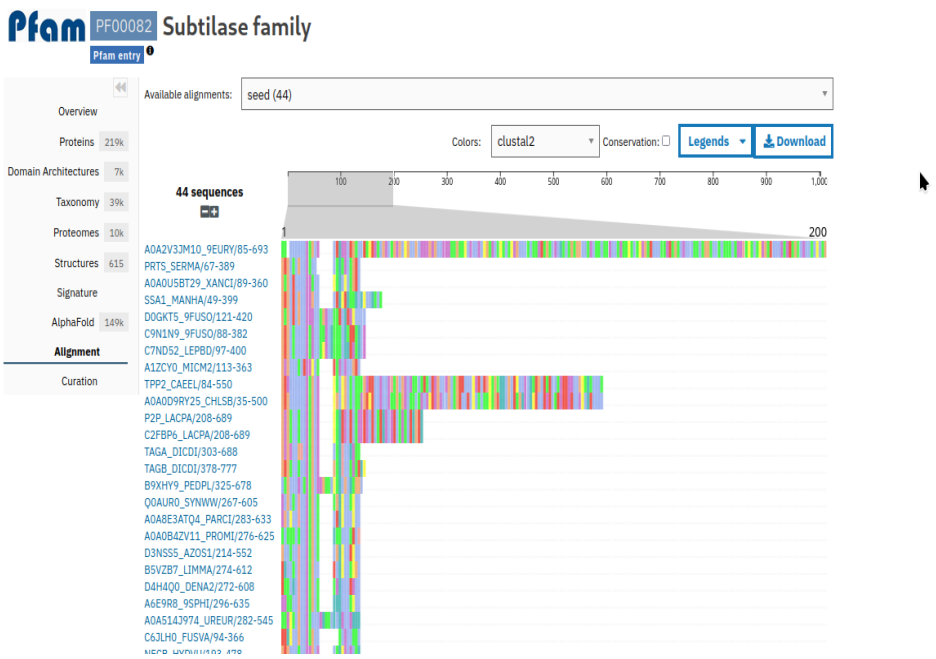
Εικόνα 1.3: *Άποψη της καταχώρησης με κωδικό "SUBTILISIN BPN' MUTANT 7186" της InterPro, η οποία δομικά απαντάει στο domain ενδιαφέροντος S8/S53. Δεδομένα προερχόμενα από κρυσταλλογραφία ακτίνων Χ. Φαίνονται στις διάφορες καρτέλες οι πληροφορίες για την πρωτεΐνη σε σχέση με τα δομικά χαρακτηριστικά, τις ομόλογες οικογένειες, τα ενεργά κέντρα καθώς και αναφορές. (πηγή: Στιγμιότυπο οθόνης από την ιστοσελίδα της InterPro μετά από αναζήτηση με τον κωδικό "IPR000209").*

Εικόνα 1.4: *Δείγμα των δυνατοτήτων της InterPro: Η ίδια αναζήτηση με την Εικόνα 1.2.3 αλλά αυτή τη φορά επιλέχθηκε με τον κέρσορα του ποντικιού το Domain Peptidase_S8 όπως φαίνεται και η επισήμανση με το πλαίσιο κειμένου "Pfam PF00082 Subtilase family 23-266". Η τρισδιάστατη δομή που παρέχεται από την PDB έχει αλλάξει χρώμα από πράσινο σε μπλε για να σηματοδοτήσει το εν λόγω domain.*

1.2.4. Pfam

Η βάση δεδομένων Pfam [21] είναι ένα ευρέως χρησιμοποιούμενο αποθετήριο πρωτεϊνικών οικογενειών, οι οποίες αντιπροσωπεύονται από πολλαπλές στοιχίσεις και μοντέλα Markov (hidden Markov models HMMs). Αποτελεί σημαντική πηγή χαρακτηρισμού των πρωτεϊνικών αλληλουχιών και των λειτουργιών που επιτελούν και η βασική ιδέα είναι η κατηγοριοποίηση σε οικογένειες βάσει κοινών χαρακτηριστικών στην αλληλουχία και στις δομικές λειτουργικές μονάδες (domains) αυτών. Αξιοποιεί την αναγνώριση καλά συντηρημένων περιοχών οι οποίες είναι βαρυσήμαντες για την πρωτεϊνική δομή και λειτουργία. Παρέχει λειτουργίες στον χρήστη όπως εργαλεία για την ανίχνευση Pfam domains που εμπεριέχονται σε πρωτεϊνικές αλληλουχίες, απεικόνιση στοίχισης για την αξιολόγηση συντηρημένων περιοχών (Εικόνα **1.5**) στην αλληλουχία και οπτικοποίηση της αρχιτεκτονικής των δομών για την απεικόνιση της οργάνωσης στο χώρο

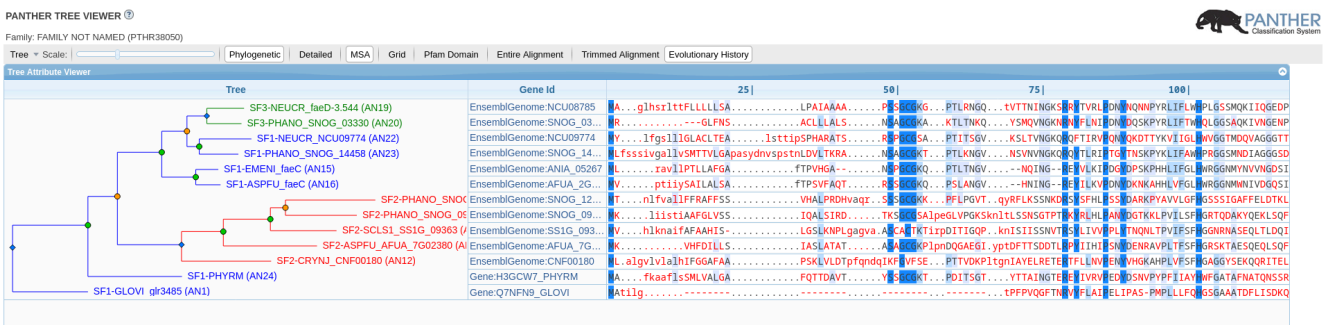
πολλαπλών αυτοτελών δομικών περιοχών (domains) εντός των πρωτεϊνών. Πλέον η Pfam φιλοξενείται στην InterPro.



Εικόνα 1.5: *Αναζήτηση του κωδικού "PF00082", domain ενδιαφέροντος της παρούσας εργασίας, και προβολή της πολλαπλής στοίχισης που αποτελεί seed alignment για την παραγωγή του profile Hmm μοντέλου για την συγκεκριμένη πρωτεϊνική οικογένεια (πηγή: Στιγμιότυπο οθόνης από την επίσημη ιστοσελίδα της InterPro https://www.ebi.ac.uk/interpro/entry/pfam/PF00082/entry_alignments/?type=seed).*

1.2.5. PANTHER

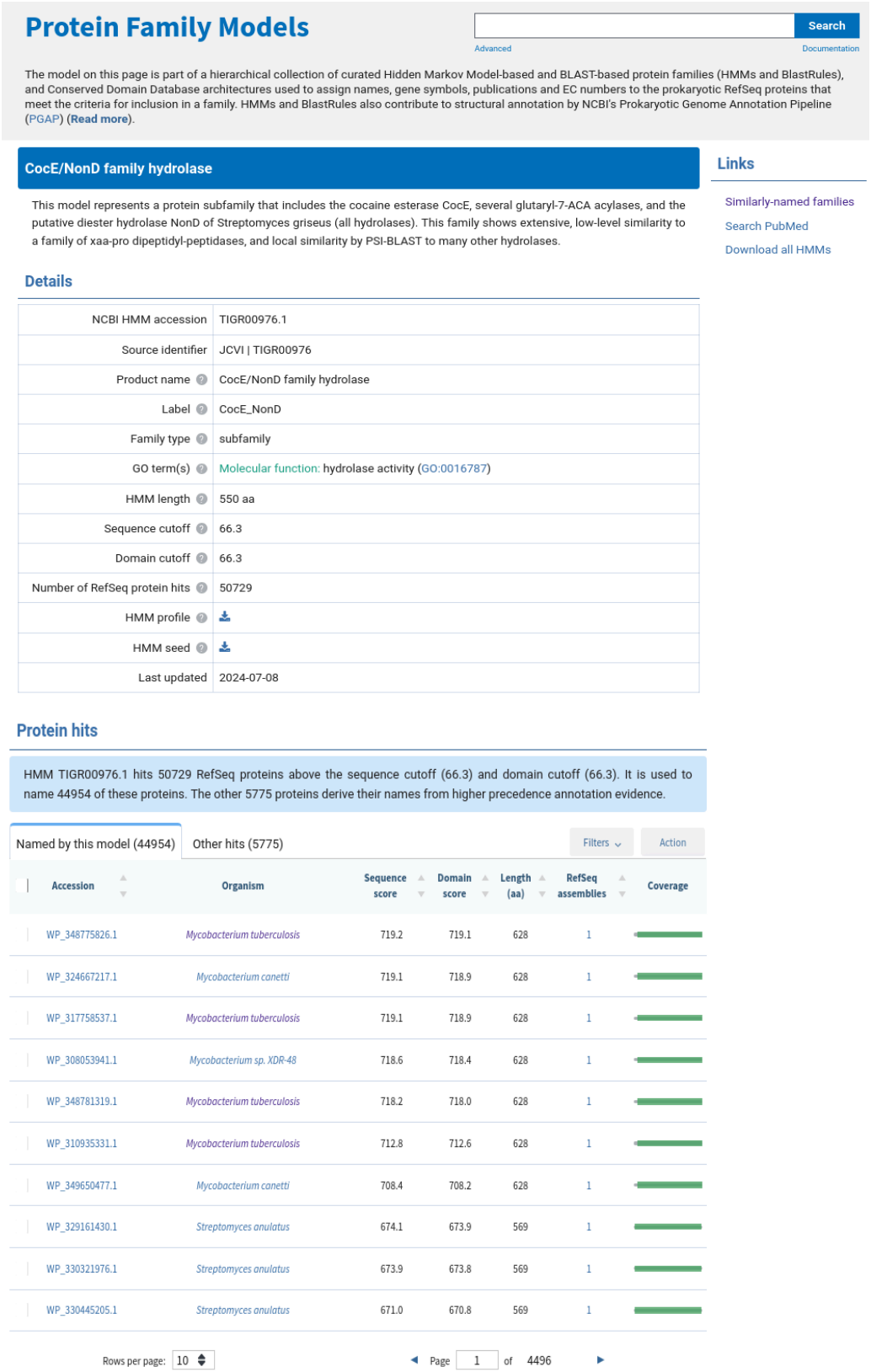
Η βάση δεδομένων PANTHER (Protein ANalysis THrough Evolutionary Relationships) παρέχει πληροφορίες σχετικά με πρωτεϊνικές λειτουργίες και εξελικτικές σχέσεις. Κατηγοριοποιεί γονίδια που κωδικοποιούν πρωτεΐνες σε οικογένειες και υπο-οικογένειες βάσει κοινών προγόνων και καλά συντηρημένων αυτοτελών λειτουργικών δομικών περιοχών (functional domains). Για το σκοπό αυτό αξιοποιεί ένα συνδυασμό χειροκίνητης επιμέλειας και αλγορίθμων για να εξασφαλίσει καλής ποιότητας χαρακτηρισμό. Η εξελικτική ταξινόμηση βασίζεται σε μια βιβλιοθήκη που απαρτίζεται από πάνω από 15,000 φυλογενετικά δέντρα και η λειτουργική ταξινόμηση περιλαμβάνει *Gene Ontology* όρους και μονοπάτια. Προσφέρει εργαλεία για στοίχιση αλληλουχιών (Εικόνα **1.6**) , κατασκευή φυλογενετικών δέντρων καθώς και ανάλυση λειτουργικού εμπλουτισμού καθιστώντας την κατάλληλη για τη μελέτη σε επίπεδο πρωτεϊνικής εξέλιξης, γονιδιακής λειτουργίας και συγκριτικής γονιδιωματικής γενικότερα.



Εικόνα 1.6: *Φυλογενετικό δέντρο και πολλαπλή στοίχιση (MSA) για την υπεροικογένεια ενδιαφέροντος (εμπεριέχει την αυτοτελή δομική περιοχή των Feruloyl Esterases) μετά από αναζήτηση στην PANTHER βάση δεδομένων με χρήση του κωδικού PTHR38050 (πηγή : στιγμιότυπο οθόνης από την επίσημη ιστοσελίδα της PANTHER https://www.pantherdb.org/panther/family.do?clsAccession=PTHR38050).*

1.2.6. TIGRFAMs

Η βάση δεδομένων TIGRFAMs [22], [23], αρχικά αποτελούσε ένα ερευνητικό εγχείρημα του Ινστιτούτου TIGR (The Institute for Genomic Research) και η συλλογή TIGRFAMs απαρτίζεται από επιμελημώς χαρακτηρισμένες πρωτεϊνικές οικογένειες που προέρχονται κυρίως από προκαρυωτικές αλληλουχίες. Η βάση αποτελείται από Μαρκοβιανά μοντέλα (HMMs), πολλαπλές στοιχίσεις αλληλουχιών, ορολογία Gene Ontology (GO), ταξινομικους αριθμους ενζύμων (EC numbers), γονιδιακά σύμβολα, ονόματα πρωτεϊνικών οικογενειών, περιγραφικά κείμενα, παραπομπές σε σχετικά μοντέλα σε άλλες βάσεις και βιβλιογραφικούς δείκτες. Από τον Απρίλιο του 2018 η διαχείριση της βάσης πέρασε στο National Center for Biotechnology Information (NCBI) και πλέον αποτελεί μέρος της ροης εργασίας του NCBI Prokaryotic Genome Annotation Pipeline για τον χαρακτηρισμό των αλληλουχιών της GenBank και RefSeq. Κάθε καταχώρηση στην TIGRFAMs περιλαμβάνει λεπτομερή χαρακτηρισμό, περιγραφή του βιολογικού ρόλου, των λειτουργικών χαρακτηριστικών και των εξελικτικών σχέσεων της πρωτεϊνικής οικογένειας, ενώ η αξιοποίηση των HMMs δίνει τη δυνατότητα για ανίχνευση ομολογίας με υψηλή ευαισθησία ακόμα και σε μακρινούς συγγενείς, καθιστώντας την βάση ιδιαίτερα χρήσιμη στον σχολιασμό νέων γονιδιωμάτων και τη μελέτη των πρωτεϊνικών διεργασιών ανάμεσα σε διαφορετικούς οργανισμούς (Εικόνα **1.7**).

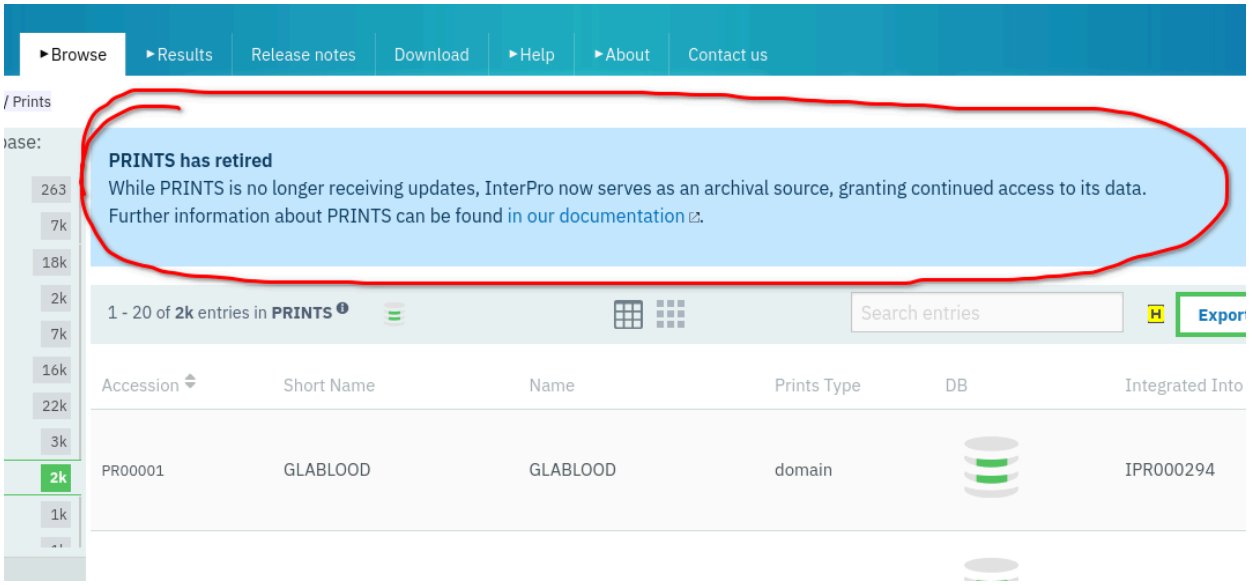


Εικόνα 1.7: *Αποτελέσματα αναζήτησης για τον κωδικό "TIGR00976" και επιλογή υπερσυνδέσμου μέσω της InterPro, οδηγώντας στην βάση δεδομένων NCBIfam (πρώην*

TIGRFAMs). Στο επάνω μέρος της ιστοσελίδας φαίνονται οι λεπτομέρειες σχετικά με το συγκεκριμένο μοντέλο ενδιαφέροντος (με εύκολη πρόσβαση και στο pHMM αρχείο), ενώ στο κάτω μέρος φαίνονται αποτελέσματα και στοιχεία της αναζήτησης ομοιοτήτων του μοντέλου έναντι της βάσης δεδομένων RefSeq. (πηγή: Στιγμιότυπο οθόνης μετά από περιήγηση στην επίσημη ιστοσελίδα της NCBIfam https://www.ncbi.nlm.nih.gov/genome/annotation_prok/evidence/NF000124/).

1.2.7. PRINTS

Η βάση δεδομένων PRINTS [24] είναι μια επιτομή πρωτεϊνικών αποτυπωμάτων. Ως αποτυπώματα αναφέρονται ομάδες καλά συντηρημένων μοτίβων τα οποία χρησιμοποιούνται για να χαρακτηρίσουν μια συγκεκριμένη πρωτεϊνική οικογένεια. Κάθε αποτύπωμα αποτελείται από πολλαπλές στοιχίσεις τέτοιων μοτίβων τα οποία προέκυψαν από συντηρημένες περιοχές πρωτεϊνικών αλληλουχιών, προσδίδοντας διακριτότητα στις συγκεκριμένες πρωτεϊνικές οικογένειες. Συνήθως τα μοτίβα δεν αλληλοεπικαλύπτονται παρά διαχωρίζονται κατα μήκος της αλληλουχίας παρόλο που δύναται να είναι συνεχόμενα στον τρισδιάστατο χώρο. Ο κύριος ρόλος της είναι η διευκόλυνση του σχολιασμού των λειτουργιών και της κατηγοριοποίησης των πρωτεϊνών, ειδικά όσων παρουσιάζουν μακρινές ομολογίες και δεν εντοπίζονται εύκολα μέσω μεμονομένων μοτίβων ή αυτοτελών δομικών περιοχών (domains). Υποστηρίζει την ανίχνευση και χαρακτηρισμών πρωτεϊνικών αλληλουχιών μέσω πολλαπλών στοιχίσεων των προαναφερθέντων μοτίβων, διερεύνηση εξελικτικών σχέσεων και λειτουργικής πρόβλεψης, ενώ ενισχύει την γενικότερη αντίληψη περί των λειτουργικών ρόλων των μελών των πρωτεϊνικών οικογενειών. Παλαιότερα φιλοξενούνταν από το University of Manchester Bioinformatics Education και τώρα έχει ενσωματωθεί στην InterPro ως ένα από τα αποθετήρια της (Εικόνα **1.8**).



Εικόνα 1.8: Αναζήτηση στην InterPro κωδικού “PR00723” που ανήκει στην βάση δεδομένων PRINTS. Φαίνεται χαρακτηριστικά το μήνυμα της InterPro που υποδηλώνει πως η PRINTS έχει πλέον ενσωματωθεί πλήρως στην InterPro. (πηγή: Στιγμιότυπο οθόνης της επίσημης ιστοσελίδας της InterPro https://www.ebi.ac.uk/interpro/entry/prints/#table).

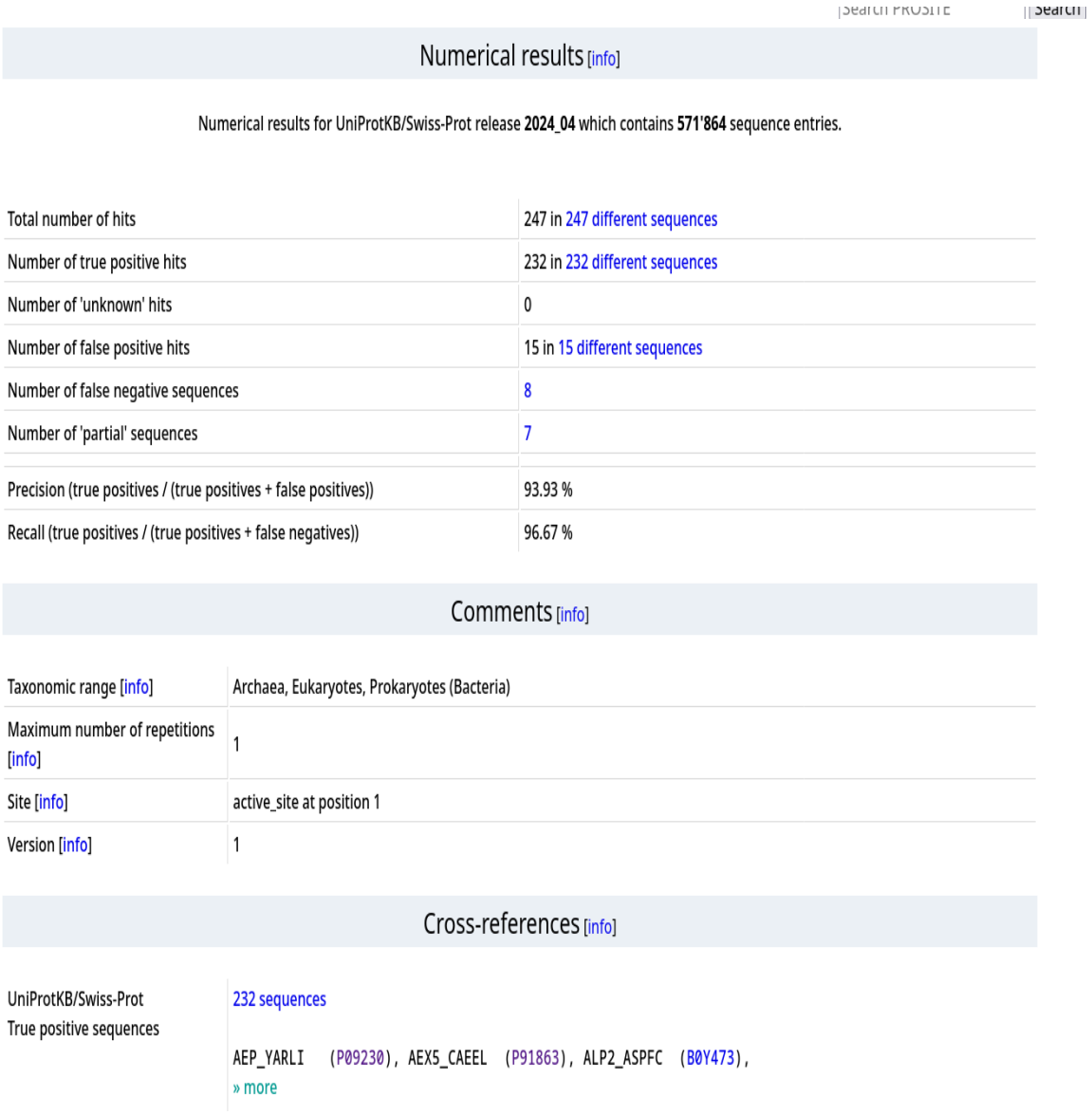
1.2.8. PROSITE

Η βάση δεδομένων PROSITE [25] είναι μια βάση πρωτεϊνικών οικογενειών και αυτοτελών δομικών περιοχών (domains) και αφορμή για την δημιουργία της αποτέλεσε η παρατήρηση ότι οι περισσότερες πρωτεΐνες μπορούν να ομαδοποιηθούν σε περιορισμένο αριθμό οικογενειών βάσει ομοιοτήτων στις αλληλουχίες τους. Επί του παρόντος η PROSITE (PROSITE profiles ως μέλος της InterPro) περιλαμβάνει μοτίβα και προφίλ σχετιζόμενα με περισσότερες από 1000 πρωτεϊνικές οικογένειες και domains, έχει τη δυνατότητα να αναγνωρίζει συγκεκριμένες αμινοξικές αλληλουχίες ως ενδεικτικές για συγκεκριμένες πρωτεϊνικές λειτουργίες και επίσης προσφέρει στον χρήστη λειτουργίες ανάλυσης, επιτρέποντας του την σάρωση δικών του αλληλουχιών έναντι των προφίλ της PROSITE με στόχο την αναγνώριση και πρόβλεψη λειτουργικά σημαντικών περιοχών (Εικόνα **1.9**). Τα πρότυπα της PROSITE ορίζονται με τις εξης συμβάσεις:

- Για την περιγραφή των αμινοξέων χρησιμοποιούνται μονογράμματοι κωδικοί από τα IUPAC.
- Όπου ‘x’ συμβολίζεται θέση στην οποία οποιοδήποτε αμινοξύ είναι αποδεκτό.
- Μέσα σε αγκύλες ‘[ ]’ συμβολίζονται όλες οι επιλογές αμινοξέων για τη συγκεκριμένη θέση, για παράδειγμα ‘[ALT]’ σημαίνει ότι σε αυτή τη θέση μπορεί να υπάρχει είτε Αλανίνη, είτε Λευκίνη, είτε Θρεονίνη (Ala / Leu / Thr).

- Κάθε στοιχείο της αλληλουχίας διαχωρίζεται από τα γειτονικά του με χρήση '-'.
- Επαναλήψεις στοιχείων μπορούν να υποδεικνύονται αν το στοιχείο ακολουθείται με αριθμούς εντός παρενθέσεων ή διαστήματα αριθμών. Για παράδειγμα: x(3) αναπαριστά x-x-x, x(2,4) αναπαριστά x-x ή x-x-x ή x-x-x-x.
- Όταν ένα μοτίβο περιορίζεται είτε στο N- ή C- άκρο μιας αλληλουχίας, τότε αυτό το μοτίβο θα αρχίζει με '<' ή θα τελειώνει με '>' αντίστοιχα. Σε σπάνιες περιπτώσεις το '>' μπορεί να εμφανιστεί και εντός αγκύλης '[ ]'.
- Το μοτίβο τελειώνει με την εμφάνιση τελείας '.'.

Η PROSITE δημιουργήθηκε και συντηρείται από το Swiss Institute of Bioinformatics (SIB).

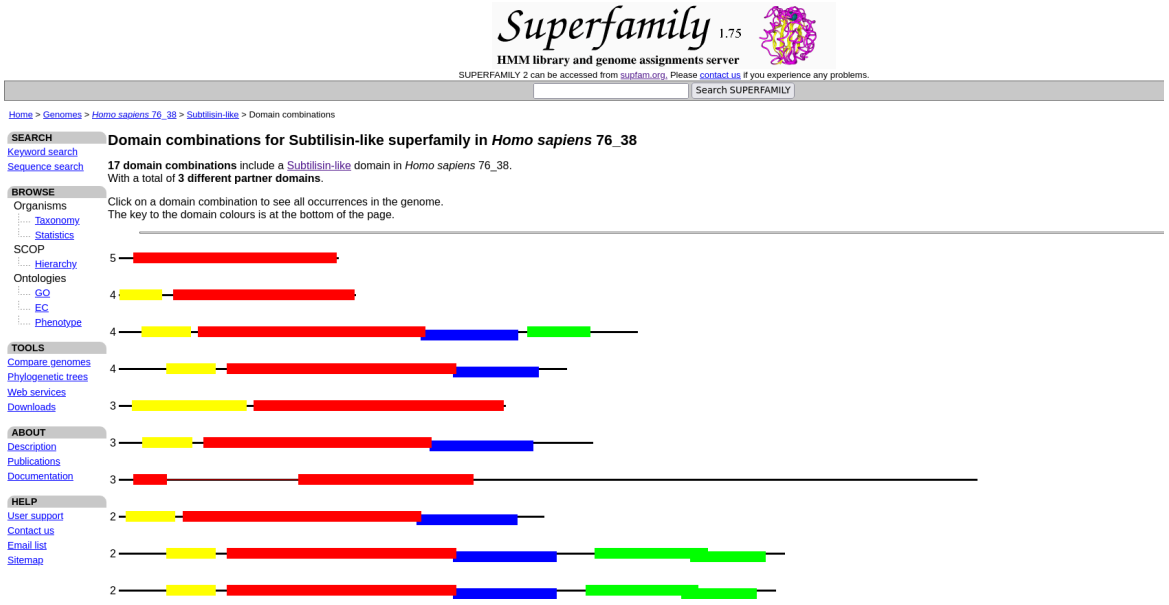


Εικόνα 1.9: *Αποτελέσματα αναζήτησης για την καταχώρηση στην PROSITE του κωδικού "PS00137" που αποτελεί μέλος της οικογένειας σουμπτιλισίνης κι αποτελεί ένζυμο ενδιαφέροντος Nattokinase. Στο επάνω μέρος της σελίδας περιήγησης φαίνεται*

χαρακτηριστικά το PROSITE pattern - ανάμεσα σε άλλες γενικές πληροφορίες για την καταχώρηση. Στο κάτω μέρος εμφανίζονται τα αποτελέσματα της αναζήτησης του συγκεκριμένου μοντέλου έναντι της ενημερωμένης έκδοσης UniProtKB/Swiss-Prot, με λεπτομερή στοιχεία ακρίβειας (λόγος αληθώς θετικών προς άθροισμα αληθώς και ψευδώς θετικών), ταξονομική πληροφορία καθώς και διασταυρούμενες αναφορές. (πηγή: Στιγμιότυπο οθόνης από το επίσημο ιστότοπο της PROSITE https://prosite.expasy.org/PS00137).

1.2.9. SUPERFAMILY

Η βάση δεδομένων SUPERFAMILY [26], [27] είναι μια βάση με χαρακτηρισμένες δομικές και λειτουργικές ιδιότητες πρωτεϊνών και γονιδιωμάτων. Ο σχολιασμός γίνεται βάσει μιας βιβλιοθήκης Μαρκοβιανών Μοντέλων (HMMs), η οποία αντιπροσωπεύει αυτοτελείς δομικές περιοχές (domains) πρωτεϊνών στο επίπεδο υπεροικογένειας (superfamily) σύμφωνα με την ταξινόμηση της βάσης SCOP [28], [29]. Στο επίπεδο υπεροικογένειας ομαδοποιούνται domains που μοιράζονται κοινά εξελικτικά μονοπάτια. Η SUPERFAMILY παρέχει εργαλεία ανάλυσης για την διάκριση υπο- ή υπερ εκπροσώπησης των domains ανάμεσα στα γονιδιώματα, κατασκευή φυλογενετικών δέντρων, ανάλυση της κατανομής των domains υπερ οικογενειών και οικογενειών κατά μήκος της βιόσφαιρας και δεδομένα οντολογίας για τα domains και τα δομικά στοιχεία που περιλαμβάνει (Εικόνα **1.10**).



Εικόνα 1.10: *Αποτελέσματα αναζήτησης του κωδικού "SSF52743" που απαντά στην οικογένεια των Nattokinase ενζύμων ενδιαφέροντος στην βάση δεδομένων Superfamily. Επιλέχθηκε η καρτέλα domain combinations και από την πληθώρα αποτελεσμάτων επιλέχθηκε ενδεικτικά η καρτέλα Homo Sapiens 76_38. (πηγή : Στιγμιότυπο οθόνης από την επίσημη ιστοσελίδα της Superfamily https://supfam.org/SUPERFAMILY/cgi-bin/allcombs.cgi?genome=hs;sf=52743;;password=;subdomain=n)*

1.3. Ένζυμα και εφαρμογές

Τα τελευταία χρόνια τα ένζυμα κατέχουν σημαντικό ρόλο στην φαρμακευτική και χημική βιομηχανία. Η ικανότητά τους να δεσμεύουν το υπόστρωμα με υψηλή συγγένεια και ειδικότητα καθώς και η ικανότητα μετατροπής του υποστρώματος στο επιθυμητό προϊόν με σχετική ταχύτητα και μηδαμινές παρενέργειες, τα ξεχωρίζουν ως υποψήφια στον σχεδιασμό φαρμάκου και τελευταία όλο και περισσότερο βρίσκουν μία ευρεία εφαρμογή στην αντιμετώπιση ασθενειών, από πεπτικές διαταραχές μέχρι καρδιαγγειακές παθήσεις και καρκίνο. Επιπρόσθετα, η πρόοδος της βιοτεχνολογίας με τις νέες τεχνικές στη γενετική τροποποίηση και το ανασυνδυασμένο DNA διευκολύνει τον σχεδιασμό και την κατασκευή ενζύμων ή την ενίσχυση - τροποποίηση αυτών και των υποστρωμάτων τους ανάλογα με τους επιθυμητούς στόχους. Η γενικότερη επίγνωση των αντιδράσεων οι οποίες είναι δόκιμες για βιοκατάλυση (όπως βιοκαταλυτική ρετροσυνθεση), συνδυασμένη με την ανακάλυψη νέων ενζύμων μέσω χρήσης σύγχρονων μεθόδων βιοπληροφορικής και υψηλής απόδοσης τεχνικών (high-throughput screening) καθώς και τεχνικές πρωτεϊνοσύνθεσης, έχουν οδηγήσει σε συνεχώς αυξανόμενα ποσοστά προβλεψιμότητας και εμπιστοσύνης στον σχεδιασμό και την εφαρμογή μικρών μορίων καθώς και την δημιουργία και χρήση πολυ-ενζυμικών σχημάτων και διαδικασιών στα συνθετικά αυτά μόρια - φαρμακευτικούς στόχους [30], [31].

Επιπλέον, λόγω της καταλυτικής τους δράσης, τα ένζυμα εδώ και δεκαετίες χρησιμοποιούνται σε μια πληθώρα χημικών εφαρμογών (πχ εδώ και 30 χρόνια χρησιμοποιούνται στα απορρυπαντικά). Είναι εξαιρετικά ισχυρά επειδή συνδυάζουν την ικανότητα ελέγχου της επιλεκτικότητας και τη λειτουργία του καταλύτη σε ένα μόνο αντιδρών, επιτρέποντας τη χρήση τους σε συνδυασμό με άλλα ένζυμα σε μία ενιαία αντίδραση. Τα τελευταία 20 χρόνια, τα συνδυασμένα συνθετικά-ενζυμικά συστήματα έχουν καθιερωθεί στη βιομηχανία, επιτρέποντας την επιτυχημένη σύνθεση πολύπλοκων μορίων. Η διαθεσιμότητα ενζύμων για χημικές συνθέσεις έχει αυξηθεί, με αποτέλεσμα οι ερευνητές να έχουν στη διάθεσή τους μια ποικιλόμορφη εργαλειοθήκη ενζύμων [32].

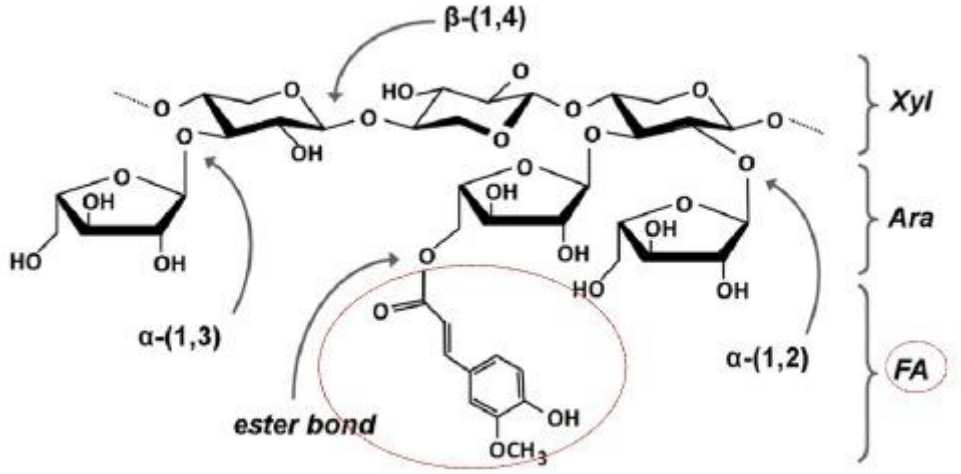
Το 2014, η ανάπτυξη μιας συνολικής ενζυμικής σύνθεσης του νουκλεοσιδίου διδανοσίνης ανέδειξε τη δυνατότητα της «βιο-ρετροσύνθεσης», η οποία περιλαμβάνει τον σχεδιασμό τεχνητών ενζυμικών αλληλουχιών για την παραγωγή επιθυμητών μορίων. Η πλήρης βιοκαταλυτική σύνθεση του αναστολέα του HIV, ισλατραβίρη, δείχνει τη δύναμη του συνδυασμού σύγχρονων τεχνικών στον σχεδιασμό νέων ενζυμικών αλληλουχιών, επαναχρησιμοποιώντας γνωστά βιοσυνθετικά μονοπάτια και εφαρμόζοντας κατευθυνόμενη εξέλιξη για αύξηση της σταθερότητας και της αποδοτικότητας των ενζύμων [31], [32].

Γενικότερα, η βιο καταλυτική δράση των ενζύμων, τα έχει φέρει στο προσκήνιο ως σημαντικούς βιοτεχνολογικούς στόχους και συνεχείς έρευνες πραγματοποιούνται για την περαιτέρω διερεύνηση της χρήσης τους. Αποτελεί μεθοδολογία τελευταίας γενιάς, η ανίχνευση ενζύμων ενδιαφέροντος σε μεταγονιδιωματικές βάσεις δεδομένων όπως αυτές που περιγράφονται στο παράρτημα 1.2, η πρόβλεψη της στερεοδιαταξης αυτών με εργαλεία όπως το AlphaFold, η προσομοίωση πρόσδεσης (docking) και τέλος η πειραματική διαδικασία επιλογής των κατάλληλων πρωτεϊνών [33], [34].

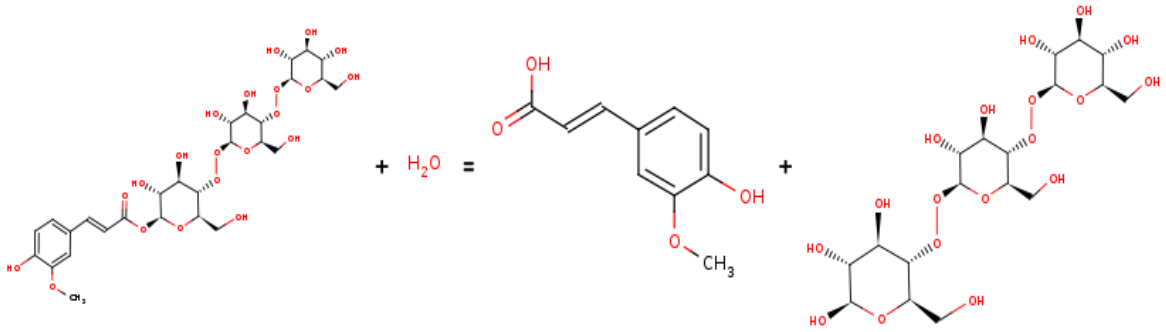
Η παρούσα εργασία, αφιερώθηκε στην ανακάλυψη τέτοιων ενζυμικών στόχων, χρησιμοποιώντας βιοπληροφορικά εργαλεία και μεθόδους, επικεντρώνοντας σε 4 ομάδες ενζύμων. Στην επόμενη ενότητα θα περιγραφούν αυτές οι 4 κατηγορίες ενζύμων.

1.3.1. Εστεράσες του Φερουλικού Οξέος - Feruloyl Esterases

Οι εστεράσες του φερουλικού οξέος (ΕΦΟ, Feruloyl Esterases, EC 3.1.1.73) είναι ένζυμα που χρησιμοποιούνται από βακτήρια, μύκητες και άλλους μικροοργανισμούς για την αποδόμηση των φυτικών πολυσακχαριτών, απελευθερώνοντας το φερουλικό οξύ (Ferulic Acid - FA) και άλλα αρωματικά οξέα από τα κυτταρικά τοιχώματα, το κυτταρόπλασμα και συνθετικά υποστρώματα των φυτών [35]. Ανήκουν στην οικογένεια των υδρολασών ειδικά εκείνων που επιδρούν σε καρβοξυλικούς εστερικούς δεσμούς. Αυτές οι εστεράσες έχουν κεντρίσει το ενδιαφέρον της βιομηχανίας, ιδίως στους τομείς των τροφίμων, του χάρτου και των βιοκαυσίμων, λόγω της ικανότητάς τους να αυξάνουν τη βιοδιαθεσιμότητα του φερουλικού οξέος (ΦΟ) κατά την αποικοδόμηση της φυτικής βιομάζας [36] καθιστώντας τες κατά αυτό τον τρόπο πολύτιμες για μια σειρά βιοτεχνολογικών εφαρμογών όπως: την εξαγωγή ΦΟ από παραπροϊόντα που προέρχονται από γεωργικά απόβλητα αγροτικής βιομηχανίας και τον περαιτέρω βιοσχηματισμό του σε αρωματικές ενώσεις ως προσθετικά υψηλής αξίας, την ενζυμική υδρόλυση για παραγωγή βιοαιθανόλης, την βιολογική αποδόμηση λιγνίνης από μη ξυλώδη φυτά ως πρώτη ύλη στην βιομηχανία χάρτου καθώς και για την κατάλυση εστεροποίησης και μετεστεροποίησης υδροξυκινναμικών οξέων [35].



Εικόνα 1.11: *Εστερικός δεσμός Φερουλικού Οξέος με δομή αραβινοξυλάνης σε πίτουρο σιταριού. Με κόκκινο επισημαίνεται το χημικό μόριο Φερουλικού Οξέος* [37].

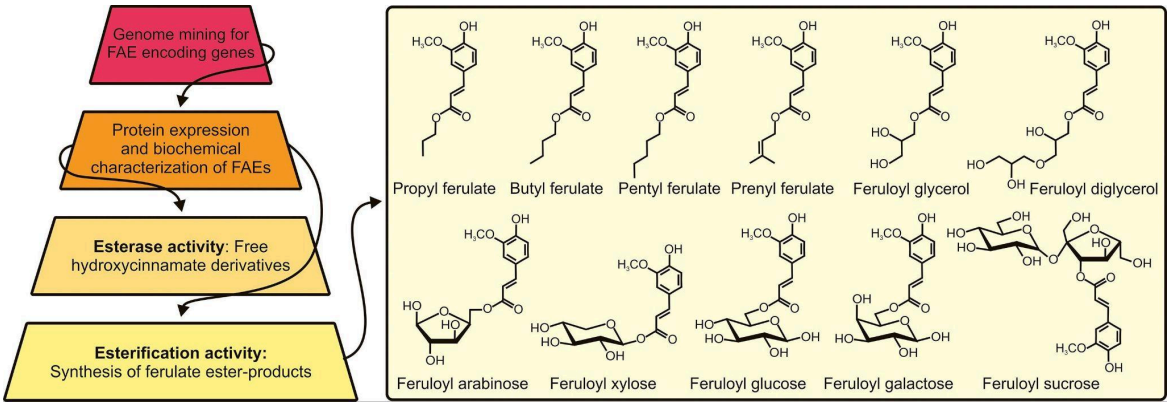


Εικόνα 1.12: *Σχηματική αναπαράσταση της αντίδρασης που καταλύουν οι Εστεράσες του Φερουλικού οξέος :* ***Feruloyl-polysaccharide + H2O = ferulate (ferulic acid) + polysaccharide*** *(πηγή:* [38]*).*

Το ΦΟ είναι ένα υδροξυκινναμικό οξύ με τη συστηματική ονομασία (3-μεθοξυ-4-υδροξυ)-3-φαινυλ-2-προπενικό οξύ ή 3-μεθοξυ-4-υδροξυ-κινναμωμικό οξύ, που απαντάται σε αφθονία στα φυτά και έχει γνωστές αντιοξειδωτικές, αντιδιαβητικές, αντιφλεγμονώδεις, αντικαρκινικές και αντιυπερτασικές ιδιότητες [37], [39]. Ο χημικός τύπος του είναι $C_{10}H_{10}O4$ και έχει Μοριακό Βάρος 94.18 g/mol. Στα φυτά, συνδέεται με εστερικούς και αιθερικούς δεσμούς στις αραβινοξυλάνες (Εικόνα **1.11**) και τη λιγνίνη των κυτταρικών τοιχωμάτων. Αυτοί οι δεσμοί καθιστούν το φερουλικό οξύ λιγότερο διαθέσιμο. Οι ΕΦΟ διασπούν αυτούς τους δεσμούς, απελευθερώνοντας το φερουλικό οξύ σε ελεύθερη μορφή (Εικόνα **1.12)** [40], [41]. Η καταλυτική δράση των ΕΦΟ επιτρέπει την

παραγωγή πολύτιμων εστεροποιημένων ενώσεων φερουλικού οξέος διευρύνοντας το πεδίο εφαρμογής τους με σύγχρονες βιοτεχνολογικές προσεγγίσεις (Εικόνα **1.13)** [35].

Η εφαρμογή των ΕΦΟ δεν περιορίζεται μόνο στη βιομηχανία τροφίμων, όπου χρησιμοποιούνται για τον εμπλουτισμό των προϊόντων με φερουλικό οξύ, το οποίο μπορεί να βελτιώσει τις αντιοξειδωτικές ιδιότητες των τροφίμων. Στον τομέα της βιοενέργειας, οι ΕΦΟ συνδυασμό με άλλα βοηθητικά ένζυμα μπορούν να συμβάλλουν στη βελτίωση της αποδοτικότητας της παραγωγής βιοκαυσίμων μέσω της αποδόμησης των φυτικών κυτταρικών τοιχωμάτων και της απελευθέρωσης υδατανθράκων. Επίσης, στη βιομηχανία χάρτου, η χρήση τους μπορεί να βελτιώσει τη διαδικασία παραγωγής χαρτοπολτού, ενώ σύγχρονες μελέτες διερευνούν την εξαγωγή του αντιοξειδωτικού ΦΟ από αραβινοξυλάνες που υπάρχουν σε περίσσεια σε βιομηχανικά απόβλητα αγροτικής παραγωγής (πχ επεξεργασία σιτηρών) και εστεροποίηση του ΦΟ για την δημιουργία παραγόντων με αντιφλεγμονώδεις, αντιοξειδωτικές, αντιβακτηριακές ιδιότητες με εφαρμογή στο πεδίο των καλλυντικών και της φαρμακοκινητικής γενικότερα [36], [41], [42], [43] .



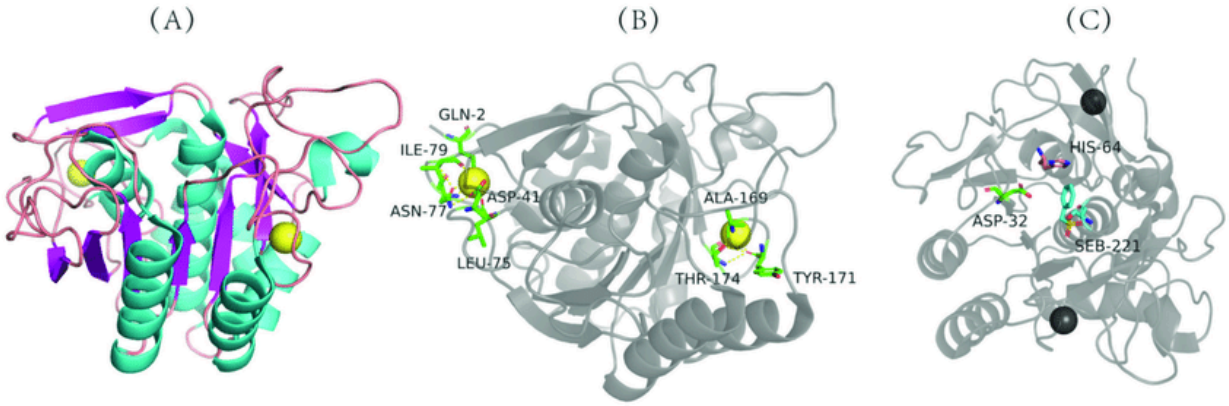
Εικόνα 1.13: *Σύγχρονες προσεγγίσεις στη Βιοτεχνολογία για την παραγωγή ΕΦΟ και την εφαρμογή τους με στόχο παράγωγα του Φερουλικού Οξέος (πηγή:* [35]*).*

1.3.2. Νατοκινάσες - Nattokinase

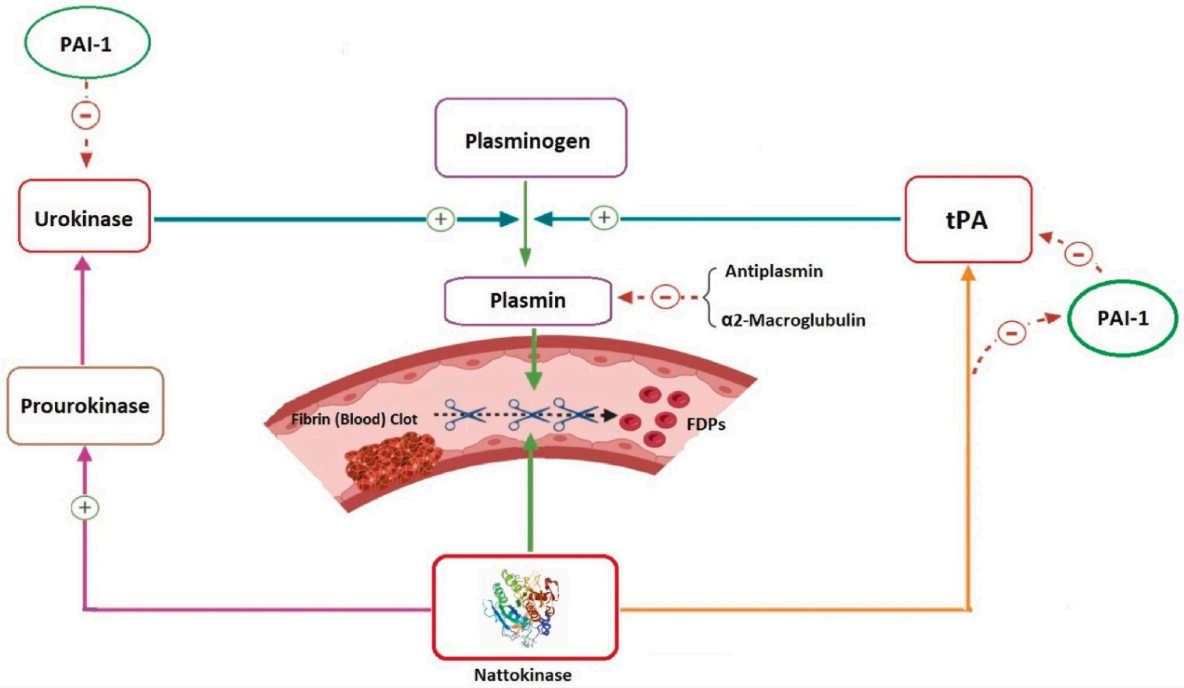
Η Νατοκινάση (NK) (Nattokinase, EC 3.4.21.62), η οποία παρά την ονομασία της δεν πρόκειται για κινάση παρά για μια πρωτεάση σερίνης (serine protease) - έχει επίσης την ονομασία σουμπτιλισίνη NAT (subtilisin NAT) - εξάγεται από το παραδοσιακό Ιαπωνικό τρόφιμο “natto”, ένα προϊόν ζύμωσης σόγιας με *Bacillus subtilis natto*. [44], [45]. Παρουσιάζει έντονη αντιθρομβωτική δράση αφού, όταν έρχεται σε επαφή με ανθρώπινο αίμα ή θρόμβους, αδρανοποιεί τον αναστολέα του ενεργοποιητή του πλασμινογόνου 1 (PAI-1) προκαλώντας κατ’ αυτό τον τρόπο λύση του ινώδους. Παρόλο που θα ήταν αναμενόμενη η πέψη και αδρανοποίηση της στο έντερο του ανθρώπου όπως οι περισσότερες πρωτεΐνες, έρευνες υποδεικνύουν ότι παραμένει ενεργή όταν λαμβάνεται εκ του στόματος [46].

Δεν παρατηρούνται δισουλφιδικοί δεσμοί στην δομή της NK, καθώς πρόκειται για πρωτεάση χωρίς κυστεΐνες ανάμεσα στα 275 αμινοξικά κατάλοιπα που αποτελούν την αλληλουχία της (MB 27.7 kDa pI 8.6). Η τρισδιάστατη δομή της έχει λυθεί επιτυχώς (PDB code: 4DWW [47], Εικόνα **1.14**) και εμφανίζει υψηλή ομοιότητα με αλκαλικές πρωτεάσες σερίνης εντός οικογένειας. Η καταλυτική της δράση παρόλα αυτά, δεν έχει ακόμη προσδιοριστεί.

Αν και η NK παρουσιάζει υψηλή ομολογία με πολλές σουμπτιλισίνες της οικογένειας πρωτεασών σερίνης, λίγες πρωτεΐνες εμφανίζουν υψηλή ειδικότητα ως προς το υπόστρωμα του ινώδους και είναι ικανές να διασπούν άμεσα τις διασυνδεδεμένες ινώδεις δομές *in vitro* και *in vivo* [44]. Αντίθετα, η NK εμφανίζει άμεση ινωδολυτική δράση, με την ικανότητα της να κόβει τους πεπτιδικούς δεσμούς μεταξύ φαινυλαλανίνης και λευκίνης της α-έλικας του ινώδους, να διεγείρει την κυτταρική απελευθέρωση του ενεργοποιητή πλασμινογόνου των ιστών για την αποδόμηση του ινώδους και επίσης ενισχύει την παραγωγή θρομβολυτικών παραγόντων [48] (Εικόνα **1.15)**.

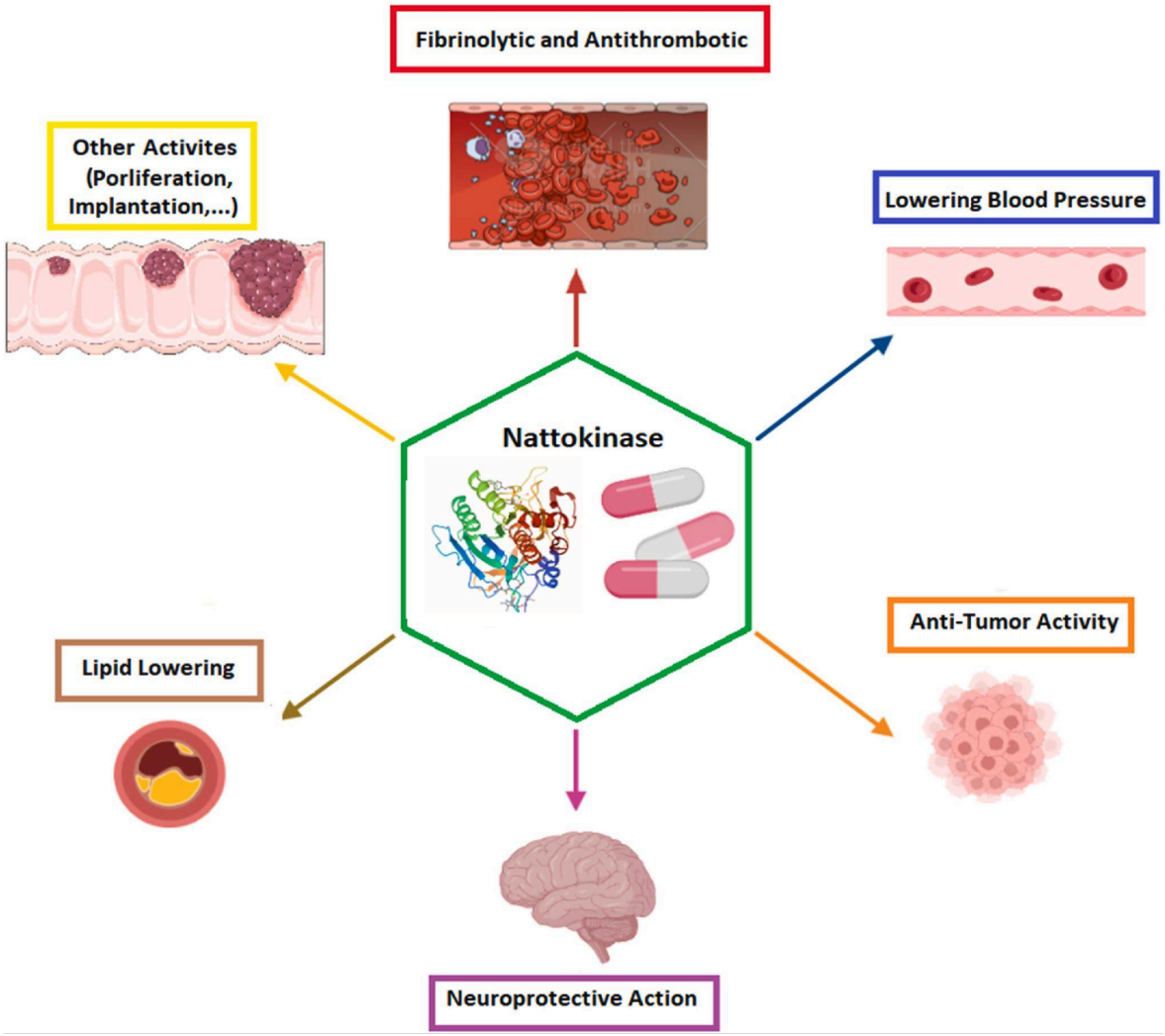


Εικόνα 1.14: *Τριτοταγής δομή της NK στον τρισδιάστατο χώρο. (A) Τρισδιάστατη δομή της NK (B) Θέσεις δέσμευσης ασβεστίου (Gln2, Asp41, Leu75, Asn77, Ile79, Val81, Ala169, Tyr171, Thr174). (C) Ενεργό κέντρο (Asp32, His64, Ser221). Πηγή: PDB code: 4DWW,* [49].



Εικόνα 1.15: *Ινωδολυτική δράση της Ναττοκινάσης. Λύση των θρόμβων του αίματος μέσω υδρόλυσης του ινώδους και του υποστρώματος της πλασμίνης απελευθερώνοντας προϊόντα αποδομής ινώδους. Μετατρέπει την ενδογενή προουροκινάση σε ουροκινάση, αποδομεί τον ενεργοποιητή του πλασμινογόνου 1 (PAI-1) και αυξάνει τα επίπεδα ιστικού πολυπεπτιδικού αντιγόνου (t-PA) διεγείροντας την παραγωγή πλασμίνης από το πλασιμονογόνο, με αποτέλεσμα ακόμη αποτελεσματικότερη αντιπηκτική δράση στο σώμα. (πηγή:* [44]*)*

Πρόσφατες και συνεχόμενες έρευνες για την εφαρμογή της NK, αναφέρουν αποτελεσματική και άνευ ανεπιθύμητων παρενεργειών θεραπεία και πρόληψη καρδιαγγειακών παθήσεων, ανάμεσα στις οποίες το εγκεφαλικό έμφραγμα, το ισχαιμικό εγκεφαλικό επεισόδιο και το έμφραγμα του μυοκαρδίου, τα οποία σχετίζονται όλα με θρόμβους προερχόμενους από σύμπλοκα του ινώδους και των αιμοπεταλίων και η μέχρι τώρα κλινική εφαρμογή θρομβολυτικών παραγόντων, συμπεριλαμβανομένης της ουροκινάσης, του ενεργοποιητή του πλασμινογόνου ιστού (t-PA) και της στρεπτοκινάσης, εμφανίζουν σοβαρές παρενέργειες όπως αιμορραγία ή γαστρικό έλκος.



Εικόνα 1.16: *Πεδία εφαρμογής της NK ως θεραπευτικό φάρμακο (πηγή:* [44]*)*

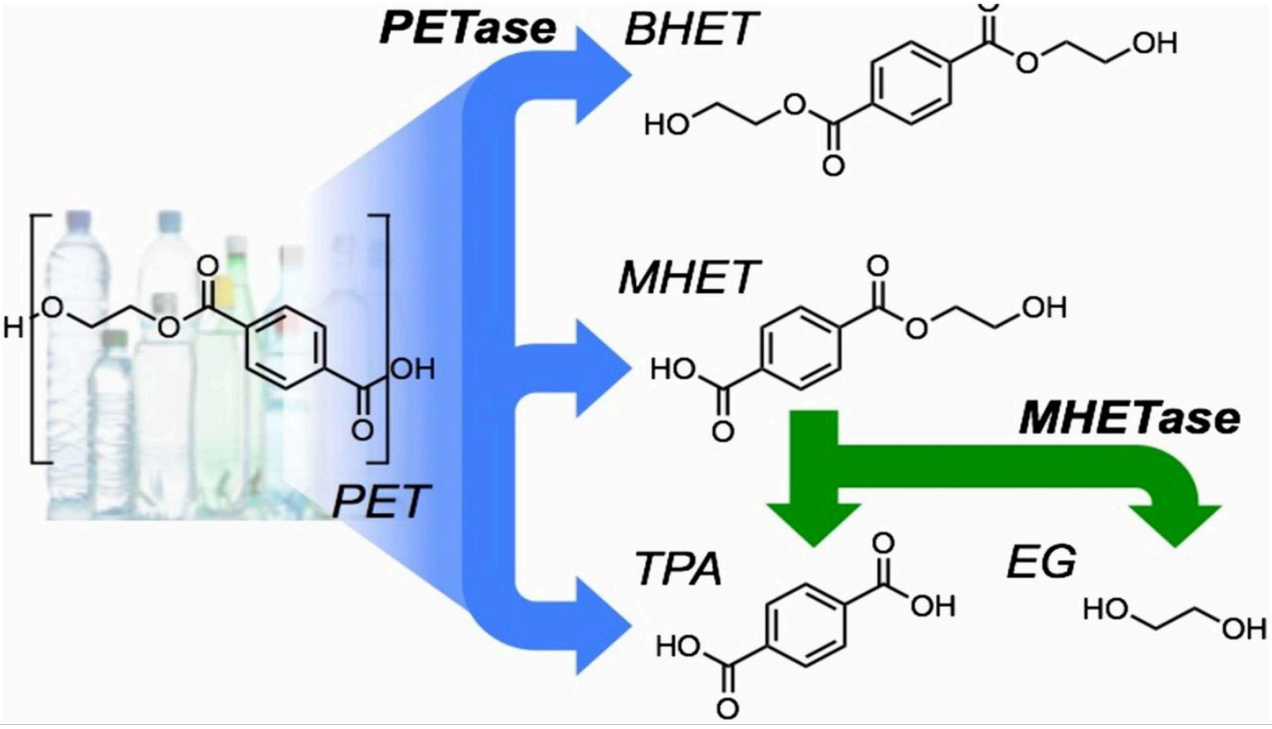
Επιπλέον από τον ρόλο που μπορεί να παίξει στην αντιμετώπιση των καρδιαγγειακών παθήσεων, η NK παρουσιάζει σημαντικές δυνατότητες και για τη μείωση λιπιδίων στο αίμα, την αντιμετώπιση του καρκίνου, τη θεραπεία της υπέρτασης, της νόσου του Αλτσχάιμερ και των πολλαπλασιαστικών διαταραχών του υαλοειδούς, καθιστώντας την

ιδανικό υποψήφιο φαρμακευτικό παράγοντα είτε μεμονωμένα είτε σε συνδυασμό με άλλα αντιθρομβωτικά φάρμακα [44], [49] (Εικόνα **1.16**).

1.3.3. Υδρολάσες του PET (Πετάσες) - Petases, Pet Hydrolases

Πετάσες (Petases), ονομάζονται μια κατηγορία εστερασών, ενζύμων τα οποία καταλύουν την υδρολυση του πλαστικού τερεφθαλικού, πολυαιθυλενίου (ή πολυτερεφθαλικό αιθυλένιο) (PET), σε μονομερές μονο-2-υδροξυαιθυλοτερεφθαλικό εστέρα (MHET). Αυτή τους η ιδιότητα καθιστά την ενζυμική αποδόμηση του PET πλαστικού, πεδίο εντατικής διερεύνησης [50]. Οι ερευνητές αναζητούν νέους βιοκαταλύτες μέσω της χρήσης βιοπληροφορικής και μηχανικής μάθησης, εξετάζουν ένζυμα από διαφορετικές φυλογενετικές ομάδες και αξιολογούν τη δραστικότητά τους σε διάφορες συνθήκες pH και θερμοκρασίας. Επίσης, εξετάζουν την εκλεκτικότητα των ενζύμων σε σχέση με τη μορφολογία του PET. Μέσω της χρήσης κρυσταλλογραφίας ακτίνων X και του AlphaFold, αποκαλύπτουν πρωτεϊνικές πτυχές και βοηθητικές περιοχές που συνδέονται με την αποδόμηση του PET.

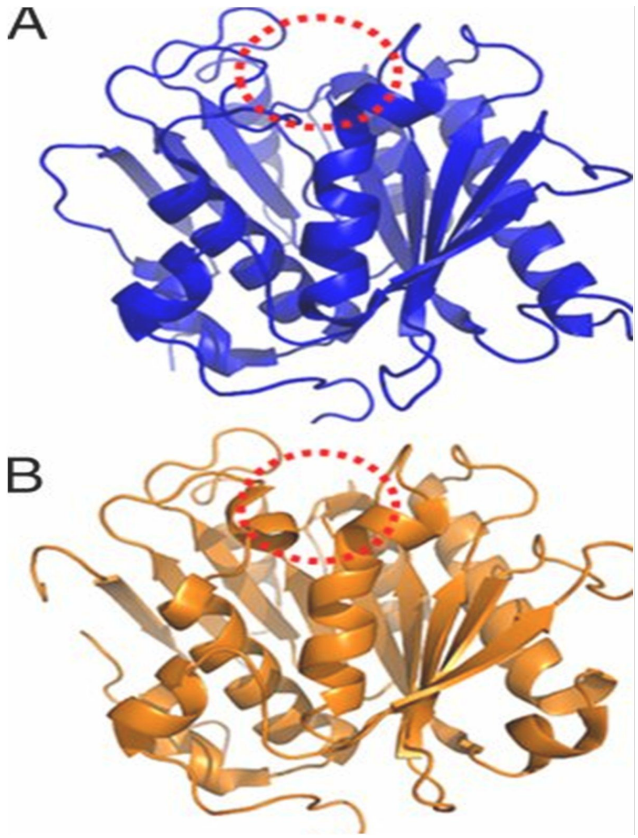
Η χρήση πλαστικού στις μέρες μας έχει εδώ και καιρό πάρει ανεξέλεγκτες διαστάσεις. Το 8% του παγκόσμιου ορυκτού καυσίμου το 2015, δαπανήθηκε για την δημιουργία συνθετικών πολυμερών, ενώ από την παραγωγή 368 εκατομμυρίων τόνων πλαστικού παγκοσμίως, 90% προήλθαν από πετρέλαιο [51]. Πρόσφατες μελέτες δίνουν ελπίδα στην διαχείριση της τεράστιας χρήσης και ως αποτέλεσμα ρύπανσης του περιβάλλοντος, με την απομόνωση βακτηρίων από μικροβιακές κοινότητες σε περιβάλλοντα τα οποία έχουν εκτεθεί σημαντικά σε PET [52]. Τέτοια βακτήρια φαίνεται πώς χρησιμοποιούν PET για ενέργεια, αποδομώντας τα συστατικά του μέσω ενζυμικής δραστηριότητας. Ένα από αυτά τα ένζυμα, η PETase μετατρέπει το PET στα 2 του μονομερή, το τερεφθαλικό οξύ (*MHET*) και αιθυλενική γλυκόλη (Εικόνα **1.17**).



Εικόνα 1.17: *Βιοκατάλυση του αποπολυμερισμού του PET στα αρχικά μονομερή του σε μονο (2-υδροξυλευλ) τερεφθαλικό οξύ (MHET), με δευτερεύοντα προϊόντα κλάσματα τερεφθαλικού οξέος (TPA) και bis(2-hydroxyethyl)-TPA. Ένα δεύτερο ένζυμο, η MHETase επιπλέον μετατρέπει το MHET σε δυο μονομερή, TPA και αιθυλενική γλυκόλη (EG) (πηγή:* [51], [53]*).*

Ο βιοκαταλυτικός αποπολυμερισμός του PET προσφέρει μια φιλική προς το περιβάλλον και ενεργειακά αποδοτική λύση για την ανακύκλωση αυτού του υλικού. Ωστόσο, η πρόοδος προς αυτή την κατεύθυνση εξαρτάται από την ανάπτυξη ενζύμων και διεργασιών που είναι κατάλληλες για βιομηχανική χρήση. Προς αυτή την κατεύθυνση, γίνονται προσπάθειες για τροποποίηση του ενζύμου, αρχικά με την λύση της κρυσταλλικής δομής του , αλλάζοντας στη συνέχεια τις θέσεις πρόσδεσης του υποστρώματος και προκαλώντας μοριακές μεταλλάξεις, με τα αποτελέσματα να παρουσιάζουν σημαντικές δυνατότητες για περαιτέρω βελτίωση της αποδοτικότητας των ενζύμων στην αποδόμηση του πλαστικού PET. Από την λύση της δομής έγινε φανερό πως παρουσιάζει ομοιότητες με τις κουτινάσες και τις λιπάσες, κατέχοντας όμως μια περισσότερο ανοιχτή ενεργή σχισμή (Εικόνα **1.18**). Ένα πρόβλημα που προκύπτει είναι η υψηλή περιεκτικότητα του υποστρώματος σε στερεά, η οποία μειώνει την αποδοτικότητα των ενζύμων στην αρχική φάση της υδρόλυσης. Επιπλέον, η αύξηση του φορτίου στερεών σε βιομηχανικά επίπεδα μπορεί να μειώσει την απόδοση του αποπολυμερισμού. Είναι σημαντικό να ληφθούν υπόψη πολλοί παράγοντες που επηρεάζουν την αποδοτικότητα των ενζύμων σε πραγματικές συνθήκες. Ένας παράγοντας που συχνά παραβλέπεται είναι

η διαφοροποίηση της ενζυμικής υδρόλυσης μεταξύ εργαστηριακών πειραμάτων και βιομηχανικής παραγωγής. Συνεπώς, απαιτείται περαιτέρω έρευνα και ανάπτυξη για την αποδοτική βιομηχανική αποικοδόμηση του PET [54].



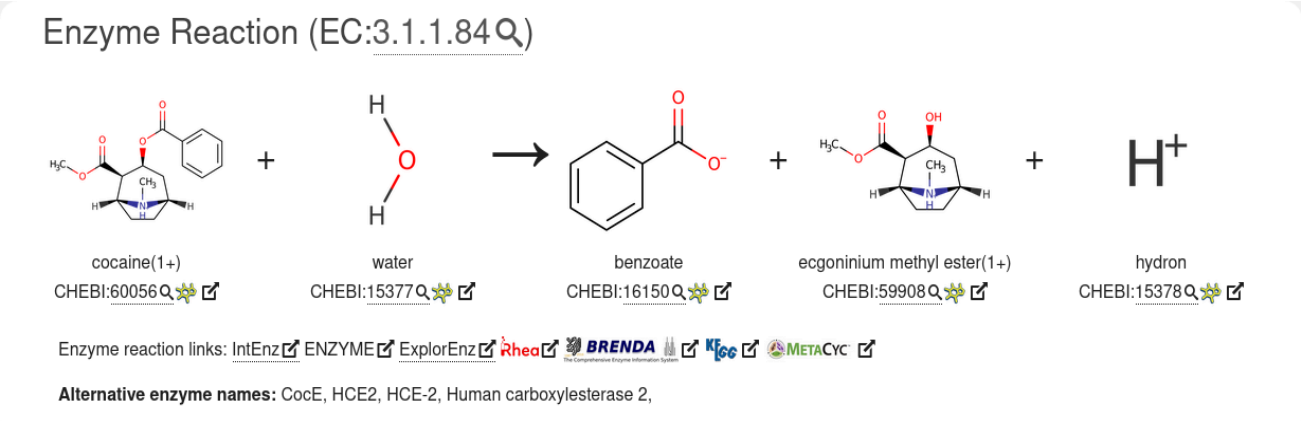
Εικόνα 1.18: *(A) Τριτοταγής δομή της PETase σε ανάλυση 0.92 Å [Protein Data Bank (PDB) ID code 6EQE]. Με κόκκινη διακεκομένη γραμμή επισημαίνεται η ενεργή σχισμή στο επάνω μέρος της δομής. (B) Τριτοταγής δομή της κουτινάσης T. fusca (PDB ID code 4CG1) πηγή:* [51].

1.3.4. Εστεράσες της Κοκαΐνης - Cocaine Esterases

Η διαταραχή της χρήσης κοκαΐνης (Cocaine Use Disorder - CUD) αποτελεί κρίσιμο δημοσιονομικό πρόβλημα με κοινωνικές, οικονομικές και φυσικά υγειονομικές πτυχές σε παγκόσμιο επίπεδο. Σύμφωνα με τα στατιστικά του 2019, 20 εκατομμύρια άνθρωποι είχαν κάνει χρήση κοκαΐνης παγκοσμίως με 6,9 εκατομμύρια από αυτούς να εντοπίζονται στη Βόρεια Αμερική. Γενικότερα, η κατάχρηση ουσιών και οι βλαβερές επιπτώσεις στην υγεία λόγω αυτής, επιδεινώθηκε περαιτέρω κατά τη διάρκεια της πανδημίας COVID-19, ενώ ειδικότερα ο αριθμός των χρηστών κοκαΐνης και των θανάτων από υπερβολική δόση αυτής παρουσιάζει σταθερή άνοδο τα τελευταία χρόνια. Δεδομένου ότι δεν υπάρχει εγκεκριμένη φαρμακοθεραπεία από τον ΠΟΦ (FDA) για τα άτομα με CUD, οι ασθενείς παραμένουν χωρίς ιδιαίτερη υγειονομική φροντίδα [55].

Οι συμβατικές στρατηγικές φαρμακοδυναμικής οι οποίες έχουν σχεδιαστεί για την αντιμετώπιση των επιπτώσεων της κοκαΐνης έχουν αποδειχθεί εξαιρετικά δυσεφάρμοστες. Η ανάπτυξη καινοτόμων θεραπευτικών προσεγγίσεων είναι απαραίτητη. Πρόσφατες μελέτες υποδεικνύουν ότι η χρήση ενός ενζύμου για την επιτάχυνση του μεταβολισμού της κοκαΐνης αποτελεί μια πιθανή τακτική [55], [56], [57].

Η εστεράση της κοκαΐνης (CocE) (EC 3.1.1.84) προέρχεται από ένα στέλεχος *Rhodococcus* που ζει στο έδαφος του φυτού *Erythroxylum coca*, το οποίο παράγει κοκαΐνη και είναι σε θέση να χρησιμοποιεί την κοκαΐνη ως μοναδική πηγή άνθρακα και αζώτου για την ανάπτυξή του.

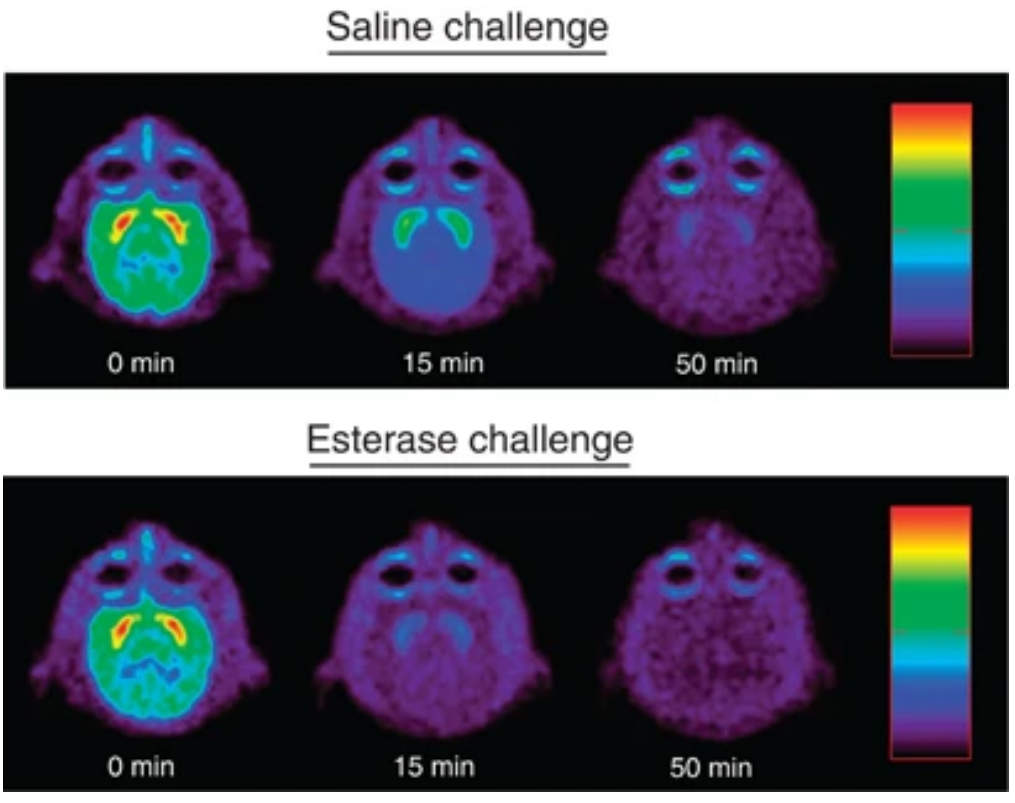


Εικόνα 1.19: *Σχηματική αναπαράσταση της ενζυμικής αντίδρασης που καταλύει η Εστεράση της Κοκαΐνης:* ***cocaine + H2O ⇌ ecgonine methyl ester + benzoate***
(πηγή: επίσημη ιστοσελίδα Mechanism and Catalytic Site Atlas, EMBL EBI https://www.ebi.ac.uk/thornton-srv/m-csa/entry/465/)

Η CocE ανήκει στην υπεροικογένεια των α/β υδρολασών, συγκεκριμένα στην οικογένεια CocE/σερινοεστεράσης. Μια περιοχή της αμινοξικής της αλληλουχίας εμφανίζει

ομοιότητα με την ενεργή περιοχή σερίνης των *X-prolyl dipeptidyl aminopeptidases* γεγονός που υποδηλώνει ότι η εστεράση της κοκαΐνης είναι μια εστεράση σερίνης. Με 15 καταχωρημένες δομές της CocE στην PDB και αρκετές πληροφορίες σχετικές με τον ενζυμικό μηχανισμό της, αποτελεί ιδανικό στόχο για ανάπτυξη φαρμάκου για τη θεραπευτική αντιμετώπιση κατάχρησης κοκαΐνης.

Η CocE υδρολύει την κοκαΐνη σε βενζοϊκό και μεθυλεστέρα εκγονίνης (Εικόνα **1.19**). Επίσης υδρολύει αποτελεσματικά το κοκααιθυλένιο, έναν πιο ισχυρό μεταβολίτη της κοκαΐνης που έχει εντοπιστεί σε ασθενείς που κάνουν ταυτόχρονα κατάχρηση κοκαΐνης και αλκοόλ. Σε πειράματα που έχουν γίνει σε ζώα, έχει φανεί πως η CocE είναι σε θέση να αποτρέψει τους σπασμούς και τη θνησιμότητα που προκαλούνται από την κοκαΐνη, επιτυγχάνοντας την απομάκρυνσή της από τον εγκέφαλο έως και τρεις φορές γρηγορότερα εαν χορηγηθεί περιφερικά CocE (Εικόνα **1.20**) . Ως εκ τούτου, η CocE παρουσιάζει ενδιαφέρον ως ανταγωνιστής της κοκαΐνης σε περιπτώσεις οξείας υπερδοσολογίας, ανακουφίζοντας ή και προλαμβάνοντας δυσμενείς επιδράσεις αυτής στο Κεντρικό Νευρικό Σύστημα, ενώ αντίστοιχες μελέτες έχουν δείξει ότι η CocE επίσης προστατεύει από καρδιαγγειακές παρενέργειες λόγω χρήσης κοκαΐνης [58].



Εικόνα 1.20: Δείγμα εικόνων από PET scan από εγκέφαλο μαϊμούς δείχνει την πρόσληψη κοκαΐνης και την επακόλουθη αποβολή της μετά από χορήγηση φυσιολογικού ορού (επάνω) και CocE (κάτω) σε διαφορετικές χρονικές στιγμές. Η συνέχεια των εικόνων παρουσιάζεται από αριστερά προς τα δεξιά στο οριζόντιο επίπεδο, ενώ με χρώμα

διακρίνεται η πυκνότητα της ουσίας στους τομείς του εγκεφάλου. CocE, cocaine esterase; PET, positron emission tomography. Πηγή: Fig. 2 [58].

1.4. Στρατηγικές ανάλυσης μεταγονιδιωματικών αλληλουχιών

1.4.1. Profile Hidden Markov Models

Από τη δεκαετία του 1990 μέχρι και σήμερα, η χρήση των προφιλ (profile) Hidden Markov Models (pHMMs) για την αναζήτηση ομοιότητας αλληλουχιών και ιδιαίτερα για τον εντοπισμό απομακρυσμένων ομολογιών (remote homologies), είναι ιδιαίτερα διαδεδομένη. Ανάμεσα σε άλλα, η δημοτικότητά τους έγκειται στο γεγονός πως μια πολλαπλή στοίχιση ενός σχετικά μικρού πλήθους αλληλουχιών, με εμφανή ομοιότητα, μπορεί να οδηγήσει στην κατασκευή μοντέλου - profile HMM, το οποίο με τη σειρά του αξιοποιείται ως μήτρα για την ανίχνευση ομοιοτήτων και ομολογίας σε μεγάλες βάσεις δεδομένων. Πληθώρα τέτοιων βάσεων δεδομένων μάλιστα, έχει υλοποιήσει αυτή τη μέθοδο για την αναπαράσταση πρωτεϊνικών οικογενειών, ανάμεσά τους οι Pfam [59], CATH-Gene3D [60], TIGRFAMs [22], Superfamily [27], PIRSF [61] και η TreeFam [62]. [59], [63], [64]

Στην συγκεκριμένη εργασία, ο μεγαλύτερος όγκος της αναζήτησης των πρωτεϊνικών domain ενδιαφέροντος από το συγκεντρωτικό αποθετήριο της UniParc, πραγματοποιήθηκε με τη χρήση εργαλείων εφαρμογής profile HMMs - κυρίως της σουίτας εργαλείων HMMER [64], [65]. Ως εκ τούτου κρίνεται αναγκαία μια συνοπτική αναφορά στο θεωρητικό υπόβαθρο των μαθηματικών Μαρκοβιανών και κρυπτομαρκοβιανών μοντέλων (Hidden Markov Models) [66] καθώς και των βασικών αλγορίθμων οι οποίοι εφαρμόζονται σε αυτά τα μοντέλα.

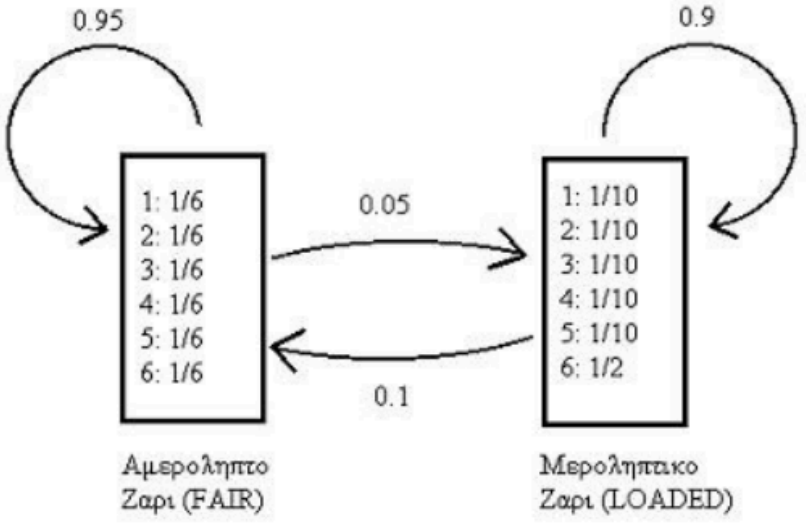
Τα στοχαστικά πιθανοθεωρητικά μοντέλα εξάρτησης Markov ή αλλιώς αλυσίδες Markov έχουν θεωρηθεί ιδανικά για την διερεύνηση και περιγραφή μεγαλομορίων (DNA, πρωτεϊνών) καθώς είναι ικανά να προσεγγίσουν την πληροφορία εντός μιας ακολουθίας [64], [66] και απόδειξη αποτελεί η ποικιλία εφαρμογών τους στο παρελθόν, όπως για εύρεση γονιδίων σε βακτήρια και ευκαρυωτικούς οργανισμούς, για την εύρεση μοτίβων σε βιολογικές αλληλουχίες, για τον εντοπισμό οριζόντιας γονιδιακής μεταφοράς, για την ταξινόμηση πρωτεϊνικών αλληλουχιών κ.α.[66].

Τα Hidden Markov Models (HMMs), επεκτείνουν την ιδέα των αλυσίδων Markov, οι οποίες περιγράφουν ένα σύστημα όπου η μελλοντική κατάσταση του συστήματος εξαρτάται μόνο από την παρούσα κατάσταση και όχι από το παρελθόν (ιδιότητα Markov), δηλαδή η πιθανότητα μετάβασης σε μια νέα κατάσταση εξαρτάται μόνο από την τρέχουσα

κατάσταση, επιτρέποντας την ύπαρξη κρυφών καταστάσεων που δεν μπορούν να παρατηρηθούν άμεσα. Αντίθετα, παρατηρούμε δεδομένα που εξαρτώνται από αυτές τις κρυφές καταστάσεις. Τα βασικά στοιχεία ενός HMM είναι:

1. **Κρυφές Καταστάσεις (Hidden States):** Αντιπροσωπεύουν την πραγματική κατάσταση του συστήματος, η οποία δεν είναι απευθείας παρατηρήσιμη.
2. **Παρατηρήσεις (Observations):** Τα δεδομένα που μπορούμε να παρατηρήσουμε, τα οποία σχετίζονται με τις κρυφές καταστάσεις.
3. **Μεταβάσεις Καταστάσεων (State Transitions):** Οι πιθανότητες μετάβασης από μία κατάσταση σε μια άλλη (παρόμοια με τις αλυσίδες Μαρκοβ).
4. **Πιθανότητες Εκπομπής (Emission Probabilities):** Οι πιθανότητες να παρατηρηθούν συγκεκριμένα δεδομένα από κάθε κατάσταση.
5. **Αρχικές Πιθανότητες (Initial Probabilities):** Οι πιθανότητες να ξεκινά το σύστημα από μια συγκεκριμένη κατάσταση.

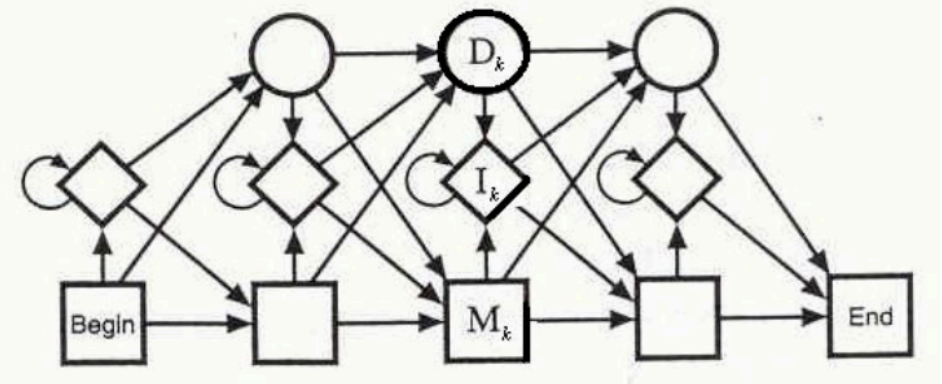
Συνοπτικά, τα HMM επεκτείνουν τις αλυσίδες Μαρκοφ προσθέτοντας μια διάσταση κρυφών καταστάσεων, επιτρέποντας την ανάλυση πιο πολύπλοκων και ρεαλιστικών μοντέλων για δεδομένα που παρουσιάζουν αλληλεξαρτήσεις. Πλέον οι παρατηρήσεις αποδεσμεύονται από τις καταστάσεις και η σύνδεση των παρατηρήσεων και των καταστάσεων, καθώς και η μετάβαση από μια κατάσταση σε μιά άλλη, υπολογίζεται μέσω πιθανοτήτων. Παράδειγμα ενός τέτοιου συστήματος HMM αποτελεί αυτό του “ανέντιμου καζίνο” (dishonest casino) [67] (Εικόνα **1.21**) όπου υποτίθεται το καζίνο χρησιμοποιεί ανά την περίσταση αμερόληπτα ζάρια ή μεροληπτικά χωρίς όμως να γνωρίζουμε πότε χρησιμοποιεί ποιό είδος (κρυφές καταστάσεις).



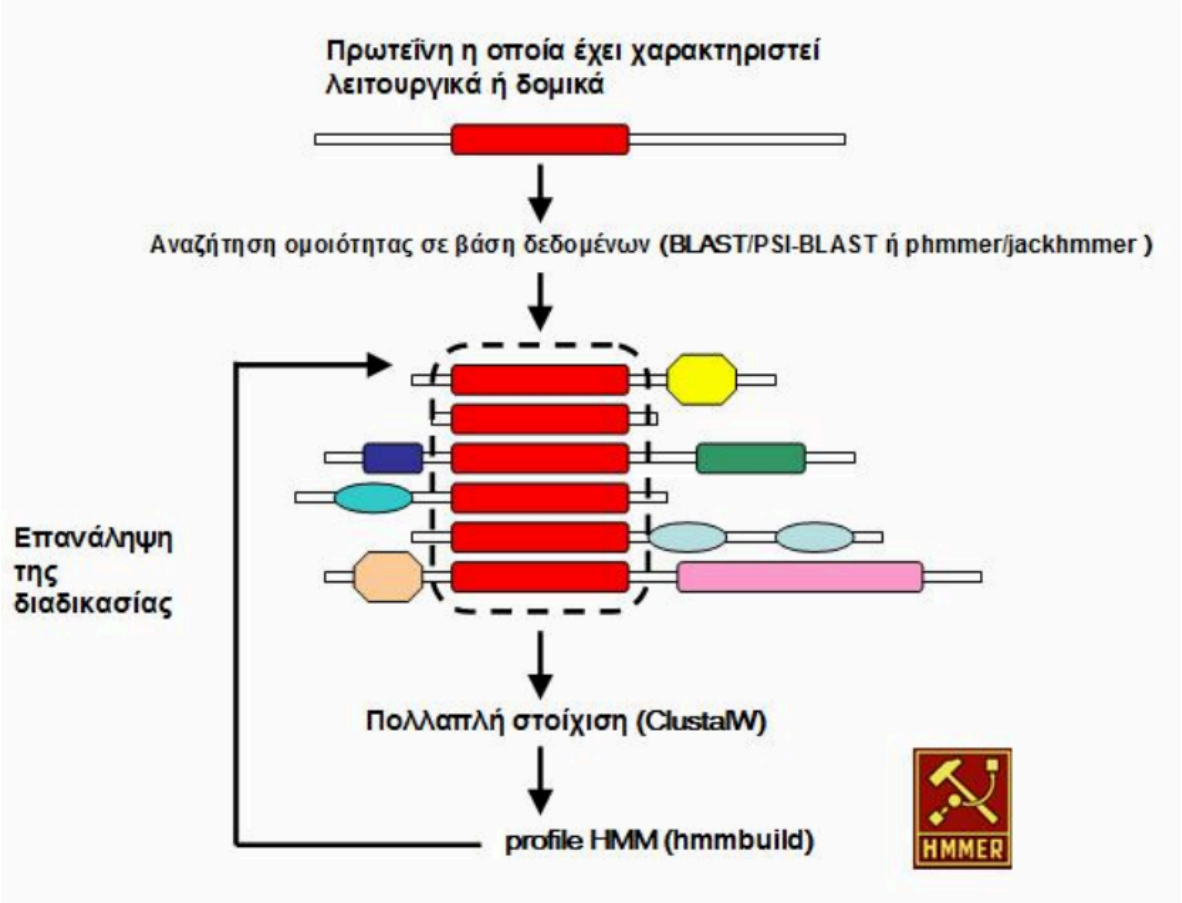
Εικόνα 1.21: *Το παράδειγμα του 'ανέντιμου καζίνο'. Τα δύο παραλληλόγραμμα συμβολίζουν τις δυο καταστάσεις του ζαριού (αμερόληπτο-μεροληπτικό), και τα βέλη τις επιτρεπτές μεταβάσεις. Μέσα σε κάθε κατάσταση, αναγράφονται οι πιθανότητες εμφάνισης των συμβόλων. (πηγή:* [66] *σελ. 277 Εικόνα 8.3).*

Οι πρότυποι αλγόριθμοι δυναμικού προγραμματισμού που χρησιμοποιούνται για το σκορ και τη στοίχιση των αλληλουχιών με τα HMMs ανεξαρτήτου πολυπλοκότητας αυτών είναι οι Forward και Viterbi αντίστοιχα, ενώ αντίστοιχα πρότυποι αλγόριθμοι εκπαίδευσης των μοντέλων υπάρχουν στη βιβλιογραφία, όπως ενδεικτικά οι Baum-Welch, Gibbs [64], [66].

Τα pHMMs, αποτελούν στην ουσία τους ένα HMM το οποίο όμως περιγράφει μια πολλαπλή στοίχιση. Συνεπώς, σε ένα pHMM (Εικόνα **1.22**) κάθε κατάσταση δεν απεικονίζει την εμφάνιση ενός συμβόλου όπως στα κλασικά μοντέλα που αναφέρονται παραπάνω, αλλά μια συγκεκριμένη στήλη της πολλαπλής στοίχισης. Επιπλέον, κάνει την εμφάνιση της και μια κατάσταση εισαγωγής κενού ("*insert*") και διαγραφής/απαλοιφής ("*delete*"), δίνοντας λύση στο πρόβλημα επιβολής ποινής για τα κενά, ενώ γενικότερα το μοντέλο αυτό εμφανίζει αυστηρά γραμμική, μονόδρομη πορεία και ονομάζονται "*left right*". Τα pHMMs μετατρέπουν μια πολλαπλή στοίχιση αλληλουχιών σε ένα σύστημα βαθμολόγησης συγκεκριμένης θέσης κατάλληλο για την αναζήτηση απομακρυσμένων ομολογιών σε βάσεις δεδομένων αλληλουχιών. Οι αναλύσεις με pHMMs συμπληρώνουν τις τυπικές μεθόδους σύγκρισης αλληλουχιών ανά ζεύγη σε αναλύσεις μεγάλης κλίμακας.



Εικόνα 1.22 : *Σχηματική αναπαράσταση ενός τυπικού pHMM. Όπου με Mk τετράγωνα απεικονίζονται οι Καταστάσεις Ταύτισης (Match states), με ρόμβους Ik οι Καταστάσεις Εισαγωγής (Insertion states) και με κύκλους Dk οι Καταστάσεις Απαλοιφής (Deletion states). (πηγή:* [66] *σελ. 301, Εικόνα 8.16).*



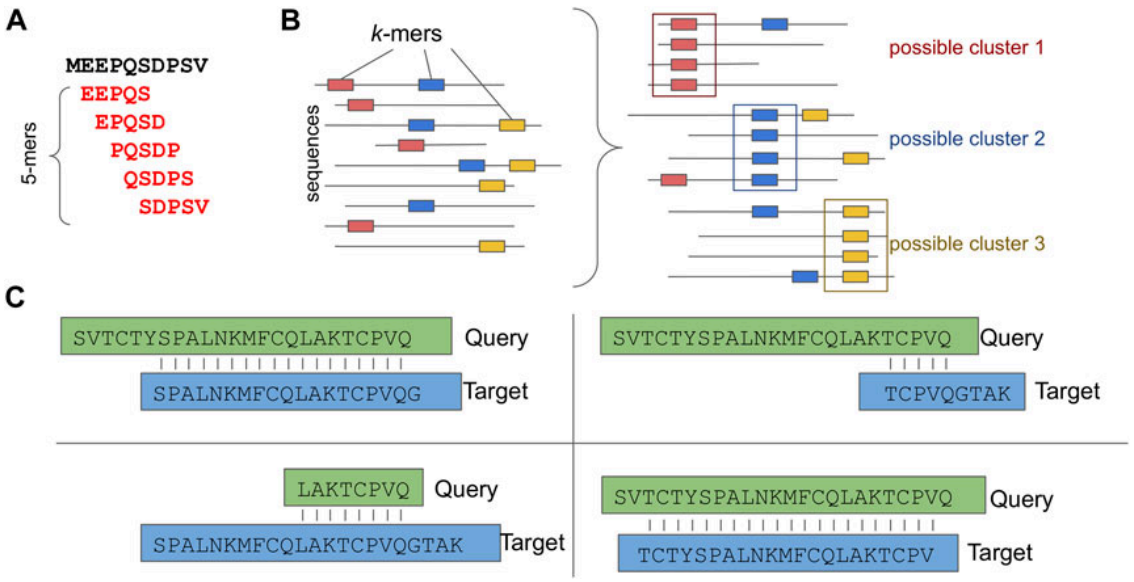
Εικόνα 1.23 : *Σχηματική αναπαράσταση της διαδικασίας χαρακτηρισμού μιας νέας πρωτεϊνικής οικογένειας. (πηγή:* [66] *σελ. 302, Εικόνα 8.17).*

Μια ποικιλία πακέτων λογισμικού υλοποιούν pHMM ή μοντέλα που ομοιάζουν με HMMs, όπως τα HMMER3, SAM, PFTOOLS, HMMpro, GENEWISE, PROBE, META-MEME, BLOCKS, PSI-BLAST, ενώ ήδη αναφέρθηκαν οι βάσεις δεδομένων που στήριξαν την απεικόνιση των πρωτεϊνικών οικογενειών τους στα pHMMS [64]. Αυτή τους η ιδιότητα, μαζί με την πρόσφατα αναβαθμισμένη ισχύ του HMMER συναρτήσει του χρόνου (υπολογίζεται οτι η αναζήτηση ενός πρωτεϊνικού pHMM έναντι ενός σετ δεδομένων 100 εκατομμυρίων πρωτεϊνικών αλληλουχιών, θα ολοκληρωθεί σε περίπου 10 λεπτά με τη χρήση μάλιστα 1 επεξεργαστή [63]) καθιστούν το HMMER ιδανικό εργαλείο στην αναζήτηση ομοιοτήτων σε επίπεδο αλληλουχίας (Εικόνα **1.23**), σε μεγάλου όγκου δεδομένων αποθετήρια όπως αυτά που αναλύθηκαν σε αυτή την εργασία.

1.4.2. Ομαδοποίηση Αλληλουχίας (Sequence Clustering)

Το στάδιο της ομαδοποίησης των αλληλουχιών (sequence clustering) αποτελεί κρίσιμο σημείο στην μελέτη μεταγονιδιωμάτων. Μεγάλης κλίμακας ομαδοποίηση συχνά βοηθάει στην μείωση του όγκου δεδομένων, εντάσσοντας τις αλληλουχίες σε συστάδες (clusters), τα μέλη των οποίων μοιράζονται κοινά χαρακτηριστικά σε επίπεδο αμινοξικής αλληλουχίας και κατά συνέπεια ενδεχομένως επιτελούν παρόμοιες βιολογικές διαδικασίες / λειτουργίες. Οι παραγόμενες ομάδες αξιοποιούνται περαιτέρω στη φυλογενετική ανάλυση και τη διερεύνηση της εξελικτικής σχέσης των μελών τους και επιπλέον η ομαδοποίηση μπορεί να αποτελέσει θεμελιώδες παράγοντα λειτουργικού σχολιασμού νέων πρωτεϊνών, ρίχνοντας φως στην μεταγονιδιωματική “μαύρη ύλη” (*metagenomic dark matter*), είτε με την συμμετοχή τους σε συστάδες με γνωστά γονίδια ή πρωτεΐνες, είτε με την χρήση των αποτελεσμάτων ομαδοποίησης σε προηγμένες αναλύσεις όπως για παράδειγμα εργαλεία δομικής πρόβλεψης [3].

Υπάρχουν 3 διακριτές προσεγγίσεις για ομαδοποίηση αλληλουχιών - βάση αλληλουχίας (sequence-based γνωστή και ως k-mer based), βάση γράφων (graph-based) και ιεραρχικής ομαδοποίησης (hierarchical clustering) (Εικόνα **1.24**) [3], οι οποίες βοηθούν να ξεπεραστεί το εμπόδιο της διαχείρισης μεγάλου όγκου αλληλουχιών όταν θέλουμε να πραγματοποιήσουμε σύγκριση όλων εναντίον όλων (all-against-all). Στην παρούσα εργασία επιλέχθηκε η ομαδοποίηση βάση αλληλουχίας με το εργαλείο MMSeqs2 [68] λόγω της ευρείας δοκιμασμένης χρήσης του και των δυνατοτήτων που προσφέρει σε επαναληψιμότητα και επεκτασιμότητα, ενώ εναλλακτικές εφαρμογές sequence-based clusterin αποτελούν ενδεικτικά το DIAMOND [69], το CD-HIT [70], uclust/usearch [71] και το πιο πρόσφατο Clusterize [72] . Αξίζει να σημειωθεί πως παρόλη την ποικιλομορφία στις μεθόδους, η πλειοψηφία αυτών των εργαλείων εφαρμόζει σύγκριση αλληλουχιών που μοιράζονται κοινά k-mers, δηλαδή ένα υποσύνολο στο μήκος της ακολουθίας νουκλεοτιδίων ή αμινοξέων, αποφεύγοντας κατ’ αυτό τον τρόπο περιττές συγκρίσεις. Για την ομοιογένεια των συστάδων (clusters) που θα προκύψουν, οι 2 συνηθέστεροι παράμετροι που παίζουν καθοριστικό ρόλο είναι το ποσοστό ομοιότητας της αλληλουχίας και το ποσοστό αλληλοεπικάλυψης κατά την στοίχιση (alignment length coverage percentage). Στο παράδειγμα της Εικόνας παρουσιάζονται 4 εναλλακτικές περιπτώσεις: α) x% επικάλυψη της μεγαλύτερης σε μήκος αλληλουχίας, β) x% επικάλυψη της αλληλουχίας στόχο, γ) x% επικάλυψη της αλληλουχίας σε αναζήτηση και δ) αμφίδρομη x% επικάλυψη στην στοίχιση και των 2 αλληλουχιών.

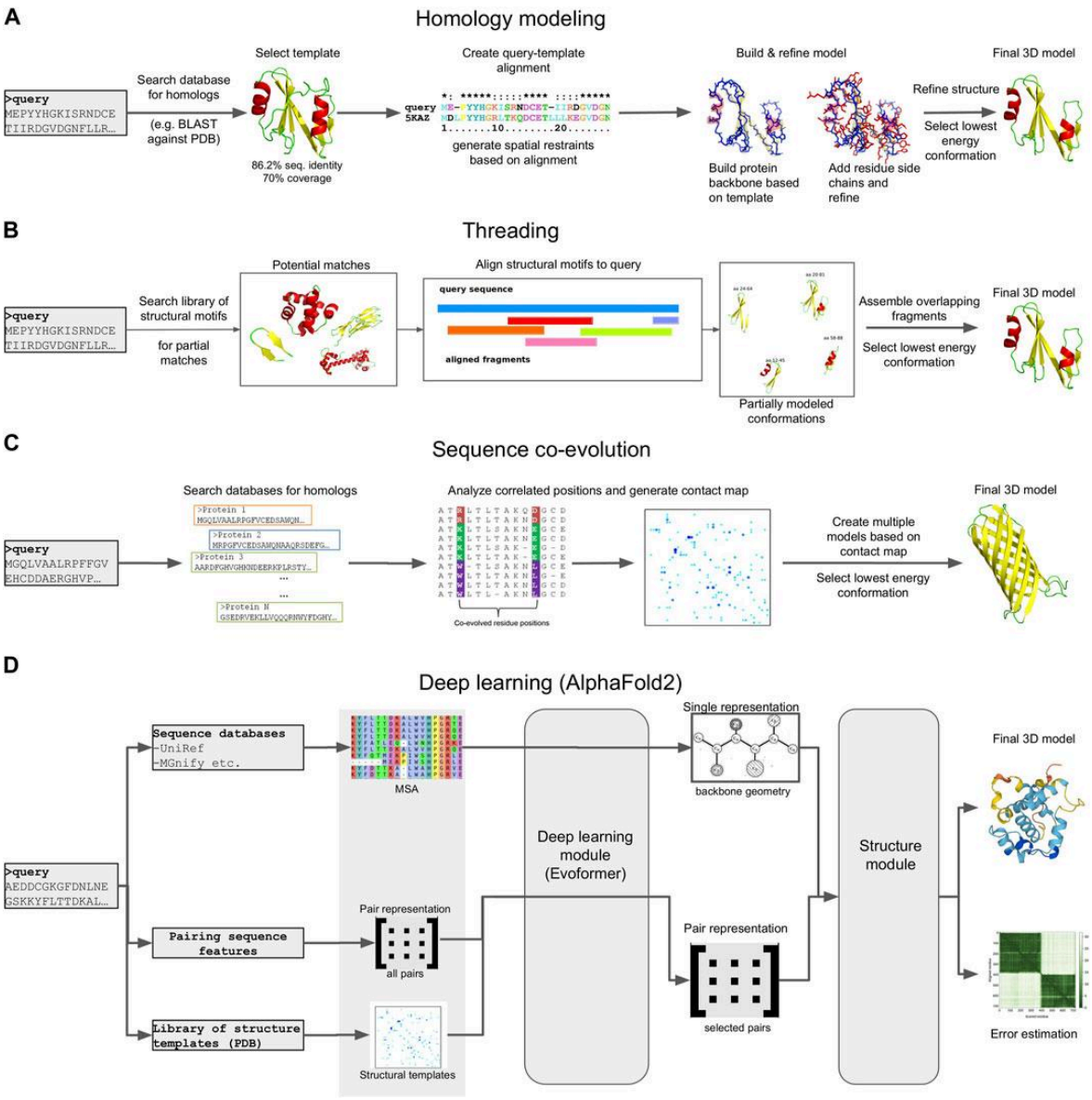


Εικόνα 1.24: *Ομαδοποίηση βασει αλληλουχίας (Sequence based Clustering).* ***(A)*** *Παράδειγμα k-mer,* ***(B)*** *Πιθανές συστάδες (clusters) βασει κοινών k-mers.* ***(C)*** *Διαφορετικές περιπτώσεις κατάταξης αλληλουχίας βασει του μήκους επικάλυψης. (Πηγή : [3] figure 3).*

1.4.3. Πρόβλεψη Πρωτεϊνικής Δομής (Structure prediction)

Είναι γνωστό από τη δομική βιολογία, πως η δομή μιας πρωτεΐνης στον τρισδιάστατο χώρο είναι άρρηκτα συνδεδεμένη με την λειτουργία που αυτή επιτελεί. Ταυτόχρονα, το πώς μια πρωτεϊνική αλληλουχία, δηλαδή μια ακολουθία αμινοξέων, θα διπλώσει στο χώρο και τι δομικό σχήμα θα προκύψει στις τρεις διαστάσεις, παραμένει άλυτο μυστήριο. Μέσω των δομικών χαρακτηριστικών τους, οι πρωτεΐνες είναι ικανές να λαμβάνουν ενεργό ρόλο σε σηματοδοτικά μονοπάτια, μεταβολικές οδούς, κυτταρικές λειτουργίες, ανοσοαπόκριση, ενζυμική λειτουργία, αντιγραφή και μεταγραφή του DNA [73]. Συνεπώς η αναγνώριση και ταυτοποίηση τέτοιων δομικών χαρακτηριστικών μπορεί να διαφωτίσει τον λειτουργικό ρόλο νέων, αχαρακτήριστων μέχρι τώρα πρωτεϊνών. Εμπόδιο σε αυτή την διαδικασία αποτελούν οι προκλήσεις (οικονομικό κόστος, χρονοβόρες διαδικασίες, δυσκολία εφαρμογής σε μαζικό επίπεδο) των σύγχρονων τεχνολογιών πειραματικού προσδιορισμού της πρωτεϊνικής δομής (κρυσταλλογραφία ακτίνων-Χ, φασματοσκοπία NMR, κρυο-ηλεκτρονική μικροσκοπία). Μια εναλλακτική οδό για την παράκαμψη τέτοιων

εμποδίων έχουν φέρει οι μέθοδοι πρόβλεψης της τρισδιάστατης δομής πρωτεϊνών μέσω μοντελοποίησης με υπολογιστικά μέσα (Εικόνα **1.25**).



Εικόνα 1.25: *Σχηματική αναπαράσταση των διαφόρων μεθόδων πρόβλεψης της τρισδιάστατης δομής πρωτεϊνών μέσω μοντελοποίησης με υπολογιστικά μέσα.* ***(A)*** *Μοντελοποίηση μέσω ομολογίας (Homology Modeling).* ***(B)*** *Threading.* ***(C)*** *Μοντελοποίηση μέσω συνεξέλιξης (Sequence coevolution).* ***(D)*** *Μοντελοποίηση μέσω βαθιάς μάθησης (Deep learning) (εμφανίζεται μοντέλο που προέκυψε από το πρόγραμμα AlphaFold2 ως παράδειγμα) πηγή: Figure 5 [3].*

Οι μέθοδοι μοντελοποίησης μέσω συνεξέλιξης (*coevolution-based modeling*) και βαθιάς μάθησης (*deep-learning models*) παρουσιάζουν σημαντικά πλεονεκτήματα όσον αφορά την πρόβλεψη νέων δομών έναντι των εναλλακτικών μεθόδων (homology modeling, sequence threading) [3] οι οποίες προϋποθέτουν ομόλογες αλληλουχίες, παρουσιάζουν πολλές φορές σφάλματα διαμόρφωσης και γενικά τα προκύπτοντα μοντέλα απαιτούν αρκετά βήματα βελτίωσης για να φτάσουν ένα αποδεκτό αποτέλεσμα. Επιπλέον, αξιοποιώντας τις 2 τελευταίες μεθόδους καταφέραμε να προβλέψουμε τις δομές νέων ενζύμων, κάτι που δεν θα ήταν δυνατό διαφορετικά.

Η αρχή της συνεξέλιξης αλληλουχίας βασίζεται στην παρατήρηση πως μια συντηρημένη λειτουργία μιας συγκεκριμένης πρωτεϊνικής οικογένειας, θεμελιώνει σαφή όρια στην ποικιλία της αλληλουχίας των αμινοξέων της και κατά κανόνα συνεπάγεται μια δομική ομοιότητα ανάμεσα στα μέλη της οικογένειας αυτής. Υπό το πρίσμα της διατήρησης ενεργειακά ευνοϊκών αλληλεπιδράσεων, αυτό μπορεί να υποδηλώνει ότι κατάλοιπα σε χωρική εγγύτητα ενδέχεται να παρουσιάσουν φαινόμενα συν-εξέλιξης και άρα οι συσχετίσεις τέτοιων καταλοίπων σε μια στοίχιση αλληλουχιών ομόλογων πρωτεϊνών μπορεί να επιφέρει συμπεράσματα σχετικά με την τρισδιάστατη δομή αυτών. [3], [74]

Η μέθοδος της συνεξέλιξης μαζί με την τελευταία πρόοδο στις υπολογιστικές τεχνολογίες με τη χρήση GPU καθώς και η ραγδαία εξελισσόμενη τεχνολογία τεχνητής νοημοσύνης (A.I.) έφεραν στο προσκήνιο τις μεθόδους deep-learning για το σχεδιασμό μοντέλων πρόβλεψης πρωτεϊνικής δομής. Οι συσχετίσεις των συν-εξελισσόμενων καταλοίπων που προκύπτουν από πολλαπλή στοίχιση αλληλουχιών όπως περιγράφηκαν παραπάνω, τροφοδοτούνται αυτή τη φορά ως δεδομένα εισόδου σε λειτουργικά μοντέλα βαθιας μάθησης τα οποία εκπαιδεύονται επαναληπτικά κι από τα οποία προκύπτουν δομικές συσχετίσεις βάσει μιας ποικιλίας παραμέτρων. Αυτή η εφαρμογή της τεχνητής νοημοσύνης έχει δώσει ανεπανάληπτα αποτελέσματα σε αντίθεση με όλες τις υπόλοιπες μεθόδους, με λαμπρότερο παράδειγμα το εργαλείο της DeepMind, AlphaFold [75] (AlphaFold2 στην τρέχουσα έκδοση του) το οποίο χρησιμοποιήθηκε και στην παρούσα μελέτη.

1.5. Στοχοι της διπλωματικης εργασιας

Η ανάλυση των μεταγονιδιωμάτων έχει προσφέρει έναν τεράστιο όγκο νέας πληροφορίας πρωτεϊνών και της λειτουργικής τους σημασίας. Η διερεύνηση αυτής της αχανής “μαύρης ύλης” και η αποκωδικοποίηση του λειτουργικού σκοπού της μπορεί να προσφέρει σημαντικά οφέλη στην ανακάλυψη βιολογικών μηχανισμών και την ενσωμάτωση και

εφαρμογή τους σε τομείς όπως η υγειονομική περίθαλψη, ο σχεδιασμός φαρμάκου και η προστασία του περιβάλλοντος.

Σκοπός της παρούσας διπλωματικής εργασίας είναι η ανακάλυψη νέων ενζύμων από τις 4 κατηγορίες που περιγράφονται στο κεφάλαιο 1.3 *Ένζυμα και Εφαρμογές*. Πρόκειται για κατηγορίες ενζύμων τα οποία αποτελούν φαρμακευτικούς στόχους ή στόχους με εφαρμογή σε τομείς προστασίας του περιβάλλοντος. Οι στρατηγικές αναζήτησης που επιστρατεύτηκαν περιελάμβαναν την σάρωση υπέρογκων βάσεων δεδομένων μεταγονιδιωματικής, με στόχο την στρατολόγηση υποψηφίων πρωτεϊνικών αλληλουχιών, άγνωστης μέχρι τώρα λειτουργίας, οι οποίες παρουσιάζουν ομοιότητες στις αυτοτελείς δομικές περιοχές (domains) με τις εν λόγων κατηγορίες ενζύμων. Η χρήση pHMMs με το εργαλείο HMMER αξιοποιήθηκε για την παραπάνω ανάλυση και πλήθος προγραμματιστικών scripts δημιουργήθηκε για την διαχείριση και το φιλτράρισμα του μεγάλου όγκου δεδομένων που προέκυψαν. Μέθοδοι ομαδοποίησης (clustering) επιστρατεύτηκαν για την δημιουργία ομάδων ομοίων αλληλουχιών και την παραγωγή μοντέλων μέσω της πολλαπλής στοίχισης αυτών (Multiple Sequence Alignment) με το εργαλείο MMseqs2. Η τρισδιάστατη μοντελοποίηση (3d modeling) πραγματοποιήθηκε με χρήση του AlphaFold2 κατόπιν βελτιστοποίησης των πολλαπλών στοιχίσεων των αλληλουχιών με τελευταίας γενιάς αλγορίθμων μηχανικής μάθησης και βιβλιοθηκών γλώσσας προγραμματισμού (*Biopython*).

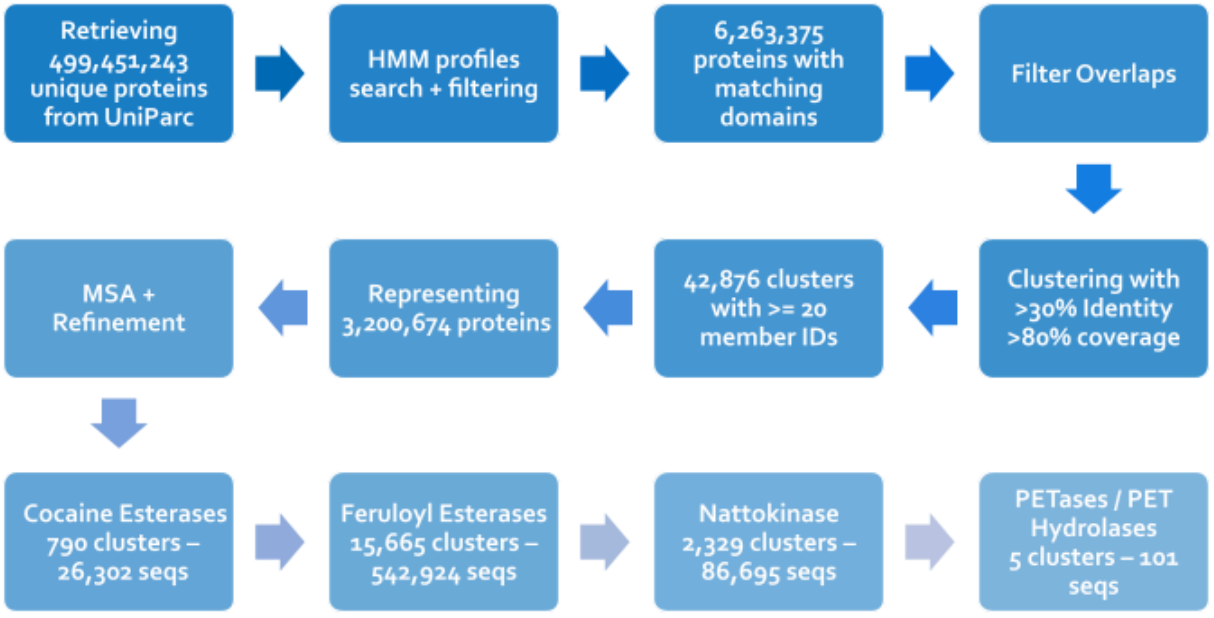
Στο σύνολο της, η ροή εργασίας της ανάλυσης θα ενσωματωθεί σε αυτοματοποιημένα προγράμματα, σε πολυεπίπεδη μορφή με σύνθετες αλληλοσυνδεόμενες επιμέρους διεργασίες. Το πρόγραμμα που θα περικλείει όλα τα παραπάνω θα ονομάζεται MetaSA-Scan και θα είναι γραμμένο σε γλώσσα προγραμματισμού Nextflow [76]. Τα επιμέρους προγράμματα θα είναι δομημένα σε διάφορες γλώσσες προγραμματισμού (όπως Python, R, bash), όμως όλα θα τα διαπερνάει ο άξονας σε Nextflow και θα συνδέονται μεταξύ τους για την επίτευξη των σύνθετων αναλύσεων. Ο απώτερος στόχος είναι να δίνεται η δυνατότητα στον χρήστη, ορίζοντας τις επιθυμητές παραμέτρους, να μπορεί με μια εντολή να φέρνει σε πέρας όλη την ανάλυση (end-to-end), με τα αποτελέσματα να αποθηκεύονται σε φιλική προς τον χρήστη μορφή αρχείων (πίνακες, γραφήματα) και με δυνατότητα επεκτασιμότητας ανάλογα με τις ανάγκες και δυνατότητες του εκάστοτε εργαστηρίου.

Τέλος, μια πλήρης ηλεκτρονική εφαρμογή αναπαράστασης των αποτελεσμάτων και των μεταδεδομένων αυτών δημιουργήθηκε χρησιμοποιώντας τις γλώσσες προγραμματισμού java, javascript και angular, υποστηρίζοντας την βάση δεδομένων σε περιβάλλον MongoDB, όλα στην πλατφόρμα spring boot, ενώ ολόκληρη η ροή εργασίας θα ενσωματωθεί σε ολοκληρωμένο πρόγραμμα (*end-to-end pipeline*) σε κώδικα Nextflow, αξιοποιώντας την πλατφόρμα nf-core.

2. ΥΛΙΚΑ ΚΑΙ ΜΕΘΟΔΟΙ

2.1. Ροή Εργασίας

Η ανάλυση που πραγματοποιήθηκε στην παρούσα εργασία μπορεί να συνοψιστεί σε 4 κύρια στάδια. Μεταφόρτωση δεδομένων από την βάση UniParc, αναζήτηση πρωτεϊνικών domain με χρήση profile HMMs, clustering των αποτελεσμάτων της αναζήτησης και μοντελοποίηση. Δευτερεύοντα στάδια σε κάθε βήμα της ανάλυσης περιελάμβαναν δημιουργία pHMMs για τα οποία δεν υπήρχαν έτοιμα, φιλτράρισμα των αποτελεσμάτων αναζήτησης για διαχωρισμό επαναληψιμότητας, βελτιστοποίηση των αποτελεσμάτων clustering πριν το βήμα της μοντελοποίησης, επιλογή υποψήφιων πρωτεϊνικών αλληλουχιών βάσει μήκους αλληλουχίας για μοντελοποίηση. Για την διαχείριση του μεγάλου όγκου δεδομένων καθώς και για την εφαρμογή σε σύνθετη αυτοματοποιημένη ροή εργασίας των προγραμματιστικών εργαλείων της ανάλυσης, συντάχθηκαν συνολικά περισσότερα από 35 προγραμματικά scripts με περισσότερες από 2000 γραμμές κώδικα γραμμένες σε γλώσσες προγραμματισμού Python, Bash, Javascript, R και Nextflow. Μια συγκεντρωτική άποψη της συνολικής ροής εργασίας που ακολουθήθηκε με τα βασικά βήματα αυτής φαίνεται στην Εικόνα **2.1**.



Εικόνα 2.1: *Σχηματική αναπαράσταση της ροής εργασίας των σταδίων της ανάλυσης για το σύνολο της μελέτης. MSA, Multiple Sequence Analysis. Seqs, Sequences.*

2.2. UniParc

Η έκδοση της UniParc που χρησιμοποιήθηκε για την ανάλυση των πρωτεϊνικών αλληλουχιών ήταν η 2023_3. Η λήψη της πραγματοποιήθηκε από την διεύθυνση *https://ftp.uniprot.org/pub/databases/uniprot/current_release/uniparc/fasta/active/* , όπου υπάρχει διαμοιρασμένη σε συμπιεσμένα αρχεία (μορφής *.gz*) για διευκόλυνση της διαδικασίας λήψεως, αφού πρόκειται για μεγάλο όγκο δεδομένων. Συγκεκριμένα, η έκδοση που λήφθηκε τοπικά, περιείχε 200 αρχεία, συνολικού μεγέθους 205 GB, τα οποία αντιστοιχούσαν σε 543,904,874 ενεργές - active καταχωρήσεις (από τις 584,211,892 συνολικές). Η λήψη εκτελέστηκε σε περιβάλλον Linux του elidek-srv 2 του BSRC Alexander Fleming με χρόνο ολοκλήρωσης 2d 9h 33m 58s με ταχύτητα 548 KB/s. Για να γίνει αντιληπτό το μέγεθος της νέας πληροφορίας που εισάγεται με κάθε ανανεωμένη έκδοση της UniParc, αξίζει να αναφερθεί ότι στην παρούσα έκδοση (2024_04) όταν γράφονταν αυτές οι γραμμές οι συνολικές καταχωρήσεις αντιστοιχούσαν σε 807,190,165 - από τις οποίες οι 747,285,870 ήταν ενεργές.

2.3. HMMER

Συνολικά, 17 πρωτεϊνικά domain επιλέχθηκαν ως στόχοι για αναζήτηση σε ολόκληρη την UniParc. Τα προφιλ των οποίων υπάρχουν ενσωματωμένα στην InterPro από τις εκάστοτε πρωτογενείς βάσεις μέλη της. Συγκεκριμένα για την οικογένεια Nattokinase έγινε αναζήτηση για τους 7 κωδικούς όπως αναγράφονται στον πίνακα **2.1**, για τις Feruloyl Esterases για 5 κωδικούς (πίνακας **2.2**), για τις Petases/Pet hydrolases έγινε αναζήτηση για 1 κωδικό (πίνακας **2.3**) και για τις Cocaine Esterases έγινε αναζήτηση για 4 κωδικούς (πίνακας **2.4**). Οι κωδικοί ταυτοποίησης της στήλης “Profile Member ID” αντιστοιχούν στις βάσεις δεδομένων ανάλογα με το πρόθεμα πριν τους αριθμούς. Δηλαδή: PF - Pfam, PS - Prosite, PR - PRINTS, SSF - Superfamily, PTHR - Panther και TIGR - TIGRFAMs (NCBIfams).

Profile Member ID	InterPro Profile ID	Families
PF00082	IPR000209	Nattokinase
PF05922	IPR010259	Nattokinase
PS00136	IPR023827	Nattokinase
PS00137	IPR022398	Nattokinase
PS00138	IPR023828	Nattokinase
PR00723	IPR015500	Nattokinase
SSF52743	IPR036852	Nattokinase

Πίνακας 2.1: *Οι 7 Profile Member και InterPro Profile κωδικοί αναγνώρισης για τους οποίους πραγματοποιήθηκε αναζήτηση σχετικά με την οικογένεια ενζύμων Nattokinase. Στην πρώτη στήλη αναγράφονται οι κωδικοί των πρωτογενών βάσεων δεδομένων που αποτελούν μέλη της InterPro. Για τις συγκεκριμένες καταχωρήσεις αναγνωρίζονται οι PF(am), PS(PROSITE), PR(PRINTS) και SSF(Superfamily)*

Profile Member ID	InterPro Profile ID	Families
PF07519	IPR011118	Feruloyl_Esterases
PF01764	IPR002921	Feruloyl_Esterases
PF10503	IPR010126	Feruloyl_Esterases
PTHR38050	IPR043595	Feruloyl_Esterases
SSF53474	IPR029058	Feruloyl_Esterases

Πίνακας 2.2: *Οι 5 Profile Member και InterPro Profile κωδικοί αναγνώρισης για τους οποίους πραγματοποιήθηκε αναζήτηση σχετικά με την οικογένεια ενζύμων Feruloyl Esterases. Για τις συγκεκριμένες καταχωρήσεις αναγνωρίζονται οι PF(am), PTHR(Panther) και SSF(Superfamily).*

Profile Member ID	InterPro Profile ID	Families
PF12740	IPR041127	Petases_Pet_Hydrolases

Πίνακας 2.3: *Ο Profile Member και InterPro Profile κωδικός αναγνώρισης για τον οποίο πραγματοποιήθηκε αναζήτηση σχετικά με την οικογένεια ενζύμων Petases Pet Hydrolases.*

Profile Member ID	InterPro Profile ID	Families
PF02129	IPR000383	Cocaine_Esterases
PF00135	IPR002018	Cocaine_Esterases
PF08530	IPR013736	Cocaine_Esterases
TIGR00976	IPR005674	Cocaine_Esterases

Πίνακας 2.4: *Οι 4 Profile Member και InterPro Profile κωδικοί αναγνώρισης για τους οποίους πραγματοποιήθηκε αναζήτηση σχετικά με την οικογένεια ενζύμων Cocaine Esterases. Για τις συγκεκριμένες καταχωρήσεις αναγνωρίζονται οι PF(am), και TIGR(TIGRFAMs).*

Για τα προφίλ προερχόμενα από την Pfam, τα raw hmm profiles ήταν διαθέσιμα και λήφθηκαν απευθείας από τον ηλεκτρονικό ιστότοπο της InterPro, στην σελίδα της εκάστοτε Pfam καταχώρησης από την καρτέλα “Curation”. Για τα προφίλ προερχόμενα από την PANTHER, την SUPERFAMILY και την TIGRFAMs, τα hmm αρχεία λήφθηκαν από τις εκάστοτε ξεχωριστές ιστοσελίδες της κάθε βάσης. Όλα τα εν λόγω αρχεία ήταν ήδη διαμορφωμένα σε hmmer format - μορφή κατάλληλα επεξεργασμένη για να αξιοποιηθούν με το πρόγραμμα HMMER.

Αντίθετα, για το προφίλ σουμπτιλισίνης το οποίο αποτελεί μέλος της MEROPS οικογένειας πεπτιδασών S8 (subtilisin family , clan SB) προερχόμενο από την PRINTS με κωδικό PR00723, για το οποίο δεν υπήρχε έτοιμο hmm προφίλ, ένα σύνολο 400 καλά σχολιασμένων και ελεγμένων πρωτεϊνικών αλληλουχιών ομόλογων με την S8/S53 domain υπεροικογένεια λήφθηκε από την InterPro και στην συνέχεια υποβλήθηκε σε πολλαπλή στοίχιση με το εργαλείο ClustalW [19] μέσω του αποστολέα εργασιών του EBI (EBI job dispatcher) [77] . Από το αποτέλεσμα της πολλαπλής στοίχισης, δημιουργήθηκε κατόπιν το hmm profile μέσω του HMMER με το επιμέρους εργαλείο hmmbuild.

Η έκδοση του πακέτου λογισμικού HMMER που χρησιμοποιήθηκε ήταν η 3.3.2 (Nov 2020). Στην πιο πρόσφατη έκδοσή του (version 3.4, August 2023) το πακέτο HMMER περιλαμβανει τα εξης εργαλεία (αναφέρονται τα πιο σημαντικά για την ανάλυση μας):

- ***hlimask*** - για τον υπολογσιμό και επικάλυψη στήλης στόχου σε μια πολλαπλή στοίχιση.

- ***hmmalign*** - για τη στοίχιση μιας αλληλουχίας σε ένα profile HMM.
- ***hmmbuild*** - για την κατασκευή profile HMM από πολλαπλές στοιχίσεις αλληλουχιών.
- ***hmmemit*** - για την εξαγωγή μιας αλληλουχίας από ένα profile HMM.
- ***hmmpress*** - για την προετοιμασία μιας βάσης δεδομένων με profile HMMs για hmmscan.
- ***hmmscan*** - για την αναζήτηση αλληλουχιών έναντι μιας βάσης δεδομένων profile HMMs.
- ***hmmsearch*** - για την αναζήτηση profile HMMs έναντι μιας βάσης δεδομένων αλληλουχιών (sequence database).
- ***jackhmmer*** - για την επαναλλειπτική αναζήτηση αλληλουχιών έναντι βάσης δεδομένων αλληλουχιών .
- ***phmmer*** - για την αναζήτηση πρωτεϊνικών αλληλουχιών έναντι βάσης δεδομένων πρωτεϊνικών αλληλουχιών.

Από την σουίτα εργαλείων του *HMMER* χρησιμοποιήθηκε το *hmmsearch* για την αναζήτηση. Οι hmm βιβλιοθήκες για κάθε κατηγορία αναζήτησης η οποία περιελάμβανε προφιλ με την πρότυπη (όπως της *Pfam*) μορφή hmm, προέκυψαν από συνένωση των μεμονωμένων προφίλ με μια απλή εντολή cat σε γραμμή εντολών. Στη συνέχεια έτρεξε το εργαλείο hmmsearch για κάθε τέτοια βιβλιοθήκη έναντι της UniParc. Για την προτιμώμενη μορφή εξόδου επιλέχθηκαν οι –tblout όπου αποθηκεύονται τα αποτελέσματα ανά στόχο σε αρχείο πίνακα με κάθε γραμμή να απεικονίζει και μια εύρεση ομόλογης αλληλουχίας και –domtblout όπου συνοψίζονται αντίστοιχα τα αποτελέσματα ανά domain. Τα προτιμώμενα κατώφλια για το ποιά αποτελέσματα να συμπεριληφθούν και να εμφανιστούν κατά την αναζήτηση, ήταν τα -cut_tc (trusted cutoffs) για τα προφίλ της Pfam, τα οποία ορίζουν τα κατώφλια ανά αλληλουχία και ανά domain (TC1 per-sequence and TC2 per-domain) και γενικά αναφέρονται στα χαμηλότερα γνωστά σκορ των αληθώς θετικών που βρίσκονται ακριβώς υψηλότερα από τα γνωστά ψευδώς θετικά κατώφλια, ενώ για τα υπόλοιπα hmm προφίλ που δεν υπάγονταν στη μορφή Pfam, επιλέχθηκαν οι παράμετροι: κατώφλι αναφοράς με ελάχιστο bit-score αλληλουχίας στόχου (target sequence) -T 25.0, κατώφλι αναφοράς με ελάχιστο bit-score domain στόχου -domT 22.0, κατώφλι συμπερίληψης για κάθε στόχο (per-target inclusion threshold) με ελάχιστο bit-score –incT 7.0 και συμπερίληψης domain (per-domain inclusion threshold) με ελάχιστο bit-score –incdomT 5.0.

Για τα 3 προφίλ της PROSITE, δημιουργήθηκε πρόγραμμα σε γλώσσα προγραμματισμού python για την ανίχνευση με τη μέθοδο αναγνώρισης μοτίβων (pattern matching), μιας και τα συγκεκριμένα προφίλ αποτελούν μοτίβα της βάσης (PROSITE patterns) και όχι ακριβώς μοντέλα hmm , ενώ τέλος τα 2 hmm προφίλ της SUPERFAMILY

(SSF53474 και SSF52743) επειδή πρόκειται για υπεροικογένειες και εγκολπώνουν μεγάλο φάσμα διαφορετικών πρωτεϊνικών domains, τροποποιήθηκαν ώστε να περιλαμβάνουν μόνο τους αριθμούς εκχώρησης (accession numbers) των αντίστοιχων domain ενδιαφέροντος, προς αποφυγή ενδεχόμενου πλεονασμού των αποτελεσμάτων (όπως θα δούμε παρακάτω).

Ο κώδικας για τα PROSITE προφίλ (prosite_matcher.py) καθώς και ο κώδικας για το τρέξιμο του hmmsearch στο σύνολο των 200 κομματιών (chunks) της UniParc (hmmer_searcher.sh), είναι διαθέσιμοι στο παράρτημα.

2.4. Filtering

Κατά τη διαδικασία συγκέντρωσης των αποτελεσμάτων από την εφαρμογή του hmmsearch στο σύνολο της UniParc, διαπιστώθηκαν επικαλύψεις ανάμεσα στις ομάδες domain ενδιαφέροντος, με σπουδαιότερη αυτή μεταξύ των Cocaine Esterases και Feruloyl Esterases. Μια ποικιλία προγραμματιστικών μεθόδων σε διάφορες γλώσσες (python, bash, awk, R) επιστρατεύτηκαν για την διύλιση των μοναδικών αποτελεσμάτων, βάσει του μοναδικού κωδικού της UniParc (UPI***) που κατέχει η κάθε καταχώρηση όπως έχει ήδη αναφερθεί στο κεφάλαιο 1.2.2 και μια αναπαράσταση των επικαλύψεων σε διάγραμμα upset με χρήση του προγραμματιστικού πακέτου UpSetR [78] παρέχεται στο κεφάλαιο 3. *ΑΠΟΤΕΛΕΣΜΑΤΑ* (**Σχήμα 3.1**). Τα αποτελέσματα με την καλύτερη βαθμολογία επιλέχθηκαν από τις επικαλυπτόμενες καταχωρήσεις των Cocaine Esterases και Feruloyl Esterases, με χρήση προγράμματος python το οποίο είναι διαθέσιμο στο παράρτημα. Στη συνέχεια οι πρωτογενείς πρωτεϊνικές ακολουθίες από τα αρχικά *.fasta* αρχεία της UniParc αντιγράφηκαν σε νέα συγκεντρωτικά *.fasta* για να αξιοποιηθούν στο επόμενο στάδιο της ομαδοποίησης (clustering). Για ακόμη μια φορά η διαδικασία επιτελέστηκε με δημιουργία προγράμματος στη γλώσσα python (διαθέσιμο στο ΠΑΡΑΡΤΗΜΑ).

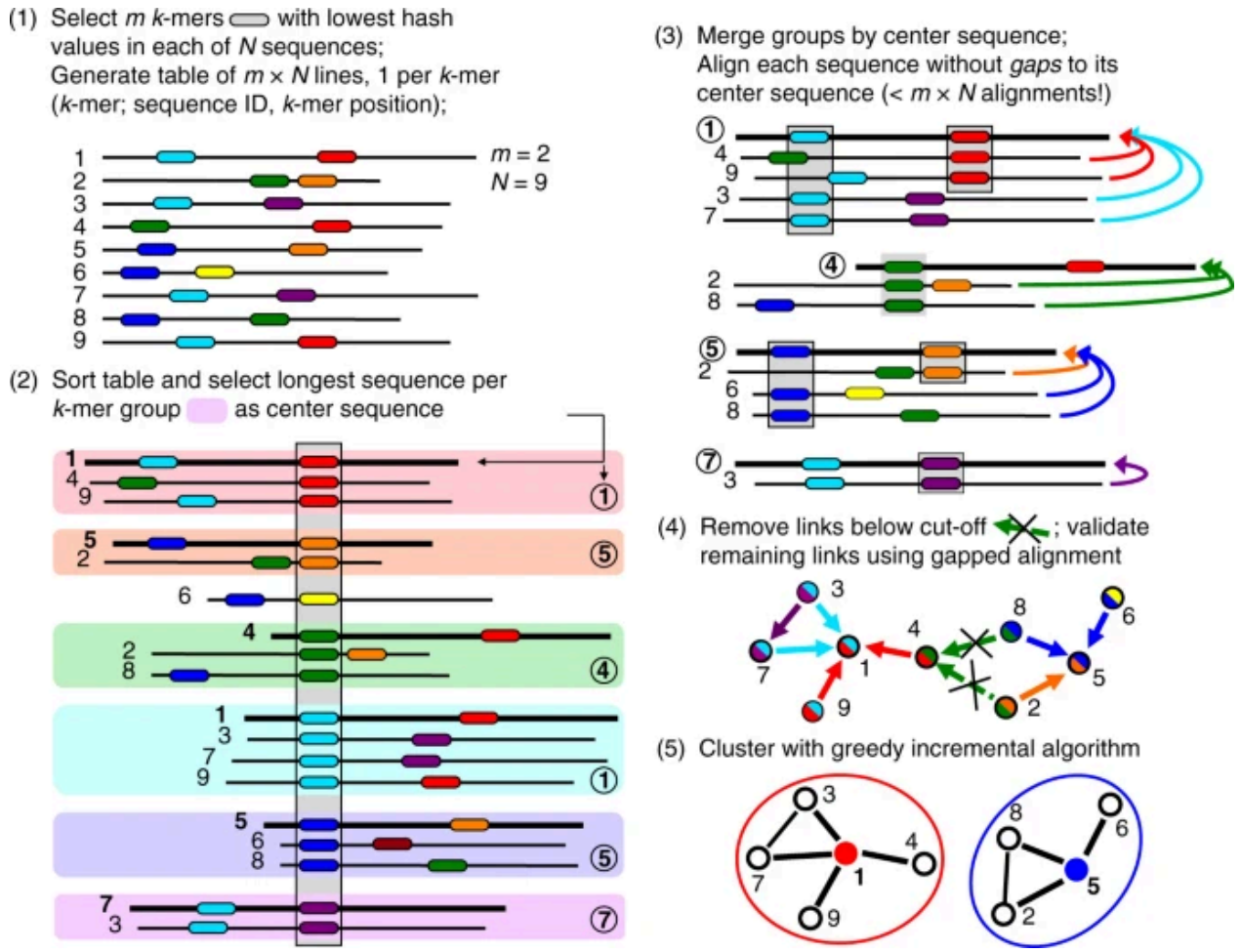
2.5. Clustering

Για το στάδιο της ομαδοποίησης επιλέχθηκε το πακέτο λογισμικού *MMseqs2 (Many-against-Many searching)* [68], [79], [80] το οποίο ενδείκνυται για αναζήτηση και ομαδοποίηση τεράστιων συνόλων αλληλουχιών. Πρόκειται για ένα πρόγραμμα ανοιχτού κώδικα με άδεια GPL (open source GPL-licensed) γραμμένο σε C++ γλώσσα προγραμματισμού και διαμορφωμένο κατάλληλα για εκτέλεση σε όλα τα λειτουργικά συστήματα (Linux, Mac OS, Windows) και σχεδιασμένο για εφαρμογή σε πολλαπλή χρήση επεξεργαστών και servers. Η έκδοση που χρησιμοποιήθηκε στην παρούσα εργασία ήταν η 13-45111+ds-2.

Αρχικά δημιουργήθηκε εσωτερική βάση δεδομένων για την κάθε κατηγορία domain ξεχωριστά, καθώς το MMseqs2 απαιτεί την μετατροπή των FASTA αρχείων αναζήτησης και στόχων (query + target sequences), με χρήση του εργαλείου *createdb*. Για αυτή την διαδικασία χρησιμοποιήθηκαν τα FASTA αρχεία των μοναδικών (non-redundant) αποτελεσμάτων που προέκυψαν από το προηγούμενο στάδιο (filtering). Στην συνέχεια επιλέχθηκε η ροή εργασιών easy-linclust (Εικόνα **2.2**) ώστε να γίνει η ομαδοποίηση. Ο αλγόριθμος linclust που χρησιμοποιεί το εργαλείο συνοψίζεται στα εξης 5 βήματα [68]:

- Βήμα 1ο : Παραγωγή πίνακα με τα k-mers.
- Βήμα 2ο : Ακριβής εύρεση ταιριαστών k-mer.
- Βήμα 3ο: Υπολογισμούς hamming απόστασης προ ομαδοποίησης (hamming distance pro-clustering).
- Βήμα 4ο: Στοίχιση αλληλουχίας τοπικά με επιτρεπόμενα κενά.
- Βήμα 5ο: Ομαδοποίηση με άπληστο αυξητικό αλγόριθμο (greedy incremental clustering).

Το συγκεκριμένο εργαλείο παρουσιάζεται ως το πιο κατάλληλο για τον χειρισμό μεγάλου όγκου δεδομένων με το πρόγραμμα linclust να πραγματοποιεί ομαδοποίηση σε γραμμικό χρόνο (linear clustering) έχοντας καλύτερους χρόνους ολοκλήρωσης με αντίτιμο λιγότερη ευαισθησία. Οι παράμετροι που δόθηκαν ήταν για ελάχιστη ταύτιση αλληλουχίας (minimum sequence identity) *--min-seq-id 0.3* , δηλαδή ομαδοποίηση όταν οι αλληλουχίες εμφανίζουν ομοιότητα >= 30% και βαθμό επικάλυψης (coverage) *-c 0.8* , δηλαδή ομαδοποίηση όταν οι αλληλουχίες στόχου και αναζήτησης (target and query) αλληλοεπικαλύπτονται σε ένα κλάσμα αμινοξικών καταλοίπων >= 80% του συνόλου αυτών των αλληλουχιών. Η μορφή των αρχείων εξόδου της διαδικασίας είναι σε μορφή πίνακα, με επικεφαλή την αλληλουχία αντιπρόσωπο (representative sequence) γύρω από την οποία συσπειρώθηκε η κάθε συστάδα (cluster).



Εικόνα 2.2: *Σχηματική αναπαράσταση του αλγορίθμου ομαδοποίησης σε γραμμικό χρόνο που χρησιμοποιεί το MMseqs2 (linclust). (1) Ο αλγόριθμος επιλέγει m k-mers από κάθε αλληλουχία (προεπιλογή: 20) με τις χαμηλότερες τιμές της συνάρτησης κατακερματισμού (lowest hash function values), με σκοπό την διαλογή κοινών k-mers ανάμεσα σε ομόλογες αλληλουχίες. Χρησιμοποιεί ένα αλφάβητο 13 γραμμάτων για τα k-mers και θέτει το k μεταξύ 10 και 14 αναλόγως το πλήθος των αλληλουχιών εισόδου και το κατώφλι ομοιότητας αλληλουχίας που έχει οριστεί. Παράγει πίνακα που περιλαμβάνει τα k-mers, τον κωδικό αναφοράς της αλληλουχίας (Sequence Identifier) και την τοποθεσία του k-mer επάνω στην αλληλουχία. (2) Διαχωρίζεται ο πίνακας ανα k-mer αναγνωρίζοντας με αυτό τον τρόπο ομάδες όμοιων k-mer (κουτιά στην Εικόνα). Για κάθε ομάδα ο αλγόριθμος επιλέγει την μεγαλύτερη σε μήκος αλληλουχία ως κέντρο, τείνοντας κατ΄αυτό τον τρόπο να επιλέγει τις ίδιες αλληλουχίες ως κέντρα ανάμεσα στις διάφορες ομάδες. (3) Συνενώνει τις ομάδες k-mer οι οποίες εμφανίζουν ίδια κεντρική αλληλουχία (1: κόκκινη με κυανή και 5: πορτοκαλί και μπλε) και συγκρίνει τα μέλη της ομάδας με την κεντρική αλληλουχία σε 2 βήματα: ανά ολική απόσταση hamming και με τοπική στοίχιση χωρίς κενά, επεκτείνοντας έτσι την ταύτιση. (4) Όσες αλληλουχίες ξεπεράσουν το κατώφλι της βαθμολογίας στο βήμα 3*

στοιχίζονται στην κεντρικη της ομάδας του με χρήση τοπικής στοίχισης με κενά. Τα ζευγάρια δε που ικανοποιούν τα κριτήρια των παραμέτρων που έχουν επιλγεί(όπως πχ E-value, ομοιότητα αλληλουχίας και ποσοστό επικάλυψης) συνδέονται μετακύ τους με μια ακμή. (5) Ο άπληστος αυξητικός αλγόριθμος βρίσκει μια ομαδοποίηση έτσι ώστε κάθε ακολουθία εισόδου να έχει μια ακμή προς την αντιπροσωπευτική ακολουθία της ομάδας της. (πηγή: [68]*, figure 5)*

2.6. MSA + Refinement

Από τα συγκεντρωτικά αρχεία υπολογιστικών φύλλων που προέκυψαν στο στάδιο της ομαδοποίησης, επιλέχθηκαν για κάθε οικογένεια μόνο τα clusters με περισσότερα από 20 μέλη για να διασφαλιστεί η ποιότητα της ανάλυσης. Με 2 προγράμματα σε Python ανακτήθηκαν οι πρωτεϊνικές αλληλουχίες των μελών των clusters και δημιουργήθηκαν εκ νέου FASTA αρχεία για την κάθε ομάδα με κορυφαία την αλληλουχία εκπροσώπησης, ακολουθούμενη από τα μέλη. Κατόπιν, για την περαιτέρω βελτίωση της ακρίβειας του τελικού σετ δεδομένων μας, πραγματοποιήθηκε πολλαπλή στοίχιση (Multiple Sequence Alignment - MSA) στα FASTA αρχεία που προέκυψαν με τη χρήση του εργαλείου *MAFFT* τοπικά. [81], [82] Η έκδοση του MAFFT που χρησιμοποιήθηκε ήταν η v7.490 (2021/Oct/30) και η ανάλυση του εργαλείου έτρεξε με τις παραμέτρους "*--retree 2*", "*--maxiterate 2*", "*--thread 64"*. Περαιτέρω βελτίωση των αποτελεσμάτων πολλαπλής στοίχισης πραγματοποιήθηκε με πρόγραμμα σε γλώσσα Python όπου εφαρμόστηκαν φίλτρα (90% ταύτιση αλληλουχίας και 75% κάλυψη στοίχισης), για την παραγωγή (μοναδικών) στοιχίσεων οδηγών (seed MSA) αξιοποιώντας τις ενότητες (modules) ProDy/Evol και Biopython [83], [84] Από αυτά τα βελτιωμένα clusters, όσα είχαν τουλάχιστον 16 μέλη επιλέχθηκαν και προωθήθηκαν στο στάδιο του 3d modeling.

2.7. Trimming + Modeling

Από το σύνολο των αλληλουχιών που προέκυψαν από την έως τώρα διαδικασία, μόνον εκείνες με μήκος αλληλουχίας λιγότερο των 1000 αμινοξικών καταλοίπων επιλέχθηκαν για το επόμενο στάδιο του 3d modeling, για λόγους υπολογιστικής ευελιξίας. Το 3d modeling πραγματοποιήθηκε με χρήση του ColabFold [85], ένα φιλικό προς τον χρήστη λογισμικό ανοιχτού κώδικα που συνδυάζει το MMseqs2 και το AlphaFold2 στην πλατφόρμα Google Colaboratory υπό τη μορφή Jupyter Notebook, με το οποίο παρέχεται υψηλή υπολογιστική ισχύς σε ερευνητές χωρίς πρόσβαση σε αντίστοιχους πόρους. Η ανάλυση με το ColabFold πραγματοποιήθηκε σε server με GPUs. Η έκδοση Alphafold που χρησιμοποιήθηκε ήταν η

2.3.2 (colabfold 1.5.5) και προέκυψαν 3 μοντέλα ανά cluster σε de novo mode (δλδ δεν χρησιμοποιήθηκαν PDB templates).

Προκειμένου να προκύψει από τα προηγούμενα MSA η αλληλουχία οδηγός με την οποία θα γίνει το 3d modeling με το ColabFold. Βασική μεθοδολογική αρχή του εργαλείου προκειμένου να γίνει επιτυχημένα η τρισδιάστατη μοντελοποίηση, προϋποθέτει η αλληλουχία οδηγος η οποία προκύπτει από την πολλαπλή στοίχιση (seed MSA), να είναι καλής ποιότητας και χωρίς πολλά κενά (gaps) και να αποτελεί τον άξονα της πολλαπλής στοίχισης, ισαπέχοντας δηλαδή από τις υπόλοιπες αλληλουχίες της στοίχισης (hamming distance).

2.8. Nextflow - MetaSA-Scan

Ο κεντρικός άξονας της ροής εργασίας μετουσιώθηκε στο αυτοματοποιημένο πρόγραμμα Metagenomic Sequence Analysis - Scanner (MetaSA-Scan), με στόχο την διευκόλυνση της πολυπλοκότητας του συνόλου των διαφορετικών διεργασιών. Επιλέχθηκε η γλώσσα προγραμματισμού Nextflow [76], η οποία παρουσιάζει τελευταία ευρεία αναγνώριση και ενδείκνυται για δημιουργία τέτοιων ροών εργασίας (pipelines) βιοπληροφορικού περιεχομένου, καθώς παρέχει πλεονεκτήματα όσον αφορά την προσαρμοστικότητα σε διάφορα επίπεδα υπολογιστικών συστημάτων (από προσωπικούς υπολογιστές μέχρι HPCs, clusters ή cloud). Το MetaSA-Scan αποτελείται από επιμέρους αυτοματοποιήσεις εργαλείων με χρήση γλώσσας προγραμματισμού Python τα οποία συνδέονται μεταξύ τους με nextflow και συνθέτουν ανώτερου επιπέδου πολυπλοκότητας ροές εργασίας. Ο κώδικας είναι διαθέσιμος στο δημόσιο αποθετήριο GitHub.

2.9. Meta-4

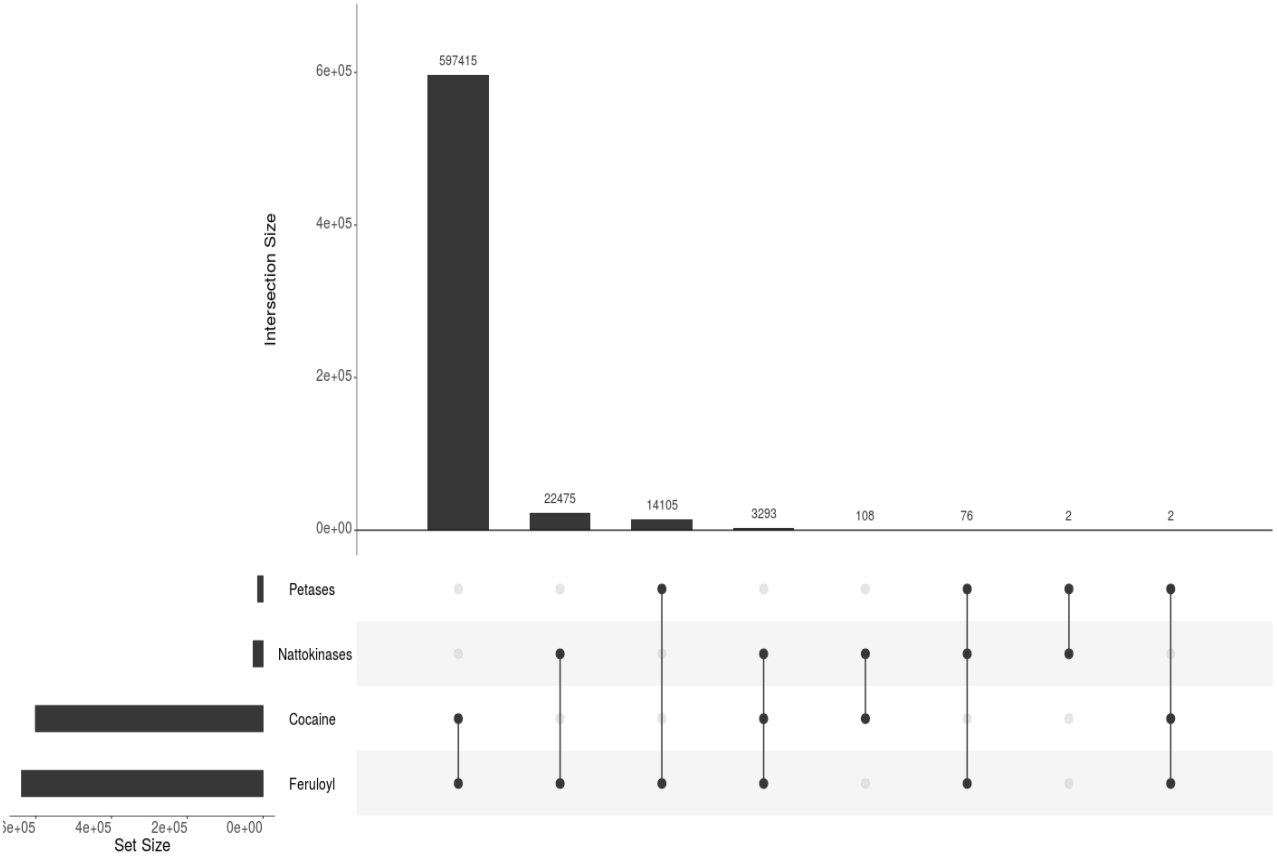
Τα αποτελέσματα της ερευνητικής ανάλυσης, δηλαδή τα ένζυμα, τα μεταδεδομένα αυτών και τα μοντέλα που προέκυψαν, έχουν καταχωρηθεί στη βάση δεδομένων Meta-4. Η Meta-4 αποτελεί πλήρες πρόγραμμα, δομημένο με τις γλώσσες προγραμματισμού java/javascript και Angular. Η βάση δεδομένων βασίστηκε στην τεχνολογία MongoDB και το πρόγραμμα εκκινήθηκε με το εργαλείο Spring Boot. Η μεταφόρτωση των δεδομένων για το πλήθος των πρωτεϊνικών καταχωρήσεων τα οποία αποτελούν αποτελέσματα της ανάλυσης, πραγματοποιήθηκε με χρήση προγράμματος σε γλώσσα Python, μέσω του API της UniProt - UniParc. Ο κώδικας για τη δημιουργία της βάσης δεδομένων Meta-4 παρατίθεται στο παράρτημα.

3. ΑΠΟΤΕΛΕΣΜΑΤΑ

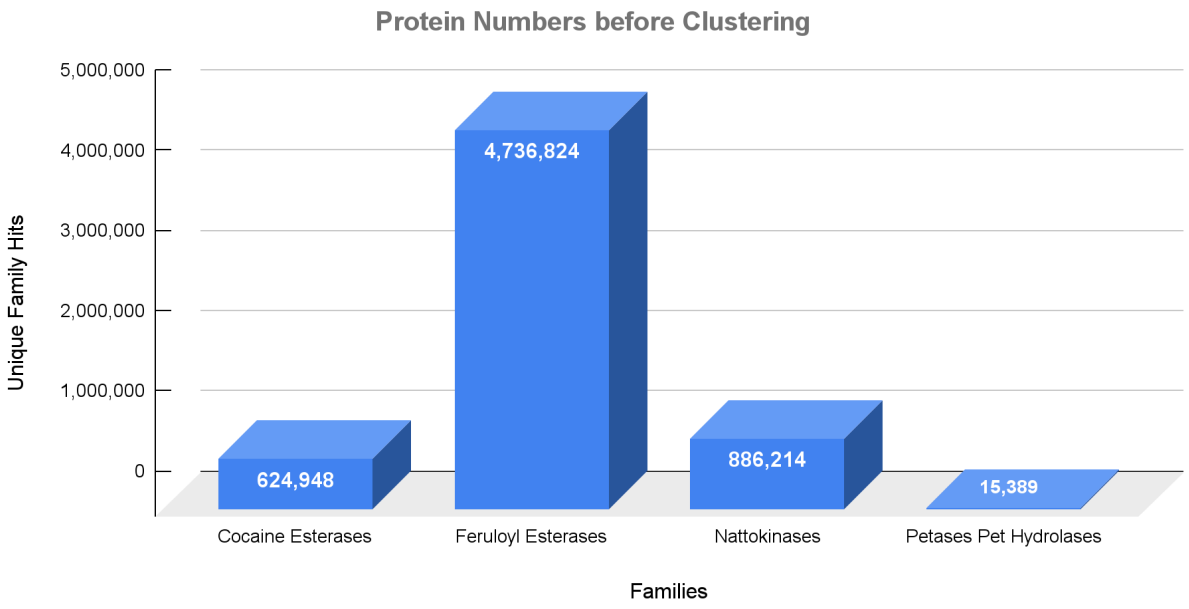
Συνολικά 543,904,874 πρωτεϊνικές αλληλουχίες περιλαμβάνονταν στην έκδοση της UniParc που λήφθηκε. Από αυτές οι 499,451,243 αναλύθηκαν με το HMMER και οι υπόλοιπες 44,453,631 με το InterProScan. Από αυτές τα 2 εργαλεία μετά την ανάλυση έφεραν 6,456,379 ταιριαστά αποτελέσματα - επιτυχίες (matching hits) στα προφίλ ενδιαφέροντος, από τα οποία μετά φιλτραρίσματος επικαλύψεων προέκυψαν 6,263,375 μοναδικά hits.

Παρατηρήθηκαν αλληλοεπικαλύψεις (overlaps) μεταξύ Cocaine Esterases και Feruloyl Esterases της τάξεως των 597,415 , μεταξύ Nattokinase και Feruloyl Esterases της τάξεως των 22,475 , μεταξύ Petases και Feruloyl Esterases της τάξεως των 14,105 , μεταξύ Nattokinase Cocaine και Feruloyl της τάξεως των 3,293 , μεταξύ Nattokinase και Cocaine Esterases 108, Petases + Nattokinase + Feruloyl Esterases 76, μεταξύ Petases + Nattokinase 2 και τέλος μεταξύ Petases + Cocaine Esterases + Feruloyl Esterases 2. Κάθε πρωτεϊνική αλληλουχία από τις αλληλοεπικαλυπτόμενες οικογένειες Feruloyl Esterases + Cocaine Esterases κατανεμήθηκε στην οικογένεια για την οποία εμφανίζει καλύτερη βαθμολογία ταιριάσματος domain από την ροή εργασίας του hmmsearch (bit score of –domtblout output). Στις υπόλοιπες περιπτώσεις οι κοινές αλληλουχίες κατανεμήθηκαν στα εκάστοτε dataset που υστερούσαν αριθμητικά για να βελτιωθούν ποσοτικά για τα επόμενα βήματα ανάλυσης.

Πληροφορίες για τα overlaps φαίνονται στο σχήμα UpSet (Σχήμα **3.1**) το οποίο δημιουργήθηκε με το πακέτο UpSetR [78] σε περιβάλλον προγραμματισμού με γλώσσα R (RStudio). Τα 6,263,375 μη πλεονάζοντα hits κατανέμονται στις 4 οικογένειες ως εξής : Cocaine Esterases 624,948 , Feruloyl Esterases 4,736,824, Nattokinase 886,214 και Petases/Pet Hydrolases 15,389 όπως φαίνονται στο Σχήμα **3.2** .



Σχήμα 3.1: *UpSet Plot για την απεικόνιση των διασταυρούμενων αποτελεσμάτων από τις αναζητήσεις που πραγματοποίηθηκαν με HMMER και InterProScan (περιγράφονται στη παράγραφο 2.3 “Filtering”) . Τα δεδομένα αντλήθηκαν από υπολογιστικά φύλλα με χρήση προγραμματιστικών μεθόδων ενώ το UpSet Plot δημιουργήθηκε σε περιβάλλον R.*



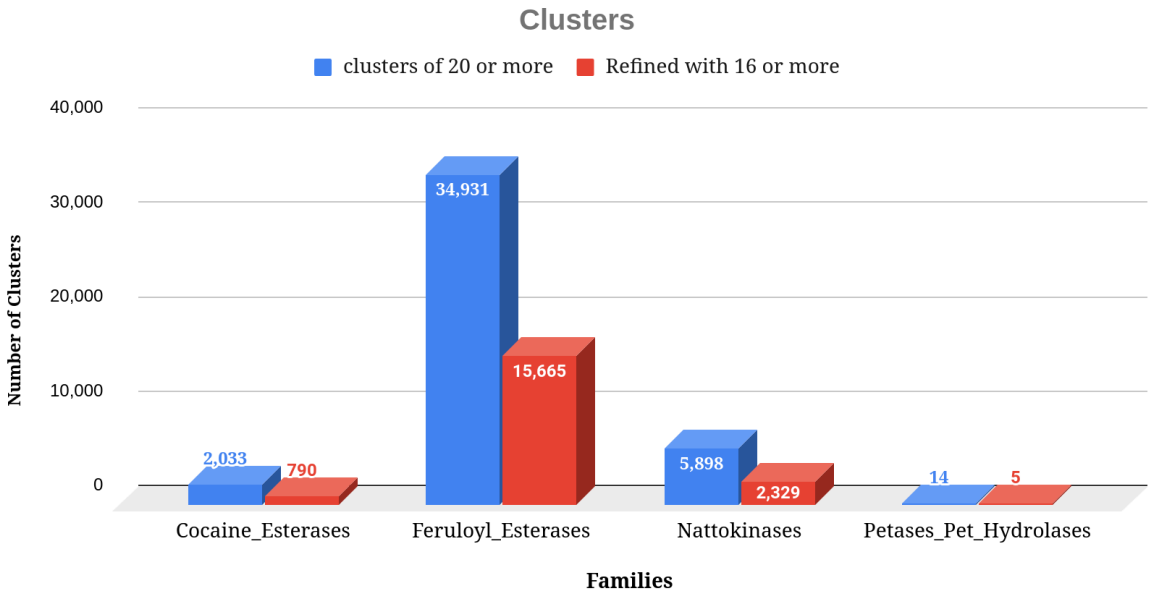
Σχήμα 3.2: *Ραβδόγραμμα απεικόνισης του πλήθους των μοναδικών πρωτεϊνικών αλληλουχιών που εντοπίστηκαν μετά το πέρας της αναζήτησης στη UniParc των domain ενδιαφέροντος. Κατανομή ανά οικογένεια ενζύμων. Τα ποσά αυτά αφορούν το στάδιο πριν γίνει ομαδοποίηση (clustering).*

Μετά το clustering με το MMseqs2 και με παραμέτρους 30% ομοιότητα αλληλουχίας (sequence identity) και 80% αλληλοεπικάλυψη σε στοίχιση κατά ζεύγη (coverage) προέκυψαν συστάδες (clusters) από τις οποίες ενδεικτικά αναφέρεται πως εκείνες με τουλάχιστον 3 μέλη ανέρχονται στις 324,565. Διασφαλίζοντας την ποιότητα της ανάλυσης επιλέχθηκαν μόνο εκείνες με τουλάχιστον 20 μέλη, οι οποίες καταμετρήθηκαν σε 42,876 και αντιστοιχούσαν σε ένα σύνολο 3,200,674 πρωτεϊνικών αλληλουχιών, με μέσο όρο μήκος αλληλουχίας να κυμαίνεται μεταξύ 300 και 700 αμινοξικών καταλοίπων.

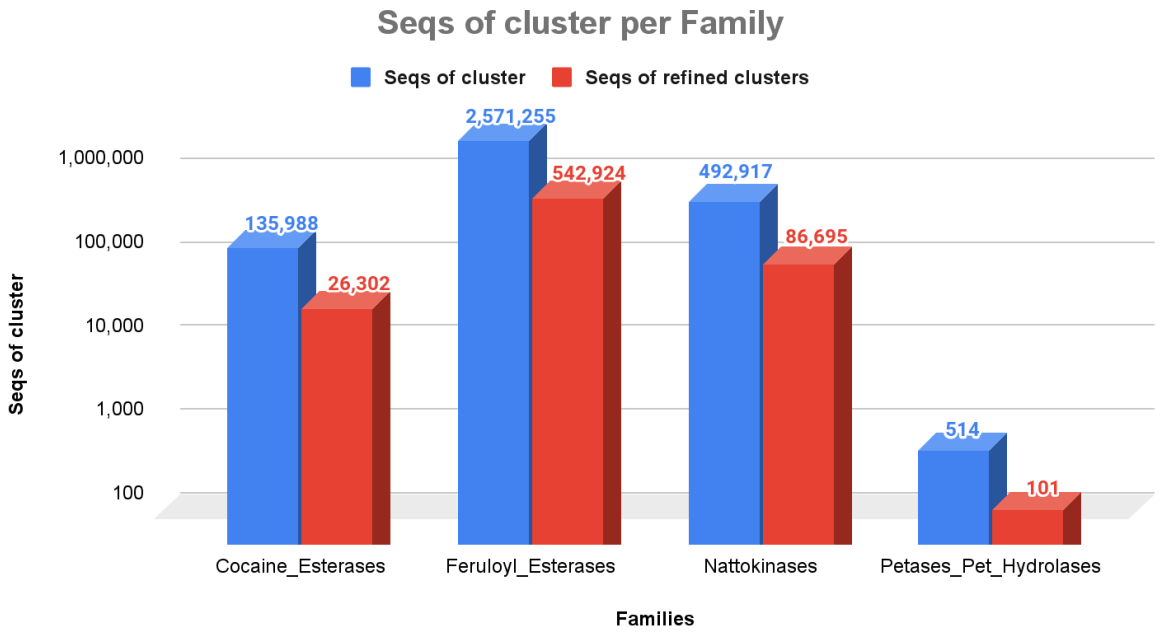
Συγκεκριμένα για κάθε οικογένεια από την ανάλυση του MMseqs2 προέκυψαν οι εξής clusters των 20 και πλέον μελών με τις αντίστοιχες αλληλουχίες στις οποίες ανταποκρίνονται:

- Cocaine Esterases 2,033 clusters με 135,988 αλληλουχίες συνολικά με μ.ο. μήκος 540 αμινοξικά κατάλοιπα.
- Feruloyl Esterases 34,931 clusters με 2,571,255 αλληλουχίες συνολικά με μ.ο. μήκος 434 αμινοξικά κατάλοιπα.
- Nattokinase 5,898 clusters με 492,917 αλληλουχίες συνολικά με μ.ο. μήκος 716 αμινοξικά κατάλοιπα.

- Petases / Pet Hydrolases 14 clusters με 514 αλληλουχίες συνολικά με μ.ο μήκος 321 αμινοξικά κατάλοιπα.



Σχήμα 3.3: *Ραβδόγραμμα αναπαράστασης του πλήθους των συστάδων (clusters) ανά οικογένεια ενζύμων, τα οποία προέκυψαν από το στάδιο ομαδοποίησης με το MMseqs2, όπως περιγράφηκε στο κεφάλαιο 2.4 "Clustering". Με μπλε χρώμα εμφανίζονται οι clusters με τουλάχιστον 20 μέλη, ενώ με κόκκινο οι clusters με τουλάχιστον 16 μέλη οι οποίοι προέκυψαν από το στάδιο Refinement, όπως περιγράφεται στο κεφάλαιο 2.5 "MSA + Refinement".*



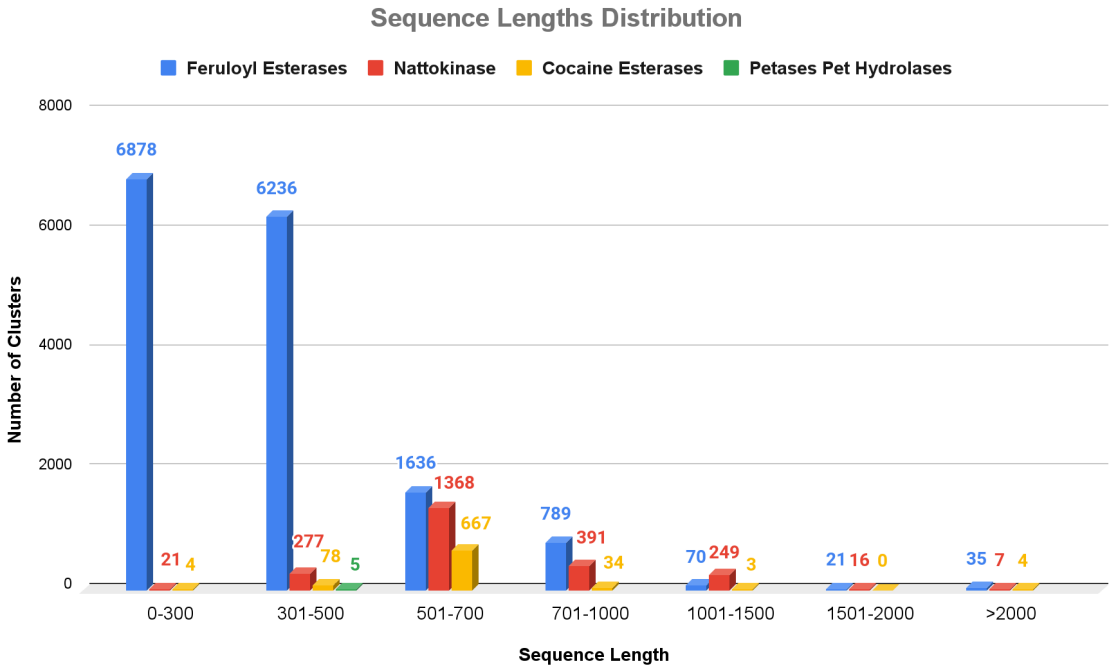
Σχήμα 3.4: *Ραβδόγραμμα αναπαράστασης του πλήθους των πρωτεϊνικών αλληλουχιών που περιλαμβάνονται στο σύνολο των clusters. Με μπλε χρώμα εμφανίζονται τα αρχικά ποσά και με κόκκινα κατόπιν του σταδίου refinement, κάνοντας εμφανή τα ποσά μείωσης του τελικού dataset για λόγους ποιότητας. Περισσότερα περιγράφονται στην παράγραφο 3.5 "MSA + Refinement".*

Μετά την πολλαπλή στοίχιση με το MAFFT και την περαιτέρω βελτίωση του σετ δεδομένων με το Python πρόγραμμα, επιλέχθηκαν οι συστάδες με τουλάχιστον 16 μέλη οι οποίες ανήλθαν σε ένα σύνολο 18,789 περιλαμβάνοντας αντίστοιχα 656,022 πρωτεϊνικές αλληλουχίες (Σχήμα **3.3**).

Οι νέοι clusters μαζί με τις συμπεριλαμβανόμενες πρωτεϊνικές αλληλουχίες που προέκυψαν από το στάδιο βελτιστοποίησης για να προχωρήσουν στο στάδιο μοντελοποίησης και περιείχαν 16 τουλάχιστον μέλη κατανέμονται ως εξής (Σχήμα **3.4** και σχήμα **3.5**):

- Cocaine Esterases 790 clusters με 26,302 αλληλουχίες συνολικά με μ.ο. Μηκος 589 αμινοξικα καταλοιπα.
- Feruloyl Esterases 15,665 clusters με 542,924 αλληλουχίες συνολικά με μ.ο. μήκος 378 αμινοξικα καταλοιπα.
- Nattokinase 2,329 clusters με 86,695 αλληλουχίες συνολικά με μ.ο. μήκος 673 αμινοξικα καταλοιπα.

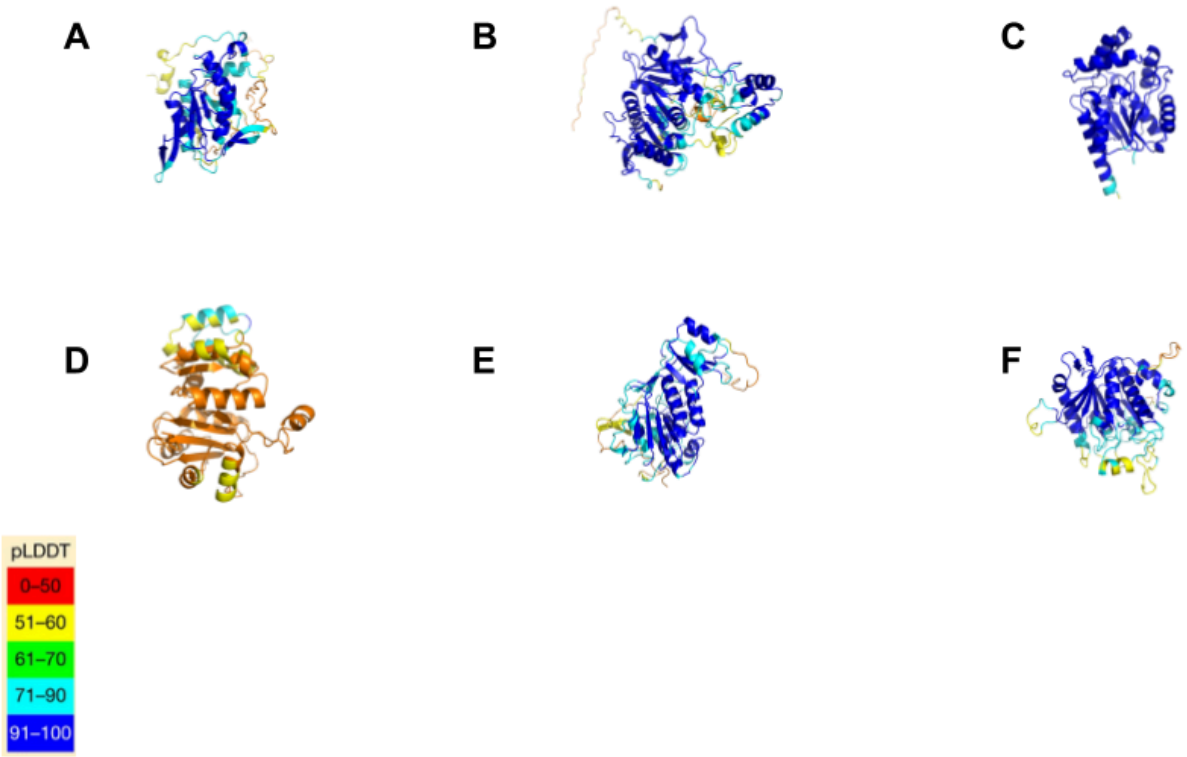
- Petases / Pet Hydrolases 5 clusters με 101 αλληλουχίες συνολικά με μ.ο. μήκος 319 αμινοξικα καταλοιπα.



Σχήμα 3.5: *Ραβδόγραμμα απεικόνισης της κατανομής του μήκους της αλληλουχίας στο σύνολο των clusters που προέκυψαν από την ανάλυση. Οι διαφορετικές ενζυμικές οικογένειες διακρίνονται με τα διαφορετικά χρώματα. Η κατανομή περιγράφεται για ορισμένα διαστήματα για συγκεκριμένο αριθμό αμινοξικών καταλοίπων: 0-300, 301-500, 501-700, 701-1000, 1001-1500, 1501-2000 και τέλος περισσότερα από 2000 κατάλοιπα.*

Μοντελοποίηση

Οι τελικοί βελτιστοποιημένοι clusters με τουλάχιστον 16 μέλη από κάθε μια από τις 4 οικογένειες, τροφοδοτήθηκαν στο AlphaFold2 για το στάδιο της μοντελοποίησης. Αναλυτικά, για τις Cocaine Esterases προέκυψαν μοντέλα για 82 clusters από τα οποία τα 4 αντιστοιχούν σε MSA με μήκη αλληλουχίας 1-300 αμινοξικά κατάλοιπα και 78 σε 301-500. Για τις Feruloyl Esterases 6878 μοντέλα συνολικά, όλα για μήκη αλληλουχίας 1-300 αμινοξικών καταλοίπων. Για τις Nattokinase 298 μοντέλα συνολικά, από τα οποία 21 αντιστοιχούν σε MSA με μήκη αλληλουχίας 1-300 αμινοξικών καταλοίπων και 277 σε 301-500. Για τις Petases προέκυψαν μοντέλα και για τους 5 clusters που αναλύθηκαν και οι οποίοι εμπίπτουν στην κατηγορία πρωτεϊνικών αλληλουχιών με μήκος 301-500 αμινοξικά κατάλοιπα (Σχήμα **3.7**). Για τα υπόλοιπα διαστήματα όσον αφορά τα μήκη των πρωτεϊνικών ακολουθιών, προτείνεται η ανάλυση να πραγματοποιηθεί σε επόμενα βήματα, κυρίως λόγω του τεράστιου όγκου των Feruloyl Esterases (15,665 clusters) ο οποίος απαιτεί αρκετά υψηλή υπολογιστική ισχύ και χρόνο για την υλοποίηση της μοντελοποίησης.



Εικόνα 3.1: *Ενδεικτικά μοντέλα που προβλέφθηκαν από την μοντελοποίηση με το πρόγραμμα AlphaFold2. Πρόκειται για αντιπροσωπευτικά αποτελέσματα για τους εκάστοτε clusters με κατάταξη 1, δηλαδή με το καλύτερο σκορ πρόβλεψης (Rank 1). Παρουσιάζονται 6 αποτελέσματα: (A) UPI000D4BE3AF Nattokinase, (B) UPI00023A05FC Cocaine*

Esterase, (C) UPI000A0CBAAA Feruloyl Esterase, (D) UPI000CE0517B Cocaine Esterase, (E) UPI000BAC425A Nattokinase, (F) UPI00146016FB Petase - Pet/Hydrolase. Οι περιοχές των μοντέλων είναι χρωματισμένες σύμφωνα με τον πίνακα που φαίνεται κάτω αριστερά, αναλόγως με το score εμπιστοσύνης που δίνει σε κάθε αμινοξικό κατάλοιπο το AlphaFold. Η κλίμακα ξεκινάει από το κόκκινο (κακό - bad) εως το μπλε χρώμα (εξαιρετικό - excellent).

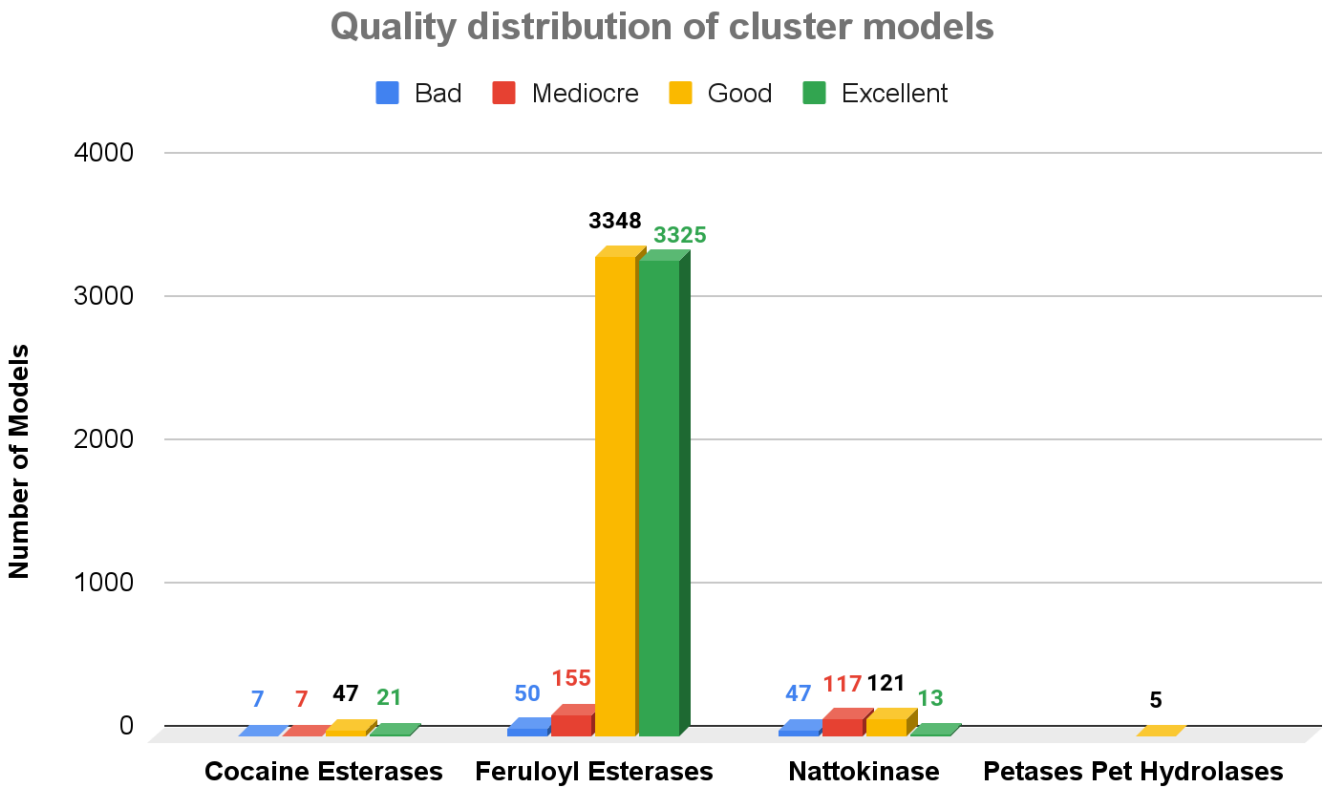
Τα παραπάνω μοντέλα κατατάσσονται στις κλίμακες αξιολόγησης της ποιότητας των μετρικών του AlphaFold2 -

- Bad: pLDDT 0-50, pTM 0-0.50
- Mediocre: pLDDT 50-70, pTM 0.50-0.70
- Good: pLDDT 70-90, pTM 0.70-0.90
- Excellent: pLDDT-90-100, pTM 0.90-1.00

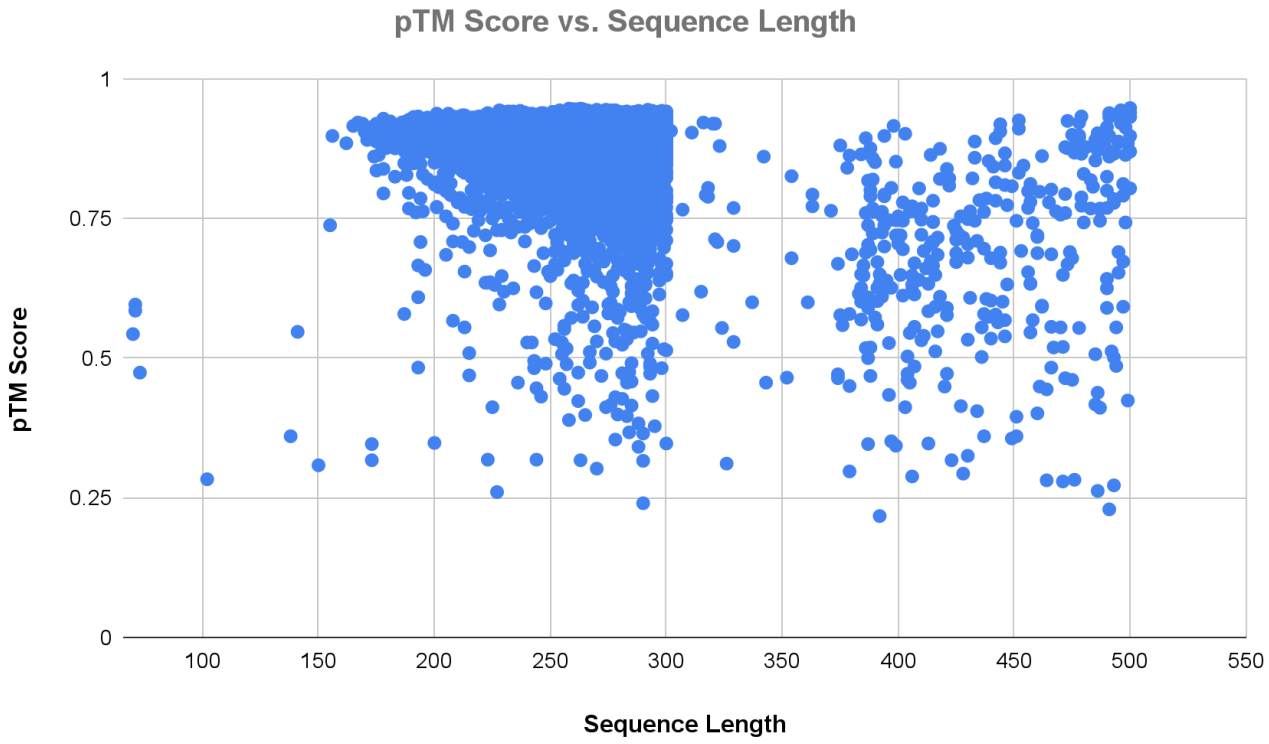
ως εξης:

- Για τις Cocaine Esterases : 7 μοντέλα στην κατηγορία Bad, 7 μοντέλα στην κατηγορία Mediocre, 47 μοντέλα στην κατηγορία Good και 21 μοντέλα στην κατηγορία Excellent.
- Για τις Feruloyl Esterases : 50 μοντέλα στην κατηγορία Bad, 155 μοντέλα στην κατηγορία Mediocre, 3348 μοντέλα στην κατηγορία Good και 3325 μοντέλα στην κατηγορία Excellent
- Για τις Nattokinase : 47 μοντέλα στην κατηγορία Bad, 117 μοντέλα στην κατηγορία Mediocre, 121 μοντέλα στην κατηγορία Good και 13 μοντέλα στην κατηγορία Excellent.
- Για τις Petases : 5 μοντέλα στην κατηγορία Good (Σχήμα **3.6**)

Όλες οι παραπάνω κατηγοριοποιήσεις αφορούν τα καλύτερα από τις τριάδες μοντέλων που προβλέφθηκαν για κάθε cluster ανά οικογένεια. Πάνω από το 85% των τρισδιάστατων μοντέλων είναι καλής ποιότητας (κατηγορία Good ή Excellent) σύμφωνα με τις μετρικές του AlphaFold, δηλαδή παρουσιάζει pTM άνω του 0.70. Ενδεικτικά μοντέλα φαίνονται στην Εικόνα **3.1**.



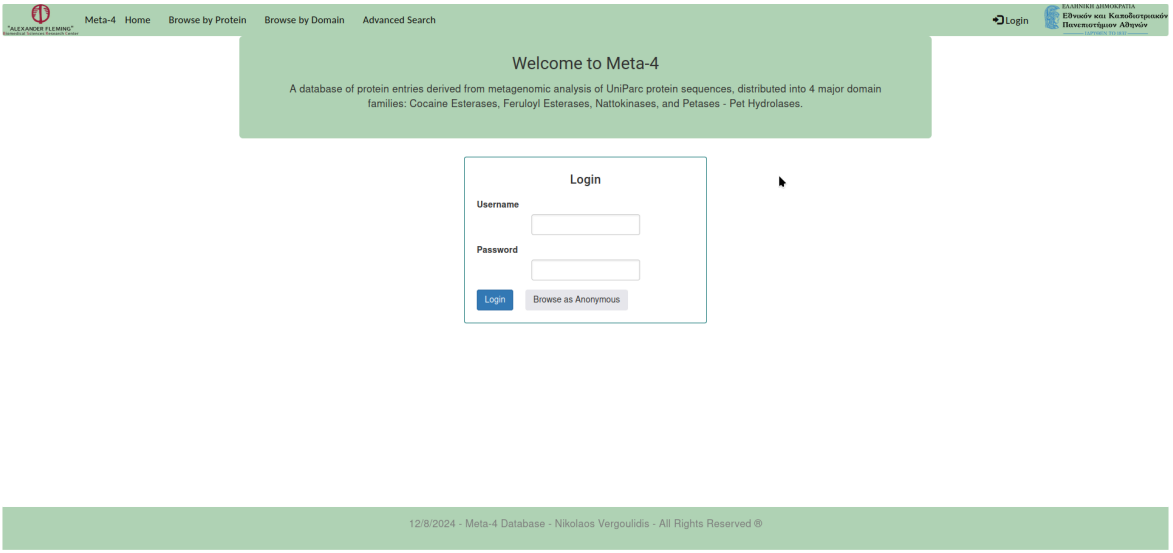
Σχήμα 3.6: *Ραβδόγραμμα απεικόνισης της κατανομής του πλήθους των τρισδιάστατων μοντέλων που προβλέφθηκαν μέσω AlphaFold2 για τους clusters κάθε οικογένειας, στην κλίμακα αξιολόγησης ποιότητας με βάση το pTM score: Bad: pLDDT 0-50, pTM 0-0.50, Mediocre: pLDDT 50-70, pTM 0.50-0.70 Good: pLDDT 70-90, pTM 0.70-0.90 Excellent: pLDDT-90-100, pTM 0.90-1.00.*



Σχήμα 3.7: *Scatter Plot απεικόνισης της κατανομής των pTM scores του συνολικού αριθμού των μοντέλων που προβλέφθηκαν από το AlphaFold ανά μήκος πρωτεϊνικής αλληλουχίας.*

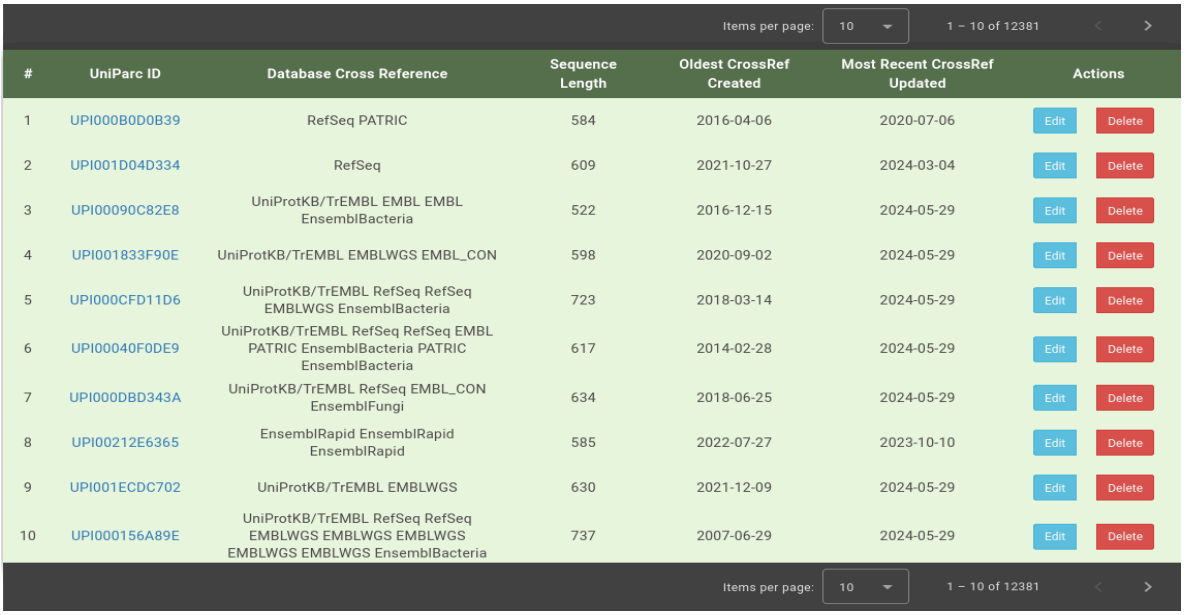
Meta-4

Η βάση δεδομένων Meta-4, ένα πλήρες πρόγραμμα χτισμένο στην τεχνολογία MongoDB σε συνδυασμό με γλώσσα προγραμματισμού java/js και με γραφικό περιβάλλον δομημένο με το εργαλείο Angular, δημιουργήθηκε και φιλοξενεί στην ολότητα τους τα αποτελέσματα της ανάλυσης της παρούσας εργασίας. Όλες οι πρωτεϊνικές αλληλουχίες που εντοπίστηκαν με τις μεθόδους αναζήτησης είναι διαθέσιμες για περιήγηση από τον χρήστη, περιλαμβάνοντας για την κάθε μια όλα τα μεταδεδομένα που παρέχονται από την UniParc (accession ID, πρωτογενής βάση δεδομένων από την οποία προέρχεται, περιβάλλον στο οποίο εντοπίστηκε πρώτη φορά, μήκος αλληλουχίας, αμινοξική αλληλουχία, Μοριακό Βάρος, κ.α.). Οι αλληλουχίες είναι ταξινομημένες και προσβάσιμες και ανά ενζυμική οικογένεια, ενώ παρέχεται πρόσβαση και σε καρτέλα με τα χαρακτηριστικά καθώς και με τα μοντέλα που προέκυψαν. Επίσης είναι δυνατή και η σύνθετη αναζήτηση από τον χρήστη, με επιλογή όλων των πεδίων μεταδεδομένων που χαρακτηρίζουν και διαφοροποιούν τις καταχωρήσεις (πχ UniParc ID, την κατηγορία domain που ανήκει, το μήκος της αλληλουχίας, το μοριακό βάρος, το CRC64 ή ακόμα και χρήση λέξης-κλειδιού) και με τη χρήση λογικών όρων (AND, OR, NOT). Υπάρχουν 2 κατηγορίες χρηστών, με τον χρήστη admin να έχει δυνατότητα εγγραφής, τροποποίησης και διαγραφής καταχώρησης (Εικόνα **3.2**). Στην παρούσα έκδοση της Meta-4 φιλοξενούνται μόνο οι representative πρωτεΐνες του κάθε cluster (Εικόνα **3.3**).



Εικόνα 3.2: *Στιγμιότυπο οθόνης από την αρχική σελίδα εισόδου στην βάση δεδομένων Meta-4. Διακρίνεται παράθυρο εισαγωγής στοιχείων διαπίστευσης για σύνδεση χρήστη με*

δικαιώματα διαχειριστή καθώς και επιλογή ανώνυμης περιήγησης. Στο επάνω μέρος φαίνονται οι καρτέλες επιλογής για περιήγηση στα δεδομένα είτε ανά πρωτεϊνικό αποτέλεσμα, είτε ανά Domain ενδιαφέροντος ή για σύνθετη αναζήτηση.



Εικόνα 3.3: *Στιγμιότυπο οθόνης μετά από επιλογή περιήγησης ανα πρωτεϊνικό αποτέλεσμα, “Browse by Protein” . Στην πρώτη στήλη με τίτλο “UniParc ID” φαίνονται τα αποτελέσματα της ανάλυσης και αποτελούν υπερσύνδεσμο για την μετάβαση στα επιμέρους στοιχεία και μεταδεδομένα του αντίστοιχου ID που θα επιλέξει ο χρήστης. Επίσης στην τελευταία στήλη φαίνονται οι επιλογές τροποποίησης και διαγραφής της καταχώρησης (Edit / Delete), οι οποίες είναι διαθέσιμες μόνο σε χρήστη με δικαιώματα Administrator.*

MetaSA-Scan

Ένα ολοκληρωμένο πρόγραμμα ενσωματώνοντας το σύνολο της ανάλυσης, είναι σε στάδιο παραγωγής. Όταν γράφονταν αυτές οι γραμμές, αποτελούνταν από 4 bin αρχεία, 2 configuration, 7 modules, 2 subworkflows και 1 main workflow αρχείο, ενώ ποσοστιαία χρησιμοποιήθηκαν οι γλώσσες προγραμματισμού Nextflow 60.8%, Python 35.1% και Shell 4.1%. Στο παρόν στάδιο πραγματοποιεί την ανάλυση με το hmmer, επεξεργάζεται τα αποτελέσματα και φέρνει τις πρωτογενείς πρωτεϊνικές αλληλουχίες, κάνει καταμέτρηση των αποτελεσμάτων της Pfam και είναι παραμετροποιήσιμο για εκτέλεση ανάλογα με τους υπολογιστικούς πόρους του χρήστη. Προς το παρόν είναι εκτελέσιμο μόνο σε περιβάλλον Unix.

4. ΣΥΖΗΤΗΣΗ - ΣΥΜΠΕΡΑΣΜΑΤΑ

Αναμφίβολα οι συνεχώς εξελισσόμενες δυνατότητες της μεταγονιδιωματικής, την καθιστούν ένα πολυδύναμο εργαλείο στο πεδίο της σύγχρονης βιολογικής έρευνας με ευρύ φάσμα εφαρμογών. Σε συνδυασμό με τις τελευταίας γενιάς τεχνολογίες για την αλληλούχιση γονιδιωμάτων, η μεταγονιδιωματική ανάλυση έχει την ικανότητα και ηδη αποδίδει υπέρογκες ποσότητες δεδομένων, με τις τελευταίες εκτιμήσεις να κάνουν λόγο για αύξηση του αριθμού των νουκλεοτιδικών ζεύγων βάσεων (base pairs - bp) σε τουλάχιστον εξαπλάσια κλίμακα (της τάξης 10^{18} bp) μέσα στην επόμενη 5ετία [14]. Η αλληλούχιση μέσω shotgun metagenomics ενός περιβαλλοντικού δείγματος αποτελεί σοβαρό πλεονέκτημα στην διερεύνηση μικροβιακών κοινοτήτων, ρίχνοντας φως στα δίκτυα μικροοργανισμών, τις συσχετίσεις μεταξύ τους και την ποικιλία διεργασιών που επιτελούν, παρακάμπτοντας την ανάγκη για καλλιέργεια αυτών των μικροβίων στο εργαστήριο.

Ανάμεσα σε άλλα πεδία, η ανακάλυψη και ο σχεδιασμός ενζύμων είναι ένας σημαντικός τομέας στον οποίο διαδραματίζει καθοριστικό ρόλο η μεταγονιδιωματική. Στην παρούσα μελέτη, επιτελέστηκε εμπλουτισμός τέτοιων αξιοσημείωτων ενζύμων - στόχων, με πρωτεϊνικές αλληλουχίες προερχόμενες από δεδομένα μεταγονιδιωματικής. Η πρόσφατη πρόοδος της τεχνολογίας και ιδιαίτερα αυτή της τεχνητής νοημοσύνης (AI) δίνει συνεχόμενα λύσεις σε ένα από τα θεμελιώδη ζητούμενα της δομικής βιολογίας, το πώς δηλαδή από την αμινοξική αλληλουχία προκύπτει συγκεκριμένη μορφή της δομής της πρωτεΐνης στον τρισδιάστατο χώρο. Μια από τις επόμενες συνεπώς προκλήσεις, αποτελεί η πρόβλεψη της λειτουργίας και η αξιοποίηση του τεράστιου όγκου πληροφορίας σε επίπεδο πρωτεϊνικών αλληλουχιών, με διερεύνηση δομικών ομοιοτήτων (domains) με ήδη χαρακτηρισμένες φαίνεται να είναι μια υποσχόμενη απάντηση.

Στην παρούσα μελέτη, αξιοποιήθηκε ένα πλήθος δεδομένων πρωτεϊνικών αλληλουχιών, καταχωρημένων σε γνωστά ανοιχτά αποθετήρια, της τάξης των δισεκατομμυρίων, το οποίο αποτελεί ένα μικρό μόνο μέρος της συνολικά διαθέσιμης πληροφορίας. Φυσικά η διαχείριση τέτοιου όγκου δεδομένων αποτελεί από μόνη της πρόκληση και η χρήση της επιστήμης των υπολογιστών και συγκεκριμένα μεθόδων προγραμματισμού αποτέλεσε μονόδρομο για τις ανάγκες της ανάλυσης.

Με τη δύναμη τέτοιων μεθόδων, έγιναν αντιληπτές αλληλοεπικαλύψεις των αποτελεσμάτων με σημαντικότερη αυτή μεταξύ δύο πρωτεϊνικών οικογενειών, των Feruloyl Esterases και Cocaine Esterases. Λόγω του ότι η αναζήτηση πραγματοποιήθηκε με βάση τα πρωτεϊνικά domains, αυτό το γεγονός υποδηλώνει πως εντοπίστηκαν πρωτεϊνικές αλληλουχίες οι οποίες μοιράζονταν αυτοτελείς δομικές περιοχές οι οποίες εμπίπτουν και στις 2 οικογένειες, κάτι διόλου παράλογο αν αναλογιστούμε ότι και οι 2

πρόκειται για οικογένειες που απαντάνε κατά κύριο λόγο σε φυτικούς οργανισμούς και κατ' επέκταση επιτελούν και παρόμοιες λειτουργίες (υδρολάσες). Ένα δεύτερο συμπέρασμα που μπορεί να εξαχθεί με ασφάλεια είναι σε σχέση με τον μικρό αριθμό προσέλκυσης (recruitment) αποτελεσμάτων από την οικογένεια των PETases, γεγονός που ενδεχομένως υποδηλώνει ανεπαρκή δεδομένα σε σχέση με την συγκεκριμένη οικογένεια και καταδεικνύει την ανάγκη για περαιτέρω διερεύνηση αυτών των πρωτεϊνικών αλληλουχιών και των domain τους.

Όσον αφορά την μοντελοποίηση, γίνεται φανερή η ανάγκη υπολογιστικών πόρων για την απαιτητική αυτή διαδικασία, ιδιαίτερα όταν πρόκειται για μεγάλο όγκο δεδομένων (στην περίπτωση μας παράδειγμα αποτελεί το dataset των Feruloyl Esterases), αλλά και όταν πρόκειται για αλληλουχίες μεγάλου μήκους. Επιπλέον παρουσιάστηκαν καλύτερα αποτελέσματα (pTM score) σε μήκη αλληλουχίας που κυμαίνονταν μεταξύ 350-500 αμινοξικών καταλοίπων, παρόλα αυτά ίσως δεν μπορεί να εξαχθεί κάποιο ασφαλές συμπέρασμα για αυτό το γεγονός προτού προκύψουν μοντέλα και για τα επόμενα διαστήματα μήκους αλληλουχίας. Περαιτέρω διερεύνηση απαιτείται πάνω σ' αυτό.

Η χρήση των ενζύμων από τον άνθρωπο έχει σημαντική αξία στην σύγχρονη κοινωνία, με πολλές πλευρές να εξαρτώνται και να καθορίζονται από την ενζυμική δραστηριότητα προερχόμενη με τεχνητά μέσα. Για την εγκαθίδρυση όσο δυνατόν βιόσιμων διεργασιών, τα αξιοποιούμε ως εύρωστους βιοκαταλύτες, κάτι που προϋποθέτει την συνεχή ανακάλυψη νέων ή και την βελτίωση των ήδη υπάρχοντων, ώστε να είναι ανθεκτικά σε ακραίες θερμοκρασίες και pH, διαλύματα και άλατα. Η μεταγονιδιωματική και η ανάλυση δεδομένων προερχόμενων από αυτήν, μπορεί να δώσει τα απαραίτητα σε βιοτεχνολογικές μεθόδους υλικά, για τον σχεδιασμό και την τροποποίηση των ενζύμων. Η παρούσα εργασία απέδωσε τρισδιάστατα μοντέλα τα οποία μελλοντικά θα αποτελέσουν την βάση για προσομοιώσεις μοριακής πρόσδεσης (molecular docking) ώστε μετέπειτα να προχωρήσουν στο στάδιο της πειραματικής ανάπτυξης in vitro των υποψήφιων πρωτεϊνών (ενζύμων).

Οι βάσεις δεδομένων αποτελούν αναπόσπαστο κομμάτι τέτοιων αναλύσεων και η συνεισφορά τους σε παγκόσμιο επίπεδο είναι πολύτιμη. Η πρόσβαση στην πληροφορία προχωράει την επιστήμη μπροστά και η παρούσα μελέτη προσφέρει τα αποτελέσματα σε μορφή βάσης δεδομένων (Meta-4) συνεχίζοντας αυτό το έργο. Μελλοντικά, η διάθεση και περαιτέρω βελτίωση του προγράμματος MetaSA-Scan, το οποίο ενσωματώνει ολόκληρη τη ροή εργασίας που πραγματοποιήθηκε στην παρούσα μελέτη, θα δίνει τη δυνατότητα στους ερευνητές και την κοινότητα, να πραγματοποιεί την ανάλυση για οποιοδήποτε domain ενδιαφέροντος, επεκτείνοντας τις 4 οικογένειες ενζύμων οι οποίες απασχόλησαν την παρούσα εργασία.

5. ΔΙΑΧΥΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Από την παρούσα μελέτη προέκυψε η παρακάτω επιστημονική δημοσίευση:

Visualizing metagenomic data: a comprehensive review

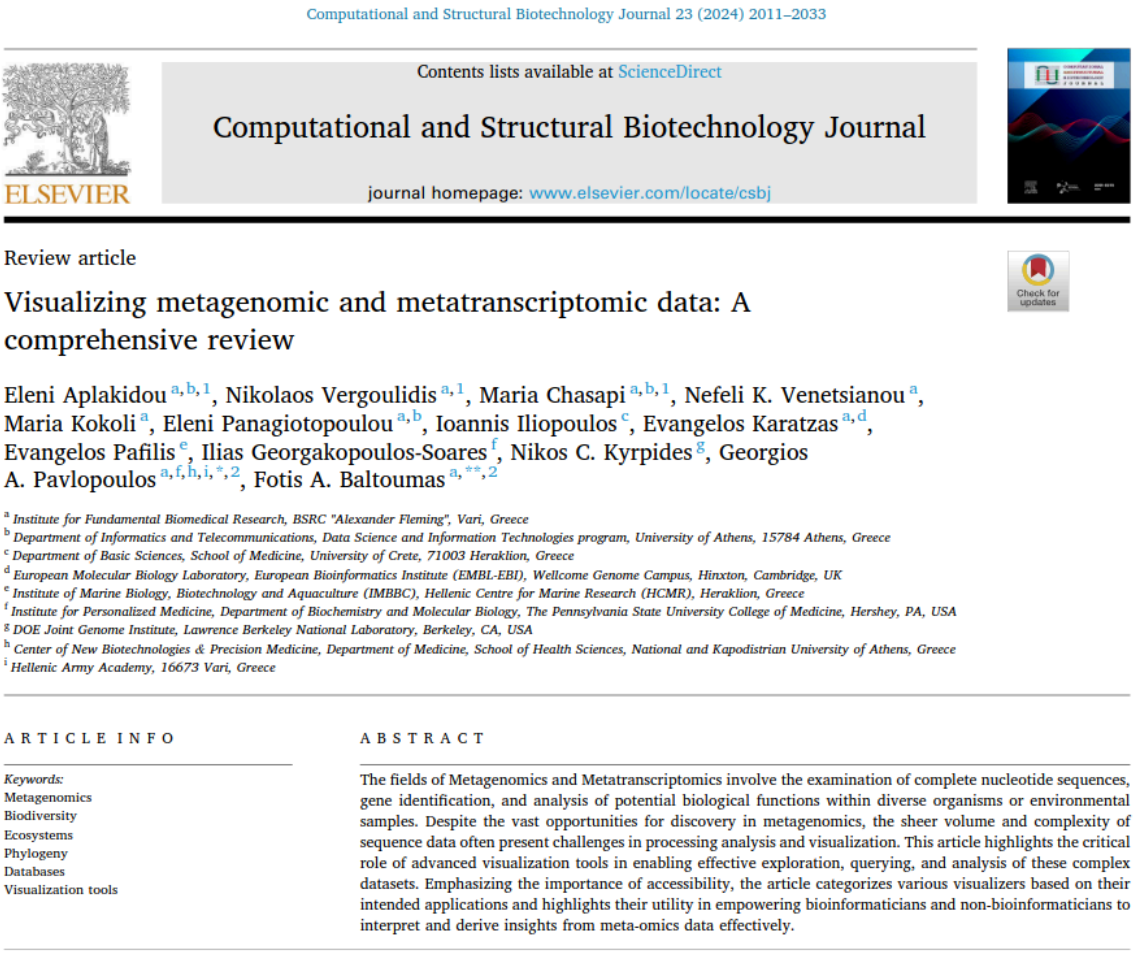
Aplakidou E., Vergoulidis N., Chasapi M., Venetsianou NK, Kokoli M., Panagiotopoulou E., Iliopoulos I., Karatzas E., Pafilis E., Georgakopoulos-Soares I., Kyrpides NC., Pavlopoulos GA.*, Baltoumas FA*.

Computational and Structural Biotechnology Journal (CSBJ) 2024 May 3;23:2011-2033.

doi: 10.1016/j.csbj.2024.04.060.

PMID:38765606

*co-corresponding



Computational and Structural Biotechnology Journal 23 (2024) 2011–2033

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Review article

Visualizing metagenomic and metatranscriptomic data: A comprehensive review

Eleni Aplakidou [a,b,1], Nikolaos Vergoulidis [a,1], Maria Chasapi [a,b,1], Nefeli K. Venetsianou [a], Maria Kokoli [a], Eleni Panagiotopoulou [a,b], Ioannis Iliopoulos [c], Evangelos Karatzas [a,d], Evangelos Pafilis [e], Ilias Georgakopoulos-Soares [f], Nikos C. Kyrpides [g], Georgios A. Pavlopoulos [a,f,h,i,*,2], Fotis A. Baltoumas [a,**,2]

[a] Institute for Fundamental Biomedical Research, BSRC "Alexander Fleming", Vari, Greece
[b] Department of Informatics and Telecommunications, Data Science and Information Technologies program, University of Athens, 15784 Athens, Greece
[c] Department of Basic Sciences, School of Medicine, University of Crete, 71003 Heraklion, Greece
[d] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge, UK
[e] Institute of Marine Biology, Biotechnology and Aquaculture (IMBBC), Hellenic Centre for Marine Research (HCMR), Heraklion, Greece
[f] Institute for Personalized Medicine, Department of Biochemistry and Molecular Biology, The Pennsylvania State University College of Medicine, Hershey, PA, USA
[g] DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[h] Center of New Biotechnologies & Precision Medicine, Department of Medicine, School of Health Sciences, National and Kapodistrian University of Athens, Greece
[i] Hellenic Army Academy, 16673 Vari, Greece

ARTICLE INFO

Keywords:
Metagenomics
Biodiversity
Ecosystems
Phylogeny
Databases
Visualization tools

ABSTRACT

The fields of Metagenomics and Metatranscriptomics involve the examination of complete nucleotide sequences, gene identification, and analysis of potential biological functions within diverse organisms or environmental samples. Despite the vast opportunities for discovery in metagenomics, the sheer volume and complexity of sequence data often present challenges in processing analysis and visualization. This article highlights the critical role of advanced visualization tools in enabling effective exploration, querying, and analysis of these complex datasets. Emphasizing the importance of accessibility, the article categorizes various visualizers based on their intended applications and highlights their utility in empowering bioinformaticians and non-bioinformaticians to interpret and derive insights from meta-omics data effectively.

Ο κώδικας για την βάση δεδομένων Meta-4 καθώς και για το πρόγραμμα MetaSA-Scan είναι διαθέσιμος στο **GitHub:**

https://github.com/IceGreb/Meta-4

https://github.com/IceGreb/MetaSA-Scan

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] R. Rappuoli, P. Young, E. Ron, S. Pecetta, and M. Pizza, ‘Save the microbes to save the planet. A call to action of the International Union of the Microbiological Societies (IUMS)’, *One Health Outlook*, vol. 5, no. 1, p. 5, Mar. 2023, doi: 10.1186/s42522-023-00077-2.

[2] E. Aplakidou *et al.*, ‘Visualizing metagenomic and metatranscriptomic data: A comprehensive review’, *Comput. Struct. Biotechnol. J.*, vol. 23, pp. 2011–2033, Dec. 2024, doi: 10.1016/j.csbj.2024.04.060.

[3] F. A. Baltoumas *et al.*, ‘Exploring microbial functional biodiversity at the protein family level—From metagenomic sequence reads to annotated protein clusters’, *Front. Bioinforma.*, vol. 3, p. 1157956, Mar. 2023, doi: 10.3389/fbinf.2023.1157956.

[4] S. Mukherjee *et al.*, ‘1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life’, *Nat. Biotechnol.*, vol. 35, no. 7, pp. 676–683, Jul. 2017, doi: 10.1038/nbt.3886.

[5] Z. Y. Kho and S. K. Lal, ‘The Human Gut Microbiome – A Potential Controller of Wellness and Disease’, *Front. Microbiol.*, vol. 9, Aug. 2018, doi: 10.3389/fmicb.2018.01835.

[6] A. C. Parte, J. Sardà Carbasse, J. P. Meier-Kolthoff, L. C. Reimer, and M. Göker, ‘List of Prokaryotic names with Standing in Nomenclature (LPSN) moves to the DSMZ’, *Int. J. Syst. Evol. Microbiol.*, vol. 70, no. 11, pp. 5607–5612, 2020, doi: 10.1099/ijsem.0.004332.

[7] N. N. Nam, H. D. K. Do, K. T. Loan Trinh, and N. Y. Lee, ‘Metagenomics: An Effective Approach for Exploring Microbial Diversity and Functions’, *Foods*, vol. 12, no. 11, Art. no. 11, Jan. 2023, doi: 10.3390/foods12112140.

[8] A. Oulas *et al.*, ‘Metagenomics: Tools and Insights for Analyzing Next-Generation Sequencing Data Derived from Biodiversity Studies’, *Bioinforma. Biol. Insights*, vol. 9, p. BBI.S12462, Jan. 2015, doi: 10.4137/BBI.S12462.

[9] B. C. Nwachukwu and O. O. Babalola, ‘Metagenomics: A Tool for Exploring Key Microbiome With the Potentials for Improving Sustainable Agriculture’, *Front. Sustain. Food Syst.*, vol. 6, Jun. 2022, doi: 10.3389/fsufs.2022.886987.

[10] B. Liu *et al.*, ‘An Optimized Metagenomic Approach for Virome Detection of Clinical Pharyngeal Samples With Respiratory Infection’, *Front. Microbiol.*, vol. 11, p. 1552, Jul. 2020, doi: 10.3389/fmicb.2020.01552.

[11] G. M. de Campos *et al.*, ‘Exploring Viral Metagenomics in Pediatric Patients with Acute Respiratory Infections: Unveiling Pathogens beyond SARS-CoV-2’, *Microorganisms*, vol. 11, no. 11, Art. no. 11, Nov. 2023, doi: 10.3390/microorganisms11112744.

[12] L. Call, S. Nayfach, and N. C. Kyrpides, ‘Illuminating the Virosphere Through Global Metagenomics’, *Annu. Rev. Biomed. Data Sci.*, vol. 4, no. 1, pp. 369–391, Jul. 2021, doi: 10.1146/annurev-biodatasci-012221-095114.

[13] G. A. Pavlopoulos *et al.*, ‘Unraveling the functional dark matter through global metagenomics’, *Nature*, vol. 622, no. 7983, pp. 594–602, Oct. 2023, doi: 10.1038/s41586-023-06583-7.

[14] S. L. Robinson, J. Piel, and S. Sunagawa, ‘A roadmap for metagenomic enzyme discovery’, *Nat. Prod. Rep.*, vol. 38, no. 11, pp. 1994–2023, Nov. 2021, doi: 10.1039/D1NP00006C.

[15] A. Popovic *et al.*, ‘Metagenomics as a Tool for Enzyme Discovery: Hydrolytic Enzymes from Marine-Related Metagenomes’, in *Prokaryotic Systems Biology*, P. Krogan Nevan J. and P. Babu Mohan, Eds., Cham: Springer International Publishing, 2015, pp. 1–20. doi: 10.1007/978-3-319-23603-2_1.

[16] S. Kim *et al.*, ‘Multidisciplinary approaches for enzyme biocatalysis in pharmaceuticals: protein engineering, computational biology, and nanoarchitectonics’, *EES Catal.*, vol. 2, no. 1, pp. 14–48, Jan. 2024, doi:

10.1039/D3EY00239J.
[17] The UniProt Consortium, ‘UniProt: the Universal Protein Knowledgebase in 2023’, *Nucleic Acids Res.*, vol. 51, no. D1, pp. D523–D531, Jan. 2023, doi: 10.1093/nar/gkac1052.
[18] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, ‘Basic local alignment search tool’, *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–410, Oct. 1990, doi: 10.1016/S0022-2836(05)80360-2.
[19] F. Sievers and D. G. Higgins, ‘The Clustal Omega Multiple Alignment Package’, in *Multiple Sequence Alignment: Methods and Protocols*, K. Katoh, Ed., New York, NY: Springer US, 2021, pp. 3–16. doi: 10.1007/978-1-0716-1036-7_1.
[20] T. Paysan-Lafosse *et al.*, ‘InterPro in 2022’, *Nucleic Acids Res.*, vol. 51, no. D1, pp. D418–D427, Jan. 2023, doi: 10.1093/nar/gkac993.
[21] J. Mistry *et al.*, ‘Pfam: The protein families database in 2021’, *Nucleic Acids Res.*, vol. 49, no. D1, pp. D412–D419, Jan. 2021, doi: 10.1093/nar/gkaa913.
[22] D. H. Haft *et al.*, ‘TIGRFAMs: a protein family resource for the functional identification of proteins’, *Nucleic Acids Res.*, vol. 29, no. 1, pp. 41–43, Jan. 2001, doi: 10.1093/nar/29.1.41.
[23] W. Li *et al.*, ‘RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation’, *Nucleic Acids Res.*, vol. 49, no. D1, pp. D1020–D1028, Jan. 2021, doi: 10.1093/nar/gkaa1105.
[24] T. K. Attwood *et al.*, ‘The PRINTS database: a fine-grained protein sequence annotation and analysis resource--its status in 2012’, *Database J. Biol. Databases Curation*, vol. 2012, p. bas019, 2012, doi: 10.1093/database/bas019.
[25] C. J. A. Sigrist *et al.*, ‘New and continuing developments at PROSITE’, *Nucleic Acids Res.*, vol. 41, no. D1, pp. D344–D347, Jan. 2013, doi: 10.1093/nar/gks1067.
[26] J. Gough, K. Karplus, R. Hughey, and C. Chothia, ‘Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure’, *J. Mol. Biol.*, vol. 313, no. 4, pp. 903–919, Nov. 2001, doi: 10.1006/jmbi.2001.5080.
[27] A. P. Pandurangan, J. Stahlhacke, M. E. Oates, B. Smithers, and J. Gough, ‘The SUPERFAMILY 2.0 database: a significant proteome update and a new webserver’, *Nucleic Acids Res.*, vol. 47, no. D1, pp. D490–D494, Jan. 2019, doi: 10.1093/nar/gky1130.
[28] A. Andreeva, D. Howorth, C. Chothia, E. Kulesha, and A. G. Murzin, ‘SCOP2 prototype: a new approach to protein structure mining’, *Nucleic Acids Res.*, vol. 42, no. D1, pp. D310–D314, Jan. 2014, doi: 10.1093/nar/gkt1242.
[29] A. Andreeva, E. Kulesha, J. Gough, and A. G. Murzin, ‘The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures’, *Nucleic Acids Res.*, vol. 48, no. D1, pp. D376–D382, Jan. 2020, doi: 10.1093/nar/gkz1064.
[30] N. Labrou, Ed., *Therapeutic Enzymes: Function and Clinical Implications*, vol. 1148. in Advances in Experimental Medicine and Biology, vol. 1148. Singapore: Springer, 2019. doi: 10.1007/978-981-13-7709-9.
[31] P. N. Devine, R. M. Howard, R. Kumar, M. P. Thompson, M. D. Truppo, and N. J. Turner, ‘Extending the application of biocatalysis to meet the challenges of drug development’, *Nat. Rev. Chem.*, vol. 2, no. 12, pp. 409–421, Nov. 2018, doi: 10.1038/s41570-018-0055-1.
[32] E. L. Bell *et al.*, ‘Biocatalysis’, *Nat. Rev. Methods Primer*, vol. 1, no. 1, pp. 1–21, Jun. 2021, doi: 10.1038/s43586-021-00044-z.
[33] R. Wohlgemuth *et al.*, ‘Discovering novel hydrolases from hot environments’, *Biotechnol. Adv.*, vol. 36, no. 8, pp. 2077–2100, Dec. 2018, doi: 10.1016/j.biotechadv.2018.09.004.
[34] K. Myrtollari, N. Katsoulakis, D. Zarafeta, I. V. Pavlidis, G. Skretas, and I. Smonou, ‘Activity and specificity studies of the new thermostable esterase EstDZ2’, *Bioorganic Chem.*, vol. 104, p. 104214, Nov. 2020, doi: 10.1016/j.bioorg.2020.104214.

[35] D. M. Oliveira *et al.*, ‘Feruloyl esterases: Biocatalysts to overcome biomass recalcitrance and for the production of bioactive compounds’, *Bioresour. Technol.*, vol. 278, pp. 408–423, Apr. 2019, doi: 10.1016/j.biortech.2019.01.064.
[36] I. Benoit, E. G. J. Danchin, R.-J. Bleichrodt, and R. P. De Vries, ‘Biotechnological applications and potential of fungal feruloyl esterases based on prevalence, classification and biochemical diversity’, *Biotechnol. Lett.*, vol. 30, no. 3, pp. 387–396, Mar. 2008, doi: 10.1007/s10529-007-9564-6.
[37] W. Zhao, H. Chen, L. Wu, W. Ma, and Y. Xie, ‘Antioxidant properties of feruloylated oligosaccharides of different degrees of polymerization from wheat bran’, *Glycoconj. J.*, vol. 35, no. 6, pp. 547–559, Dec. 2018, doi: 10.1007/s10719-018-9847-2.
[38] A. Chang *et al.*, ‘BRENDA, the ELIXIR core data resource in 2021: new developments and updates’, *Nucleic Acids Res.*, vol. 49, no. D1, pp. D498–D508, Jan. 2021, doi: 10.1093/nar/gkaa1025.
[39] D. W. S. Wong, ‘Feruloyl Esterase: A Key Enzyme in Biomass Degradation’, *Appl. Biochem. Biotechnol.*, vol. 133, no. 2, pp. 87–112, 2006, doi: 10.1385/ABAB:133:2:87.
[40] Y. Li *et al.*, ‘Fermentation of *Lactobacillus fermentum* NB02 with feruloyl esterase production increases the phenolic compounds content and antioxidant properties of oat bran’, *Food Chem.*, vol. 437, p. 137834, Mar. 2024, doi: 10.1016/j.foodchem.2023.137834.
[41] S. kumaran Palani Swamy and V. Govindaswamy, ‘Therapeutical properties of ferulic acid and bioavailability enhancement through feruloyl esterase’, *J. Funct. Foods*, vol. 17, pp. 657–666, Aug. 2015, doi: 10.1016/j.jff.2015.06.013.
[42] K. Juhnevica-Radenkova *et al.*, ‘Highly-Efficient Release of Ferulic Acid from Agro-Industrial By-Products via Enzymatic Hydrolysis with Cellulose-Degrading Enzymes: Part I–The Superiority of Hydrolytic Enzymes Versus Conventional Hydrolysis’, *Foods*, vol. 10, no. 4, Art. no. 4, Apr. 2021, doi: 10.3390/foods10040782.
[43] I. Antonopoulou, E. Sapountzaki, U. Rova, and P. Christakopoulos, ‘Ferulic Acid From Plant Biomass: A Phytochemical With Promising Antiviral Properties’, *Front. Nutr.*, vol. 8, Feb. 2022, doi: 10.3389/fnut.2021.777576.
[44] N. Jamali *et al.*, ‘Nattokinase: Structure, applications and sources’, *Biocatal. Agric. Biotechnol.*, vol. 47, p. 102564, Jan. 2023, doi: 10.1016/j.bcab.2022.102564.
[45] Y. Wang, H. Wang, Y. Zhang, F. Xu, J. Wang, and F. Zhang, ‘Stepwise Strategy to Identify Thrombin as a Hydrolytic Substrate for Nattokinase’, *J. Chem. Inf. Model.*, vol. 62, no. 22, pp. 5780–5793, Nov. 2022, doi: 10.1021/acs.jcim.2c00978.
[46] Y. Weng, J. Yao, S. Sparks, and K. Wang, ‘Nattokinase: An Oral Antithrombotic Agent for the Prevention of Cardiovascular Disease’, *Int. J. Mol. Sci.*, vol. 18, no. 3, p. 523, Feb. 2017, doi: 10.3390/ijms18030523.
[47] Y. Yanagisawa *et al.*, ‘Purification, crystallization and preliminary X-ray diffraction experiment of nattokinase from Bacillus subtilis natto’, *Acta Crystallograph. Sect. F Struct. Biol. Cryst. Commun.*, vol. 66, no. 12, pp. 1670–1673, Dec. 2010, doi: 10.1107/S1744309110043137.
[48] F. Dabbagh *et al.*, ‘Nattokinase: production and application’, *Appl. Microbiol. Biotechnol.*, vol. 98, no. 22, pp. 9199–9206, Nov. 2014, doi: 10.1007/s00253-014-6135-3.
[49] L. Yuan, C. Liangqi, T. Xiyu, and L. Jinyao, ‘Biotechnology, Bioengineering and Applications of Bacillus Nattokinase’, *Biomolecules*, vol. 12, no. 7, p. 980, Jul. 2022, doi: 10.3390/biom12070980.
[50] E. Erickson *et al.*, ‘Sourcing thermotolerant poly(ethylene terephthalate) hydrolase scaffolds from natural diversity’, *Nat. Commun.*, vol. 13, no. 1, p. 7850, Dec. 2022, doi: 10.1038/s41467-022-35237-x.
[51] L. S. Lens-Pechakova, ‘Recent studies on enzyme-catalysed recycling and biodegradation of synthetic polymers’, *Adv. Ind. Eng. Polym. Res.*, vol. 4, no. 3, pp. 151–158, Jul. 2021, doi: 10.1016/j.aiepr.2021.06.005.
[52] S. Yoshida *et al.*, ‘A bacterium that degrades and assimilates poly(ethylene

terephthalate)', *Science*, vol. 351, no. 6278, pp. 1196–1199, Mar. 2016, doi: 10.1126/science.aad6359.
[53] H. P. Austin *et al.*, 'Characterization and engineering of a plastic-degrading aromatic polyesterase', *Proc. Natl. Acad. Sci.*, vol. 115, no. 19, pp. E4350–E4357, May 2018, doi: 10.1073/pnas.1718804115.
[54] Y. Cui *et al.*, 'Computational redesign of a hydrolase for nearly complete PET depolymerization at industrially relevant high-solids loading', *Nat. Commun.*, vol. 15, no. 1, p. 1417, Feb. 2024, doi: 10.1038/s41467-024-45662-9.
[55] H. Wei, J. E. LeSaint, Z. Jin, C.-G. Zhan, and F. Zheng, 'Long-lasting blocking of interoceptive effects of cocaine by a highly efficient cocaine hydrolase in rats', *Sci. Rep.*, vol. 14, no. 1, p. 927, Jan. 2024, doi: 10.1038/s41598-023-50678-0.
[56] L. Shang, Z. Jin, H. Wei, S. Park, C.-G. Zhan, and F. Zheng, 'Catalytic activities of a highly efficient cocaine hydrolase for hydrolysis of biologically active cocaine metabolites norcocaine and benzoylecgonine', *Sci. Rep.*, vol. 13, no. 1, p. 640, Jan. 2023, doi: 10.1038/s41598-022-27280-x.
[57] G. T. Collins *et al.*, 'Cocaine Esterase Prevents Cocaine-Induced Toxicity and the Ongoing Intravenous Self-Administration of Cocaine in Rats', *J. Pharmacol. Exp. Ther.*, vol. 331, no. 2, pp. 445–455, Nov. 2009, doi: 10.1124/jpet.108.150029.
[58] L. L. Howell *et al.*, 'A thermostable bacterial cocaine esterase rapidly eliminates cocaine from brain in nonhuman primates', *Transl. Psychiatry*, vol. 4, no. 7, pp. e407–e407, Jul. 2014, doi: 10.1038/tp.2014.48.
[59] R. D. Finn *et al.*, 'The Pfam protein families database: towards a more sustainable future', *Nucleic Acids Res.*, vol. 44, no. D1, pp. D279–D285, Jan. 2016, doi: 10.1093/nar/gkv1344.
[60] I. Sillitoe *et al.*, 'CATH: increased structural coverage of functional space', *Nucleic Acids Res.*, vol. 49, no. D1, pp. D266–D273, Jan. 2021, doi: 10.1093/nar/gkaa1079.
[61] C. H. Wu, 'PIRSF: family classification system at the Protein Information Resource', *Nucleic Acids Res.*, vol. 32, no. 90001, pp. 112D – 114, Jan. 2004, doi: 10.1093/nar/gkh097.
[62] H. Li *et al.*, 'TreeFam: a curated database of phylogenetic trees of animal gene families', *Nucleic Acids Res.*, vol. 34, no. suppl_1, pp. D572–D580, Jan. 2006, doi: 10.1093/nar/gkj118.
[63] S. C. Potter, A. Luciani, S. R. Eddy, Y. Park, R. Lopez, and R. D. Finn, 'HMMER web server: 2018 update', *Nucleic Acids Res.*, vol. 46, no. W1, pp. W200–W204, Jul. 2018, doi: 10.1093/nar/gky448.
[64] S. R. Eddy, 'Profile hidden Markov models.', *Bioinformatics*, vol. 14, no. 9, pp. 755–763, Jan. 1998, doi: 10.1093/bioinformatics/14.9.755.
[65] S. R. Eddy, 'Accelerated Profile HMM Searches', *PLOS Comput. Biol.*, vol. 7, no. 10, p. e1002195, 2011, doi: 10.1371/journal.pcbi.1002195.
[66] 'Βιοπληροφορική.pdf'.
[67] 'Durbin κ.α. - 1998 - Biological Sequence Analysis Probabilistic Models.pdf'. Accessed: Aug. 10, 2024. [Online]. Available: http://www.mcb111.org/w06/durbin_book.pdf
[68] M. Steinegger and J. Söding, 'Clustering huge protein sequence sets in linear time', *Nat. Commun.*, vol. 9, no. 1, p. 2542, Jun. 2018, doi: 10.1038/s41467-018-04964-5.
[69] B. Buchfink, C. Xie, and D. H. Huson, 'Fast and sensitive protein alignment using DIAMOND', *Nat. Methods*, vol. 12, no. 1, pp. 59–60, Jan. 2015, doi: 10.1038/nmeth.3176.
[70] W. Li and A. Godzik, 'Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences', *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, Jul. 2006, doi: 10.1093/bioinformatics/btl158.
[71] R. C. Edgar, 'Search and clustering orders of magnitude faster than BLAST', *Bioinformatics*, vol. 26, no. 19, pp. 2460–2461, Oct. 2010, doi: 10.1093/bioinformatics/btq461.
[72] E. Wright, 'Accurately clustering biological sequences in linear time by relatedness

sorting', *Nat. Commun.*, vol. 15, no. 1, p. 3047, Apr. 2024, doi: 10.1038/s41467-024-47371-9.
[73] J. Skolnick, J. S. Fetrow, and A. Kolinski, 'Structural genomics and its importance for gene function analysis', *Nat. Biotechnol.*, vol. 18, no. 3, pp. 283–287, Mar. 2000, doi: 10.1038/73723.
[74] D. Altschuh, A. M. Lesk, A. C. Bloomer, and A. Klug, 'Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus', *J. Mol. Biol.*, vol. 193, no. 4, pp. 693–707, Feb. 1987, doi: 10.1016/0022-2836(87)90352-4.
[75] J. Jumper *et al.*, 'Highly accurate protein structure prediction with AlphaFold', *Nature*, vol. 596, no. 7873, pp. 583–589, Aug. 2021, doi: 10.1038/s41586-021-03819-2.
[76] P. Di Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, and C. Notredame, 'Nextflow enables reproducible computational workflows', *Nat. Biotechnol.*, vol. 35, no. 4, pp. 316–319, Apr. 2017, doi: 10.1038/nbt.3820.
[77] F. Madeira *et al.*, 'Search and sequence analysis tools services from EMBL-EBI in 2022', *Nucleic Acids Res.*, vol. 50, no. W1, pp. W276–W279, Jul. 2022, doi: 10.1093/nar/gkac240.
[78] J. R. Conway, A. Lex, and N. Gehlenborg, 'UpSetR: an R package for the visualization of intersecting sets and their properties', *Bioinformatics*, vol. 33, no. 18, pp. 2938–2940, Sep. 2017, doi: 10.1093/bioinformatics/btx364.
[79] M. Mirdita, M. Steinegger, and J. Söding, 'MMseqs2 desktop and local web server app for fast, interactive sequence searches', *Bioinformatics*, vol. 35, no. 16, pp. 2856–2858, Aug. 2019, doi: 10.1093/bioinformatics/bty1057.
[80] M. Steinegger and J. Söding, 'MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets', *Nat. Biotechnol.*, vol. 35, no. 11, pp. 1026–1028, Nov. 2017, doi: 10.1038/nbt.3988.
[81] K. Katoh, J. Rozewicki, and K. D. Yamada, 'MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization', *Brief. Bioinform.*, vol. 20, no. 4, pp. 1160–1166, Jul. 2019, doi: 10.1093/bib/bbx108.
[82] S. Kuraku, C. M. Zmasek, O. Nishimura, and K. Katoh, 'aLeaves facilitates on-demand exploration of metazoan gene family trees on MAFFT sequence alignment server with enhanced interactivity', *Nucleic Acids Res.*, vol. 41, no. W1, pp. W22–W28, Jul. 2013, doi: 10.1093/nar/gkt389.
[83] P. J. A. Cock *et al.*, 'Biopython: freely available Python tools for computational molecular biology and bioinformatics', *Bioinforma. Oxf. Engl.*, vol. 25, no. 11, pp. 1422–1423, Jun. 2009, doi: 10.1093/bioinformatics/btp163.
[84] A. Bakan *et al.*, 'Evol and ProDy for bridging protein sequence evolution and structural dynamics', *Bioinformatics*, vol. 30, no. 18, pp. 2681–2683, Sep. 2014, doi: 10.1093/bioinformatics/btu336.
[85] M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, and M. Steinegger, 'ColabFold: making protein folding accessible to all', *Nat. Methods*, vol. 19, no. 6, pp. 679–682, Jun. 2022, doi: 10.1038/s41592-022-01488-1.

ΠΑΡΑΡΤΗΜΑ

prosite_matcher.py

```
import re
import csv
from Bio import SeqIO

# Define your pattern
pattern_dict = {
      'PS00136':
r'[STAIV][^ERDL][LIVMF][LIVM]D[DSTA]G[LIVMFC].{2,3}[DNH]',
      'PS00137': r'HG[STM].[VIC][STAGC][GS].[LIVMA][STAGCLV][SAGM]',
      'PS00138': r'GTS.[SA].P.[^L][STAVC][AG]'
}

pattern = r'[STAIV][^ERDL][LIVMF][LIVM]D[DSTA]G[LIVMFC].{2,3}[DNH]'

with open('matches_output.tsv', 'w', newline='') as tsvfile:
      tsv_writer = csv.writer(tsvfile, delimiter='\t')
      tsv_writer.writerow(['PROTEIN', 'DOMAIN','PATTERN', 'Start',
'End'])
      total_matches = 0

      for prot_record in
SeqIO.parse("/home/nikoverg/Documents/bioinformatics/projects/Thesis/d
ata/uniparc_active_p200.fasta", "fasta"):
      for pattern_name, pattern in pattern_dict.items():
            matches = re.finditer(pattern, str(prot_record.seq))
            for match in matches:
                  match_data = [prot_record.id, pattern_name,
match.group(), match.start(), match.end()]
                  tsv_writer.writerow(match_data)
                  total_matches += 1
      tsv_writer.writerow(['Summary', f'Total Matches:
{total_matches}'])
```

hmmer_parser.py

```
import csv
import os
import argparse
from Bio import SearchIO

parser = argparse.ArgumentParser()
parser.add_argument("-t","--table", help = "Input domtblout hits")
      # Directory containing .domtblout files
parser.add_argument("-o","--output", help = "Output file to print
results")
args = parser.parse_args()
file_name = args.table

if file_name.endswith('.domtblout'):
      output_file = f"{file_name.split('.')[0]}_output.tsv"

      with open(output_file, 'w', newline='') as csvfile:
            writer = csv.writer(csvfile, delimiter='\t')
            writer.writerow(['Index', 'Protein', 'Bit_Score',
'E-value', 'Start', 'Stop'])
            line_count = 1
            for result in SearchIO.parse(file_name,
'hmmsearch3-domtab'):
                  for hit in result.hits:
                  for hsp in hit:
                        writer.writerow([line_count, hit.id,
hit.bitscore, hit.evalue, hsp.hit_start, hsp.hit_end])
                        line_count += 1
```

sequence_fetcher.py

```
from Bio import SeqIO
import argparse
from numpy import rec
import pandas as pd

parser = argparse.ArgumentParser()
parser.add_argument("-t","--table", help = "Input tsv hits")
parser.add_argument("-f","--fa", help = "Input FASTA query")
parser.add_argument("-o","--output", help = "Output file to print
results")
args = parser.parse_args()
tsv_hits = args.table
fa_query = args.fa
output_file = f"{fa_query.split('.')[0]}_SeqCatches.fasta"

table_hits = pd.read_csv(tsv_hits,header=0, sep="\t")
queries = []
for q in table_hits.index:
      queries.append(table_hits["Protein"][q])



catches=[]
for prot_record in SeqIO.parse(fa_query, "fasta"):
      if prot_record.id in queries:
      catches.append(prot_record)
SeqIO.write(catches, f"{output_file}", "fasta")
```

seq_len_plot.py

```
import os
import csv
from Bio import SeqIO
import pylab

fasta_path =
"/home/nikoverg/Documents/bioinformatics/projects/Thesis/results/mafft
_results"
len_sum = 0
sum_all = 0
sizes = []


for file in os.listdir(fasta_path):
      if file.endswith(".fasta"):
      fasta_file = os.path.join(fasta_path,file)
      print(fasta_file)
      size = ([len(rec) for rec in SeqIO.parse(fasta_file, "fasta")])
      sizes.append(max(size))
print(sizes)
pylab.hist(sizes, bins=20)
pylab.title("%i  clusters of Petases\nLengths %i to %i" % (len(sizes),
min(sizes), max(sizes)))
pylab.xtics(ticks=sizes)
pylab.xlabel("Sequence length (bp)")
pylab.ylabel("Count")
pylab.savefig("Petases_Plot", format='png')
```

fetch_metadata.py

```
import json
import csv
import re
import requests
from requests.adapters import HTTPAdapter, Retry

# Regular expression to find the 'next' link in the response headers
re_next_link = re.compile(r'<(.+)>; rel="next"')
retries = Retry(total=5, backoff_factor=0.25, status_forcelist=[500,
502, 503, 504])
session = requests.Session()
session.mount("https://", HTTPAdapter(max_retries=retries))

def get_next_link(headers):
      if "Link" in headers:
      match = re_next_link.match(headers["Link"])
      if match:
            return match.group(1)
      return None

def get_batch(batch_url):
      while batch_url:
      response = session.get(batch_url)
      response.raise_for_status()
      total = response.headers["x-total-results"]
      yield response, total
      batch_url = get_next_link(response.headers)
# Read IDs from the external file
ids = set()
with
open('/home/nikoverg/Documents/bioinformatics/projects/Thesis/results/
Pet_Repr_metadata.tsv', 'r') as ids_tsv:
      reader = csv.reader(ids_tsv, delimiter='\t')
      next(reader)
      for row in reader:
      ids.add(row[1])

# Flag to indicate if headers have been written
headers_written = False
with open('petases_Repr.json', 'w') as f:
      all_data = []
      for interpro_ID in ids:
      url =
f"https://rest.uniprot.org/uniparc/{interpro_ID}/databases?fields=data
base%2Caccession%2Corganism%2Cfirst_seen%2Clast_seen%2Cactive%2Cgene%2
Corganism_id%2Cprotein&format=json&size=500"
      for batch, total in get_batch(url):
            data = batch.json()
            for entry in data["results"]:
                  entry["ID"] = interpro_ID
                  all_data.append(entry)
            print(f'{len(all_data)} / {total} entries fetched')
```

```
        json.dump(all_data, f)

print("Data fetching completed and saved to interpro_data.json")
```

prepare_MSA_for_alphafold.py

```
import os
import subprocess as sp
import argparse
import string, sys, getopt

#numpy and sklearn
import numpy as np
from sklearn.metrics import pairwise_distances

# Sergei's custom script utils.py
#from utils import *
alpha_1 = list("ARNDCQEGHILKMFPSTWYV-")
states = len(alpha_1)
alpha_3 =
['ALA','ARG','ASN','ASP','CYS','GLN','GLU','GLY','HIS','ILE',

'LEU','LYS','MET','PHE','PRO','SER','THR','TRP','TYR','VAL','GAP']
aa_1_N = {a:n for n,a in enumerate(alpha_1)}
aa_3_N = {a:n for n,a in enumerate(alpha_3)}
aa_N_1 = {n:a for n,a in enumerate(alpha_1)}
aa_1_3 = {a:b for a,b in zip(alpha_1,alpha_3)}
aa_3_1 = {b:a for a,b in zip(alpha_1,alpha_3)}

def AA_to_N(x):
  # ["ARND"] -> [[0,1,2,3]]
  x = np.array(x);
  if x.ndim == 0: x = x[None]
  return [[aa_1_N.get(a, states-1) for a in y] for y in x]

def N_to_AA(x):
  # [[0,1,2,3]] -> ["ARND"]
  x = np.array(x);
  if x.ndim == 1: x = x[None]
  return ["".join([aa_N_1.get(a,"-") for a in y]) for y in x]

def parse_fasta(filename, a3m=False):
  '''function to parse fasta file'''
  if a3m:
      # for a3m files the lowercase letters are removed
      # as these do not align to the query sequence
      rm_lc = str.maketrans(dict.fromkeys(string.ascii_lowercase))
  header, sequence = [],[]
  lines = open(filename, "r")
  for line in lines:
      line = line.rstrip()
      if len(line) > 0:
      if line[0] == ">":
      header.append(line[1:])
      sequence.append([])
      else:
      if a3m: line = line.translate(rm_lc)
      else: line = line.upper()
```

```
            sequence[-1].append(line)
    lines.close()
    sequence = [''.join(seq) for seq in sequence]
    return header, sequence

def mk_msa(seqs):
    '''one hot encode msa'''
    alphabet = list("ARNDCQEGHILKMFPSTWYV-")
    states = len(alphabet)

    alpha = np.array(alphabet, dtype='|S1').view(np.uint8)
    msa = np.array([list(s) for s in seqs], dtype='|S1').view(np.uint8)
    for n in range(states):
        msa[msa == alpha[n]] = n
    msa[msa > states] = states-1

    return np.eye(states)[msa]
```

cluster_fetcher.py

```
import os
import csv
from Bio import SeqIO

members = []
clus=set()
queries = []
output_path =
"/home/nvergoulidis/scripts/results/mmseq/Cocaine/Cocaine_New/Cocaine_
20_or_more.fasta"
fasta_path =
"/home/nvergoulidis/scripts/results/mmseq/Cocaine/Cocaine_New/Cocaine_
new_res_all_seqs.fasta"
with open
("/home/nvergoulidis/scripts/results/mmseq/Cocaine/Cocaine_New/Cocaine
_20ormore_rep.tsv", newline='') as tsv_file:
      reader = csv.reader(tsv_file, delimiter='\t')
      for row in reader:
      clus.add(row[0])

with open
("/home/nvergoulidis/scripts/results/mmseq/Cocaine/Cocaine_New/Cocaine
_new_res_cluster.tsv") as res_file:
      reader = csv.reader(res_file, delimiter='\t')

      for row in reader:
      rep = row[0]
      memb = row[1]

      if rep in clus:
            members.append(memb)

len_sum = 0
sum_all = 0
for record in SeqIO.parse(fasta_path, "fasta"):
      if record.id in members:
      queries.append(record)
      len_sum+=len(record)
      sum_all+=1
avg = len_sum / sum_all
print("Average lenght of sequences: ", avg)



SeqIO.write(queries, output_path, "fasta")
```

cluster_separator.py

```
def write_clusters(input_file):
      with open(input_file, 'r') as f:
      lines = f.readlines()

      num_lines = len(lines)
      cluster_num = 1
      cluster_file = None

      for i in range(num_lines - 1):            line1 = lines[i]
      line2 = lines[i + 1]

      if line1.startswith(">") and line2.startswith(">"):
            # Start of a new cluster
            if cluster_file:
                  cluster_file.close()

            cluster_name = line1.strip().lstrip(">")
            cluster_file = open(f"{cluster_name}.fasta", 'w')
            cluster_file.write(line1)

      elif cluster_file:
            cluster_file.write(line1)

      # Close the last cluster file
      if cluster_file:
      cluster_file.close()

input_file = "Cocaine_20_or_more.fasta"
write_clusters(input_file)
```

fasta_fetcher_new.py

```
import csv
import os
from Bio import SearchIO
from Bio import SeqIO

skipped_IDs = set()
query_IDs = set()
uniparc_fastas_1 =
"/home/nvergoulidis/scripts/data/uniparc_actives/uniparc_1st_set"
uniparc_fastas_2 =
"/home/nvergoulidis/scripts/data/uniparc_actives/uniparc_2nd_set"
out_file =
"/home/nvergoulidis/scripts/results/Unique_Uniparc_Results/Unique_Fast
a_Results/Petases_Ultimate_unique.fasta"
catches = []
with open("Feruloyl_vs_Petases_Ultimate_Best_Scores.tsv", newline='')
as skipped_1:
      reader = csv.reader(skipped_1,delimiter='\t')
      for row in reader:
      skipped_IDs.add(row[0])
with open("Cocaine_vs_Petases_Ultimate_Best_Scores.tsv", newline='')
as skipped_2:
      reader = csv.reader(skipped_2, delimiter='\t')
      for row in reader:
      skipped_IDs.add(row[0])
with open("Nattokinases_vs_Petases_Ultimate_Best_Scores.tsv",
newline='') as skipped_3:
      reader = csv.reader(skipped_3, delimiter='\t')
      for row in reader:
      skipped_IDs.add(row[0])
with open("Petases_Ultimate_Best_Scores.tsv", newline='') as ids_tsv:
      reader = csv.reader(ids_tsv, delimiter='\t')
      for row in reader:
      if row[0] not in skipped_IDs:
            query_IDs.add(row[0])
print(len(query_IDs))

for fasta_file in os.listdir(uniparc_fastas_1):
      fasta_path = os.path.join(uniparc_fastas_1, fasta_file)
      print("parsing fasta file:\t"+fasta_path)
      for prot_record in SeqIO.parse(fasta_path, "fasta"):
            if prot_record.id in query_IDs:
                  catches.append(prot_record)
for fasta_file in os.listdir(uniparc_fastas_2):
      fasta_path = os.path.join(uniparc_fastas_2, fasta_file)
      print("parsing fasta file:\t"+fasta_path)
      for prot_record in SeqIO.parse(fasta_path, "fasta"):
            if prot_record.id in query_IDs:
                  catches.append(prot_record)
SeqIO.write(catches, out_file, "fasta")
```

scores_filtering.py

```
import os
import csv
import argparse

parser = argparse.ArgumentParser()
parser.add_argument("-a")
parser.add_argument("-b")
args = parser.parse_args()
a = args.a
b = args.b

output_a_FILE =
f"{a.split('_')[0]}_vs_{b.split('_')[0]}_Ultimate_Best_Scores.tsv"
output_b_FILE =
f"{b.split('_')[0]}_vs_{a.split('_')[0]}_Ultimate_Best_Scores.tsv"
output_common_FILE = f"{a.split('_')[0]} +
{b.split('_')[0]}_Ultimate_Best_Scores.tsv"
out_sum = f"{a.split('_')[0]} +
{b.split('_')[0]}_Ultimate_OVERLAP_with_both_scores.tsv"
output_a_descend_FILE =
f"{a.split('_')[0]}_vs_{b.split('_')[0]}_Descending.tsv"
output_b_descend_FILE =
f"{b.split('_')[0]}_vs_{a.split('_')[0]}_Descending.tsv"

print(f"{a} vs {b}")
IDs_a = {}
IDs_b = {}
with open(a,"r",newline='') as a_scores_tsv:
      reader = csv.reader(a_scores_tsv, delimiter='\t')
      for line in reader:

      id = line[0]
      a_score = float(line[1])
      IDs_a[id] = [a_score]


with open(b,"r",newline='') as b_scores_tsv:
      reader = csv.reader(b_scores_tsv, delimiter='\t')
      for line in reader:

      id = line[0]
      b_score = float(line[1])

      if id in IDs_a:

            IDs_b[id] = IDs_a[id]
            IDs_b[id].append(b_score)

a_sortedIDs = dict(sorted(IDs_b.items(), key=lambda item: item[1][0],
reverse=True))

b_sortedIDs = dict(sorted(IDs_b.items(), key=lambda item: item[1][1],
```

```
reverse=True))

with open(output_a_FILE,"w", newline='') as output_file:
      writer = csv.writer(output_file, delimiter='\t')
      for i in a_sortedIDs:
      if a_sortedIDs[i][0]>a_sortedIDs[i][1]:
            writer.writerow([i,a_sortedIDs[i][0], a_sortedIDs[i][1]])

with open(output_b_FILE,"w", newline='') as output_file:
      writer = csv.writer(output_file, delimiter='\t')
      for i in b_sortedIDs:
      if b_sortedIDs[i][1]>b_sortedIDs[i][0]:
            writer.writerow([i,b_sortedIDs[i][1], b_sortedIDs[i][0]])

with open(output_common_FILE, "w", newline='') as output_file:
      writer = csv.writer(output_file, delimiter='\t')
      for i in b_sortedIDs:
      if b_sortedIDs[i][0]==b_sortedIDs[i][1]:
            writer.writerow([i,b_sortedIDs[i][1], b_sortedIDs[i][0]])


with open(out_sum,"w", newline='') as output_sum:
      writer = csv.writer(output_sum, delimiter='\t')

writer.writerow(["ID",f"{a.split('_')[0]}_Score",f"{b.split('_')[0]}_S
core"])
      for i in IDs_b:
      writer.writerow([i,IDs_b[i][0],IDs_b[i][1]])



with open(output_a_descend_FILE,'w', newline='') as out:

print("ID",f"{a.split('_')[0]}",f"{b.split('_')[0]}",sep='\t',file=out
)
      for i in a_sortedIDs:
      print(i,a_sortedIDs[i][0],a_sortedIDs[i][1],sep='\t', file= out)

with open(output_b_descend_FILE,'w', newline='') as out2:

print("ID",f"{a.split('_')[0]}_Score",f"{b.split('_')[0]}_Score",sep='
\t',file=out2)
      for j in b_sortedIDs:
      print(j,b_sortedIDs[j][0],b_sortedIDs[j][1],sep='\t', file=
out2)
```

scoring.py

```
import os
import csv
from Bio import SearchIO
IDs = {}
in_folder =
"/home/nvergoulidis/scripts/results/hmm_results/hmm_summary/Mixed_Summ
ary/Cocaine_all"
output =
"/home/nvergoulidis/scripts/results/hmm_results/hmm_summary/Mixed_Summ
ary/Feruloyl_all/Cocaine_PFAM_Ultimate_Best_Scores.tsv"
for file in os.listdir(in_folder):
    print(f"processing file : {file}")
    with open(file, newline='') as dom:
        reader = SearchIO.parse(file, 'hmmsearch3-domtab')
        for result in reader:

            for hit in result:

                for hsp in hit:

                    if result.accession ==
"PF02129.22"or"PF00135.32"or"PF08530.14" and hit.id not in IDs:
                        IDs[hit.id] = [float(hit.bitscore),
hit.evalue, hsp.hit_start, hsp.hit_end]

                    elif result.accession ==
"PF02129.22"or"PF00135.32"or"PF08530.14" and hit.id in IDs:
                        if float(hit.bitscore) >
float(IDs[hit.id][0]):

                            IDs.update([(hit.id,[float(hit.bitscore),
hit.evalue, hsp.hit_start, hsp.hit_end])])

                        else:
                            pass


with open(output,'w', newline='') as out:
    writer = csv.writer(out, delimiter='\t')
    for i in IDs:

        writer.writerow([i,IDs[i][0],IDs[i][1],IDs[i][2],IDs[i][3]])
```

pfam_counter.sh

```
#!/bin/bash
folder_path="/home/nvergoulidis/MetaSA-scan/testing_results"
for file in "$folder_path"/uniparc_active_p*.domtblout; do
      awk '!seen[$1]++ {
      if ($5 == "PF00082.26" || $5 == "PF05922.20") {
            natt += 1
      } else if ($5 == "PF00135.32" || $5 == "PF02129.22" || $5 ==
"PF08530.14") {
            cocaine += 1
      } else if ($5 == "PF01764.29" || $5 == "PF07519.15" || $5 ==
"PF10503.13") {
            feruloyl += 1
      } else if ($5 == "PF12740.11") {
            petases += 1
      }
      }
      END {
      print "Natt = " natt
      print "Cocaine = " cocaine
      print "Feruloyl = " feruloyl
      print "Petases = " petases
      }' "$file" > "${file}_output.txt"
done
```

hmmer_searcher.sh

```
#!/bin/bash

folder_path="/home/nvergoulidis/scripts/data/uniparc_actives/uniparc_1
st_set"
folder_path2="/home/nvergoulidis/scripts/data/uniparc_actives/uniparc_
2nd_set"

for file in "$folder_path"/uniparc_active_p*.fasta; do
   echo "Processing file: $file"


   hmmsearch  -T 25.0 --domT 22.0 --incT 7.0 --incdomT 5.0 --tblout
"${file%.fasta}_Feruloyl_results_per_sequence.tblout" \
      --domtblout
"${file%.fasta}_Feruloyl_results_per_domain.domtblout" \
      -o "${file%.fasta}_Feruloyl_raw_output.hmmout" --cpu 32
Feruloyl.hmm "$file"
done

for file in "$folder_path2"/uniparc_active_p*.fasta; do
   echo "Processing file: $file"


   hmmsearch  -T 25.0 --domT 22.0 --incT 7.0 --incdomT 5.0 --tblout
"${file%.fasta}_Feruloyl_results_per_sequence.tblout" \
      --domtblout
"${file%.fasta}_Feruloyl_results_per_domain.domtblout" \
      -o "${file%.fasta}_Feruloyl_raw_output.hmmout" --cpu 32
Feruloyl.hmm "$file"
done
```

UniParc_Fetcher.py

```
import requests
import json
import pymongo
from pymongo import MongoClient
import pandas as pd

client = MongoClient("mongodb://localhost:27017/backend")
db = client.backend
collection = db["Feruloyl Esterases"]


tsv_file =
"/home/nikoverg/Documents/bioinformatics/projects/Thesis/results/ferul
oyl_metadata.tsv"
ids_df = pd.read_csv(tsv_file, sep='\t')
ids = ids_df['ID'].tolist()

def fetch_data_from_uniparc(id):
      url = f"https://rest.uniprot.org/uniparc/{id}.json"
      response = requests.get(url)
      if response.status_code == 200:
      return response.json()
      else:
      print(f"Failed to fetch data for ID {id}:
{response.status_code}")
      return None


for uniparc_id in ids:
      # Check if the data for the ID already exists in the collection
      if collection.find_one({"uniParcId": uniparc_id}):
      print(f"Data for ID {uniparc_id} already exists in the
collection. Skipping fetch.")
      continue

      data = fetch_data_from_uniparc(uniparc_id)
      if data:

      if not collection.find_one({"uniParcId": data["uniParcId"]}):
            collection.insert_one(data)
      else:
            print(f"Record with ID {data['uniParcId']} already exists.
Skipping insertion.")
client.close()
```