ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

Ανάπτυξη Διαδικτυακής Βάσης Δεδομένων και Εργαλείων Εκτίμησης Αντιμικροβιακών Ιδιοτήτων για Πεπτίδια και Τμήματα Πρωτεϊνών

Αλεξανδρίδης Ευθύμιος

ΑΜ 00545

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ
Υπεύθυνος
Παντελής Μπάγκος
Αναπληρωτής Καθηγητής

Λαμία, Φεβρουάριος 2016

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

Ανάπτυξη Διαδικτυακής Βάσης Δεδομένων και Εργαλείων Εκτίμησης Αντιμικροβιακών Ιδιοτήτων για Πεπτίδια και Τμήματα Πρωτεϊνών

Αλεξανδρίδης Ευθύμιος

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Επιβλέπων
Παντελής Μπάγκος
Αναπληρωτής καθηγητής

Λαμία, Φεβρουάριος 2016

Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις [(1)], που προβλέπονται από της διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

1. *Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί* ***χωρίς να τα περικλείω σε εισαγωγικά*** *και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάσθηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.*

2. *Δέχομαι ότι η αυτολεξεί* ***παράθεση χωρίς εισαγωγικά****, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.*

3. *Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια*

4. *Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.*

Ημερομηνία: ……/…../20……

Ο – Η Δηλ.

(Υπογραφή)

(1) «Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται ή αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση του άρθρου 8 παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων σκόπευε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκόπευε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ετών.

Αλεξανδρίδης Ευθύμιος

Τριμελής Επιτροπή:

Παντελής Μπάγκος , Αναπληρωτής Καθηγητής

Βασίλειος Πλαγιανάκος , Αναπληρωτής Καθηγητής

Μαρία Αδάμ , Επίκουρος Καθηγητής

Ευχαριστίες

Θα ήθελα να ευχαριστήσω την οικογένεια μου για την υποστήριξη τους όλα αυτά τα χρόνια.
Ακόμη, θα ήθελα να ευχαριστήσω τον επιβλέπων καθηγητή μου κ. Παντελή Μπάγκο για την συνεργασία και την κατανόηση του καθώς και τον Ειδικό Λειτουργικό Επιστήμονα Α' του Ιδρύματος Ιατροβιολογικών Ερευνών της Ακαδημίας Αθηνών, κ. Γεώργιο Σπύρου για την ακατάπαυστη καθοδήγηση και υποστήριξη του, καθ' όλη την διάρκεια διεκπεραίωσης της παρούσας πτυχιακής εργασίας καθώς η ολοκλήρωση της πραγματοποιήθηκε υπό την επίβλεψη του.

Περιεχόμενα

Περίληψη

Στα πλαίσια κατανόησης των αντιμικροβιακών πεπτιδίων υπολογίσαμε ορισμένα από τα χαρακτηριστικά των αντιμικροβιακών πεπτιδίων που συλλέξαμε από διαφορές βάσεις δεδομένων (camp, adp, lamp και άλλες) . Τα χαρακτηριστικά αυτά τα υπολογίσαμε με την χρήση υπολογιστικών εργαλείων και μετέπειτα δημιουργήσουμε μία βάση δεδομένων στην οποία εισαγάγαμε τα δεδομένα μας. Πιο συγκεκριμένα στο πρώτο κεφάλαιο πραγματευόμαστε το θεωρητικό υπόβαθρο για την πρωτεΐνη, τις δομές πρωτεϊνών, καθώς και για τα αντιμικροβιακά πεπτίδια. Στο επόμενο κεφάλαιο ορίζουμε τα χαρακτηριστικά των αντιμικροβιακών πεπτιδίων, ενώ μετέπειτα στο τρίτο κεφάλαιο δημιουργούμε την βάση δεδομένων, υπολογίζουμε τα αντιμικροβιακά χαρακτηριστικά και παράλληλα δημιουργούμε τον ιστότοπο που συνδέεται με την βάση δεδομένων μας. Τέλος προσπαθούμε να ομαδοποιήσουμε τα δεδομένα μας και να κατανοήσουμε κατά πόσο τελικά οι ομάδες που δημιουργούμε ταιριάζουν με κάποιες από τις κατηγορίες των αντιμικροβιακών πεπτιδίων που γνωρίζουμε, με μελλοντικό στόχο την δημιουργία ευφυούς συστήματος.

ΚΕΦΑΛΑΙΟ 1ο

Εισαγωγή

Τα αντιμικροβιακά πεπτίδια ανιχνεύονται σε όλα τα είδη του ζωικού βασιλείου, καθώς και σε μερικά είδη φυτών, ως ισχυροί παράγοντες κατά των μικροοργανισμών. Η κύρια δράση τους ασκείται μέσω της αλληλεπίδρασής τους με τη μεμβράνη των μικροοργανισμών, επιτυγχάνοντας τελικά την εξόντωσή τους. Συνεπώς κρίνεται σπουδαία η αναγνώριση των αντιμικροβιακών πεπτιδίων αφού αποτελούν άμεσους θεραπευτικούς παράγοντες κατά των ιών, των βακτηρίων, των καρκινικών κυττάρων, καθώς και άλλων μικροοργανισμών. Αρκετές βάσεις δεδομένων βιοπληροφορικής(CAMP, LAMP, C-PAMP κλπ) περιέχουν αντιμικροβιακά πεπτίδια. Συλλέξαμε λοιπόν όλα τα αντιμικροβιακά πεπτίδια από όλα τα είδη σε επίπεδο τόσο ακολουθίας όσο και δομής (εφόσον φυσικά υπάρχει) ψάχνοντας τις αντίστοιχες πρωτεΐνες στην PDB και ανακαλύπτοντας έτσι την τρισδιάστατη δομή της συγκεκριμένης πρωτεΐνης. Μετέπειτα αφού δημιουργήσαμε τα δεδομένα μας (dataset), αναπτύξαμε μία βάση δεδομένων εισάγοντας όλα τα αναγνωρισμένα αντιμικροβιακά πεπτίδια που υπάρχουν και έχουμε βρει. Στη συνέχεια, εφαρμόσαμε αλγορίθμους ανάλυσης χαρακτηριστικών ακολουθίας προκειμένου να εξάγουμε τα χαρακτηριστικά αυτών των πεπτιδίων και να τα αποθηκεύσουμε στην βάση δεδομένων που δημιουργήσαμε, αντιστοιχίζοντας την κάθε ιδιότητα με το αντίστοιχο αντιμικροβιακό πεπτίδιο. Ακολούθως, με την χρήση υπολογιστικών εργαλείων καταφέραμε εισάγοντας μία ακολουθία στον ιστότοπο, ο χρήστης να παίρνει τα αποτελέσματα των χαρακτηριστικών για την συγκεκριμένη ακολουθία. Ο χρήστης έχει την δυνατότητα λοιπόν είτε να αναζητά όλες τις ακολουθίες με συγκεκριμένο αριθμό αμινοξέων είτε να εισάγει μια ακολουθία μέσω του δικτυακού εργαλείου και να υπολογίζονται επιτόπου τα χαρακτηριστικά αυτά εμφανίζοντας τα στον χρήστη μέσω του δικτυακού εργαλείου.

Αντιμικροβιακά Πεπτίδια (θεωρητικό υπόβαθρο)

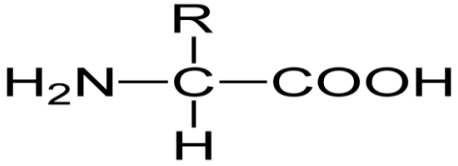
1.1 Πρωτεΐνες

Οι πρωτεΐνες είναι μία σημαντική κατηγορία μεγαλομορίων και αποτελούνται από πολλά (περισσότερα από 100) αμινοξέα συνδεδεμένα μεταξύ τους με πεπτιδικούς δεσμούς. Ανάλογα με τα προϊόντα υδρόλυσης τους διακρίνονται σε απλές και σε σύνθετες πρωτεΐνες. Μπορεί να έχουν σφαιρική ή ινώδη μορφή, ενώ η διαλυτότητά τους στο νερό ποικίλλει και μπορεί να είναι από εντελώς αδιάλυτες έως ευδιάλυτες. Από την άποψη των χημικών ιδιοτήτων οι πρωτεΐνες, όπως και τα αμινοξέα, εμφανίζουν αμφολυτική συμπεριφορά, ενώ υπάρχει για κάθε πρωτεΐνη ένα χαρακτηριστικό ισοηλεκτρικό σημείο. Οι πρωτεΐνες μπορούν να υδρολυθούν προς αμινοξέα, ενώ δίνουν χαρακτηριστικές χρωστικές αντιδράσεις, όπως η αντίδραση διουρίας. Ο βιολογικός ρόλος των πρωτεϊνών ποικίλλει. Οι πρωτεΐνες μπορεί να είναι ένζυμα, δομικές πρωτεΐνες, συσταλτικές πρωτεΐνες, πρωτεΐνες μεταφοράς, ορμονικές πρωτεΐνες, αμυντικές πρωτεΐνες. Απαραίτητη προϋπόθεση για την κατανόηση της φύσης των πρωτεϊνών είναι η περιγραφή των δομικών τους λίθων, δηλαδή, των αμινοξέων.

1.1.1 Αμινοξέα

Ορισμός Αμινοξέων

Τα αμινοξέα είναι μόρια αποτελούμενα από ένα κεντρικό άτομο άνθρακα, που ονομάζεται α-άνθρακας, ενωμένο με μια αμινομάδα ή αμινική ομάδα (-NH2), μια καρβοξυλομάδα (-COOH) και μια πλευρική ομάδα, η οποία συνδέεται μέσω ομοιοπολικού δεσμού με αυτό [1].



Εικόνα 1.1 *Η γενική δομή ενός α-αμινοξέος, με την αμινομάδα στα αριστερά και την καρβοξυλομάδα στα δεξιά*

Η πλευρική ομάδα συμβολίζεται συνήθως με το γράμμα R και αναφέρεται μόνο λεκτικά ως υπόλειμμα (residue). Η πλευρική ομάδα είναι διαφορετική για κάθε αμινοξύ και του προσδίδει μοναδικές χημικές ιδιότητες. Συνεπώς, τα αμινοξέα κατατάσσονται σε κατηγορίες σύμφωνα με το είδος της πλευρικής ομάδας, η οποία τα κάνει να συμπεριφέρονται ως ασθενή οξέα, ως ασθενείς βάσεις, ως υδρόφιλα, αν είναι πολικά, ή ως υδροφοβικά, αν είναι μη πολικά.

•Μη πολικά αμινοξέα, όπως η λευκίνη, συχνά έχουν πλευρικές ομάδες οι οποίες περιέχουν —CH2 ή —CH3.

•Πολικά μη φορτισμένα αμινοξέα, όπως η θρεονίνη, έχουν πλευρικές ομάδες οι οποίες περιέχουν οξυγόνο (ή μόνο —H).

•Φορτισμένα αμινοξέα, όπως το γλουταμικό οξύ, έχουν πλευρικές ομάδες οι οποίες περιέχουν οξέα ή βάσεις.

•Αρωματικά αμινοξέα, όπως η φαινυλαλανίνη, έχουν πλευρικές ομάδες που περιέχουν έναν οργανικό δακτύλιο με εναλλασσόμενους απλούς και διπλούς δεσμούς.

•Αμινοξέα που επιτελούν ειδικές λειτουργίες έχουν ξεχωριστές ιδιότητες, όπως για παράδειγμα η μεθειονίνη, η οποία έχει την τάση να καταλαμβάνει την πρώτη θέση σε μία αλληλουχία αμινοξέων [1].

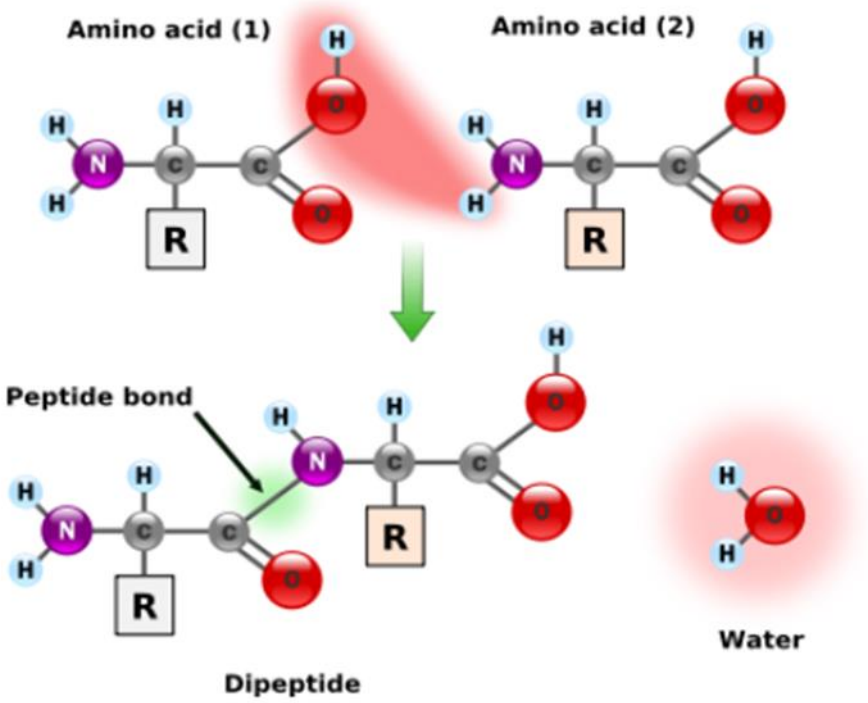
Παρακάτω παρουσιάζονται τα 20 αμινοξέα που συνθέτουν τις πρωτεΐνες των ζωντανών οργανισμών:

Ελληνική ονομασία	Διεθνής σύντμηση	Ελληνική ονομασία	Διεθνής σύντμηση
Αλανίνη	Ala	Λευκίνη*	Leu
Αργινίνη	Arg	Λυσίνη*	Lys
Ασπαραγίνη	Asn	Μεθειονίνη*	Met
Ασπαρτικό οξύ	Asp	Φαινυλαλανίνη*	Phe
Κυστεΐνη	Cys	Προλίνη	Pro
Γλουταμίνη	Gln	Σερίνη	Ser
Γλουταμικό οξύ	Glu	Θρεονίνη*	Thr
Γλυκίνη	Gly	Τρυπτοφάνη*	Trp
Ιστιδίνη	His	Τυροσίνη	Tyr
Ισολευκίνη*	Ile	Βαλίνη*	Val

Εικόνα 1.2 *Τα αμινοξέα κατ' αλφαβητική διεθνή ονομασία. Τα φερόμενα με αστερίσκο (*) είναι τα 8 βασικά αμινοξέα.*

Τα αμινοξέα που συνθέτουν μία πρωτεΐνη ενώνονται μεταξύ τους με δεσμούς πεπτιδίων, σχηματίζοντας μία αλυσίδα πολυπεπτιδίων (peptide chain). Ειδικότερα, η καρβοξυλομάδα του ενός αμινοξέος αντιδρά με την αμινομάδα του γειτονικού

του, απελευθερώνοντας ένα μόριο νερού, καθώς δημιουργείται ο πεπτιδικός δεσμός.



Εικόνα 1.3 *Αντίδραση αμινοξέων προς σχηματισμό πεπτιδικού δεσμού*

1.1.2 Ορισμός πρωτεΐνης

Οι πρωτεΐνες είναι οργανικές αζωτούχες ενώσεις, οι οποίες αποτελούνται κυρίως από C, H, O, N, S . Κάθε πρωτεΐνη χαρακτηρίζεται από την τρισδιάστατη δομή της. Συμμετέχουν ενεργά σε πολλές λειτουργίες και μπορούν να κατηγοριοποιηθούν κυρίως σε τρείς μεγάλες κατηγορίες, τις δομικές πρωτεΐνες, τις πρωτεΐνες με βιολογική δράση, και τις διατροφικές πρωτεΐνες.

Οι δομικές πρωτεΐνες είναι ινώδεις πρωτεΐνες όπως η κερατίνη και το κολλαγόνο και βρίσκονται σε όλους τους ιστούς όπως τους μυς. Αντίστοιχα οι πρωτεΐνες με βιολογική δράση είναι ορμόνες που ρυθμίζουν μεταβολικές αντιδράσεις. Μερικές εκ των οπίων είναι οι συστολικές πρωτεΐνες (μυοσίνη), οι πρωτεΐνες μεταφοράς (αιμοσφαιρίνη), τοξικές πρωτεΐνες (τοξίνη), πρωτεΐνες με προστατευτική δράση

(θρομβίνη) καθώς και αντιβιοτικές πρωτεΐνες. Οι διατροφικές τώρα πρωτεΐνες είναι πρωτεΐνες με βιολογική δράση οι οποίες μπορούν να μεταβολιστούν με την πέψη.

Πιο αναλυτικότερα οι θεμελιώδεις λειτουργίες των πρωτεϊνών είναι οι εξής:

•Ενζυμική κατάλυση: Τα ένζυμα είναι μία κατηγορία πρωτεϊνών, που αποτελούν βιολογικούς καταλύτες, οι οποίοι υποβοηθούν συγκεκριμένες χημικές αντιδράσεις διαδραματίζοντας καθοριστικό ρόλο για την εξέλιξη της ζωής.

•Άμυνα: Μια μερίδα σφαιρικών πρωτεϊνών χρησιμοποιούν τη μορφή τους για να αναγνωρίσουν ξένα μικρόβια και καρκινικά κύτταρα. Αυτοί οι επιφανειακοί κυτταρικοί υποδοχείς διαμορφώνουν τον πυρήνα του ορμονικού και ανοσοποιητικού συστήματος.

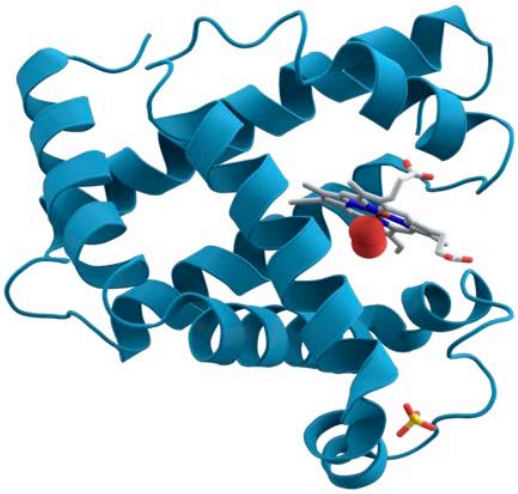
•Μεταφορά: Ποικίλες σφαιρικές πρωτεΐνες μεταφέρουν συγκεκριμένα μικρά μόρια και ιόντα. Για παράδειγμα, η πρωτεΐνη αιμοσφαιρίνη μεταφέρει οξυγόνο στο αίμα και η μυοσφαιρίνη, μία παρόμοια πρωτεΐνη, μεταφέρει οξυγόνο στους μύες.

•Στήριξη: Οι ινώδεις πρωτεΐνες παίζουν δομικό ρόλο στο κύτταρο. Χαρακτηριστικά παραδείγματα είναι η κερατίνη, που είναι συστατικό των μαλλιών και των νυχιών και το κολλαγόνο, που αποτελεί το κυρίαρχο συστατικό των συνδετικών ιστών.

•Κίνηση: Οι μύες συσπώνται μέσω της κίνησης δύο ειδών πρωτεϊνικών μυονηματίων, της ακτίνης και της μυοσίνης.

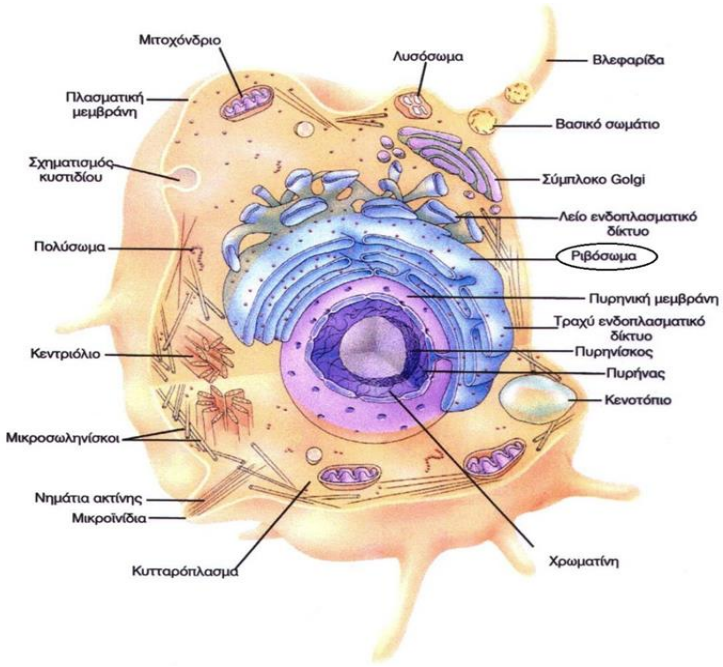
•Ρύθμιση: Κάποιες μικρού μήκους πρωτεΐνες που ονομάζονται ορμόνες λειτουργούν ως διακυτταρικοί αγγελιοφόροι στα ζώα. Γενικά, οι πρωτεΐνες έχουν ποικίλους ρυθμιστικούς ρόλους μέσα στο κύτταρο, ενεργοποιώντας και απενεργοποιώντας, για παράδειγμα, γονίδια κατά τη διάρκεια της ανάπτυξης. Επιπλέον, οι πρωτεΐνες λαμβάνουν πληροφορίες, λειτουργώντας ως επιφανειακοί κυτταρικοί υποδοχείς [2].

Μέσω της υδρόλυσης των πρωτεϊνών λαμβάνονται τα αμινοξέα που προαναφέραμε, τα οποία αποτελούν τις δομικές μονάδες (μονομερή) των πρωτεϊνών. Ο βιολογικός τους ρόλος καθορίζεται κάθε φορά από την τρισδιάστατη δομή τους που είναι συνέπεια της αλληλουχίας των αμινοξέων, η οποία και ξεκινά από την πρωτοταγή δομή [3].



Εικόνα 1.4 *Αναπαράσταση της τρισδιάστατης δομής της μυοσφαιρίνης, που παρουσιάζεται με χρωματισμένες τις άλφα έλικες.*

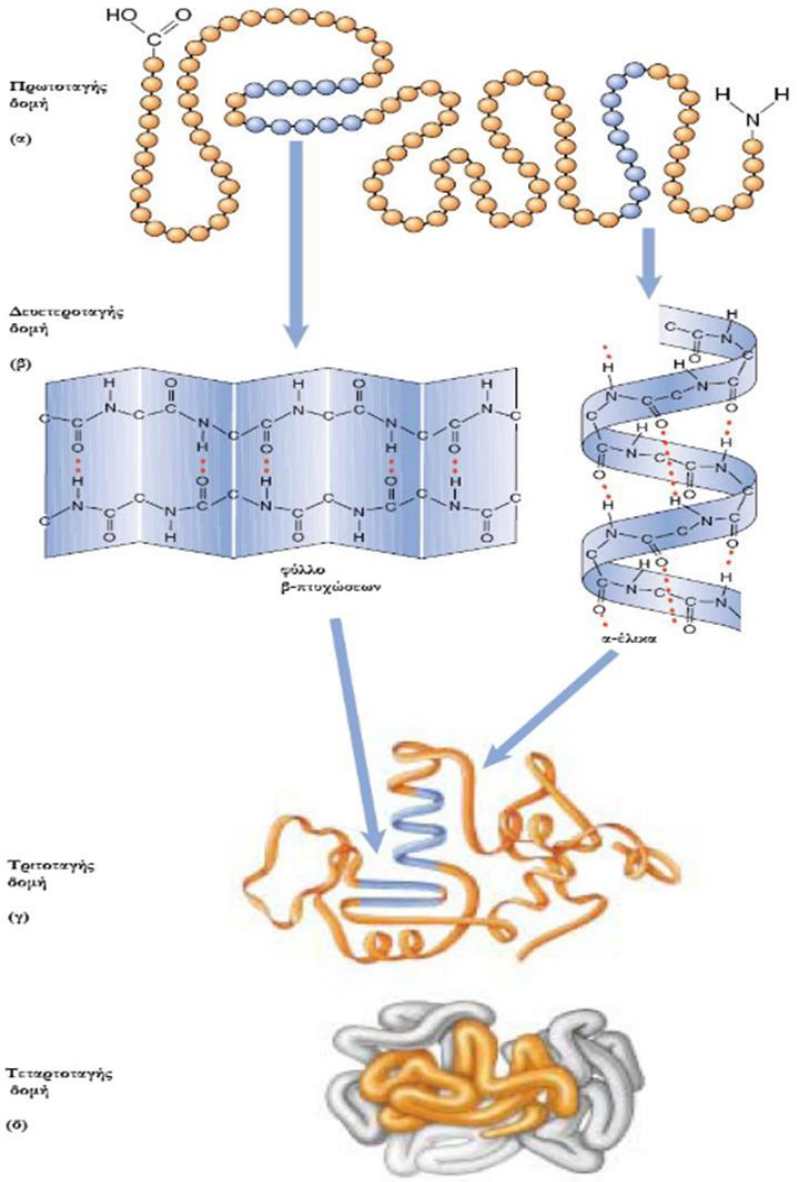
Οι πρωτεΐνες παράγονται από τα ριβοσώματα που βρίσκονται μέσα στο κυτταρόπλασμα και αρχικά εμφανίζονται ως απλές μη διακλαδωμένες αλληλουχίες αμινοξέων, δηλαδή πεπτιδίων ή πολυπεπτιδίων, σχηματίζοντας την "πρωτοταγή δομή", για την οποία καθοριστικοί παράγοντες είναι τα νουκλεϊκά οξέα, τα οποία και φέρονται να ελέγχουν όλες τις λειτουργίες αλλά και τα κληρονομικά γνωρίσματα των οργανισμών.



Εικόνα 1.5 *Εσωτερικό κυττάρου-Ριβοσώματα μέσα στο κύτταρο*

1.1.3 Δομές πρωτεϊνών

Αναλυτικά, η δομή των πρωτεϊνών αναφέρεται συνήθως στα πλαίσια τεσσάρων επιπέδων, πρωτοταγής, δευτεροταγής, τριτοταγής και τεταρτοταγής δομής όπως απεικονίζονται στο παρακάτω σχήμα.



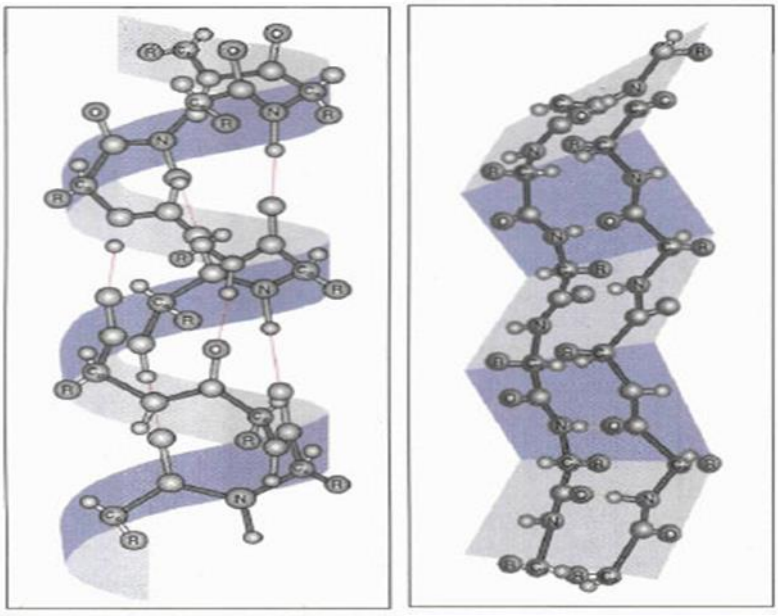
Εικόνα 1.6 *Τα επίπεδα πρωτεϊνικής δομής. (α) Η ακολουθία αμινοξέων μιας πρωτεΐνης αναφέρεται ως πρωτοταγής δομή (β) Δεσμοί μεταξύ γειτονικών αμινοξέων σχηματίζουν φύλλα β-πτυχώσεων και α-έλικες που συνιστούν τη δευτεροταγή δομή (γ) Οι πρωτεΐνες αναδιπλώνονται σχηματίζοντας μια τρισδιάστατη δομή, την τριτοταγή δομή (δ) Η συσσωμάτωση πρωτεϊνών με άλλες πεπτιδικές αλυσίδες δημιουργεί την τεταρτοταγή δομή της πρωτεΐνης.*

Πρωτοταγής δομή

Για να χαρακτηρίσουμε μία πρωτεΐνη ή ένα πεπτίδιο, δεν αρκεί να γνωρίζουμε μόνο από ποιά και από πόσα κατά περίπτωση αμινοξέα αποτελείται. Πρέπει επιπλέον να προσδιοριστεί και η σειρά με την οποία βρίσκονται συνδεδεμένα τα αμινοξέα αυτά. Και τούτο γιατί η σειρά αυτή, δηλαδή η αλληλουχία των αμινοξέων, καθορίζει και τις ιδιότητες του πεπτιδίου ή της πρωτεΐνης. Η αλληλουχία των αμινοξέων μίας πρωτεΐνης αποτελεί την πρωτοταγή της δομή.

Δευτεροταγής δομή

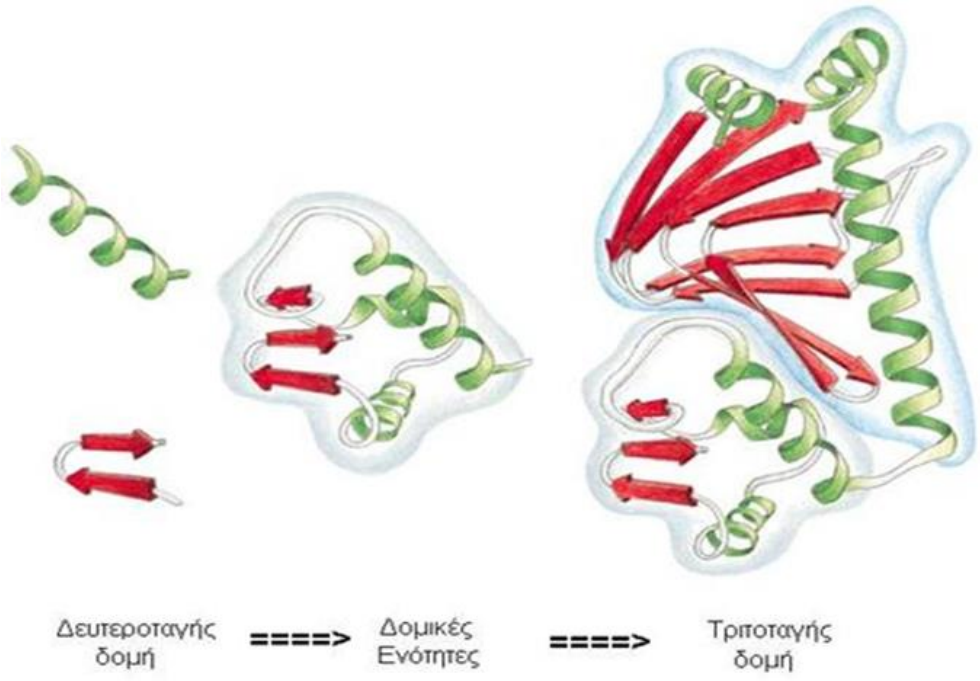
Η αλυσίδα των αμινοξέων που αποτελούν την πρωτεΐνη δεν είναι ευθεία αλλά πραγματοποιεί ορισμένες αναδιπλώσεις προσδίνοντας συγκεκριμένο σχήμα στο χώρο. Η δευτεροταγής δομή αναφέρεται στις αναδιπλώσεις που μπορεί να έχουν τα διάφορα τμήματα μίας πολυπεπτιδικής αλυσίδας. Η μελέτη της δευτεροταγούς δομής πραγματοποιείται με τη βοήθεια κρυσταλλογραφίας ακτίνων Χ. Με αυτό λοιπόν τον τρόπο βρέθηκε ότι οι αλυσίδες των πρωτεϊνών μπορεί να έχουν δύο διαφορετικές μορφές και συγκεκριμένα: α) την μορφή α-έλικας, β) την μορφή β-πτυχωτής επιφάνειας [3].



Εικόνα 1.7 *Δευτεροταγής δομής α-έλικας, και β-πτυχωτής επιφάνειας.*

Τριτοταγής δομή

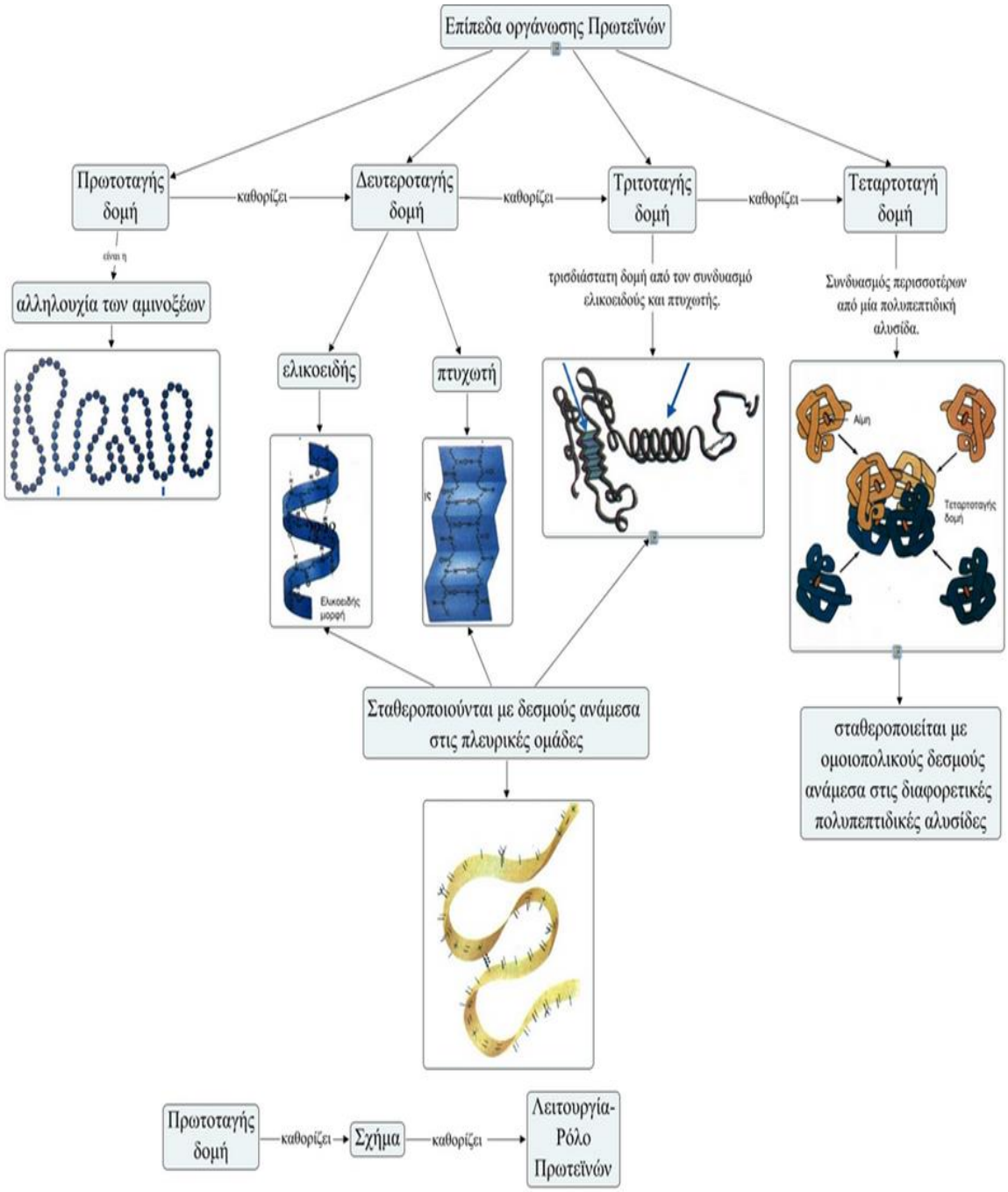
Η αναδιπλωμένη έλικα μίας πρωτεϊνικής αλυσίδας αναδιπλώνεται σε διάφορα τμήματά της προσδίνοντας στο πρωτεϊνικό μόριο συνολικά, ένα συγκεκριμένο σχήμα. Ο τρόπος αναδίπλωσης για ολόκληρη την πρωτεϊνική αλυσίδα αποτελεί την τριτοταγή δομή των πρωτεϊνών όπως φαίνεται στο παρακάτω σχήμα. Πιο συγκεκριμένα η αναδιπλωμένη μορφή της πρωτεΐνης που περιλαμβάνει διάφορα μοτίβα, αναδιπλώνοντας τις μη πολωμένες πλευρικές ομάδες στο εσωτερικό της, αποτελεί την τριτοταγή δομή της πρωτεΐνης. (Μοτίβα αποτελούν τα στοιχεία της δευτεροταγούς δομής που μπορούν να συνδυαστούν στις πρωτεΐνες με συγκεκριμένους τρόπους που ονομάζονται διαφορετικά υπερδευτεροταγείς δομές (super secondary structure). Οι πρωτεΐνες οδηγούνται στην τριτοταγή τους δομή εξαιτίας υδροφοβικών αλληλεπιδράσεων με το νερό. Η τελική αναδίπλωση μίας πρωτεΐνης καθορίζεται από τη χημική φύση των πλευρικών της ομάδων και συνεπώς από την πρωτοταγή της δομή. Είναι χαρακτηριστικό ότι πολλές πρωτεΐνες αναπτύσσονται και επαναδιπλώνονται στη χαρακτηριστική τους δομή αυτενεργώς. Επίσης, στο εσωτερικό των αναδιπλωμένων πρωτεϊνών δεν παρουσιάζονται κενά ή κοιλότητες. Με τον τρόπο αυτό μπορεί να εξηγηθεί και η πληθώρα των μη πολωμένων αμινοξέων (αλανίνη, βαλίνη, λευκίνη, ισολευκίνη). Είναι προφανές ότι η αλλαγή ενός μη πολωμένου αμινοξέος στο εσωτερικό της πρωτεΐνης σε ένα άλλο διαταράσσει πολύ συχνά την ευστάθεια της πρωτεΐνης και είναι δυνατό να οδηγήσει σε αλλαγή ή απώλεια της λειτουργικότητας της.



Εικόνα 1.8 *Απεικόνιση τριτοταγούς δομής*

Τεταρτοταγής δομή

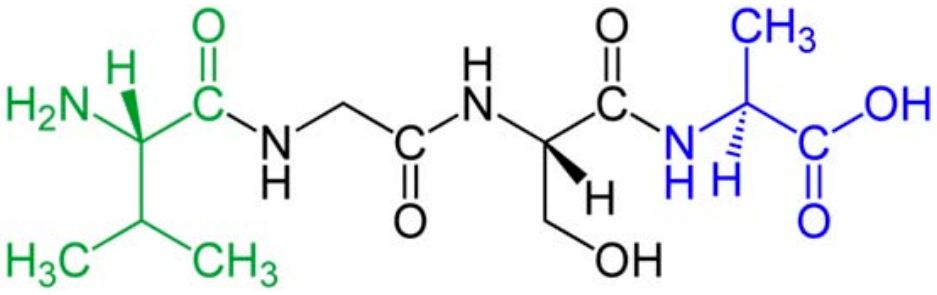
Όμοιες ή διαφορετικές πολυπεπτιδικές αλυσίδες που έχουν αναδιπλωθεί μπορούν συχνά να συνενώνονται μεταξύ τους σχηματίζοντας μεγαλύτερα πρωτεϊνικά σύμπλοκα. Το τελικό σχήμα που αποκτά το πρωτεϊνικό σύμπλοκο στο χώρο αποτελεί την τεταρτοταγή δομή της πρωτεΐνης, ενώ οι ανεξάρτητες πεπτιδικές αλυσίδες που συνθέτουν το πρωτεϊνικό σύμπλοκο αποτελούν τις υπομονάδες.



Εικόνα 1.9 *Συνολική αναπαράσταση των δομών των πρωτεϊνών*

1.2 Πεπτίδια

Με τον όρο πεπτίδιο εννοούμε μια ομάδα οργανικών ενώσεων που έχουν ως βάση τους δύο ή περισσότερα αμινοξέα τα οποία συνδέονται μεταξύ τους με χημικό δεσμό ο οποίος ονομάζεται πεπτιδικός δεσμός. Ανάλογα με τον αριθμό των πεπτιδιών συγκαταλέγονται σε διπεπτίδια (2 αμινοξέα) τριπεπτίδια (3 αμινοξέα) και ούτω καθεξής. Τα πεπτίδια λοιπόν αποτελούν το δομικό στοιχείο των πρωτεϊνών.



Εικόνα1.10 *Απεικόνιση Πεπτιδίου*

Όπως γνωρίζουμε, οι πρωτεΐνες αποτελούνται από αμινοξέα, που συνδέονται μεταξύ τους και σχηματίζουν αλυσίδες διαφορετικού μήκους, μακρές, που αποτελούνται από δεκάδες αμινοξέα και κοντές, που έχουν από 2 έως 20 αμινοξέα. Αυτά είναι λοιπόν είναι τα πεπτίδια. Στον οργανισμό, τα πεπτίδια δρουν σαν κλειδιά «πληροφοριών». Αυτά μεταδίδουν και αποκωδικοποιούν την πληροφορία σε αυστηρά συγκεκριμένα κύτταρα [4].

1.2.1 Ορισμός αντιμικροβιακών πεπτιδίων

Μια κατηγορία πεπτιδίων αποτελούν τα αντιμικροβιακά πεπτίδια, τα οποία ανιχνεύονται σε όλα τα είδη των οργανισμών, δρώντας ως χημικοί παράγοντες κατά των μικροοργανισμών. Μέσω της αλληλεπίδρασης τους με την μεμβράνη των μικροοργανισμών καταφέρνουν τελικά να τα καταστρέψουν και να τα οδηγήσουν στο «θάνατο». Συνεπώς η λειτουργία τους είναι να αντιμετωπίζουν και να καταστρέφουν έναν αριθμό βακτηριδίων, ιών μυκήτων καθώς ακόμα και καρκινικών κυττάρων [5]. Σε πολλά είδη ζώων αποτελεί την πρώτη γραμμή άμυνας καθώς αντιμετωπίζουν τα βακτήρια την στιγμή της εισβολής ενώ ο οργανισμός προσπαθεί να δημιουργήσει αντισώματα για την καταπολέμηση του «εισβολέα».

Ορισμένες ουσίες με αντιμικροβιακή δράση στον ανθρώπινο οργανισμό είναι: [6]

«Ιντερφερόνες: Στην περίπτωση των ιών δρα ένας επιπλέον μηχανισμός μη ειδικής άμυνας. Όταν κάποιος ιός μολύνει ένα κύτταρο, προκαλεί την παραγωγή ειδικών πρωτεϊνών, των ιντερφερονών. Σε ένα πρώτο στάδιο οι ιντερφερόνες ανιχνεύονται στο κυτταρόπλασμα του μολυσμένου κυττάρου. Σε επόμενο όμως στάδιο οι ιντερφερόνες απελευθερώνονται στο μεσοκυττάριο υγρό και από εκεί απορροφούνται από τα γειτονικά υγιή κύτταρα. Με την εισαγωγή των ιντερφερονών στα υγιή κύτταρα ενεργοποιείται η παραγωγή άλλων πρωτεϊνών, οι οποίες έχουν την ικανότητα να παρεμποδίζουν τον πολλαπλασιασμό των ιών. Έτσι τα υγιή κύτταρα προστατεύονται, γιατί ο ιός, ακόμη και αν κατορθώσει να διεισδύσει σ' αυτά, είναι ανίκανος να πολλαπλασιαστεί.

Συμπλήρωμα: Πρόκειται για ομάδα είκοσι πρωτεϊνών στον ορό του αίματος με αντιμικροβιακή δράση.

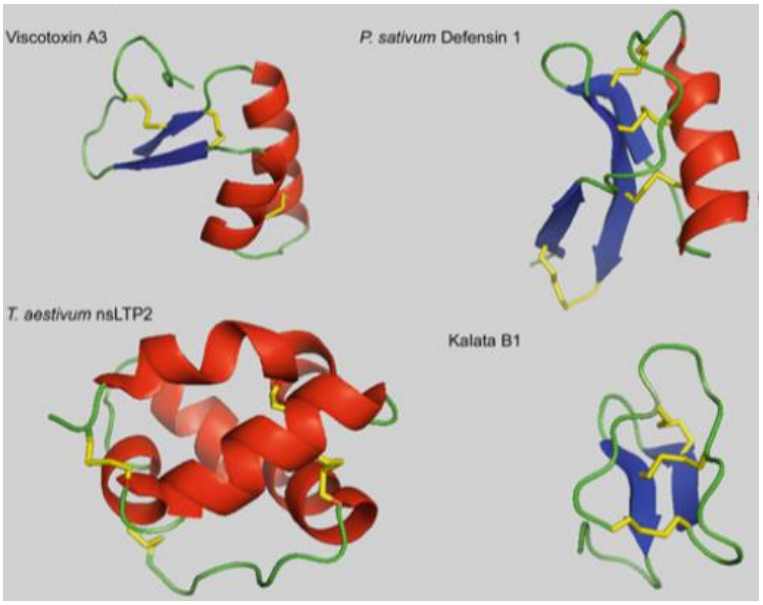
Προπερδίνη: Είναι μια ομάδα τριών πρωτεϊνών στον ορό του αίματος που δρα σε συνδυασμό με τις πρωτεΐνες του συμπληρώματος για την καταστροφή των μικροβίων.»

Όσο αναφορά τα φυτά οι πιο σημαντικές κατηγορίες είναι οι θειονίνες, αμυντοσίνες, οι πρωτεΐνες μεταφοράς λιπιδίων (LTPs) [7].

Θειονίνες: Αποτελούνται από ένα μείγμα πουροθειονίνης Α και πουροθειονίνης Β δρώντας ως αντιμικροβιακές ουσίες στον οργανισμό των φυτών. Η δομη τους είναι γ-έλικα αποτελούμενη από δύο αντιπαράλληλες α-έλικες .

Αμυντοσίνες: Οι περισσότερες αμυντίνες αποτελούνται από μια ακολουθία σήματος που στοχεύουν τα πεπτίδια στο ER.

Πρωτεΐνες μεταφοράς λιπιδίων (LTPs): Αποτελούνται από δύο είδη οικογενειών τις LTP1 και LTP2, σε μέγεθος 90-95 αμινοξέων.

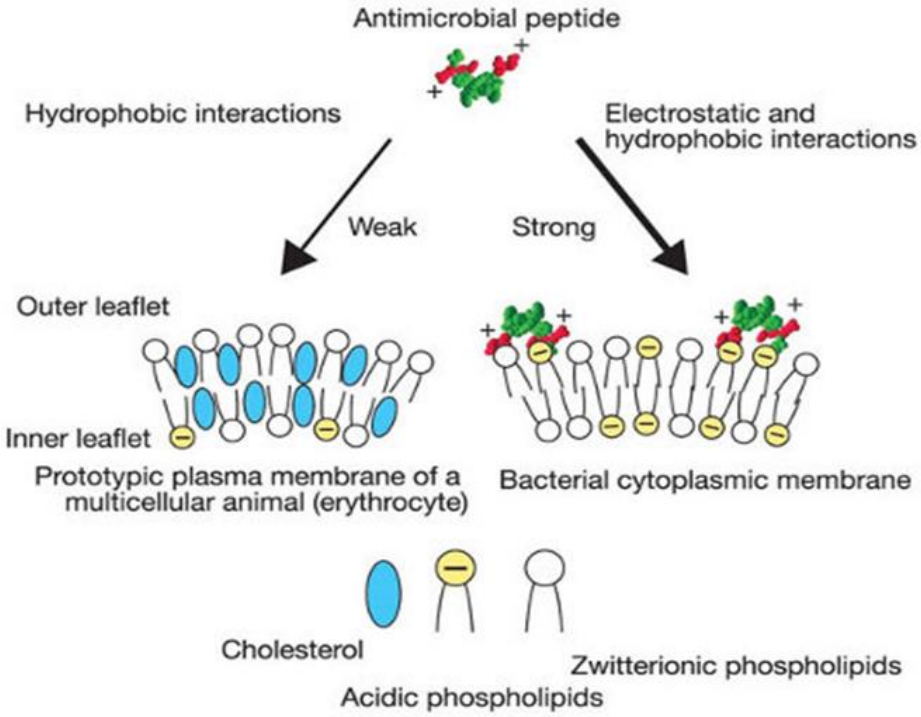


Εικόνα 1.11 *Δομές αντιμικροβιακών πεπτιδίων από τα φυτά. Viscotoxin A3 (PDB entry 1ED0), defensin 1 (PDB entry 1JKZ), η πρωτεΐνη μεταφοράς λιπιδίων nsLTP2 (PDB entry 1TUK), και Kalata B1 (PDB entry 1NB1)*

1.2.2 Ιδιότητες αντιμικροβιακών πεπτιδίων

Κύριο χαρακτηριστικό των αντιμκροβιακών πεπτιδίων είναι να εξουδετερώνουν μικροβιακά καρκινικά κύτταρα και ιούς [8].

Η «επαφή» μεταξύ του κυττάρου-στόχου και του οργανισμού είναι ηλεκτροστατική καθώς τα περισσότερα αντιμικροβιακά πεπτίδια είναι ανιονικά ή υδρόφοβα. Μπορούν να εισχωρήσουν μέσα στο κύτταρο μετά από την επαφή και να προκαλέσουν τελικά το θάνατο του κυττάρου αυτού [9]. Η μεμβράνη των βακτηρίων είναι πλούσια σε φωσφολιπίδια τα οποία είναι αρνητικά φορτισμένα, με αποτέλεσμα να προσελκύουν τα θετικά φορτισμένα αντιμικροβιακά πεπτίδια και να προκαλλούνται ηλεκτροστατικές δυνάμεις. Παράλληλα δημιουργούνται υδρόφοβες αλληλεπιδράσεις μεταξύ των υδρόφοβων περιοχών των αντιμικροβιακών πεπτιδίων και των φωσφολιπιδίων στην επιφάνεια των βακτηριακών μεμβρανών, όπως φαίνεται στην ακόλουθη εικόνα [10]. Όσο αναφορά τα φυτά και τα θηλαστικά τα αρνητικά φορτισμένα λιπίδια είναι λιγότερα ενώ οι μεμβράνες είναι συνήθως κατασκευασμένες με zwitterionic phosphatidy choline και sphingomyelin καθιστώντας πιο αδύναμη την υδρόφοβη αλληλεπίδραση με το αντιμικροβιακο πεπτίδιο [11]. Παρακάτω φαίνεται η αλληλεπίδραση των πεπτιδίων.



Εικόνα 1.12 *Αλληλεπίδραση επιφάνειας βακτηρίου με το αντιμικροβιακό πεπτίδιο*

Αφότου προσδεθούν στο κύτταρο αναστέλλουν την σύνθεση κυτταρικού τοιχώματος του DNA και άλλων ορισμένων ενζύμων του κυττάρου οδηγώντας το τελικά σε θάνατο. Ωστόσο δεν έχει αποδειχθεί με απόλυτη ακρίβεια ο τρόπος δράσης των αντιμικροβιακών πεπτιδίων, αφότου προσδεθούν στο βακτήριο στόχος.

1.3 Βάσεις δεδομένων αντιμικροβιακών πεπτιδίων

Για την διαχείριση το πεπτιδίων έχουν δημιουργηθεί διάφορες βάσεις παγκοσμίως που περιέχουν τα αντιμικροβιακά πεπτίδια. Ορισμένες όπως η phytamp [12] που ειδικεύεται στα αντιμικροβιακά πεπτίδια φυτών ενώ άλλες όπως η camp [18] η Apd [13] και η Lamp [14] περιέχουν πεπτίδια από όλα τα είδη των οργανισμών.

Στον πίνακα παρακάτω παρουσιάζονται οι πιο σημαντικές βάσεις δεδομένων αντιμικροβιακών πεπτιδίων καθώς και κάποιες πιο γενικές βάσεις πεπτιδίων/πρωτεϊνών με τους ιστοτόπους τους.

Ονομασία Βάσης	***Ιστότοπος***
Phytamp	http://phytamp.pfba-lab-tun.org
APD	http://aps.unmc.edu/AP/
Lamp	http://biotechlab.fudan.edu.cn/database/lamp/
Uniprot	http://www.uniprot.org
Ncbi	http://www.ncbi.nlm.nih.gov
Pdb	http://www.rcsb.org/pdb/home/home.do
Camp	*http://www.camp.bicnirrh.res.in*

Ας δούμε αναλυτικότερα τις κύριες βάσεις αντιμικροβιακών πεπτιδίων:

Camp (Collection of Anti-Microbial Peptides)

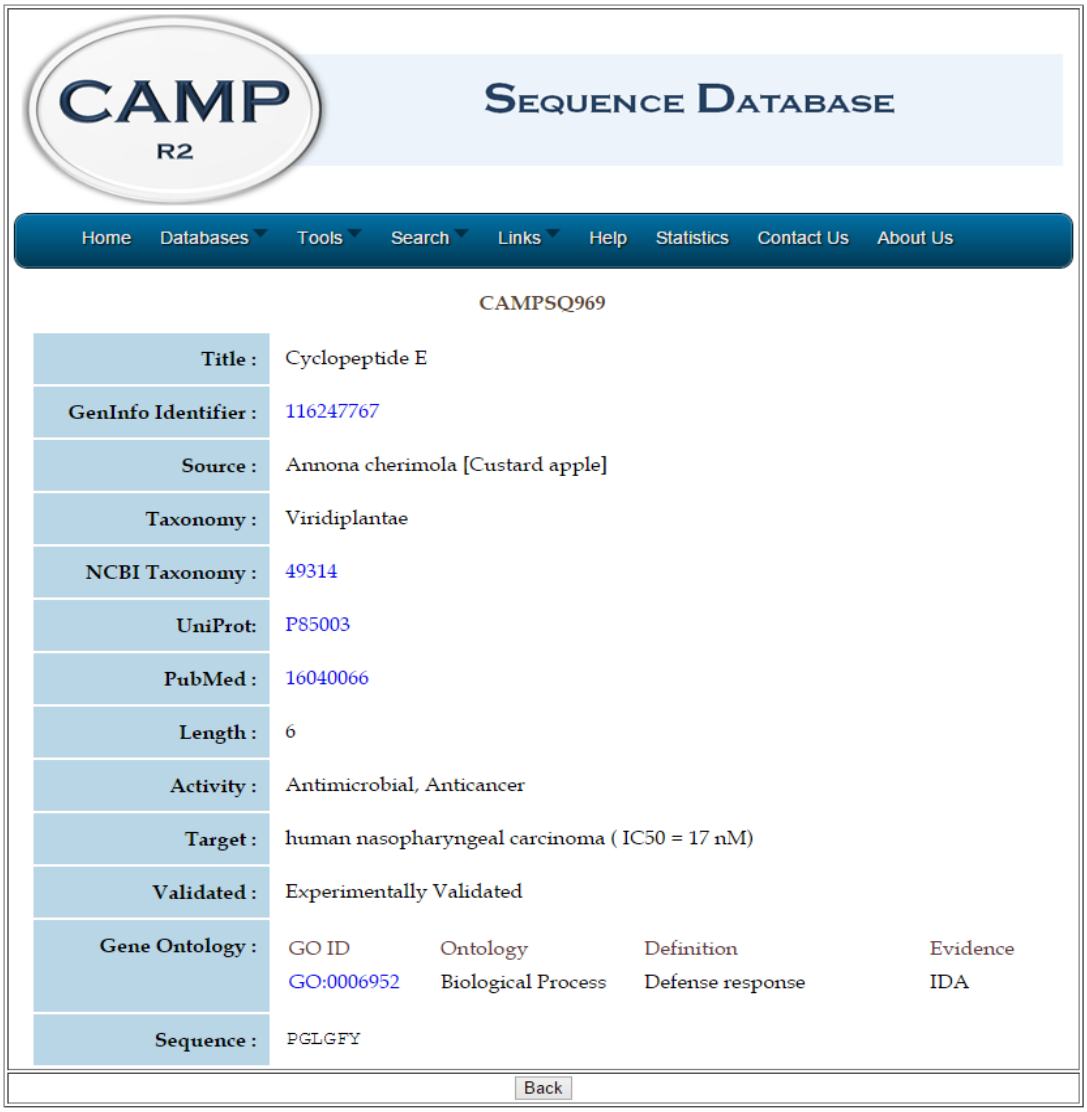
Περιέχει πεπτίδια με αντιμικροβιακή δράση, επιλέγοντας το μήκος ακολουθίας ή τον τρόπο εύρεσης (πειραματικός ή υπολογιστικός) ενώ πατώντας πάνω στην πρωτεΐνη μπορεί να δει κανείς περισσότερες πληροφορίες σχετικά με το αντιμικροβιακό πεπτίδιο όπως uniprot id (κωδικός στην uniprot) [15]. Επίσης δίνει την δυνατότητα πρόβλεψης της αντιμικροβιακότητας κάποιας ακολουθίας εισάγοντας την σε μορφή fasta (μορφή που περιέχει μία ή περισσότερες αλληλουχίες νουκλεοτιδίων στο DNA). Οι πληροφορίες των αντιμικροβιακών πεπτιδίων στην βάση εισήχθησαν από τις βάσεις δεδομένων NCBI [16] ,UniProtKB και PDB [17] εισάγωντας σε αυτές τις βάσεις, βασικές λέξεις κλειδιά όπως: 'anti-microbial', 'antibacterial', 'antifungal', 'antiviral' και 'antiparasitic' προκειμένου να επιστρέψουν τα αποτελέσματα των αντιμικροβιακών πεπτιδίων. Χειροκίνητα εισήχθησαν πληροφορίες για την δομή, για την δραστικότητα, τις πρωτεϊνικές οικογένειες καθώς και άλλες πληροφορίες που συνδέουν το αντιμικροβιακό πεπτίδιο με άλλες αντιμικροβιακές βάσεις παρέχοντας αυτές τις πληροφορίες. [18]

Όλα τα μοντέλα πρόβλεψης των SVM, RF και KNN χρησιμοποιούν πακέτα στην R (version 2.15.3) υπολογίζοντας τα αποτελέσματα. Συνολικά περιλαμβάνει 6756 αντιμικροβιακές ακολουθίες (experimentally validated (2602), pre-dicted (2438) andpatents (1716)), ενώ περιέχει 682 αντιμικροβιακές δομές [18].

Όπως προαναφέραμε δίνει την δυνατότητα πολλαπλής αναζήτησης των αντιμικροβιακών ακολουθιών όπως φαίνεται και στην ακόλουθη εικόνα με βάση πολλαπλά χαρακτηριστικά που διακρίνουν τα αντιμικροβιακά πεπτίδια όπως για παράδειγμα αν είναι πειραματικά ή ανάλογα με το μήκος της ακολουθίας τους κλπ.

Εικόνα 1.13 *Αναζήτηση αντιμικροβιακής ακολουθίας.*

Επιλέγοντας τυχαία μια ακολουθία παρέχονται περισσότερες πληροφορίες γι αυτήν, ενώ συνδέεται με την αντίστοιχη βάση δεδομένων (Uniprot, pubmed) [20] απ' όπου μπορεί να αποκομίσει κανείς περισσότερες πληροφορίες για την πρωτεΐνη.

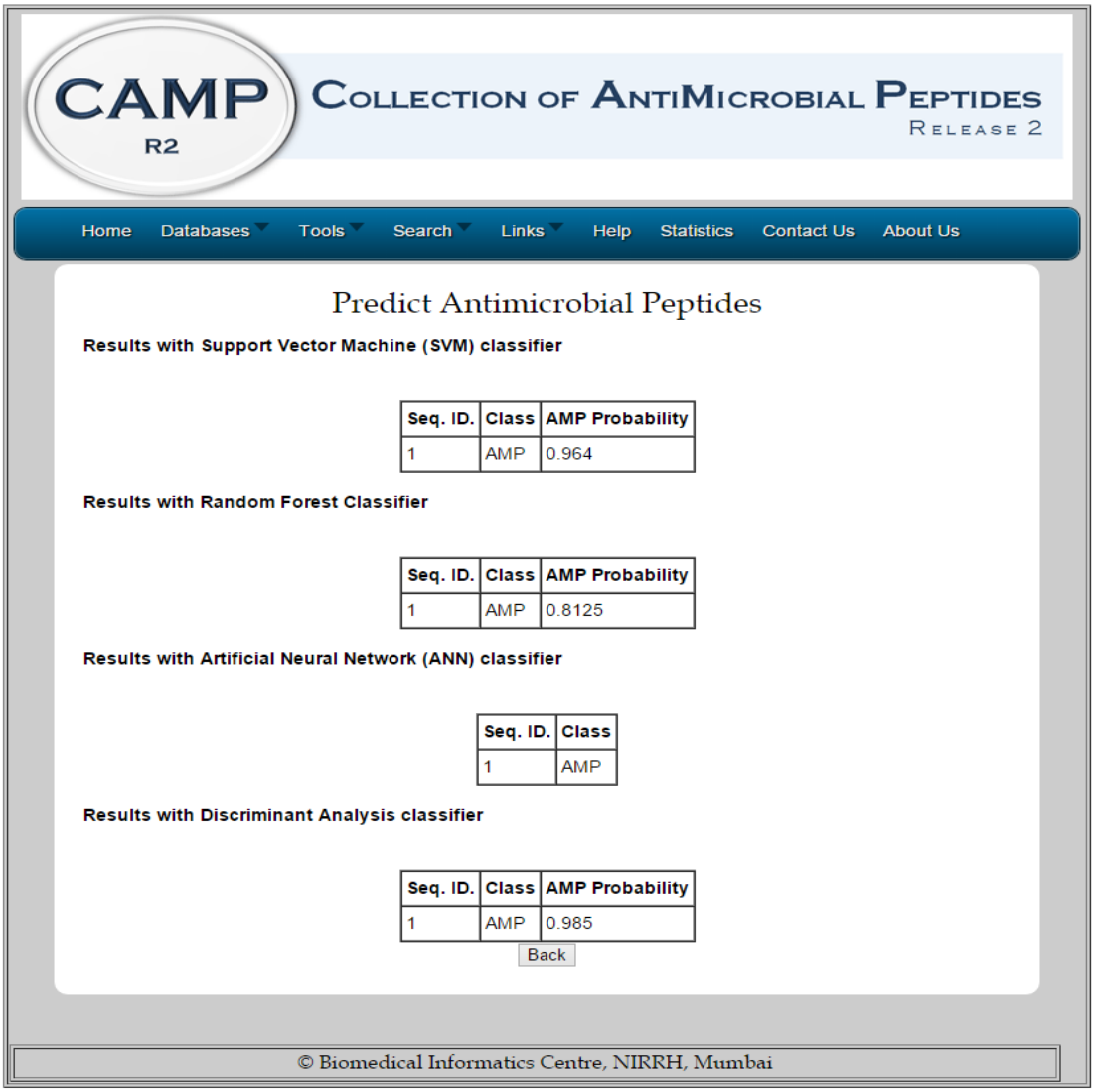


Εικόνα 1.14 *Αναζήτηση αντιμικροβιακού πεπτιδίου cyclopeptide E*

Παρακάτω φαίνονται τα αποτελέσματα από την αναζήτηση μια αντιμικροβιακής ακολουθίας (UniProtKB - Q3SAX6 (Q3SAX6_CARDV)) με την ακόλουθη FASTA μορφή, επιλέγοντας όλους τους αλγόριθμους πρόβλεψης.

>tr|Q3SAX6|Q3SAX6_CARDV Divergicin A OS=Carno bacterium divergens GN=dvnA PE=4 SV=1

MKKQILKGLVIVVCLSGATFFSTPQQASAAAPKITQKQKNCVNGQLGGMLAGALGGPGGVVL GGIGGAIAGGCFN

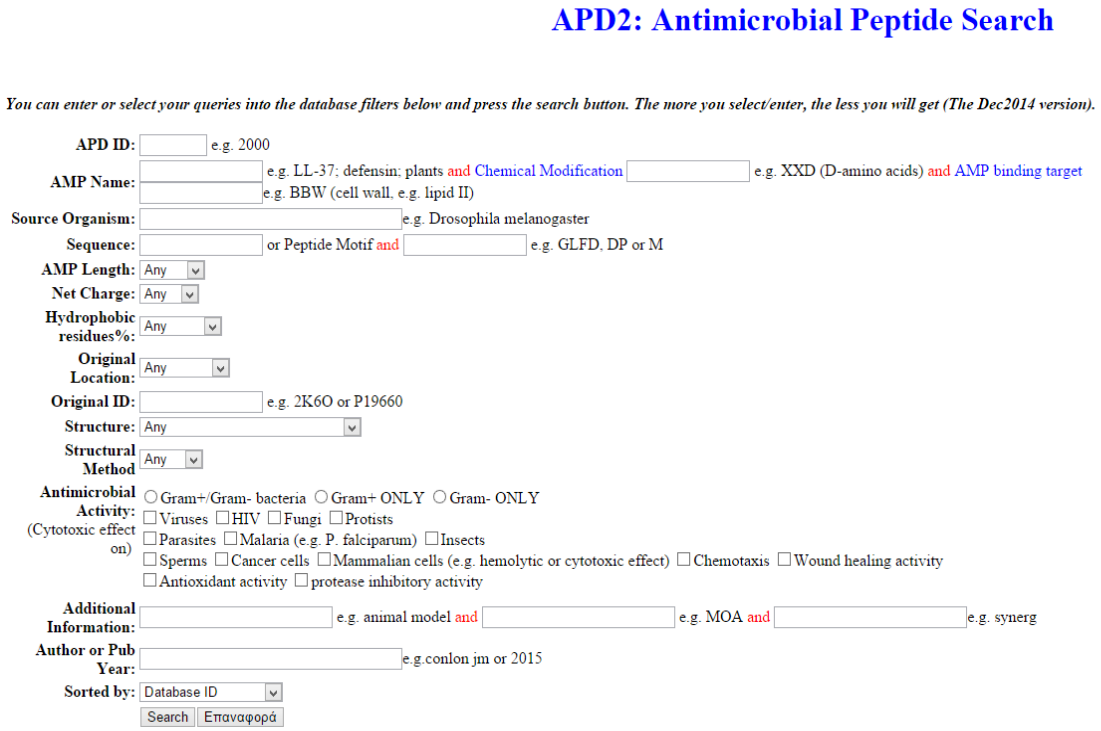


Εικόνα 1.15 *Αναζήτηση αντιμικροβιακότητας ακολουθίας*

Όπως φαίνεται παραπάνω ταξινομούν σωστά την ακολουθία ως αντιμικροβιακή με μεγάλο ποσοστό κάθε αλγόριθμος ξεχωριστά, με μικρότερο το ~0.82 του Random Forest Classifier.

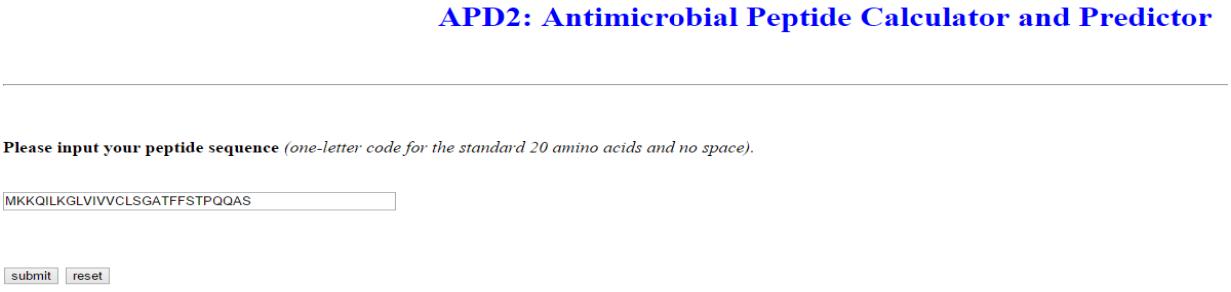
APD (antimicrobial peptide database)

Περιέχει συνολικά 525 πεπτίδια εκ των οποίων τα (498 antibacterial, 155 antifungal, 28 antiviral και 18 antitumor) από όλους τους οργανισμούς πειραματικά ή υπολογιστικά δίνοντας στο χρήστη να κάνει αναζήτηση με πολλά περισσότερα χαρακτηριστικά του πεπτιδίου όπως: Hydrophobic residues %, Structure, Antimicrobial Activity ή μέθοδο (NMR / XRA) κλπ [21].



Εικόνα 1.16 *Αναζήτηση αντιμικροβιακού πεπτιδίου*

Η συλλογή των δεδομένων έγινε από την Pubmed [20] ενώ επίσης δίνεται η δυνατότητα αναζήτησης αντιμικροβιακότητας εισάγοντας την ακολουθία (όχι σε μορφή fasta).

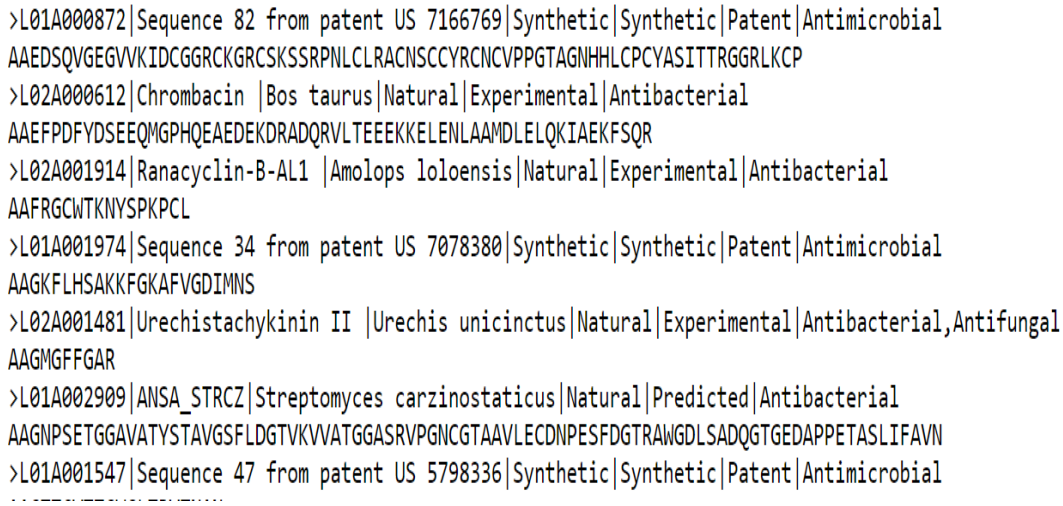


Εικόνα 1.17 *Αναζήτηση αντιμικροβιακότητας σε τυχαία ακολουθία*

Lamp (A database linking antimicrobial peptide)

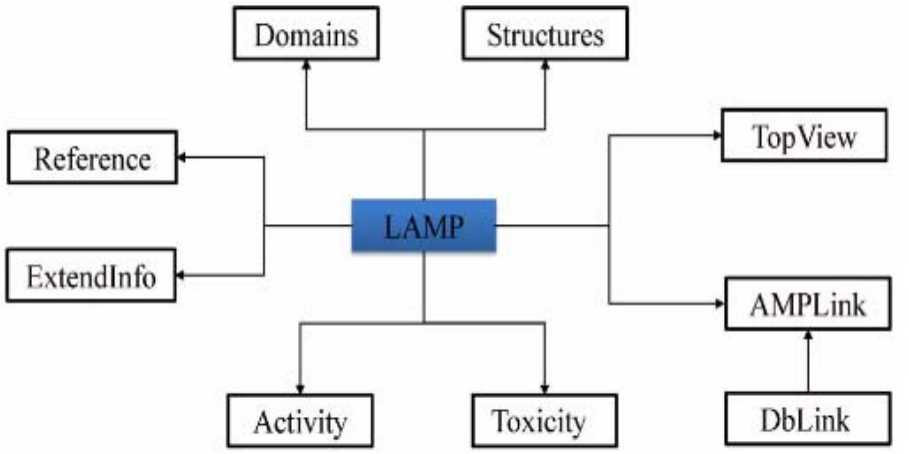
Περιλαμβάνει συνολικά 5547 ακολουθίες αντιμικροβιακών πεπτιδίων. Επιλέγοντας και εδώ συγκεκριμένα χαρακτηριστικά και κάνοντας την αναζήτηση, επιστρέφει τις αντιμικροβιακές ακολουθίες. Τα αντιμικροβιακά πεπτίδια έχουν χωριστεί σε 5547 ακολουθίες experimental, predicted και patent. Αξιοσημείωτο είναι ότι δίνει την δυνατότητα στον χρήστη να κατεβάσει όλες μαζί τις ακολουθίες των αντιμικροβιακών πεπτιδίων σε μορφή FASTA ενώ περιλαμβάνει ακολουθίες από την

camp και την apd που προαναφέραμε νωρίτερα. Συγκεκριμένα η ταξινόμηση των αντιμικροβιακών πεπτιδίων στην Lamp είναι ίδια με τον τρόπο στην camp (3, 201 experiment-validated, 863 predicted, and 1, 491 patents). Ωστόσο περιέχει περισσότερες ακολουθίες σε σχέση με την camp [15].



Εικόνα 1.18 *Αντιμικροβιακά πεπτίδια –lamp id –protein name –sequence-organism –group-collection and activity*

Παρέχει βασικές πληροφορίες όπως protein definition, accession numbers, brief activity όπως περιλαμβάνονται στο παρακάτω σχήμα της Lamp, ενώ παράλληλα παρέχονται και δομικές πληροφορίες του κάθε πεπτιδίου. Παράλληλα περιλαμβάνει τα δέκα ομοιότερα αντιμικροβιακά πεπτίδια της Lamp. Επιπρόσθετα στο Amplink περιλαμβάνονται διασυνδέσεις με άλλες βάσεις δεδομένων όπως η Uniprot. Τέλος το DB-Link περιλαμβάνει πληροφορίες από άλλες βάσεις αντιμικροβιακών πεπτιδίων.

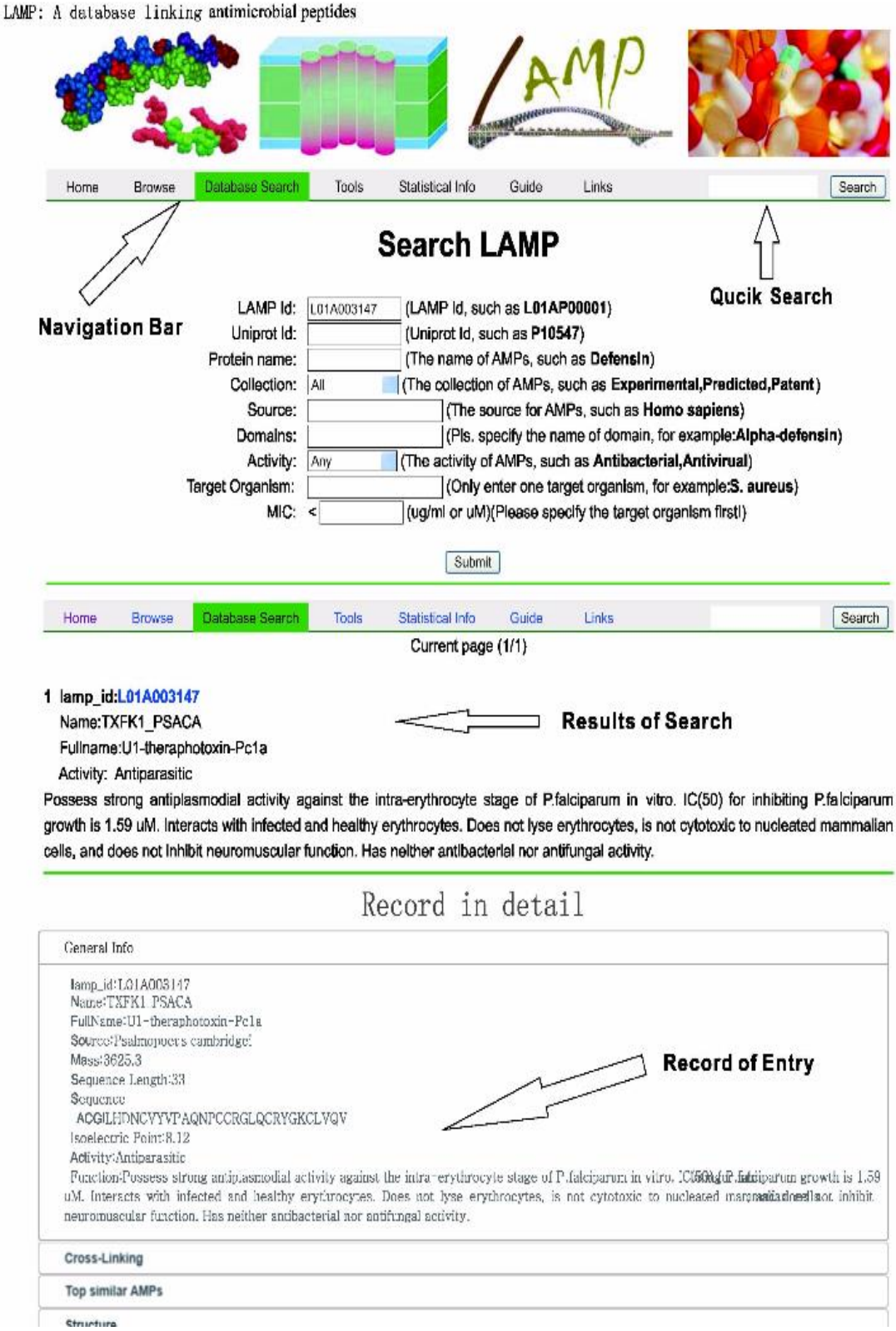


Εικόνα 1.19 *Σχήμα της βάσης δεδομένων Lamp*

Όλες οι αντιμικροβιακές ακολουθίες έχουν συλλεχθεί χειροκίνητα ενώ η επιλογή τους έγινε μέσω της βιβλιογραφίας καθώς και από άλλες βάσεις δεδομένων όπως την Uniprot και άλλες αντιμικροβιακές βάσεις δεδομένων.

Η Lamp έχει φιλικό περιβάλλον προς τον χρήστη και μπορεί εύκολα κανείς να αναζητήσει πληροφορίες για κάποιο συγκεκριμένο αντιμικροβιακό πεπτίδιο. Περιλαμβάνει διάφορα εργαλεία, στατιστικές πληροφορίες, οδηγό και Link σε άλλες βάσεις δεδομένων όπως PDB [17], Uniprot [15], καθώς και άλλες AMP (αντιμικροβιακών πεπτιδίων) βάσεις δεδομένων. Το Interface δίνει την δυνατότητα αναζήτησης χρησιμοποιώντας μία μόνο λέξη ή κάνοντας πολλαπλή αναζήτηση εισάγοντας όνομα, οργανισμό, LAMP ID (κωδικός αντιμικροβιακού πεπτιδίου στην Lamp), UniProtKB ID, όνομα πρωτεΐνης, ή άλλων χαρακτηριστικών του αντιμικροβιακού πεπτιδιου.

Παρακάτω φαίνεται πώς μπορεί να κάνει κανείς αναζήτηση στην βάση δεδομένων της lamp με πολλαπλούς τρόπους:



Εικόνα 1.20 *Πολλαπλή αναζήτηση ακολουθίας και αποτελέσματα αναζήτησης*

ΚΕΦΑΛΑΙΟ 2ο

Μέτρηση Χαρακτηριστικών Πεπτιδίων

2.1 Χαρακτηριστικά Αντιμικροβιακών Πεπτιδίων

Υπάρχουν κάποια συγκεκριμένα χαρακτηριστικά σε κάθε πεπτίδιο που το καθιστούν διαφορετικό από κάποιο άλλο επηρεάζοντας έτσι την δράση και την λειτουργία του. Θα πρέπει λοιπόν να εντοπίσουμε μέσω της βιβλιογραφίας τα κύρια αυτά χαρακτηριστικά που κατατάσσουν ένα πεπτίδιο ως αντιμικροβιακό. Μετέπειτα από τον εντοπισμό αυτών των αντιμικροβιακών χαρακτηριστικών θα προσπαθήσουμε μέσω εργαλείων να τα βρούμε για κάθε πεπτίδιο που έχουμε εισάγει στην βάση δεδομένων καθώς και να υπολογίζονται ξεχωριστά για κάθε εισαγόμενη ακολουθία.

Κυρίαρχο χαρακτηριστικό των αντιμικροβιακών πεπτιδίων στα φυτά είναι η ύπαρξη μικρών πεπτιδίων με μοριακή μάζα 2–10 kDa .Οι δομές αυτές των αντιμικροβιακών πεπτιδίων σταθεροποιούνται μέσω 2-6 δισουλφιδικών γεφυρών [23].

Επιπρόσθετα τα αντιμικροβιακά πεπτίδια χαρακτηρίζονται από τον υδρόφοβο χαρακτήρα τους, ο οποίος τους δίνει την δυνατότητα να διαχωρίζονται εντός των λιπιδίων της κυτταρικής μεμβράνης των βακτηρίων διαταράσσοντας τις δομές τους και καθιστώντας αυτές περισσότερο διαπερατές. Με αυτό λοιπόν τον τρόπο προκαλείται απώλεια ιόντων και άλλων κυτταρικών συστατικών με αποτέλεσμα η εκτεταμένη απώλεια του κυτταρικού περιεχομένου, των σημαντικών μορίων καθώς και των ιόντων να οδηγήσει τελικά το κύτταρο σε θάνατο [24].

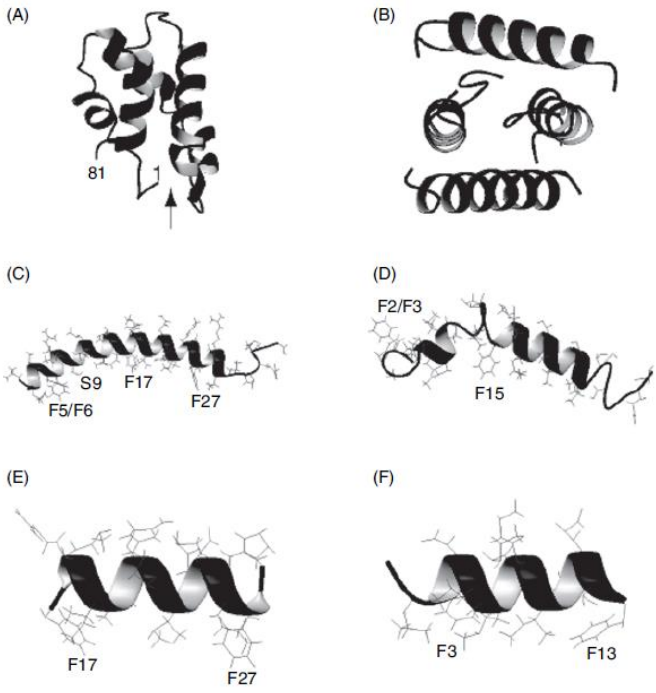
Μελέτες έχουν δείξει ότι η δράση των φαινολικών ουσιών είναι υπεύθυνη για την πρόκληση δομικών και λειτουργικών διαταραχών στην κυτταρική μεμβράνη [25].

Η έκθεση λοιπόν σε θυμόλη και καρβακρόλη προκαλεί την αποδόμηση της κυτταρικής μεμβράνης. Ωστόσο τα Gram- θετικά βακτήρια δεν δείχνουν τον ίδιο βαθμό αλλαγών στην μορφολογία του κυτταρικού τοιχώματος κάτι που πιθανώς οφείλεται στην διαλυτότητα των λιποσακχαριτών της εξωτερικής μεμβράνης των Gram- αρνητικών βακτηρίων στις φαινολικές ουσίες [24].

2.2 Δομικά χαρακτηριστικά πεπτιδίων

Οι δομές των αντιμικροβιακών πεπτιδίων χωρίζονται κυρίως σε τέσσερις ομάδες: α-έλικες, β-πτυχωτές, αβ δομές και μη αβ δομές.

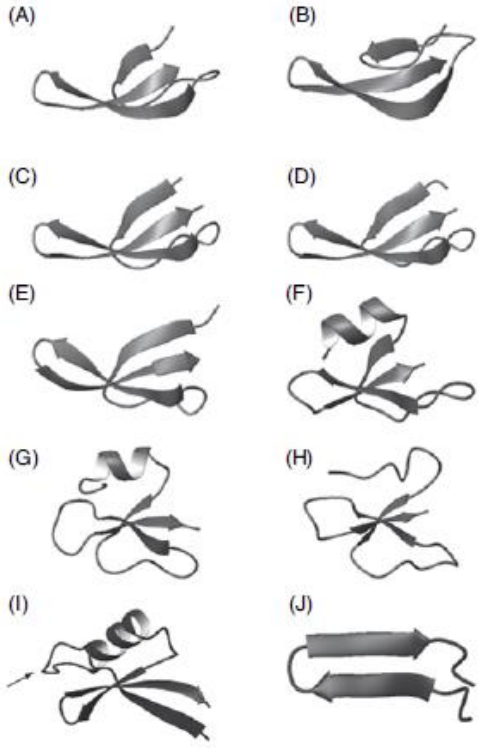
Τα περισσότερα αντιμικροβιακά πεπτίδια είναι κυρίως της μορφής α-έλικας .Η δομή της α-έλικας σταθεροποιείται με παραπάνω από έναν δισουλφιδικούς δεσμούς. Μετά την ένωση με βακτηριακές μεμβράνες, η δομή της έλικας μπορεί να ανοίξει σε κάποιο σημείο όπου είναι εκτεθειμένες αρκετές υδρόφοβες πλευρικές ομάδες. Ορισμένα αντιμικροβιακά πεπτίδια έχουν βρεθεί ότι υιοθετούν ένα μοτίβο helix-hinge-helix όπως για παραδειγμα η cecropine A, όπου η περιοχή σύνδεσης αυτών των πεπτιδίων έχει σημαντική αντιμικροβιακή δράση επιτρέποντας την βέλτιστη πρόσδεση αυτών των ελικών στις βακτηριακές μεμβράνες.



Εικόνα 2.1 *Απεικόνιση δομών αντιμικροβιακών πεπτιδίων α-έλικας (A) caenopore-5 from Caenorhab ditise legans (Protein Data Bank (PDB) ID: 2JS9) B) distinctin, at wo-chain AMP from frogs (PDBID: 1XKM); (C) LL-37, acathelicidin from humans (PDBID: 2K6O); (D) pardaxin 4 from fish (PDBID: 1XC0); (E) FK-13, anLL-37 antimicrobial core peptide (PDBID: 2FBS); and (F) aurein 1.2(PDBID:1VM5*

Γενικότερα στην οικογένεια των ελικοειδών δομών τα αντιμικροβιακά πεπτίδια με μεγαλύτερες πολυπεπτιδικές αλυσίδες τείνουν να εμφανίζουν μεγαλύτερη τοξικότητα σε κύτταρα θηλαστικών [26].

Αρκετά αντιμικροβιακά πεπτίδια όπως το RTD-1, έχουν δύο αντιπαράλληλες β-επιφάνειες σταθεροποιώντας τις με τρείς δεσμούς S–S bonds ή δύο S–S bonds ,όπως φαίνεται στην ακόλουθη εικόνα. Αυτά λοιπόν τα πεπτίδια περιέχουν την thanatin η οποία περιέχει μόνο S–S bonds. Αυτή είναι απαραίτητη για την αλληλεπίδραση του πεπτιδίου με την E. Coli ,αλλά όχι τόσο σημαντική για την σύνδεση με μεμβράνες Gram-θετικών βακτηρίων.



Εικόνα 2.2 *Δομές β- πτυχωτής επιφάνειας , (A) HNP-1(Protein Data Bank (PDB) ID: 3GNY); (B) HNP-3(PDB ID: 1DFN); (C) HNP-4 (PDB ID: 1ZMM); (D)HD-5 (PDB ID: 1ZMP); (E) HD-6 (PDB ID: 1ZMQ);(F) hBD-1 (PDB ID: 1IJV); (G) hBD-2 (PDB ID:1FD3); (H) hBD-3 (PDB ID: 1KJ5); (I) Psd1 (PDB ID: 1JKZ) ; και (J)RTD-1 (PDB ID: 1HVZ).*

Επιπρόσθετα έχει αποδειχθεί ότι πολλά πεπτίδια πλούσια σε τριπτοφάνη υιοθετούν μη αβ δομές ενώ η τριπτοφάνη έχει αποδειχθεί ότι βοηθάει ενεργά στην πρόσδεση στο κύτταρο.

Συνοπτικά τα αντιμικροβιακά πεπτίδια μπορούν να υιοθετήσουν διάφορες δομές όπως για παράδειγμα α έλικας ή β πτυχωτής επιφάνειας. Ως εκτούτου τα αντιμικροβιακά πεπτίδια δεν έχουν κάποια ειδική τρισδιάστατη δομή αλλά η αμφιπαθητική φύση τους παίζει σημαντικό ρόλο στην στόχευση των μεμβρανών.

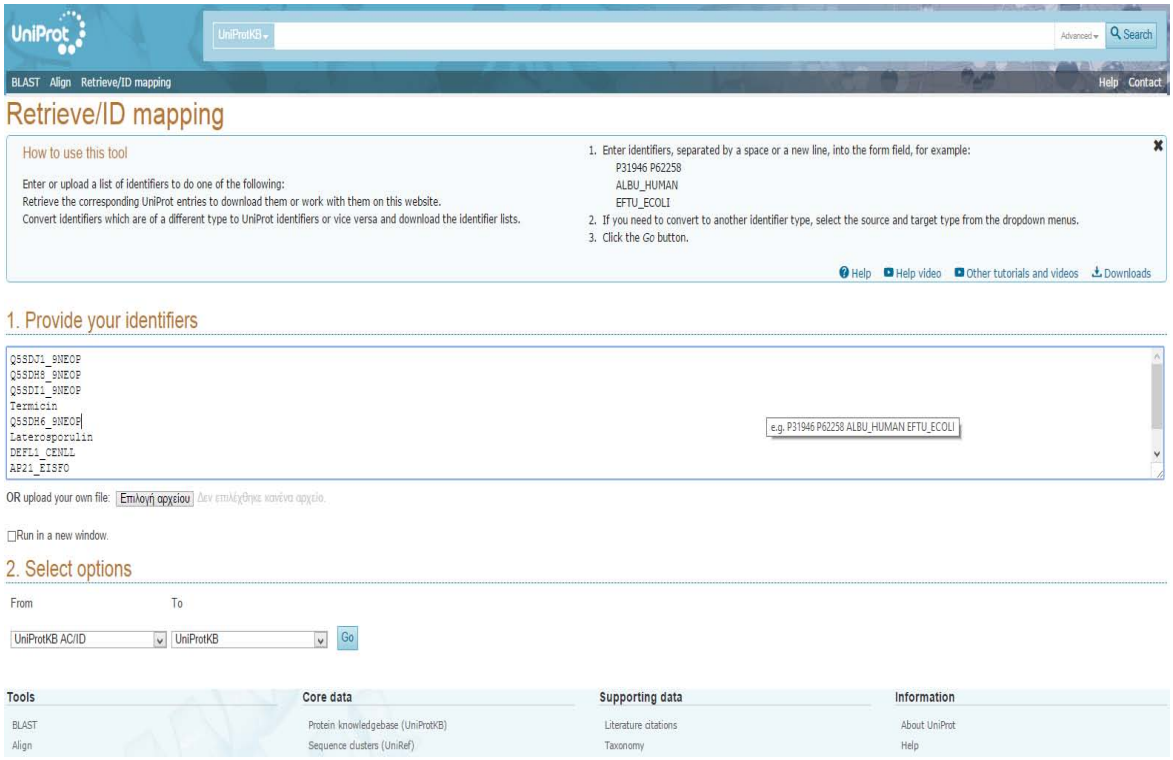
Δημιουργία Βάσης Δεδομένων και Ιστότοπου Αντιμικροβιακών Πεπτιδίων

3.1 Dataset αντιμικροβιακών πεπτιδίων

Όπως προαναφέραμε όλα τα αντιμικροβιακά πεπτίδια τα συλλέγουμε από την Lamp καθώς δίνει την δυνατότητα να κατεβάσουμε μαζικά όλα τα αντιμικροβιακά πεπτίδια ενώ περιλαμβάνει πεπτίδια και από άλλες αντιμικροβιακές βάσεις δεδομένων όπως αναφέρεται [22].

Αφότου συλλέξαμε τα δεδομένα τα εισάγαγαμε σε ένα excel ορίζοντας τις ακόλουθες στήλες:

Αρχικά την ακολουθία του αντιμικροβιακού πεπτιδίου, τον οργανισμό, το είδος αντιμικροβιακότητας (πχ αντικαρκινικό πεπτίδιο), το κωδικό στην lamp καθώς και το όνομα της πρωτεΐνης που αποτελεί τμήμα της. Απ όλα τα αντιμοκροβιακά πεπτίδια επιλέγουμε μόνο τα πειραματικά. Μετέπειτα προκειμένου να βρούμε το uniprot id πηγαίνουμε στον ιστότοπο της uniprot και κάνουμε id mapping καταχωρώντας όλα τα ονόματα των πρωτεϊνών όλων των αντιμικροβιακών πεπτιδίων που έχουμε κατεβάσει. Πιο συγκεκριμένα επιλέγουμε στο site **Retrieve/ID mapping**→Στο **Provide your identifiers** εισάγοντας όλα τα ονόματα των πρωτεϊνών → Ενώ στο **Select options** επιλέγουμε από uniprot KBAC/ID σε uniprot KB. Επίσης σε αυτό το στάδιο όπως φαίνεται στην παρακάτω εικόνα παρέχεται η δυνατότητα από το site της Uniprot να εισάγει κανείς κατευθείαν ένα αρχείο.



Εικόνα3.1 *Εύρεση uniprot id*

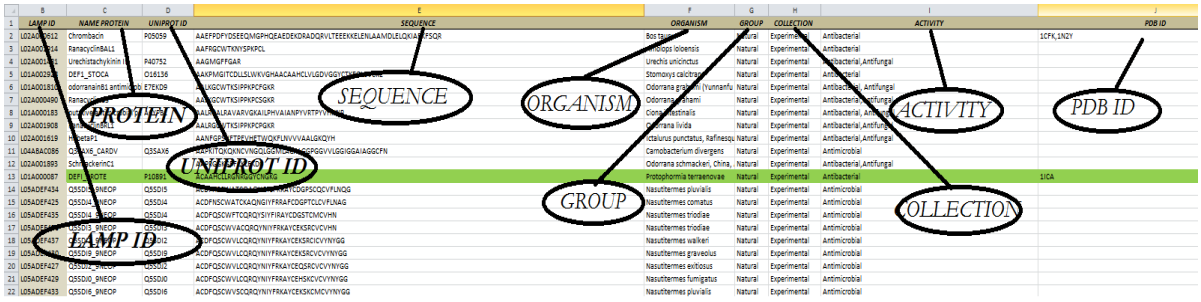
Αφότου εισάγουμε τις ονομασίες των πρωτεϊνών πατάμε το go και εμφανίζει όλες τις πληροφορίες για την κάθε πρωτεΐνη ξεχωριστά καθώς και το uniprotid που μετέπειτα μέσω αυτού θα βρούμε και το pdb id(κωδικός pdb)της πρωτεΐνης εφόσον φυσικά υπάρχει η τρισδιάστατη δομή της πρωτεΐνης.

Your list:...YWDKC	Entry	Entry name	Protein names	Gene names	Organism	Length
Q5SDJ1_9NEOP	Q5SDJ1	Q5SDJ1_9NEOP	Termicin		Nasutitermes exitiosus	62
Q5SDH8_9NEOP	Q5SDH8	Q5SDH8_9NEOP	Termicin		Tumulitermes pastinator	62
Q5SDI1_9NEOP	Q5SDI1	Q5SDI1_9NEOP	Termicin		Nasutitermes walkeri	62
Q5SDH6_9NEOP	Q5SDH6	Q5SDH6_9NEOP	Termicin		Drepanotermes rubriceps	62
DEFL1_CENLL	Q6GU94	DEFL1_CENLL	Defensin-1		Centruroides limpidus limpidus (Mexican scorpion)	56
AP21_EISFO	P84182	AP21_EISFE	Antimicrobial peptide OEP3121		Eisenia fetida (Red wiggler worm)	5
DEF4_RAT	Q62714	DEF4_RAT	Neutrophil antibiotic peptide NP-4	Np4	Rattus norvegicus (Rat)	93

Εικόνα 3.2 *Εύρεση uniprot id πρωτεΐνης.*

Εμφανίζει λοιπόν όλα τα αποτελέσματα στα οποία μας παρέχεται η δυνατότητα να τα κατεβάσουμε μαζικά σε συμπιεσμένη ή ασυμπίεστη μορφή. Εισάγουμε συνεπώς όλα τα αποτελέσματα στο uniprot id και πηγαίνουμε να βρούμε και το Pdb της κάθε πρωτεΐνης εφόσον υπάρχουν. Υπάρχουν αρκετοί τρόποι να βρούμε την pdb μορφή

της πρωτεΐνης. Είτε μέσω της Uniprot εισάγοντας όλα τα uniprotid που βρήκαμε νωρίτερα κάνοντας Id mapping από μορφή Uniprot σε μορφή pdb είτε αντίστοιχα πάλι μέσα από το site της Pdb, αναζητώντας την δομή της πρωτεΐνης μέσω του Uniprot id. Ωστόσο δεν παίρνουμε αποτελέσματα για όλες τις πρωτεΐνες παρά μόνο για τις περίπου 200 έναντι των 3044 που είναι συνολικά καθώς δεν έχει βρεθεί η τρισδιάστατη δομή για όλες τις πρωτεΐνες. Εισάγουμε συνεπώς στο αρχείο όλα τα αντίστοιχα pdb id που έχουμε βρει για την κάθε πρωτεΐνη. Συνεπώς το excel που δημιουργείται από τα data είναι της ακόλουθης πλέον μορφής.



Εικόνα 3.3 *Dataset*

3.2 Υπολογισμός χαρακτηριστικών ιδιοτήτων των πεπτιδίων

Σε αυτό το στάδιο μέσω του R studio [19] θα καταφέρουμε να υπολογίσουμε τα χαρακτηριστικά των αντιμικροβιακών πεπτιδίων προκειμένου μετέπειτα να τα εισάγουμε στην βάση δεδομένων μας. Ας δούμε αρχικά μερικές από τις βασικές λειτουργίες στο R studio μέσω του οποίου θα καταφέρουμε να υπολογίσουμε τα χαρακτηριστικά.

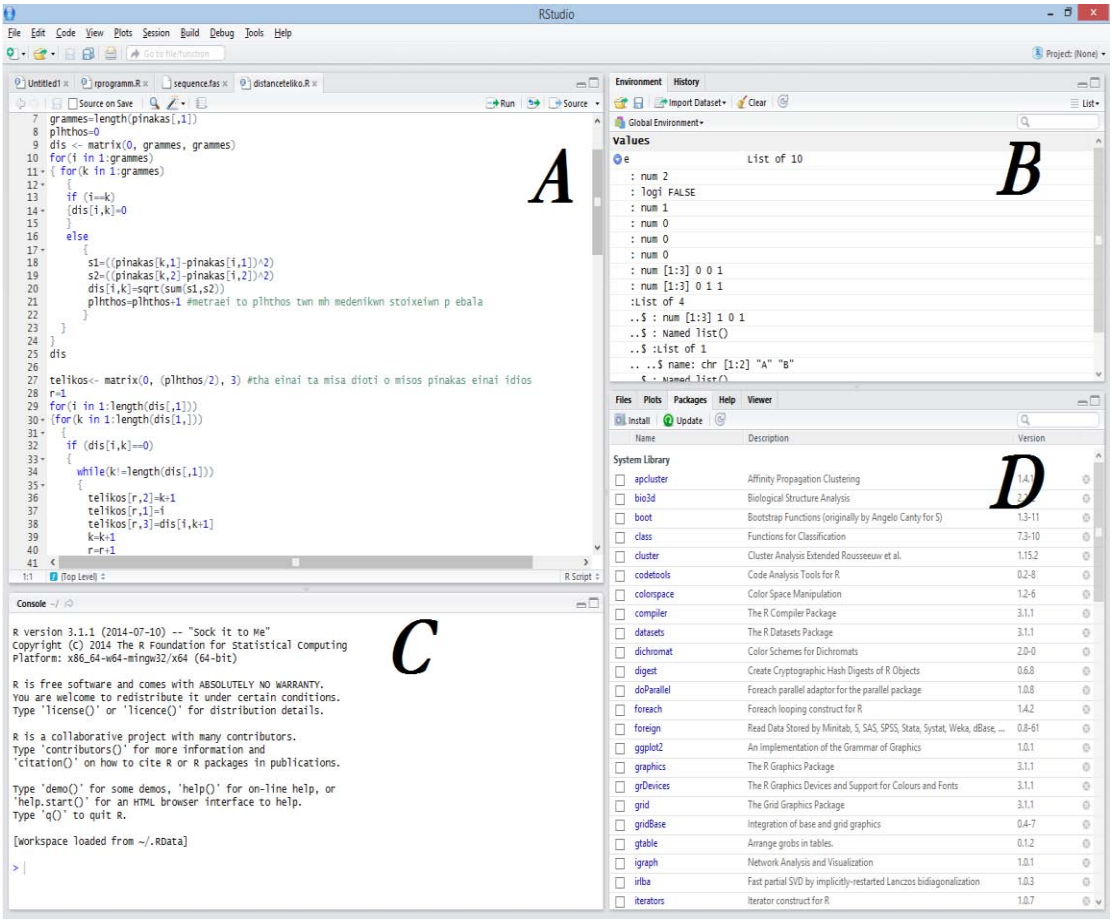
3.2.1 Εισαγωγή στην R

Η R είναι ένα υπολογιστικό πακέτο που προσφέρει δυνατότητες διαχείρισης και στατιστικής ανάλυσης δεδομένων καθώς και δυνατότητες κατασκευής γραφημάτων. Κατεβάζοντας και εγκαθιστώντας διάφορα πακέτα δίνει την δυνατότητα στο χρήστη να υπολογίσει διάφορες παραμέτρους σχετικές με το συγκεκριμένο πακέτο. Πρόκειται για λογισμικό ανοιχτού κώδικα που διατίθεται δωρεάν όπως και τα αντίστοιχα πακέτα.

Εμείς για την υλοποίηση των παρακάτω θα χρησιμοποιήσουμε το R studio το οποίο αποτελεί ουσιαστικά ένα γραφικό περιβάλλον της R.

Όπως φαίνεται παρακάτω ουσιαστικά υπάρχουν τέσσερα κεντρικά τμήματα στο R studio. Το κύριο τμήμα Α που αποτελεί το σημείο που γράφουμε τον κώδικα,

αποτελεί το λεγόμενο script. Το τμήμα B που μας δείχνει τις αποθηκευμένες μεταβλητές καθώς και τις τιμές τους μετά από την εκτέλεση του κώδικα, το τμήμα C όπου σε αυτό το τμήμα εκτελούνται οι εντολές από το script, καθώς και το τμήμα D που περιλαμβάνει όλες τις βιβλιοθήκες που έχουμε κατεβάσει στην συγκεκριμένη τρέχουσα έκδοση.



Εικόνα 3.4 *Απεικόνιση του RStudio (A) script, (B) μεταβλητές, (C) κονσόλα, (D) εγκατεστημένα πακέτα*

Οι εντολές προς την R δίνονται μέσω της R-Console: μετά το σύμβολο ">" γράφουμε την εντολή (ή το σύνολο των εντολών) που θέλουμε να εκτελεστούν και πατώντας ENTER λαμβάνουμε το αποτέλεσμα στην επόμενη γραμμή. Εναλλακτικά, μπορούμε να ανοίξουμε ένα παράθυρο script (R-editor από το menu: file/open script) (εικόνα – τμήμα A) όπου γράφουμε όσες εντολές επιθυμούμε και μετά μαρκάρουμε αυτές που θέλουμε να εκτελεστούν και πατάμε control+R. (αν δεν μαρκάρουμε τίποτε, με control+R θα εκτελεστούν μόνο οι εντολές της γραμμής που βρίσκεται ο cursoras).

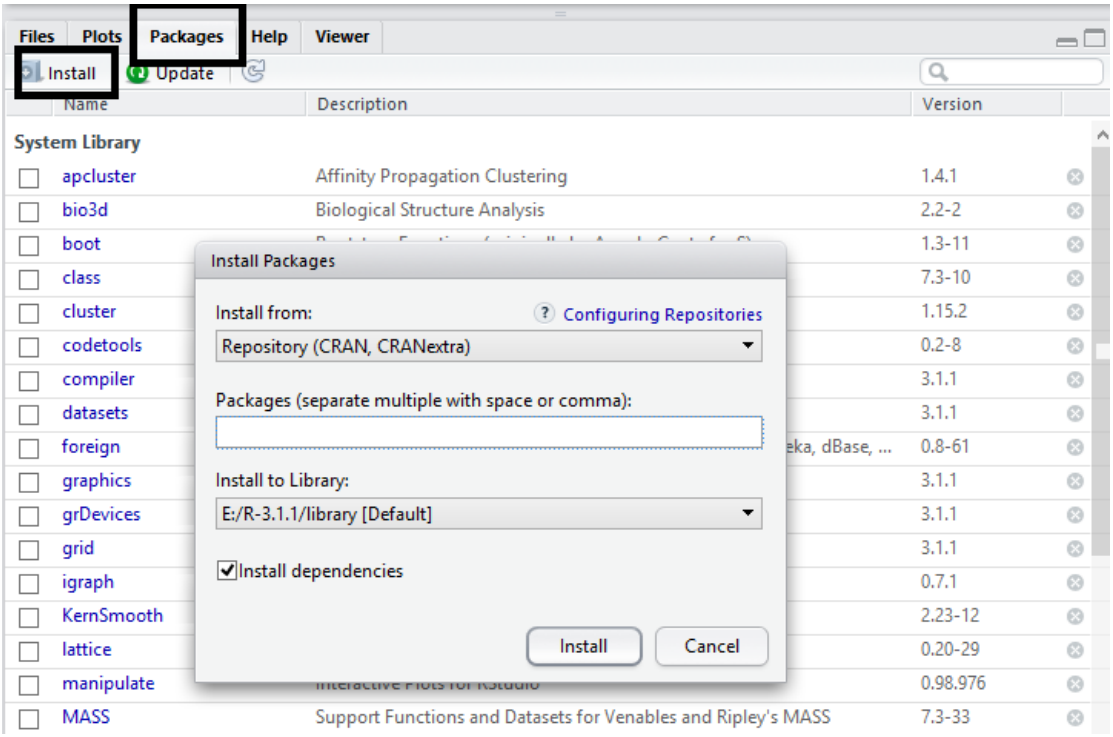
Στο τέλος ενός session μπορούμε να αποθηκεύσουμε τις μεταβλητές που έχουμε δημιουργήσει καθώς και τις εντολές που έχουμε χρησιμοποιήσει. Αυτό γίνεται από: File/save workspace, save history. Μπορούμε να τα ξαναφορτώσουμε από: File/load workspace, load history [27].

3.2.2 Μέθοδος υπολογισμού χαρακτηριστικών ιδιοτήτων

Με την χρήση της R θα υπολογίσουμε τις χαρακτηριστικές ιδιότητες των αντιμικροβιακών πεπτιδίων. Όπως προαναφέραμε, εγκαθιστούμε το Rstudio που προσφέρει καλύτερο γραφικό περιβάλλον παρόμοιο με της matlab (υπολογιστικό πρόγραμμα).

Μέχρι τώρα, έχουμε δημιουργήσει το excel το οποίο περιλαμβάνει όλα τα αντιμικροβιακά πεπτίδια τα οποία έχουμε κατεβάσει από την lamp, επιλέγοντας μόνο τα πεπτίδια τα οποία έχουν χαρακτηριστεί ως αντιμικροβιακά μέσω πειραμάτων (δηλ πειραματικά και όχι υπολογιστικά), τον αντίστοιχο κωδικό στην Uniprot καθώς και τον κωδικό της στην pdb για την τρισδιάστατη δομή της εφόσον υπάρχει. Ξεκινάμε να βρούμε μέσω της R τα χαρακτηριστικά των πεπτιδίου αυτού. Για να εξάγουμε τα αποτελέσματα από τα αντιμικροβιακά πεπτίδια απαραίτητη προϋπόθεση αποτελεί να έχουμε κατεβάσει το πακέτο «seqinr» [28] καθώς και το πακέτο «Peptides»[29] καθώς μέσω αυτών τον πακέτων θα καλέσουμε συγκεκριμένες συναρτήσεις.

Συνεπώς πηγαίνοντας στο **Rstudio→packages→install→** Αναζητάμε το συγκεκριμένο πακέτο → πατάμε **Install** και κατευθείαν στο script μέσα γράφουμε **library("seqinr")** ή απλά βάζοντας ένα τικ στο πακέτο που κατεβάσαμε.



Εικόνα 3.5 *Εγκατάσταση πακέτου στο Rstudio*

Το πακέτο «seqinr» έχει μία συνάρτηση η οποία καλείται ως read.fasta() και δίνει την δυνατότητα στο χρήστη να διαβάσει ένα fasta αρχείο, δηλ ένα αρχείο της μορφής:

```
>tr|Q5SDJ1|Q5SDJ1_9NEOP Termicin OS=Nasutitermesexitiosus PE=2 SV=1
MKTLAFLLCFLVVVCVFIAQHPADAACNFQSCWAICKAHYGIYFRRAYCDGPNCQCVHLI
```

Συνεπώς παίρνουμε όλες τις ακολουθίες που έχουμε κατεβάσει των αντιμικροβιακών πεπτιδίων και δημιουργούμε ένα αρχείο το οποίο να είναι μορφής fasta μέσω τροποποιήσεων στο txt. Δημιουργούμε λοιπόν ένα αρχείο της ακόλουθης μορφής έχοντας μέσα τις 3044 ακολουθίες:

>1

AAEFPDFYDSEEQMGPHQEAEDEKDRADQRVLTEEEKKELENLAAMDLELQKIAEKFSQR

>1

AAFRGCWTKNYSPKPCL

Ο λόγος που το κάνουμε αυτό είναι διότι πρέπει να βρούμε ένα τρόπο προκειμένου να υπολογίσουμε μαζικά τα χαρακτηριστικά των αντιμικροβιακών πεπτιδίων. Συνεπώς η καλύτερη λύση είναι να δημιουργήσουμε έναν πίνακα και να βάλουμε μέσα στο στοιχείο κάθε πίνακα την ακολουθία του κάθε πεπτιδίου. Με αυτό τον

τρόπο διαβάζοντας όλο τον πίνακα θα καταφέρουμε να εξάγουμε τα χαρακτηριστικά καλώντας τις κατάλληλες συναρτήσεις για όλες τις ακολουθίες μου.

Μετέπειτα αντίστοιχα φορτώνουμε την βιβλιοθήκη «peptides» [29] και για κάθε μία ακολουθία ξεχωριστά καλώ τις συναρτήσεις του πακέτου για να εξάγω aindex, pI, lengthpep, boman, h και KD. Οι συγκεκριμένες συναρτήσεις είναι της μορφής :

x=aindex(seq) δηλ

x=aindex("AAEFPDFYDSEEQMGPHQEAEDEKDRADQRVLTEEEKKELENLAAMDLELQ")

Και στην μεταβλητή x εισάγεται ένας αριθμός που αποτελεί στην συγκεκριμένη περίπτωση το aindex του συγκεκριμένου αντιμικροβιακού πεπτιδίου.

Συνεπώς για να καταφέρουμε να εξάγουμε τα συγκεκριμένα χαρακτηριστικά θα πρέπει να χρησιμοποιήσουμε τις ακόλουθες βιβλιοθήκες στην R:

library("seqinr") [28]

Με την βοήθεια της παραπάνω βιβλιοθήκης θα καταφέρουμε να διαβάσουμε όλες τις ακολουθίες των 3044 πεπτιδίων που έχουμε καλώντας την συνάρτηση read.fasta() αποθηκεύοντας μετέπειτα όλες τις ακολουθίες σε έναν πίνακα.

seq<- read.fasta(file = arxeio, forceDNAtolower = FALSE, seqonly=TRUE)

(Προϋπόθεση φυσικά αποτελεί πάντα να έχουμε ορίσει το σωστό μονοπάτι του αρχείου μας μέσα από το Rstudio) [27]

library("Peptides") [29]

Περιέχει συναρτήσεις που χρησιμοποιήσαμε προκειμένου να εξάγουμε και να υπολογίσουμε τα χαρακτηριστικά των αντιμικροβιακών πεπτιδίων. Μερικές εκ των οποίων είναι οι ακόλουθες:

- x=aindex()
  Υπολογισμός αλειφατικού δείκτη (Ο αλειφατικός δείκτης θεωρείται παράγοντας σταθερότητας των σφαιρικών πρωτεϊνών)

- x=pI()
  Υπολογισμός ισοηλεκτρικού σημείου ακολουθίας. Ως ισοηλεκτρικό σημείο pI ενός αμινοξέος ορίζεται το pH στο οποίο το συνολικό φορτίο του αμινοξέος είναι μηδέν, έχουμε δηλαδή ίσο αριθμό θετικά και αρνητικά φορτισμένων ιόντων.

- x=lengthpep
  Υπολογισμός αμινοξέων ακολουθίας. Καλώντας την συνάρτηση υπολογίζεται το συνολικό άθροισμα των αμινοξέων που περιέχει η πρωτεΐνη.

- x=boman()
  Υπολογισμός δυναμικού δέσμευσης πρωτεΐνης, ο δείκτης είναι ίσος με το άθροισμα των τιμών διαλυτότητας για όλα τα υπολείμματα σε μία ακολουθία.

- x=h()
  Υπολογισμός κλίμακας υδροφοβικότητας. Η υδροφοβικότητα αντιπροσωπεύει την τάση των μορίων ή των ατόμων να απωθούνται από το νερό όταν έρχονται σε επαφή με αυτό. Με ανάλογο τρόπο και τα αμινοξέα, σύμφωνα με την πολικότητα των πλευρικών τους ομάδων μπορεί να εμφανίσουν υδροφοβική (μη πολικά αμινοξέα) ή υδροφιλική (πολικά αμινοξέα) συμπεριφορά. Η κατανομή των υδροφοβικών και υδροφιλικών αμινοξέων μίας πρωτεΐνης καθορίζει την τριτοταγή δομή της, αφού αποτελεί την κινητήριο δύναμη για την αναδίπλωσή της. Προκειμένου, δηλαδή, οι μη πολικές πλευρικές ομάδες να αποφύγουν την επαφή με το νερό, συμπιέζονται στο εσωτερικό της πρωτεΐνης, διαμορφώνοντας τη δομή της και προσφέροντας της σταθερότητα.

- x=KD() (Kyte-Doolittle)
  Υπολογισμός δείκτη υδροφοβικότητας αντιμικροβιακής ακολουθίας [29]. Χρησιμοποιείται ως επί το πλείστον κατά τον εντοπισμό υδροφοβικών τμημάτων σε πρωτεΐνες, τόσο για τμήματα πρωτεϊνών στην επιφάνεια της κυτταρικής μεμβράνης, όσο και για διαμεμβρανικά τμήματα. Θετικές τιμές αντιστοιχούν σε υδροφοβικές πρωτεϊνικές περιοχές [30].

ΑΜΙΝΟΞΥ	ΚΛΙΜΑΚΑ ΥΔΡΟΦΟΒΙΚΟΤΗΤΑΣ			
	Kyte-Doolittle (KD)	Engelmann	Hopp-Woods	Cornette
Λευκίνη (L)	3.80	2.80	-1.80	5.70
Ισολευκίνη (I)	4.50	3.10	-1.80	4.80
Ασπαραγίνη (N)	-3.50	-4.80	0.20	-0.50
Γλυκίνη (G)	-0.40	1.00	0.00	0.00
Βαλίνη (V)	4.20	2.60	-1.50	4.70
Γλουταμικό οξύ (E)	-3.50	-8.20	3.00	-1.80
Προλίνη (P)	-1.60	-0.20	0.00	-2.20
Ιστιδίνη (H)	-3.20	-3.00	-0.50	0.50
Λυσίνη (K)	-3.90	-8.80	3.00	-3.10
Αλανίνη (A)	1.80	1.60	-0.50	0.20
Τυροσίνη (Y)	-1.30	-0.70	-2.30	3.20
Τρυπτοφάνη (W)	-0.90	1.90	-3.40	1.00
Γλουταμίνη (Q)	-3.50	-4.10	0.20	-2.80
Μεθειονίνη (M)	1.90	3.40	-1.30	4.20
Σερίνη (S)	-0.80	0.60	0.30	-0.50
Κυστεΐνη (C)	2.50	2.00	-1.00	4.10
Θρεονίνη (T)	-0.70	1.20	-0.40	-1.90
Φαινυλαλανίνη (F)	2.80	3.70	-2.50	4.40
Αργινίνη (R)	-4.50	-12.3	3.00	1.40
Ασπαρτικό οξύ (D)	-3.50	-9.20	3.00	-3.10

Εικόνα 3.6 *Οι αριθμητικές τιμές υδροφοβικότητας των 20 αμινοξέων για τις πιο διαδεδομένες και χρησιμοποιούμενες κλίμακες.*

library("xlsx")

Τέλος μία ακόμη βιβλιοθήκη που χρησιμοποιήσαμε μέσω της οποίας καταφέραμε να δημιουργήσουμε ένα τελικό excel με όλα τα data που δημιουργήσαμε και υπολογίσαμε με την χρήση των παραπάνω και αυτόματα μέσω της παρακάτω συνάρτησης, όποιο χαρακτηριστικό υπολογίζαμε αποθηκευόταν απευθείας σε ένα excel με την ονομασία που είχαμε ορίσει εμείς.

write.xlsx(x = x1, file = "features.xlsx", sheetName = "AINDEX", row.names = FALSE)

Με την χρήση λοιπόν όλων των παραπάνω καταφέραμε να υπολογίσουμε μέσω της R τα σημαντικότερα χαρακτηριστικά των αντιμικροβιακών πεπτιδίων που είχαμε συλλέξει και να δημιουργήσουμε αυτόματα ένα excel που να τα περιέχει για όλο το σύνολο των πεπτιδίων μας (3044) .

	A	B	C	D	E	F
1	AINDEX	PI	LENGTHPEP	BOMAN	H	KD
2	55,5	4,29	60	3,35	-0,3	-1,31
3	34,71	9,39	17	1,26	-0,04	-0,46
4	30	9,8	10	-0,16	0,38	0,72
5	97,61	6,91	46	-0,18	0,26	0,6
6	49	10,07	20	0,99	-0,06	-0,42
7	49	10,07	20	1,31	-0,13	-0,6
8	113,51	11,57	37	0,98	0,09	0,18
9	49	10,33	20	1,61	-0,16	-0,67
10	79,7	6,96	33	0,5	0,21	0,02
11	91,3	9,39	46	-0,49	0,33	0,43
12	34,71	9,39	17	1,03	0	-0,1
13	54	9,16	20	1,48	-0,05	-0,3
14	39,73	6,89	37	1,37	0,05	-0,04
15	52,97	7,81	37	0,93	0,13	0,28
16	37,35	6,73	34	1,51	0,06	0
17	45,88	8,65	34	2,2	-0,14	-0,36
18	52,7	8,63	37	1,74	-0,06	-0,28
19	50	8,63	37	1,77	-0,07	-0,28
20	50	8,32	37	1,77	-0,05	-0,27
21	50	8,32	37	1,74	-0,04	-0,26
22	31,62	8,62	37	1,78	-0,09	-0,45
23	31,62	8,32	37	2,15	-0,1	-0,52
24	42,16	7,82	37	1,73	0,01	-0,07
25	81,62	9,5	37	1,09	0,1	0,21
26	42,16	8,84	37	1,64	-0,08	-0,39
27	74,33	6,98	67	0,93	0,09	-0,07
28	88,48	8,35	33	0,88	0,06	0,22
29	18,85	8,78	26	1,9	-0,07	-0,34
30	55,41	8,89	37	0,8	0,06	-0,1
31	39	8,52	30	0,64	0,07	0,17
32	60,81	8,34	37	1,01	0,09	0,02
33	58,11	7,82	37	1,24	0,06	-0,13
34	39,73	7,82	37	1,68	-0,02	-0,35

Εικόνα 3.7 *Αποτελέσματα υπολογισμού αντιμικροβιακών πεπτιδίων για τα πρώτα 34 πεπτίδια από τα 3044 συνολικά.*

3.3 Δημιουργία βάσης δεδομένων

Για την δημιουργία της βάσης δεδομένων και του site χρησιμοποιήθηκαν τα ακόλουθα εργαλεία και οι ακόλουθες γλώσσες προγραμματισμού:

Xampp

Το XAMPP είναι ένα πακέτο ελεύθερου λογισμικού, λογισμικού ανοικτού κώδικα και ανεξαρτήτου πλατφόρμας το οποίο περιέχει τον εξυπηρετητή ιστοσελίδων http Apache, την βάση δεδομένων MySQL και ένα διερμηνέα για κώδικα γραμμένο σε γλώσσες προγραμματισμού PHP και Perl. [31] Είναι ανεξάρτητο πλατφόρμας και τρέχει σε Microsoft Windows, Linux, Solaris, και Mac OS X και χρησιμοποιείται ως πλατφόρμα για την σχεδίαση και ανάπτυξη [32].

R_Studio

Αποτελεί ένα προγραμματιστικό περιβάλλον που παρέχει την δυνατότητα να προγραμματίσει κανείς σε γλώσσα R και να κατεβάσει συγκεκριμένα πακέτα που περιέχουν κάποιες συναρτήσεις που αυτοματοποιούν τις διαδικασίες [27].

PHP

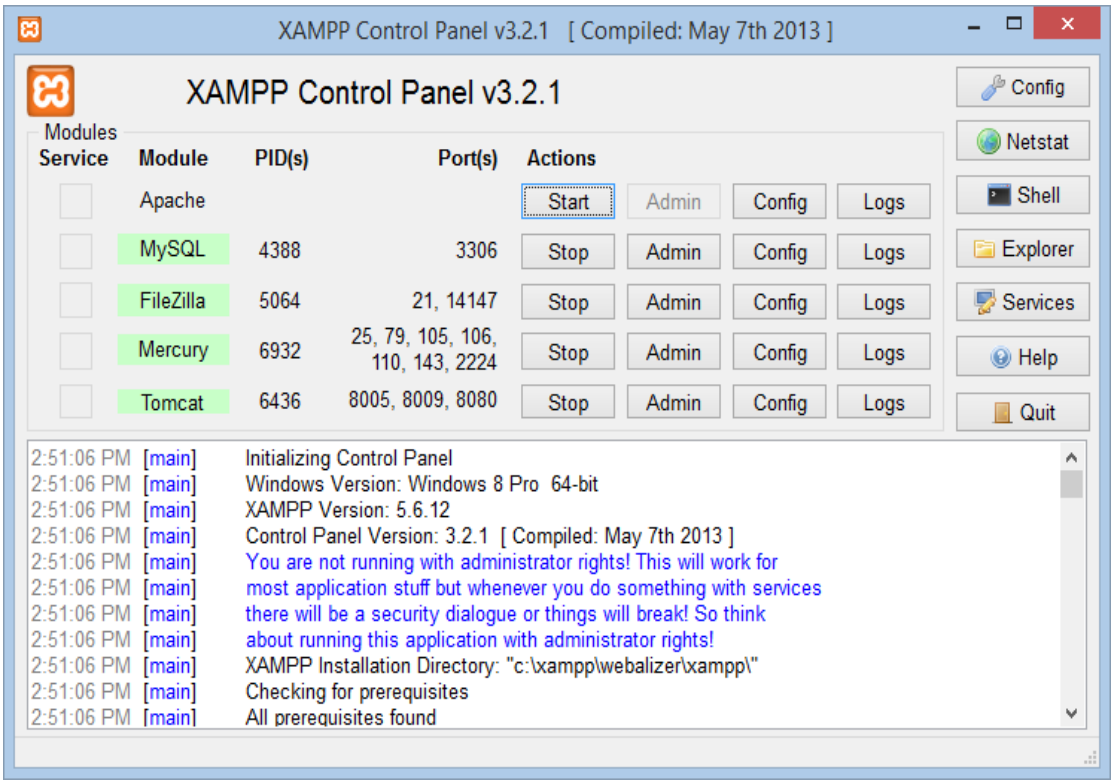
Η PHP είναι μια γλώσσα προγραμματισμού για τη δημιουργία σελίδων web με δυναμικό περιεχόμενο. Μια σελίδα PHP περνά από επεξεργασία από ένα συμβατό διακομιστή του Παγκόσμιου Ιστού (π.χ. Apache), ώστε να παραχθεί σε πραγματικό χρόνο το τελικό περιεχόμενο, που θα σταλεί στο πρόγραμμα περιήγησης των επισκεπτών σε μορφή κώδικα HTML.Τα αρχικά σημαίνουν Hyper text Pre Processor, και είναι μια γλώσσα συγγραφής σεναρίων (scripting language) που ενσωματώνεται μέσα στον κώδικα της HTML και εκτελείται στην πλευρά του server (server-side scripting).

Html

Αποτελεί μια βασική γλώσσα για την δόμηση σελίδων. Είναι μία γλώσσα σήμανσης κειμένου. Αυτό γίνεται με την βοήθεια "HTML tags". Οι ετικέτες HTML συνήθως λειτουργούν ανά ζεύγη (για παράδειγμα <h1> και </h1>), με την πρώτη να ονομάζεται ετικέτα έναρξης και τη δεύτερη ετικέτα λήξης (ή σε άλλες περιπτώσεις ετικέτα ανοίγματος και ετικέτα κλεισίματος αντίστοιχα). Ανάμεσα στις ετικέτες μπορούν να τοποθετηθούν κείμενα, πίνακες, εικόνες κλπ.

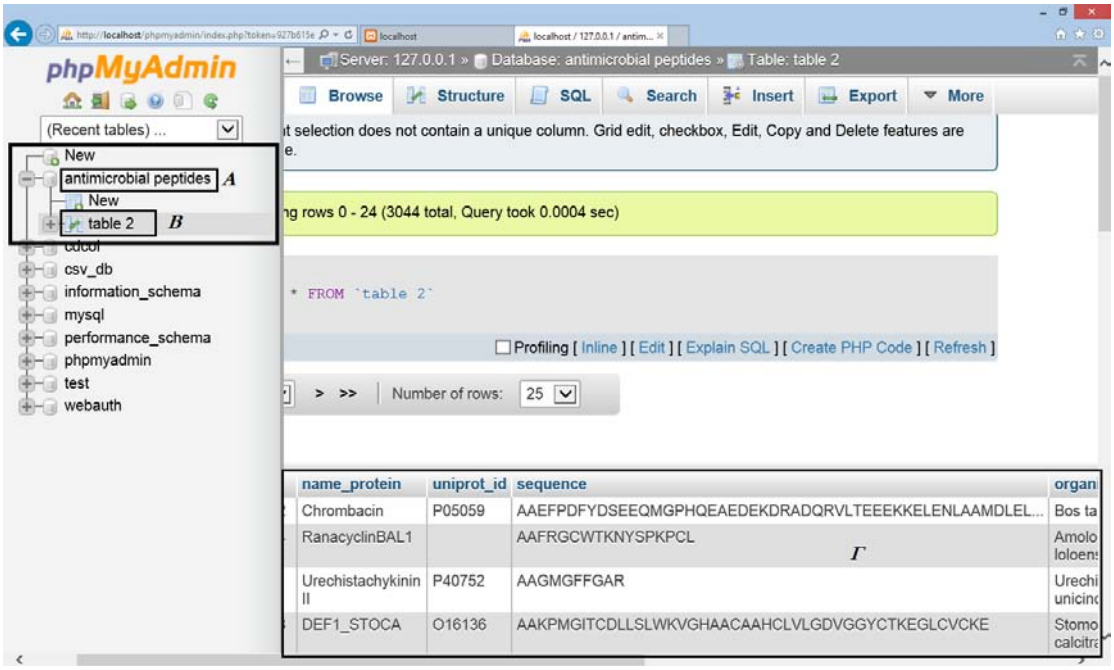
Με την χρήση των παραπάνω θα δημιουργήσουμε την βάση δεδομένων που θα επικοινωνεί με το site και θα υπολογίζει τα αντιμικροβιακά χαρακτηριστικά.

Έχουμε δημιουργήσει το dataset μας όπως προαναφέραμε και προσπαθούμε να το εισάγουμε στο xampp. Το XAMPP είναι ένα ελεύθερο λογισμικό το οποίο περιέχει ένα εξυπηρετητή ιστοσελίδων το οποίο μπορεί να εξυπηρετεί και δυναμικές ιστοσελίδες τεχνολογίας PHP/MySQL. Φορτώνουμε συνεπώς το csv αρχείο που έχω δημιουργήσει νωρίτερα, στο xampp. Σε πρώτο στάδιο ανοίγουμε το xampp και αρχίζουμε να τρέχουμε τον τοπικό server στον υπολογιστή.



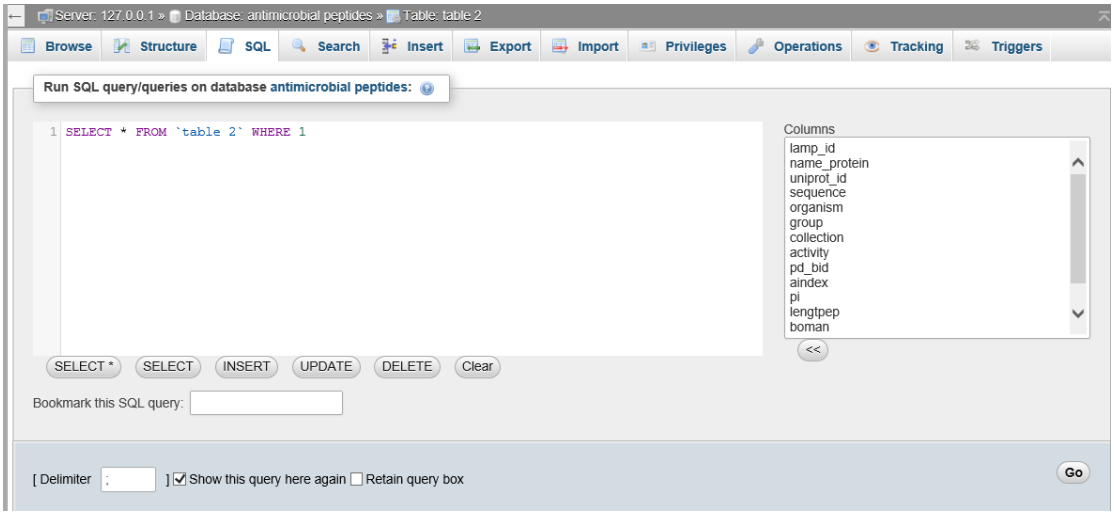
Εικόνα 3.8 *Ανοίγουμε το xampp προκειμένου να τρέξουμε τον τοπικό server*

Σε επόμενο στάδιο για να εισάγουμε τα δεδομένα στην τοπική βάση δεδομένων πληκτρολογούμε localhost στην μπάρα διεύθυνσης επιλέγοντας μετέπειτα το phpmyadmin. Μετέπειτα δημιουργούμε μια νέα βάση δεδομένων με την ονομασία antimicrobial peptides και των πίνακα table2 που περιέχει τα δεδομένα μας που τα εισάγαμε σε μορφή csv μέσω του import.

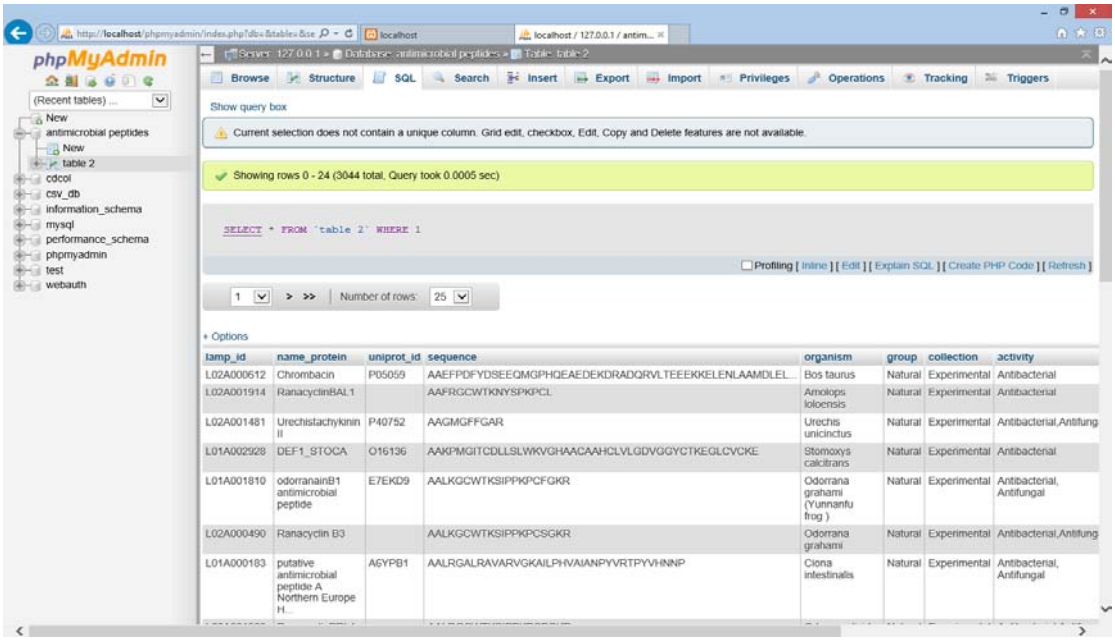


Εικόνα 3.9 *(Α) Δημιουργία νέας βάσης δεδομένων (antimicrobial peptides), (Β) Δημιουργία πίνακα δεδομένων στην συγκεκριμένη βάση, (Γ) Δεδομένα που εισήχθησαν.*

Παρατηρούμε λοιπόν ότι ορίζοντας ένα ερώτημα στην Sql (SELECT * FROM 'table2' WHERE 1) μας επιστρέφει από τον πίνακα που εισάγαμε τα συγκεκριμένα δεδομένα.



Εικόνα 3.10 *Sql ερώτημα στην βάση δεδομένων*



Εικόνα 3.11 *Αποτέλεσμα ερωτήματος στην βάση δεομένων*

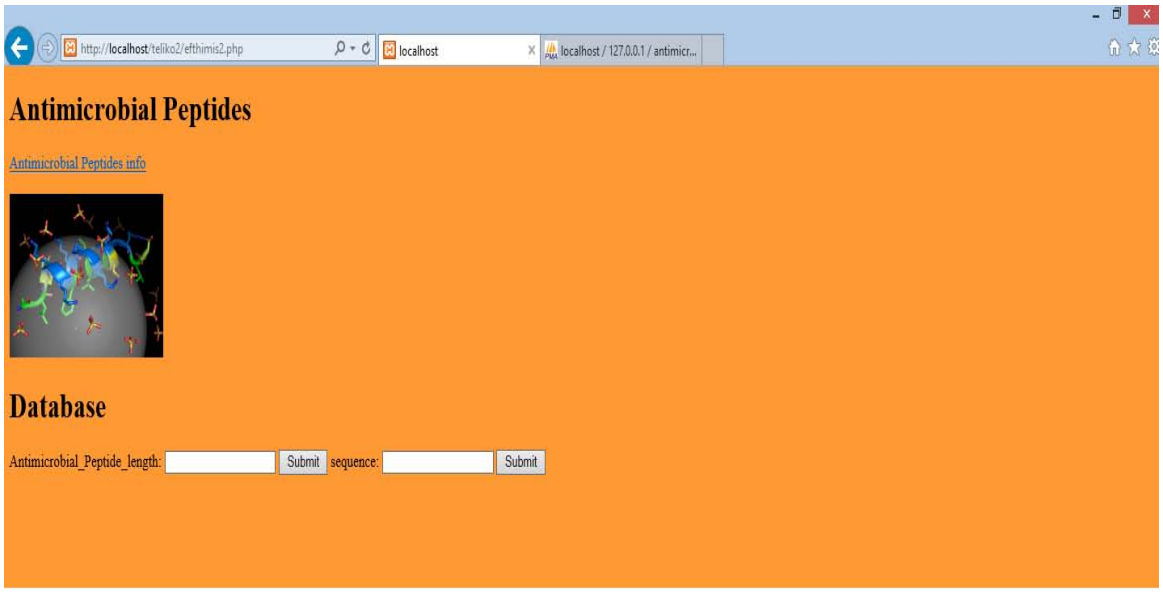
Για να καταφέρουμε να φορτώσουμε μια σελίδα θα πρέπει να έχουμε βάλει το αρχείο μου (δηλ. τον κώδικα της Php/html) μέσα στο φάκελο htdocs προκειμένου να καταφέρει να τρέξει κανονικά η σελίδα. Για να επιτευχθεί η επικοινωνία της Php και του xampp θα πρεπει να ορίσουμε το localhost. Πιο συγκεκριμένα μέσα στην Php ορίζω :

```
<?php

$link = mysql_connect('localhost',  'mysql_user',  'mysql_password');

if (!$link) {

die('Could not connect: ' . mysql_error());

}

echo 'Connected successfully';

mysql_close($link);

?>
```

Εισάγουμε τον κώδικα μας που έχουμε γράψει σε γλώσσα html και php μέσα στον φάκελο htdocs. Για να φορτώσουμε συνεπώς την σελίδα αυτή θα πρέπει να βάλουμε στην μπάρα διεύθυνσης την συγκεκριμένη ονομασία που έχουμε δώσει στο αρχείο που βάλαμε μέσα στο htdocs και μπροστά από αυτό να γράψουμε localhost.

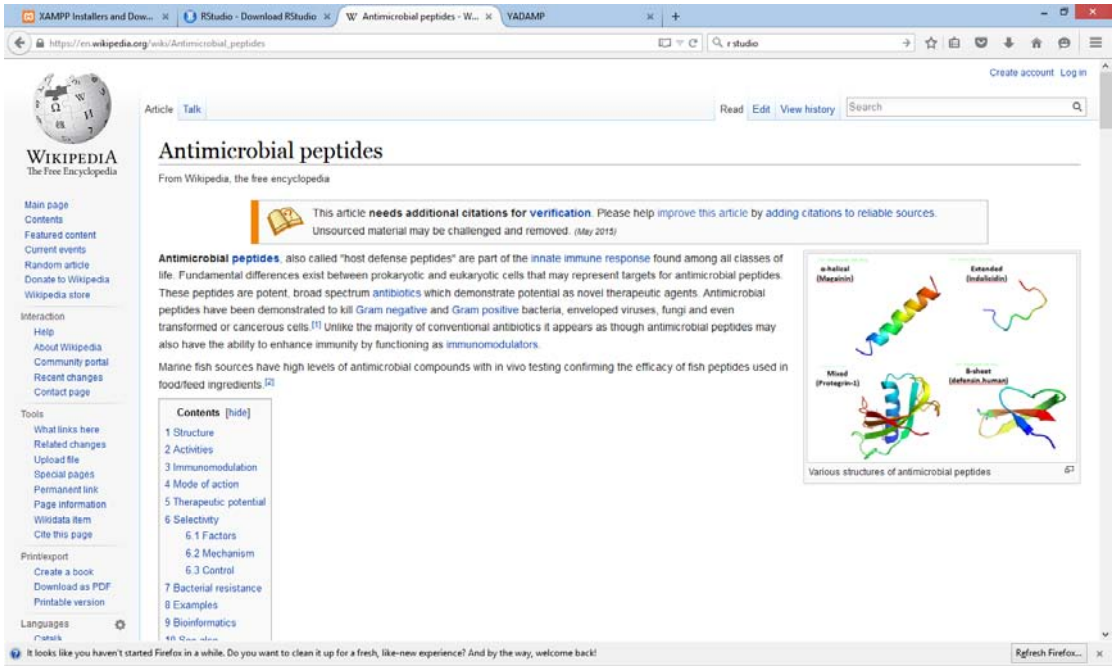
Δηλ: localhost/data/code.php

Ο κώδικας που έχουμε δημιουργήσει δίνει την δυνατότητα μέσω του Site ο χρήστης να αναζητά ένα αντιμικροβιακό πεπτίδιο ενώ παρέχει την δυνατότητα να επιστρέφει μαζικά αποτελέσματα από το dataset που εισαγάγαμε στην βάση δεδομένων.



Εικόνα 3.12 *Φόρτωση σελίδας*

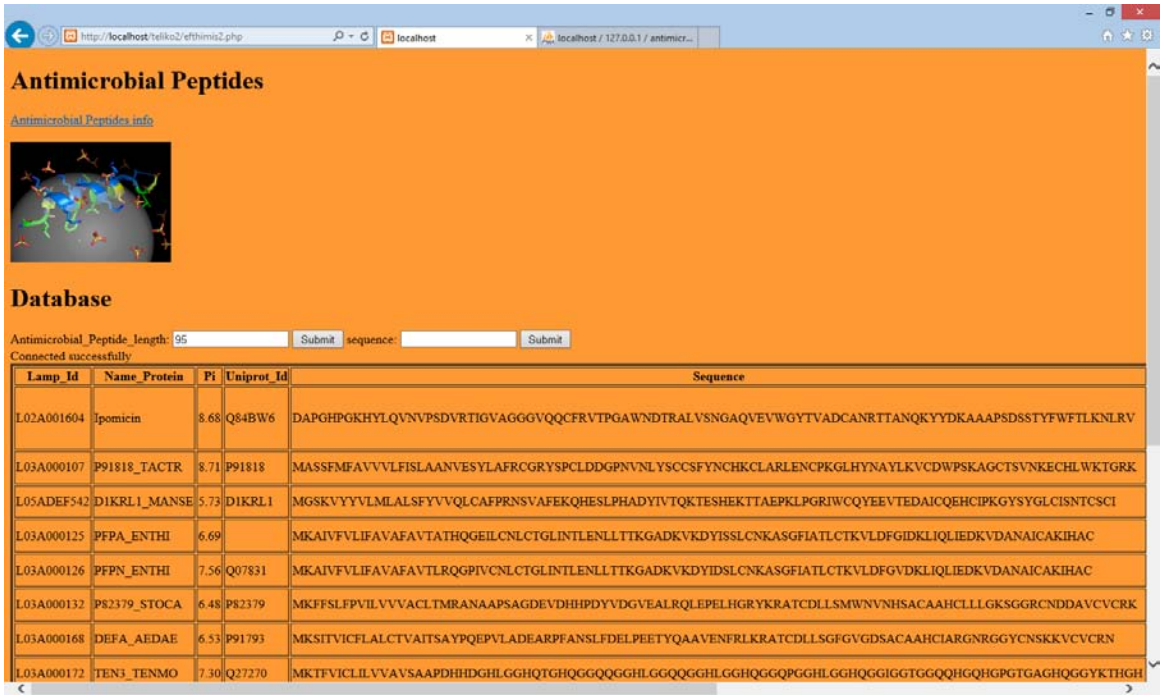
Κάνοντας κανείς κλικ πάνω στο antimicrobial peptides info αντίστοιχα ανοίγει σε νέα σελίδα το Site της Wikipedia που αναφέρει όλες τις πληροφορίες για τα αντιμικροβιακά πεπτίδια όπως απεικονίζεται στην ακόλουθη εικόνα.



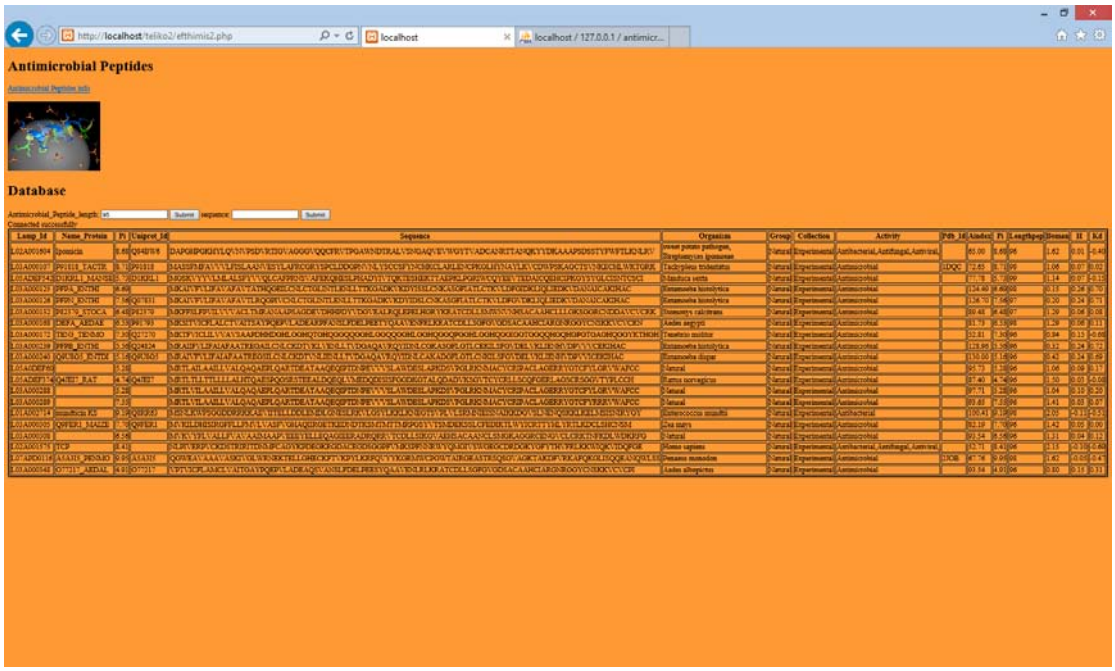
Εικόνα 3.13 *Φόρτωση σελίδας Wikipedia*

Υπάρχει δυνατότητα αναζήτησης είτε με βάση το μέγεθος του πεπτιδίου επιστρέφοντας τα αποτελέσματα από την βάση δεδομένων είτε αναζήτησης κάποιας συγκεκριμένης ακολουθίας. Αν δεν εισάγει ο χρήστης κανένα δεδομένο και κάνει κλικ στο κουμπί, τότε επιστρέφει όλα τα δεδομένα που περιέχει η βάση δεδομένων.

Πιο συγκεκριμένα για παράδειγμα αν επιθυμούμε να μας επιστρέψει όλα τα πεπτίδια που έχουν περισσότερα από 95 αμινοξέα, παίρνουμε τα ακόλουθα αποτελέσματα. Ουσιαστικά γίνεται αναζήτηση στην βάση δεδομένων αυτόματα και επιστρέφονται στο χρήστη τα χαρακτηριστικά κάθε πεπτιδίου με περισσότερα από 95 αμινοξέα.



Εικόνα 3.14 *Αναζήτηση πεπτιδίων με βάση τον αριθμό των αμινοξέων*



Εικόνα 3.15 *Αναζήτηση πεπτιδίων με βάση τον αριθμό των αμινοξέων (ολικά αποτελέσματα)*

3.4 Επικοινωνία βάσης δεδομένων με την R

Σε μετέπειτα στάδιο διαμορφώνουμε τον κώδικα που έχουμε δημιουργήσει στο htdocs αρχείο ώστε σε περίπτωση που εισάγει ο χρήστης μια δική του ακολουθία να καλείται αυτόματα το script της R και να υπολογίζει εκ νέου τα ποσοστά των αντιμικροβιακών χαρακτηριστικών για την συγκεκριμένη ακολουθία. Άρα όταν λοιπόν ο χρήστης εισάγει μια νέα ακολουθία υπολογίζονται εκ νέου τα χαρακτηριστικά καλώντας την R.

Μετέπειτα λοιπόν συνδέουμε την php και την R προκειμένου προκειμένου να καλείται το Sript της R και να βγάζει αποτελέσματα. Χρησιμοποιούμε την εντολή:

$cmd ="Rscripttest2.R $variable";

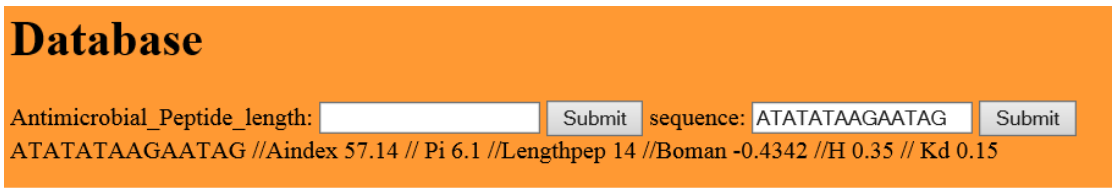
Όπου *$variable* αποτελεί την ακολουθία που έχει δώσει ο χρήστης στο site και θα αποτελεί την είσοδο στο script της R προκειμένου να υπολογίσουμε εκ νέου τα χαρακτηριστικά του πεπτιδίου που εισήγαγε ο χρήστης.

Στο Sript της R ορίζουμε την εντολή sink("data.txt") μέσω της οποίας θα δημιουργήσουμε ένα νέο αρχείο Txt μέσα στο htdocs που θα περιέχει τα αποτελέσματα.

Αντίστοιχα λοιπόν βάζουμε το Sript της R μέσα στο htdocs όπως έχουμε και τον υπόλοιπο κώδικα. Ορίζουμε μέσα στων κώδικα της Php τα αποτελέσματα από το Script να τα βάζει σε ένα αρχείοtxt όπως προαναφέραμε και αντίστοιχα μέσω της ακόλουθης εντολής τα εμφανίζουμε στο Site.

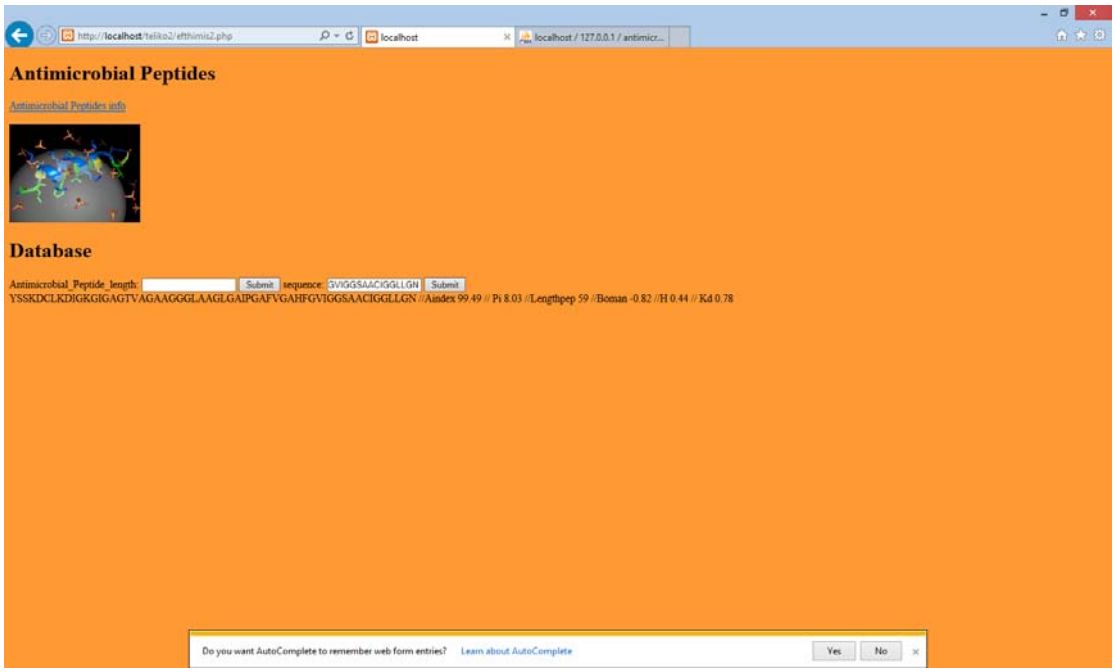
$file = file_get_contents('C:/xampp/htdocs/teliko2/data.txt', true);

print $file;



Εικόνα 3.16 *Υπολογισμός αντιμικροβιακών χαρακτηριστικών.*

Παραπάνω λοιπόν εισάγαμε μία συγκεκριμένη ακολουθία η οποία δεν βρίσκεται μέσα στην βάση δεδομένων οπότε και υπολογίζει εκ νέου τα χαρακτηριστικά του πεπτιδίου αυτού.



Εικόνα 3.17 *Υπολογισμός χαρακτηριστικών.*

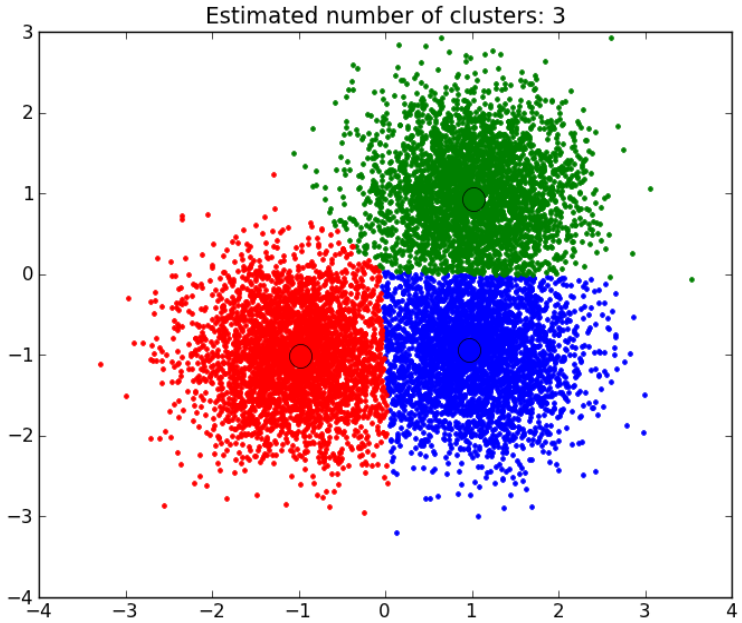
3.5 Αποτελέσματα

Δημιουργήσαμε λοιπόν τον ιστότοπο ο οποίος επικοινωνεί με την βάση δεδομένων μας προκειμένου να τραβάει τα αποτελέσματα της αναζήτησης του χρήστη με την δημιουργία κώδικα σε php. Αντίστοιχα με την δημιουργία κώδικα σε R καταφέραμε να υπολογίζουμε εκ νέου τα αποτελέσματα (χαρακτηριστικά πεπτιδίου) για κάθε νέα εισαγόμενη ακολουθία από τον χρήστη.

Ομαδοποίηση Αντιμικροβιακών Πεπτιδίων

4.1 Ομαδοποίηση Δεδομένων

Ομαδοποίηση ή αλλιώς συσταδοποίηση, είναι η οργάνωση μιας συλλογής από αντικείμενα σε ομάδες (clusters) με βάση κάποιο μέτρο ομοιότητας των στοιχείων αυτών. Τα στοιχεία αποτελούν διανύσματα ή σημεία στον πολυδιάστατο χώρο. Τα στοιχεία συνεπώς που ανήκουν στην ίδια ομάδα εμφανίζουν ομοιότητα [33]. Ένα ιδιαίτερο χαρακτηριστικό της ομαδοποίησης, σε αντίθεση με την κατηγοριοποίηση, είναι ότι η δομή και το πλήθος των ομάδων είναι καταρχάς άγνωστα και καθορίζονται δε από τον εκάστοτε αλγόριθμο συσταδιοποίησης. Αυτοί οι αλγόριθμοι βασίζονται στο σύνολο τους στην αρχή της μεγιστοποίησης της ομοιότητας ανάμεσα στα αντικείμενα την ίδιας ομάδας (intra-class similarity) και την ταυτόχρονη αρχή της ελαχιστοποίησης της ομοιότητας μεταξύ των αντικειμένων διαφορετικών ομάδων (inter-class similarity). Αξίζει να σημειωθεί ότι η ερμηνεία των ομάδων που προκύπτουν από την ανωτέρω διαδικασία καθορίζεται από τον εκάστοτε χρήστη.



Εικόνα 4.1 *Ομαδοποίηση δεδομένων σε 3 ομάδες*

Από τον παραπάνω ορισμό προκύπτει άμεσα και η βασική διαφορά μεταξύ κατηγοριοποίησης και συσταδοποίησης. Στην κατηγοριοποίηση ο αριθμός και η ουσία των συστάδων αποτελεί πληροφορία εκ των προτέρων γνωστή. Εξαιτίας αυτού, στη συσταδιοποίηση εφαρμόζεται πάντα μη εποπτευόμενη μάθηση, εν αντιθέση με την κατηγοριοποίηση όπου λόγω της πρότερης γνώσης των κλάσεων κάνουμε χρήση της εποπτευόμενης μάθησης. Στην συσταδοποίηση δεν υπάρχουν προκαθορισμένες κατηγορίες ομαδοποίησης αλλά οι εγγραφές συγκεντρώνονται σε ομάδες με βάση το κριτήριο που θέτει ο χρήστης για κάθε συστάδα όπως για παράδειγμα, η ομαδοποίηση πελατών που αγοράζουν παρόμοια αγαθά. Σκοπός είναι η δημιουργία συστάδων με όσο το δυνατόν περισσότερα κοινά χαρακτηριστικά εντός της εκάστοτε ομάδας, ενώ ταυτόχρονα η μία ομάδα από την άλλη θα πρέπει να διαφοροποιείται ικανοποιητικά ώστε να μη συγχέονται. Δηλαδή θα πρέπει να δημιουργηθούν διακριτές ομάδες με βάση ξεκάθαρα χαρακτηριστικά που περιγράφουν την κάθε ομάδα και την κάνουν να ξεχωρίζει από τις υπόλοιπες.

Η συσταδιοποίηση διακρίνεται σε τρεις βασικές μεθόδους:

1. Μέθοδοι διαχωρισμού (partitioning methods): Δημιουργούν ομάδες από ένα δεδομένο αρχικό σύνολο αντικειμένων με κάθε ομάδα να αντιπροσωπεύει ένα cluster και να ικανοποιούνται οι εξής δύο συνθήκες: (α) κάθε cluster περιέχει τουλάχιστον ένα αντικείμενο και (β) κάθε αντικείμενο ανήκει σε ένα μόνο cluster.

2. Ιεραρχικές μέθοδοι (hierarchical methods): Διασπούν το αρχικό σύνολο δεδομένων δημιουργώντας μια ιεραρχική δομή από clusters και διακρίνονται σε agglomerative (bottom-up) ή divisive (top-down) ανάλογα με τον τρόπο που γίνεται η διάσπαση.

3. Μέθοδοι βασισμένες σε μοντέλα (model-based methods): Υποθέτουν ότι καθένα από τα clusters περιγράφεται από ένα μαθηματικό μοντέλο και εντοπίζουν τα αντικείμενα που ανήκουν σε κάθε cluster, ώστε να ικανοποιούν το αντίστοιχο μοντέλο.

Προσπαθήσαμε λοιπόν να ομαδοποιήσουμε τα δεδομένα μας μέσω της R. Πρωταρχικό μας μέλημα αποτελούσε η κανονικοποίηση των δεδομένων μας στην μονάδα ενώ μετέπειτα φορτώσαμε το πακέτο apcluster στην R [33], και δημιουργήσαμε τον πίνακα αποστάσεων που θα χρησιμοποιούσαμε για την ομαδοποίηση των δεδομένων.

4.1.1 Κανονικοποίηση δεδομένων

Σκοπός της κανονικοποίησης είναι η αντιστοίχιση των τιμών των δεδομένων από το διάστημα [minA, maxA] στο διάστημα [new_minA, new_maxA], δηλαδή αποτελεί την κλιμάκωση σε ένα μικρό περιορισμένο εύρος. Αναφέρουμε ενδεικτικά κάποιες μεθόδους κανονικοποίησης:

- *min-max κανονικοποίηση*

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- *z-score κανονικοποίηση*

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

- κανονικοποίηση με δεκαδική κλίμακα

$$v' = \frac{v}{10^j}$$ (Όπου j είναι ο μικρότερος ακέραιος τέτοιος ώστε Max(| |)<1)

Παράδειγμα Min-Max Τεχνικής Κανονικοποίησης

Εμείς θα χρησιμοποιήσουμε τον αλγόριθμο MIN-MAX που αναφέρει όπως είπαμε τα ακόλουθα προκειμένου να κανονικοποιήσουμε τα δεδομένα μας. Έστω ότι τα δεδομένα μας έχουν κλίμακα από 30-50 και έστω ότι θέλουμε να τα μετασχηματίσουμε ώστε να κυμαίνονται από 0-1, χρησιμοποιώντας τον αλγόριθμο Min-max normalization.

Το στοιχείο s=30 αντιστοιχίζεται ως εξής:

S(30)=(30-30)/(50-30) = 0

Το στοιχείο s=50 αντιστοιχίζεται ως εξής:

S(50)=(50-30)/(50-30) = 1

Το ενδιάμεσο στοιχείο s=35αντιστοιχίζεται ως εξής:

S(35)= (35-30)/(50-30) = 5/20 = 0.25

4.1.2 Δημιουργία Πινάκα Αποστάσεων

Δημιουργούμε συνεπώς τον κώδικα στην R ο οποίος κανονικοποιεί τα δεδομένα μας (τα χαρακτηριστικά που έχουμε εξάγει για κάθε πεπτίδιο), κανονικοποιώντας όλα τα δεδομένα μας μεταξύ 0-1. Με αυτό τον τρόπο έχουμε δημιουργήσει έναν νέο πίνακα που περιέχει τα χαρακτηριστικά του κάθε πεπτιδίου κανονικοποιημένα όλα στην μονάδα, υπολογίζοντας μετέπειτα τον πίνακα αποστάσεων για τα χαρακτηριστικά αυτά.

Η συνάρτηση της απόστασης D μεταξύ δύο στοιχείων, πρέπει να ικανοποιεί την τριγωνική ανισότητα, δηλ:

D(x, x) = 0

D(x, y) = D(y, x)

D(x, y))<= D(x, z) + D(z, y)

Εξισώσεις εύρεσης πίνακα αποστάσεων:

Ευκλείδεια απόσταση:

$$d(\mathbf{x},\mathbf{y}) = \sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2 + \cdots + (y_n - x_n)^2} = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2}$$

Απόσταση Manhattan:

$$\mathrm{Md(S,F)} = |\, \mathrm{X_S} - \mathrm{X_F} \,| + |\, \mathrm{Y_S} - \mathrm{Y_F} \,|$$

Μέγιστο της διαφοράς σε κάθε διάσταση:

$$D(x,y) = \max_{i=1}^{k} |\, x_i - y_i \,|$$

Για τον υπολογισμό του πίνακα αποστάσεων η εξίσωση που χρησιμοποιούμε είναι η εξής : (Ευκλείδεια απόσταση)

dist (p1, p2) = sqrt(((F11-F21)^2) + ((F12-F22)^2) ...+... ((F1n-F2n)^2)), όπου F αποτελούν τα χαρακτηριστικά για κάθε πεπτίδιο. Πιο συγκεκριμένα:

Peptides	feature 1	feature 2		feature n
P1	F11	F12	...	F1n
P2	F21	F22	...	F2n
P3	F31	F32	...	F3n
P4	F41	F42	...	F4n

Η συνάρτηση S=negDistMat(A) που υπάρχει μέσα στο πακέτο υπολογίζει κατευθείαν τον πίνακα αποστάσεων. Μετέπειτα με την χρήση της apr<-apcluster(S) και της plot απεικονίζω γραφικά την ομαδοποίηση των δεδομένων. Αφότου έχουμε κανονικοποιήσει τα δεδομένα μας και έχουμε βρει τον πίνακα αποστάσεων καλούμε τον αλγόριθμο affinity propagation [34] (αλγόριθμος ομαδοποίησης δεδομένων) μέσω της συνάρτησης apcluster. Υπολογίσαμε συνεπώς τον πίνακα αποστάσεων για όλα τα πεπτίδια μας και για όλα τα χαρακτηριστικά που έχουμε βρει και ομαδοποιούμε πλέον τα πεπτίδια μας.

- *affinity propagation*

Αποτελεί αλγόριθμο ομαδοποίησης δεδομένων . Κυρίαρχο χαρακτηριστικό του AP είναι ότι δεν απαιτείται ο αριθμός των cluster να είναι προσδιορισμένος πριν την εκτέλεση του αλγορίθμου , σε αντιθεση με τον K - means. Οι δύο κύριες παράμετροι που ελέγχουν την ομαδοποίηση είναι το κλάσμα των σημείων των δεδομένων που θα πρέπει να επιλεγεί για την ομαδοποίηση των δεδομένων, καθώς και ο αριθμός των βημάτων ή των επαναλήψεων. Αρχικά επιλέγεται ένα τυχαίο δείγμα ενώ για τις επόμενες επαναλήψεις τα υποδείγματα του προηγούμενου τρεξίματος διατηρούνται στο υποσύνολο του δείγματος και από τα υπόλοιπα δείγματα στο υποσύνολο η επιλογή γίνεται εκ νέου τυχαία. Το καλύτερο αποτέλεσμα του συνόλου των ερευνών αυτών με την υψηλότερη ομοιότητα διατηρείται ως το τελικό αποτέλεσμα για την ομαδοποίηση του συνόλου [35].

4.2 Αποτελέσματα

Παρατηρούμε ότι συνολικά για όλα τα δεδομένα μας (3044 πεπτίδια) ο αλγόριθμος δημιουργεί συνολικά 165 clusters. Θα προσπαθήσουμε να περιορίσουμε σε λιγότερες τις ομάδες που δημιουργούνται. Παρατηρούμε ότι στην συνάρτηση που καλούμε apc<-apcluster(S, K) μπορούμε εάν επιθυμούμε να ορίσουμε εμείς το K που ουσιαστικά αποτελεί τον αριθμό των ομάδων που εμείς θέλουμε να δημιουργήσουμε. Για παράδειγμα αν στην παραπάνω συνάρτηση ορίσουμε K=25 τότε ο αλγόριθμος θα ταξινομήσει όλα τα πεπτίδια που έχουμε με βάση τα χαρακτηριστικά που έχουμε βρει σε 25 ομάδες πεπτιδίων. Μελετώντας όλα τα πεπτίδια που έχουμε συλλέξει στην βάση δεδομένων μας παρατηρούμε ότι συνολικά υπάρχουν οι ακόλουθες ομάδες πεπτιδίων :

1) Antifungal

2) Antimicrobial

3) Anibacterial

4) Antiparasitic

5) Antiviral

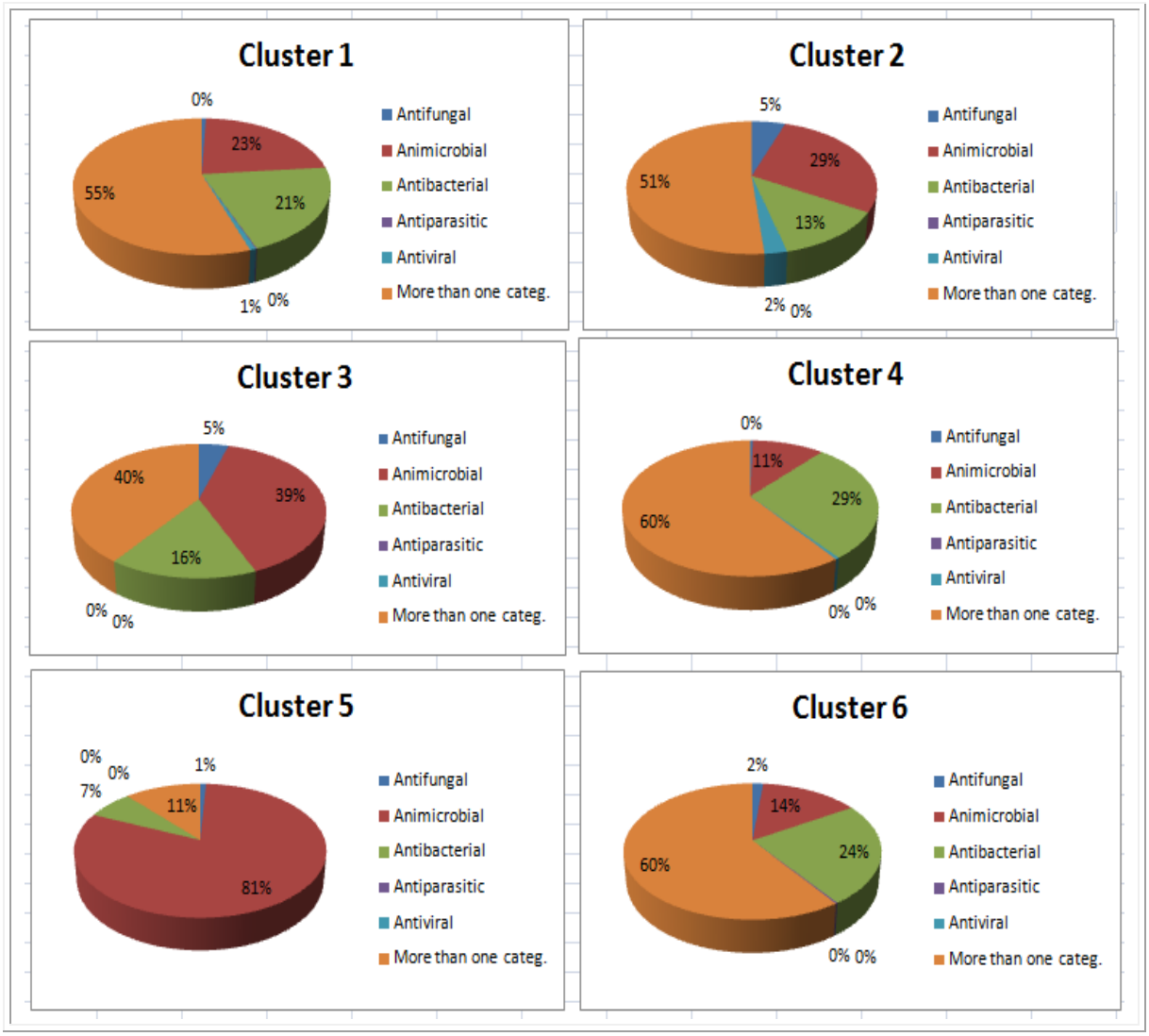
6) Πεπτίδια που ανήκουν σε περισσότερες από μια από τις παραπάνω κατηγορίες

Ομαδοποιούμε λοιπόν όλα τα δεδομένα μας σε 6 μεγάλες κατηγορίες. Ο affinity propagation δημιουργεί λοιπόν αυτές τις 6 ομάδες και πάμε τώρα στα συνολικά data που έχουμε προκειμένου να δούμε τι ποσοστό κάθε κατηγορίας βρίσκεται μέσα σε κάθε ομάδα. Δηλ ο αλγόριθμος κατατάσσει για παράδειγμα από το 3 έως το 34 μια ομάδα (cluster 1) από τις ομάδες αυτές που προκύπτουν υπολογίζουμε λοιπόν τι ποσοστό από αυτά τα πεπτίδια που ανήκουν στα antifungal, antimicrobial και ούτω καθεξής. Συνεπώς υπολογίζουμε τα ακόλουθα μετά από την ομαδοποίηση των δεδομένων μας.

Παρακάτω φαίνονται τα αποτελέσματα από την ομαδοποίηση των δεδομένων μας σε 6 συνολικά ομάδες.

	R	CLUSTERS	6	RESULTS						
cluster 01										
	SUM1	SUM2	SUM3	SUM4	SUM5	SUM6		ALL		
	5	180	164	1	5	436		791		
cluster 02										
	SUM 1	SUM 2	SUM 3	SUM 4	SUM 5	SUM6		ALL		
	21	116	51	0	10	209		407		
cluster 03										
	SUM 1	SUM 2	SUM 3	SUM 4	SUM 5	SUM 6		ALL		
	20	175	71	0	0	179		445		
cluster 04										
	SUM 1	SUM 2	SUM 3	SUM 4	SUM 5	SUM 6		ALL		
	2	53	141	0	2	295		493	Antifungal	1
									Antimicrobial	2
cluster 05									Antibacterial	3
	SUM 1	SUM 2	SUM 3	SUM 4	SUM 5	SUM 6		ALL	Antiparasitic	4
	4	376	32	0	0	53		465	Antiviral	5
									more than one categ.	6
cluster 06										
	SUM 1	SUM 2	SUM 3	SUM 4	SUM 5	SUM 6		ALL		
	7	64	107	1	0	264		443		
								3044		
SUM	59	964	566	2	17	1436		3044		

Εικόνα 4.2 *Ομαδοποίηση δεδομένων σε 6 ομάδες*



Εικόνα 4.3 *Γραφική απεικόνιση ομαδοποίησης δεδομένων σε 6 ομάδες*

Παρατηρούμε λοιπόν ότι κάποια από τα clusters ταιριάζουν κατά πλειοψηφία με κάποια κατηγορία των αντιμικροβιακών πεπτιδίων και αυτό χρήζει προς μελέτη ώστε να διερευνηθούν κάποιες μετρικές (για παράδειγμα δομικά χαρακτηριστικά των πεπτιδίων) για την εύρεση της καλύτερης ομαδοποίησης των δεδομένων.

Συμπεράσματα

Σύμφωνα με όλα τα παραπάνω καταφέραμε να υπολογίσουμε τα χαρακτηριστικά των 3044 αντιμικροβιακών πεπτιδίων με την χρήση πακέτων στην R και να τα εισάγουμε στην βάση δεδομένων μας. Παράλληλα δημιουργήσαμε έναν ιστότοπο προκειμένου να κάνει ο χρήστης αναζήτηση των χαρακτηριστικών ενός αντιμικροβιακού πεπτιδίου από την βάση δεδομένων. Επιπρόσθετα μπορεί να εισάγει μια συγκεκριμένη ακολουθία στο site και να υπολογίζει εκ νέου τα χαρακτηριστικά του πεπτιδίου αυτού. Σε μελλοντικό στάδιο εκμεταλλευόμενοι το γεγονός ότι έχουμε βρει το pdb id για κάθε πρωτεΐνη (εφόσον υπάρχει και έχει εντοπισθεί η δομή της) και το έχουμε εισάγει στην βάση δεδομένων μας θα μπορούσαμε να δημιουργήσουμε ένα γραφικό περιβάλλον προκειμένου μέσω του pdb id να εμφανίζει την δομή της συγκεκριμένης ακολουθίας στον ιστότοπο. Επιπρόσθετα θα μπορούσαμε για την ακολουθία που θα εισάγει ο χρήστης κρίνοντας και εκπαιδεύοντας το σύστημα μας να κατατάξουμε τελικά την ακολουθία μας ως αντιμικροβιακή ή όχι.

Η βάση δεδομένων που δημιουργήσαμε διαφέρει από τις υπάρχουσες βάσεις δεδομένων καθώς έχουμε υπολογίσει συγκεκριμένες μετρικές για κάθε πεπτίδιο ενώ αντίστοιχα περιέχεται μέσα στην βάση και η δομή του κάθε πεπτιδίου, παρέχοντας μελλοντικό στόχο την τρισδιάστατη απεικόνιση του πεπτιδίου όπως προαναφέραμε. Με την ομαδοποίηση των δεδομένων μας προσπαθήσαμε να διερευνήσουμε κατά πόσο μπορούμε να πλησιάσουμε στην λειτουργική/βιολογική ομαδοποίηση των αντιμικροβιακών πεπτιδίων. Χρειάζεται ακόμα περισσότερη προσπάθεια και δοκιμές για να επιτευχθεί η βέλτιστη ομαδοποίηση σε αυτήν την κατεύθυνση. Ωστόσο αυτή η προσπάθεια καθώς και οποιαδήποτε άλλη στο πεδίο της μηχανικής μάθησης είναι πιο εύκολο να πραγματοποιηθεί με την βάση που παρουσιάσαμε στην παρούσα εργασία.

Βιβλιογραφία

[1]"amino acid" ,Cambridge Dictionaries Online / Cambridge University Press. 2015. Retrieved 2015-07-03.

[2] P. H. Raven and G. B. Johnson, Biology, 6th ed. Dubuque, IA: McGraw-Hill, 2002.

[3] Αλεξάνδρα Πατρινέλη, «Η αναδίπλωση των πρωτεϊνών: Ένα φαινόμενο με πληθώρα εφαρμογών», Περισκόπιο της Επιστήμης, τεύχος 216 (Απρίλιος 1998), σσ. 70-76

[4] Γιαλούρης Παρασκευάς, Μποσινάκου Κατερίνα, Σιδέρης Διαμαντής, Βιοχημεία Γ Γενικού Λυκείου, Τεχνολογική κατεύθυνση (κεφάλαιο 3.1, 3.2, 3.3 ,3.4), Οργανισμός εκδόσεων διδακτικών βιβλίων.

[5] ReddyKV, YederyRD, AranhaC (2004). "Antimicrobial peptides: premises and promises". International Journal of Antimicrobial Agents 24 (6): 536–547.

[6] Αδαμαντιάδου Σ., Γεωργάτου Μ., Γιαπιτζάκης Χ., Λακκά Λ., Νοταράς Δ., φλωρεντίν Ν., Χατζηγεωργίου Γ, Χαντηκωντή Ολ, Βιολογία Γενικής Παιδείας (Γ΄ Τάξης Γενικού Λυκείου.), Κεφάλαιο 1.3 ,Μηχανισμοί άμυνας του ανθρώπινου οργανισμού - Βασικές αρχές ανοσίας.

[7] H.U. Stotz, F. Waller, and K. Wang, Innate Immunity in Plants: The Role of Antimicrobial Peptides.

[8] Nguyen LT, Haney EF, Vogel HJ. "The expanding scope of antimicrobial peptide structures and their modes of action". Trends in Biotechnology 29 (9): 464–472. doi:10.1016/j.tibtech.2011.05.001. PMID 21680034, 2011.

[9] Brogden, K.A., "Antimicrobial peptides: pore formers or metabolic inhibitors in bacteria?", Nature Reviews Microbiology 3 (3): 238–250., doi:10.1038/nrmicro1098, PMID 15703760, March 2005.

[10] Kuo HH1, Chan C, Burrows LL, Deber CM, Hydrophobic interactions in complexes of antimicrobial peptides with bacterial polysaccharides.

[11] Hancock, Robert E.W.; Sahl, Hans-Georg, "Antimicrobial and host-defense peptides as new anti-infective therapeutic strategies", Nature Biotechnology 24 (12): 1551–1557, doi:10.1038/nbt1267, PMID 17160061, 2006.

[12] Riadh Hammami,1,2,3 Jeannette Ben Hamida,1 Gérard Vergoten, 2 and Ismail Fliss3, PhytAMP: a database dedicated to antimicrobial plant peptides.

[13] Wang Z1, Wang G.,APD: the Antimicrobial Peptide Database.

[14] Xiaowei Zhao ,Hongyu Wu ,Hairong Lu,Guodong Li,Qingshan Huan,Published: June 18, 2013 DOI: 10.1371/journal.pone.0066557LAMP,A Database Linking Antimicrobial Peptides.

[15] Ιστότοπος: http://www.uniprot.org

[16] Ιστότοπος: http://www.ncbi.nlm.nih.gov

[17]Ιστότοπος: http://www.rcsb.org/pdb/home/home.do

[18] Waghu FH1, Gopi L, Barai RS, Ramteke P, Nizami B, Idicula-Thomas S, CAMP: Collection of sequences and structures of antimicrobial peptides, 21 Nov. 2013

[19] R Development Core Team. R: A Language and Environment for Statistical Computing, Vienna, Austria, 2009.

[20] Ιστότοπος: http://www.ncbi.nlm.nih.gov/pubmed

[21] Zhe Wang and Guangshun Wang, APD: the Antimicrobial Peptide Database, September 3, 2003.

[22] HongyuWu ,Hairong Lu,Guodong Li,Qings han Huang, LAMP: A Data base Linking Antimicrobial Peptides, June 18, 2013, DOI: 10.1371/journal.pone.0066557

[23] Stotz, F.Waller, and K.Wang, Innate Immunity in Plants: The Role of Antimicrobial Peptides. H.U , April 13, 2011.

[24] Σακκάς Η, Η μελέτη της αντιμικροβιακής δράσης αιθέριων ελαίων, Διδακτορική διατριβή, Πανεπιστήμιο Ιωαννίνων, 2007.

[25] Walsh SE, Maillard JY, Russel AD, Catrenich CE, Charboneau DL, Bartolo RG. Activity mechanisms of action of selected biocidal agents on Gramm-positive and negative bacteria.JApplMicrobiol 2003 ;94:240-247

[26] Guangshun Wang, Book “Antimicrobial Peptides Advances in Molecular and Cellular Biology”, Series(1) , 2010.

[27] D.M. Smith and the R development core team,An intro-duction to R, W.N. Venables

[28] Delphine Charif and Jean R. Lobry and AnamariaNecsulea and Leonor Palmeira and Si-mon Penel and Guy Perriere, Package ‘seqinr’, Biological Sequences Retrieval and Analysis.

[29] Daniel Osorio, PaolaRondon-Villarreal and Rodrigo Torres, Package “Peptides”, Calculate Indices and Theoretical Properties of Protein Sequences, November 27, 2015.

[30] J. Kyte and R. F. Doolittle, “A simple method for displaying the hydropathic character of a protein,” J. Mol. Biol., vol. 157, no. 1, pp. 105–132, May 1982.

[31] «Interview with Kai Seidler from the XAMPP project». MySQL AB, 22 August, 2009.

[32] Michele E. Davis, Jon A. Phillips. Learning PHP & MySQL. O'Reilly, 2007,σελ.33-38. ISBN 0-596-51401-8.

[33] Jain et al, 1999, Larose, 2004.

[34] Brendan J. Frey; Delbert Dueck, "Clustering by passing messages between data points". Science 315 (5814): 972–976. 2007.

[35] Ulrich Bodenhofer, Johannes Palme, Chrats Melkonian, and Andreas Kothmeier , APCluster-An R Package for Affinity Propagation Clustering, Version 1.4.2, December 24, 2015.

ΠΑΡΑΡΤΗΜΑ

Code in PHP

Μέσω του παρακάτω κώδικα καταφέραμε να δημιουργήσουμε το site που προαναφέραμε παραπάνω το οποίο δίνει την δυνατότητα στο χρήστη να αναζητάει εύκολα αντιμικροβιακά πεπτίδια στην βάση δεδομένων (που δημιουργήθηκε μέσω του xampp), καθώς και να υπολογίζει εκ νέου τα χαρακτηριστικά του τυχαίου πεπτιδίου που πιθανώς να εισάγει ο χρήστης.

Αρχική προϋπόθεση για την ορθή λειτουργία είναι η εισαγωγή και η δημιουργία της βάσης δεδομένων στο xampp.

```
<!DOCTYPE html>
 <html>
  <body>
   <h1>  Antimicrobial   Peptides   </h1>
   <a href="http://en.wikipedia.org/wiki/Antimicrobial_peptides"> Antimicrobial Peptides info </a>
<br>
<br>
<img src = "http://biologia.dip.unina.it/wp-content/themes/itheme2/themify/img.php?src=http://biologia.dip.u
   <style>
      body {background-color:#FF9933 ;}
      p {color:yellow;}
   </style>
<h1> Database </h1>
<form  method= "post">
Antimicrobial_Peptide_length: <input type="number" name= "variable" id="variable" >
<input type="submit" name="mysubmit" id="mysubmit" value="Submit">
sequence: <input type="char" name= "variable2" id="variable2" >
<input type="submit" name="mysubmit2" id="mysubmit2" value="Submit">
</form>
<?php
//Επικοινωνία-σύνδεση ιστοτόπου με την βάση δεδομένων μας στο xampp
if(isset($_POST['mysubmit']))
{
$link = mysql_connect('127.0.0.1','root', '');

if (!$link) {
    die('Could not connect: ' . mysql_error());
}
echo 'Connected successfully';
mysql_select_db ('antimicrobial peptides') or die('Could not select database');

if (isset($_POST['variable'])){
    $variable = $_POST['variable'];
}
settype($variable, "integer");
//Αναζήτηση μέσα στην βάση δεδομένων ,
//σύμφωνα με όσα ζήτησε ο χρήστης στο κουμπί και αποθηκεύτηκε στην μεταβλητή variable
$sql= " SELECT * FROM `table 2` WHERE `lengtpep` > $variable";
$result=mysql_query($sql)or die($sql."<br/><br/>".mysql_error());
$invest_rows_num=mysql_num_rows($result);
```

```
echo "<table border='3'>";
echo "<tr> <th>Lamp_Id</th> <th>Name_Protein</th> <th>Pi</th> <th>Uniprot_Id</th> <th>Sequence</
for($i=0; $i<$invest_rows_num; $i++)
 {
  while($row = mysql_fetch_array( $result ))
    {
// Εμφάνιση αποτελεσμάτων σε κατάλληλη μορφή πίνακα
   echo "<tr><td>";
          echo $row['lamp_id'];
          echo "</td><td>";
          echo $row['name_protein'];
          echo "</td></td><td>";
          echo $row['pi'];
          echo "</td></td></td><td>";
          echo $row['uniprot_id'];
          echo "</td></td></td></td><td>";
          echo $row['sequence'];
          echo "</td></td></td></td></td><td>";
          echo $row['organism'];
          echo "</td></td></td></td></td></td><td>";
          echo $row['group'];
          echo "</td></td></td></td></td></td></td><td>";
          echo $row['collection'];
          echo "</td></td></td></td></td></td></td></td><td>";
          echo $row['activity'];
          echo "</td></td></td></td></td></td></td></td></td><td>";
          echo $row['pd_bid'];
          echo "</td></td></td></td></td></td></td></td></td></td><td>";
          echo $row['aindex'];
          echo "</td></td></td></td></td></td></td></td></td></td></td><td>";
          echo $row['pi'];
          echo "</td></td></td></td></td></td></td></td></td></td></td></td><td>";
          echo $row['lengtpep'];
          echo "</td></td></td></td></td></td></td></td></td></td></td></td></td><td>";
          echo $row['boman'];
          echo "</td></td></td></td></td></td></td></td></td></td></td></td></td></td><td>";
          echo $row['h'];
          echo "</td></td></td></td></td></td></td></td></td></td></td></td></td></td></td><td>";
          echo $row['kd'];
}
```

Παρακάτω φαίνεται η σύνδεση με την R στην είσοδο άγνωστης ακολουθίας.

```
echo "</table>";
}
mysql_close($link);
} //Υπολογισμός των νέων αντιμικροβιακών χαρακτηριστικών καλώντας το πρόγραμμα features.R
if(isset($_POST['mysubmit2']))
{
if (isset($_POST['variable2'])){
    $variable2 = $_POST['variable2'];
}
$cmd ="Rscript features.R $variable2";
exec($cmd,$b);
//Δημιουργεί ενα νέο αρχίο μορφής .txt και αποθηκεύει εκει μέσα τα χαρακτηριστικά αυτα
$file = file_get_contents('C:/xampp/htdocs/teliko2/data.txt', true);
print $file; //Εμφανιζει τα χαρακτηριστικά που εχει μέσα το αρχείο στον ιστότοπο
}
?>
  </body>
 </html>
```

R Code

Μέσω του παρακάτω κώδικα προσπαθούμε να βρούμε τα αξιόλογα χαρακτηριστικά των πεπτιδίων που αναφέραμε.

Απαραίτητη προϋπόθεση για τα παρακάτω αποτελεί η εγκατάσταση στο R studio της βιβλιοθήκης seqinr, Peptides και xlsx καθώς και ο ορισμός του σωστού μονοπατιού στο R studio για την αναγνώριση του αρχείου.

```
library("seqinr")
#Οριζώ το μονοπάτι στο οποίο έχω αποθηκεύσει το αρχείο με τα πεπτίδια
arxeio <- dir("C:/Users/Efthimis/Desktop/experimental-predicted",full.names=TRUE)
seq <- read.fasta(file = arxeio, forceDNAtolower = FALSE,seqonly=TRUE)
#Δημιουργώ τον μονοδιάστατο όπου κάθε γραμμή είναι μία ακολουθία πεπτιδίου

library("Peptides")#Κατεβάζω και φορτώνω την βιβλιοθήκη στην R
x1=x2=x3=x4=x5=x6=0
for(i in 1:length(seq)) #Τρέχει για όλο τον πίνακα μέχρι το 3044
{
  x1[i]=0 #Μηδενισμός πίνακα
  x1[i]=aindex(unlist(seq[i])) #Πρέπει πρώτα να το κάνουμε unlist
  x2[i]=0
  x2[i]=pI(unlist(seq[i]))
  x3[i]=0
  x3[i]=lengthpep(unlist(seq[i]))
  x4[i]=0
  x4[i]=boman(unlist(seq[i]))
  x5[i]=0
  x5[i]=h(unlist(seq[i]))
  x6[i]=0
  x6[i]=KD(unlist(seq[i]))
}
library("xlsx")
#Δημιουργία excel για την εισαγωγή των αποτελεσμάτων
write.xlsx(x = x1, file = "AINDEX.excelfile.xlsx",sheetName = "AINDEX", row.names = FALSE)
write.xlsx(x = x2, file = "pI.excelfile.xlsx",sheetName = "pI", row.names = FALSE)
write.xlsx(x = x3, file = "lengthpep.excelfile.xlsx",sheetName = "lengthpep", row.names = FALSE)
write.xlsx(x = x4, file = "BOMAN.excelfile.xlsx",sheetName = "BOMAN", row.names = FALSE)
write.xlsx(x = x5, file = "h.excelfile.xlsx",sheetName = "h", row.names = FALSE)
write.xlsx(x = x6, file = "KD.excelfile.xlsx",sheetName = "KD", row.names = FALSE)

```

Ο παρακάτω κώδικας αποτελεί το αρχείο features.R το οποίο καλείται μέσω της php στον παραπάνω κώδικα (σελ 66 code in php) προκειμένου να υπολογιστούν τα χαρακτηριστικά για την άγνωστη εισαγόμενη ακολουθία του χρήστη.

Απαραίτητη προϋπόθεση για την ορθή λειτουργία αποτελεί το σημείο τοποθέτησης του συγκεκριμένου κώδικα. Θα πρέπει να βρίσκεται στον ίδιο φάκελο με το αρχείο της php μέσα στο αρχείο htdocs του xampp.

```
args <- commandArgs(TRUE)  #Αποθήκευση ακολουθίας που εισηγάγαμε απο τον ιστότοπο
seq<-args[1]

print(seq) #Εμφανιση εισαγόμενης ακολουθίας
library("seqinr")  #Φόρτωση βιβλιοθήκης
library("Peptides")

x1=x2=x3=x4=x5=x6=0
#Υπολογισμός χαρακτηριστικών ακολουθίας
x1=aindex(unlist(seq))
x2=pI(unlist(seq))
x3=lengthpep(unlist(seq))
x4=boman(unlist(seq))
x5=h(unlist(seq))
x6=KD(unlist(seq))
print("aindex")
print(x1)
print("pI")
print(x2)
print("lengthpep")
print(x3)
print("boman")
print(x4)
print("h")
print(x5)
print("kd")
print(x6)
#Αποθήκευση χαρακτηριστικών σε αρχείο .txt
sink("data.txt")
#Εμφάνιση αποτελεσμάτων στον ιστότοπο
print(x1)
print(x2)
print(x3)
print(x4)
print(x5)
print(x6)
sink()
```

Κανονικοποιούμε τα δεδομένα μας πριν προβούμε στην επεξεργασία αυτών, προκειμένου βρίσκονται όλα τα δεδομένα στην ίδια κλίμακα.

```
library("igraph")  #Φορτώνω την βιβλιοθήκη
pinakas=read.csv("E:/csv/DATA-only-experimental-features.csv")
sthles=length(pinakas[1,])
grammes=length(pinakas[,1])
min=1000;
max=-1000;
#Παω να υπολογίψω ποιό ειναι το min - max στον πίνακα
for(k in 1:sthles)
{
  max[k]=-1000;
  min[k]=1000;
}
for(i in 1:sthles)
{ for(k in 1:grammes)
   {
    if (pinakas[k,i]>max[i])
        {max[i]=pinakas[k,i];
        }
    if (pinakas[k,i]<min[i])
       {min[i]=pinakas[k,i];
       }
   }
}
#Πάω να υπολογίσω normalization min-max
new_max<-1;
new_min<-(-1);
norm <- matrix(0, grammes, sthles)


for(i in 1:sthles)
{ for(k in 1:grammes)
  { #Υπολογίζω Min-Max σύμφωνα με τον μαθηματικό τύπο
   norm[k,i]<-((pinakas[k,i]-min[i])/(max[i]-min[i]))*(new_max-new_min)+new_min;
  }
}
  norm;
#Δημιουργώ το αρχείο με τα κανονικοποιημένα δεδομένα
library("xlsx")
write.xlsx(x = norm, file = "normalization_test.xlsx",sheetName = "sample", row.names = FALSE)
```

Προσπαθήσαμε λοιπόν να ομαδοποιήσουμε τα δεδομένα μας μέσω της R. Πρωταρχικό μας μέλημα αποτελούσε η κανονικοποίηση των δεδομένων μας στην μονάδα ενώ μετέπειτα φορτώσαμε το πακέτο apcluster στην R, και δημιουργήσαμε τον πίνακα αποστάσεων που θα χρησιμοποιούσαμε για την ομαδοποίηση των δεδομένων μας.

Απαραίτητη προϋπόθεση αποτελεί η εγκατάσταση της βιβλιοθήκης apcluster στο R studio.

```
library("apcluster")
S=negDistMat(norm) #Αυτόματος υπολογισμός πίνακα αποστάσεων
apr<-apcluster(S)
plot(apr,norm)
#---------------------------------
#---------------------------------clustering pre-defined number of clusters
library("apcluster")
S=negDistMat(norm)
apc=apclusterK(S,K=5);
apc
plot(apc,norm)

apc2=apclusterK(S,K=25);
apc2
plot(apc2,norm)

apc3=apclusterK(S,K=50);
apc3
plot(apc3,norm)

#---------------------------------
typeof(apc)      #Γράφω τα αποτελέσματα σε ένα excel
apc2<-apc@clusters
apcnew<-unlist(apc2)
typeof(apcnew)
write.xlsx(apcnew, file = "clusters-5.xlsx",sheetName = "sample", row.names = FALSE)
#---------------------------------
apclus<-apclusterK(S,K=20, prc=10, bimaxit=20, exact=FALSE,
                   maxits=1000, convits=100, lam=0.9, includeSim=FALSE, details=FALSE,
                   nonoise=FALSE, seed=NA, verbose=TRUE)
apclus
plot(apclus,norm)
```