ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

ΚΑΤΑΣΚΕΥΗ ΒΑΣΗΣ ΔΕΔΟΜΕΝΩΝ ΒΙΟΛΟΓΙΚΩΝ ΑΚΟΛΟΥΘΙΩΝ

Δήμητρα Ποχάνη

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ
Επιβλέπων
Παντελής Μπάγκος
Αναπληρωτής Καθηγητής

Λαμία, 2017

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: 04/10/2017

Η Δηλούσα

(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».

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα της παρούσας πτυχιακής εργασίας, αναπληρωτή καθηγητή Παντελή Μπάγκο για την υποστήριξη, την υπομονή και την εμπιστοσύνη που μου έδειξε.

Θα ήθελα επίσης, να ευχαριστήσω ιδιαίτερα τον επίκουρο καθηγητή Βασίλειο Προμπονά του Πανεπιστημίου Κύπρου για τη βοήθεια και την καθοδήγηση που μου προσέφερε κατά τη διαμονή μου στην Κύπρο.

Ακόμη θα ήθελα να ευχαριστήσω την οικογένειά μου που βρίσκεται στο πλευρό μου όλα αυτά τα χρόνια και υποστηρίζει με αγάπη και κατανόηση την κάθε μου επιλογή.

Τέλος, θα ήθελα να ευχαριστήσω τους φίλους μου που ήταν δίπλα μου, τόσο στις εύκολες, όσο και στις δύσκολες στιγμές της ζωής μου.

ΠΕΡΙΛΗΨΗ

Εξαιτίας του όγκου πληροφορίας που παρουσιάζεται στις βάσεις δεδομένων βιολογικών ακολουθιών, η ανάπτυξη αλγορίθμων που θα μπορούσαν να ταξινομήσουν και να προβλέψουν τις ιδιότητες μιας ακολουθίας είναι σημαντική. Ουσιαστικό εργαλείο για την υλοποίηση των αλγορίθμων αυτών αποτελεί η δημιουργία συνόλων δεδομένων, που αποτελούνται από ακολουθίες νουκλεϊνικών οξέων ή αμινοξέων. Επειδή τα σύνολα δεδομένων είναι διάσπαρτα στη βιβλιογραφία και τον ιστό, η συγκέντρωσή τους σε μια βάση δεδομένων κάνει πιο εύκολη την πρόσβαση σε αυτά. Σκοπός της παρούσας πτυχιακής εργασίας είναι η δημιουργία μιας βάσης δεδομένων με σύνολα δεδομένων βιολογικών ακολουθιών και συγκεκριμένα πρωτεϊνικών ακολουθιών, που έχουν δημιουργηθεί και χρησιμοποιηθεί για την εκπαίδευση και τον έλεγχο αλγορίθμων πρόγνωσης και έχουν δημοσιευθεί κατά καιρούς στη βιβλιογραφία.. Χρησιμοποιώντας σαν λέξεις κλειδιά τους όρους «*dataset*», *«protein prediction/classification», «membrane proteins», «signal peptides», «secondary structure prediction/classification»* κι άλλες λέξεις κλειδιά που αναφέρονται σε κάποιο συγκεκριμένο βιολογικό πρόβλημα, πραγματοποιήθηκε αναζήτηση στη βιβλιογραφία για την ανάκτηση των συνόλων δεδομένων πρωτεϊνικών ακολουθιών. Στη συνέχεια δημιουργήθηκε ένας πίνακας στον οποίο και καταχωρήθηκαν το όνομα του συνόλου δεδομένων, ένας μοναδικός κωδικός κι άλλες πληροφορίες για το κάθε σύνολο δεδομένων. Ακολούθως έγινε επεξεργασία των συνόλων δεδομένων, για τη δημιουργία μιας κοινής μορφής. Με την αναζήτηση που πραγματοποιήθηκε έγινε ανάκτηση εκατόν ογδόντα συνόλων δεδομένων πρωτεϊνικών ακολουθιών για GPCRs, σηματοδοτικές ακολουθίες, διαμεμβρανικές πρωτεΐνες, PTS, τη δευτεροταγή δομή των πρωτεϊνών, PTMs, την έκκριση των πρωτεϊνών, την αναδίπλωση των πρωτεϊνών, την αλλαγή στην σταθερότητα των πρωτεϊνών κατά τις μεταλλάξεις, τη δέσμευση των πρωτεϊνών στο DNA, τη δέσμευση των πρωτεϊνών στο RNA, πρόβλεψη υποκκυταρικού εντοπισμού της πρωτεΐνης, και για διάφορες άλλες κατηγορίες. Κατόπιν δημιουργήθηκε ο ιστότοπος στον οποίο και αναρτήθηκαν οι πληροφορίες από τη βάση, δίνοντας στους χρήστες την ευκαιρία να αναζητήσουν και να εξάγουν πληροφορίες για τα σύνολα δεδομένων.

Λέξεις κλειδιά: σύνολο δεδομένων, πρωτεϊνικές ακολουθίες, βιολογικές ακολουθίες, βιολογική βάση δεδομένων

ABSTRACT

Due to the amount of information in biological sequence databases, develop algorithms that could classify and predict the properties of a sequence is important. Datasets of nucleic acid or amino acids sequences are important for the implementation of these algorithms. Because datasets are scattered in the bibliography and the web, a database of dataset collection makes it easier to access. The scope of this thesis is to create a database of biological datasets and specifically protein datasets, which have been used for training and testing prediction algorithms and previously published in literature. The protein datasets was collected during the search on citation and biological databases using the keywords "dataset", "protein prediction or classification", "membrane proteins", "signal peptides", "secondary structure prediction or classification" and other keywords that refer to a particular biological problem. Then a table was created with the name of the dataset, a unique code and other information for each dataset. Because of the different format of the datasets were created a common format. The results give one hundred and eighty protein datasets for GPCRs, signal peptides, OMPs, transmembrane proteins, PTS, protein secondary structure, PTMs, protein secretion, protein folding, protein stability changes upon mutations, DNA-binding protein, RNA-binding protein, subcellular localization of proteins and for many other categories. The information about the datasets is available in a website, giving to users the choice of downloading the datasets.

Keywords: dataset, protein sequences, biological sequences, biological database

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΜΕΡΟΣ 2ο

ΜΕΘΟΔΟΙ ΠΡΟΒΛΕΨΗΣ

ΜΕΡΟΣ 3ο

ΣΥΝΟΛΑ ΔΕΔΟΜΕΝΩΝ

ΜΕΡΟΣ 4ο

ΥΛΟΠΟΙΗΣΗ

ΕΙΣΑΓΩΓΗ

Εξαιτίας της τρομακτικής αύξησης των δεδομένων που παρουσιάζεται στις βιολογικές επιστήμες και κυρίως στην μοριακή βιολογία οδήγησαν στην δημιουργία των βιολογικών βάσεων δεδομένων που περιέχουν δεδομένα που παράγονται από υπολογιστική ανάλυση, ανάλυση βιολογικών ακολουθιών, αναγνώριση γονιδίων, τον καθορισμό της δομής των πρωτεϊνών, τον προσδιορισμό του μηχανισμού διπλώματος των πρωτεϊνών, την κατανόηση γενετικών ασθενειών κ.α. Μια βιολογική βάση δεδομένων μπορεί να περιλαμβάνει βιβλιογραφίες, ακολουθίες πρωτεϊνών, τρισδιάστατες δομές πρωτεϊνών και νουκλεονικών οξέων, νουκλεοτιδικές ακολουθίες, δεδομένα γονιδιακής έκφρασης και δεδομένα γενετικής ποικιλομορφίας.

Αντικείμενο της παρούσας πτυχιακής εργασίας είναι η δημιουργία μιας βάσης δεδομένων με σύνολα δεδομένων βιολογικών ακολουθιών. Μια βιολογική ακολουθία μπορεί να είναι μια αλληλουχία νουκλεϊνικών οξέων, (Δεσοξυριβονουκλεϊκού οξέος (DNA) ή Ριβονουκλεϊκού οξέος (RNA)), ή μια αλληλουχία αμινοξέων.

Τα σύνολα δεδομένων βιολογικών ακολουθιών ως ψηφιακή πληροφορία, εφαρμόζονται σε αλγόριθμους για την επεξεργασία τους και την παραγωγή χρήσιμων αποτελεσμάτων, όσον αφορά τη δομή και τη λειτουργικότητα των ακολουθιών. Η δημιουργία συνόλων δεδομένων αποτελεί ουσιαστικό εργαλείο υλοποίησης των αλγορίθμων πρόβλεψης και ταξινόμησης βιολογικών ακολουθιών.

Προκειμένου να αναπτυχθεί ένας αποτελεσματικός αλγόριθμος για την πρόβλεψη και την ταξινόμηση των ακολουθιών, είναι απαραίτητη η κατασκευή ενός όσο το δυνατό γίνεται μεγάλου και περιεκτικού συνόλου δεδομένων από ακολουθίες, με το οποίο θα γίνει η εκπαίδευση και η εκτέλεση του αντίστοιχου αλγορίθμου. Για τη δημιουργία του συνόλου δεδομένων γίνεται ανάκτηση των ακολουθιών από τις βάσεις δεδομένων με τη χρήση περιορισμών που ορίζονται από τη μέθοδο που θα επιλέξει ο χρήστης.

ΜΕΡΟΣ 1ο
ΒΙΟΛΟΓΙΚΕΣ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ

1. Εισαγωγή

Η μεγάλη αύξηση των βιολογικών δεδομένων που παρουσιάστηκε στη μοριακή βιολογία εξαιτίας της ραγδαίας τεχνολογικής ανάπτυξης, οδήγησε στη δημιουργία των βιολογικών βάσεων δεδομένων. Μια βιολογική βάση δεδομένων αποτελεί ένα οργανωμένο σύστημα δεδομένων, που επιτρέπει την αποδοτική εκμετάλλευση, επεξεργασία, οργάνωση και αποθήκευση των δεδομένων που εξάγονται από την ανάλυση βιολογικών ακολουθιών, την αναγνώριση γονιδίων, τον καθορισμό της δομής των πρωτεϊνών, τον προσδιορισμό του μηχανισμού διπλώματος των πρωτεϊνών, την κατανόηση γενετικών ασθενειών κ.α.

Η πρόσβαση στις βάσεις αυτές είναι εύκολη μέσω της χρήσης του Διαδικτύου. Ο χρήστης μπορεί να επισκεφτεί την ιστοσελίδα που είναι δημοσιευμένη η βάση και να κάνει αναζητήσεις αντλώντας τα δεδομένα που χρειάζεται.

2. Βάσεις δεδομένων νουκλεοτιδικών ακολουθιών

Οι βάσεις δεδομένων νουκλεοτιδικών αλληλουχιών αποτελούν τις μεγαλύτερες βιολογικές βάσεις δεδομένων, τόσο από άποψη του όγκου πληροφορίας που περιέχουν, όσο κι από την άποψη του εκθετικού ρυθμού συσσώρευσης δεδομένων που εμφανίζουν.

Οι μεγαλύτερες βάσεις δεδομένων νουκλεοτιδικών αλληλουχιών είναι οι GENBANK (NCBI), DNA Data Bank of Japan (DDBJ) και EMBL Nucleotide Sequence Database (EBI) και σε συνεργασία έχουν δημιουργήσει την International Nucleotide Sequence Database Collaboration (INSDC). Η συνεργασία μεταξύ των βάσεων περιλαμβάνει την ανταλλαγή εγγραφών που κατατίθενται ανεξάρτητα σε κάθε βάση δεδομένων δίνοντας την δυνατότητα ταξινόμησης και σχολιασμού των δεδομένων [1].

2.1. GenBank

Η GENBANK είναι μια ολοκληρωμένη βάση δεδομένων που περιέχει νουκλεοτιδικές αλληλουχίες για σχεδόν 260 000 είδη και παρέχεται στην

επιστημονική κοινότητα χωρίς να τίθεται κανένας περιορισμός όσον αφορά τη χρήση ή τη διανομή των δεδομένων της [2]. Βρίσκεται υπό την αιγίδα του Εθνικού Ινστιτούτου Υγείας των Η.Π.Α και η κύρια πηγή πληροφορίας προέρχεται από απευθείας υποβολές δεδομένων, όπως προκύπτουν από πειραματικές διεργασίες διαφόρων ερευνητικών ομάδων. Τα δεδομένα επεξεργάζονται και σχολιάζονται (annotation) για τη διευκόλυνση των ερευνητών. Ανά τακτά χρονικά διαστήματα τα ήδη κατατεθειμένα δεδομένα επανεξετάζονται και γίνονται διορθώσεις αν προκύπτουν νέα δεδομένα σχετικά με τις εγγραφές τους. Η διαδικασία κατάθεσης των δεδομένων μπορεί να πραγματοποιηθεί πολύ γρήγορα μέσω του Διαδικτύου με τη συμπλήρωση κατάλληλης φόρμας και στη συνέχεια οι υπεύθυνοι της βάσης αναλαμβάνουν τον σχολιασμό της εγγραφής και τη δημοσιοποίησή της στη βάση.

2.2. EMBL

Η EMBL Nucleotide Sequence Database αποτελεί τη μεγαλύτερη βάση νουκλεοτιδικών αλληλουχιών στην Ευρώπη και βρίσκεται υπό την αιγίδα του Ευρωπαϊκού Εργαστηρίου Μοριακής Βιολογίας (EMBL). Έχει έδρα το Ευρωπαϊκό Ινστιτούτο Βιοπληροφορικής (EBI) στο Cambridge, UK. Τα δεδομένα προέρχονται από ανεξάρτητα ερευνητικά εργαστήρια, καθώς και από ομάδες που ασχολούνται με τον προσδιορισμό των γονιδιωμάτων διαφόρων οργανισμών. Η κατάθεση ακολουθιών στην EMBL-Bank είναι μια διαδικασία απλή και πραγματοποιείται μέσω του Διαδικτύου κατ' αντίστοιχο τρόπο με αυτό της GENBANK. Στη συνέχεια οι νέες ακολουθίες επεξεργάζονται και σχολιάζονται από τους υπευθύνους της βάσης προτού γίνουν διαθέσιμες στην επιστημονική κοινότητα. Επιπλέον μέσω του Διαδικτύου παρέχονται εργαλεία ανάλυσης ακολουθιών όπως το Fasta και το BLAST [3].

2.3. DDBJ

Η DNA Databank of Japan (DDBJ) ιδρύθηκε το 1986 στο Εθνικό Ινστιτούτο Γενετικής (NIG). Βρίσκεται υπό την αιγίδα του Υπουργείου Παιδείας, Επιστημών και Αθλητισμού της Ιαπωνίας και αποτελεί τη μοναδική διεθνώς αναγνωρισμένη βάση νουκλεοτιδικών αλληλουχιών στην Ιαπωνία, ενώ η κύρια πηγή δεδομένων της είναι οι εργασίες Ιαπώνων ερευνητών. Επιπλέον, στην DDBJ είναι διαθέσιμα διάφορα εργαλεία ανάλυσης νουκλεοτιδικών αλληλουχιών [4].

3. Βάσεις δεδομένων πρωτεϊνικών ακολουθιών

Οι βάσεις δεδομένων πρωτεϊνικών ακολουθιών, αποτελούν το δεύτερο μεγαλύτερο σε όγκο τμήμα του συνόλου των βιολογικών βάσεων δεδομένων, αλλά ίσως το σημαντικότερο, καθώς οι πρωτεϊνικές ακολουθίες παρουσιάζουν μεγάλη ποικιλομορφία, τόσο στη δομή, όσο και στη λειτουργία τους. Κατά συνέπεια, μεγάλο μέρος της βιοπληροφορικής ανάλυσης αναφέρεται σε πρωτεϊνικές ακολουθίες και υπάρχει τεράστιος όγκος λειτουργικών δεδομένων που παράγονται συνεχώς πειραματικά και τα οποία αποτελούν μέρος της πληροφορίας που περιέχεται σε αυτές τις βάσεις.

3.1. UniProtKB

Η UniProtKB (UniProt Knowledgebase) αποτελεί μια ενιαία παγκόσμια βάση δεδομένων πρωτεϊνικών ακολουθιών. Αποτελείται από δύο τμήματα, το UniProtKB/Swiss-Prot και το UniProtKB/TrEMBL. Η UniProtKB/Swiss-Prot περιέχει σχολιασμένα αρχεία υψηλής ποιότητας με πληροφορίες πρωτεϊνικών ακολουθιών που προέρχονται από αξιολογούμενη υπολογιστική ανάλυση και βιβλιογραφία. Η UniProtKB/TrEMBL περιέχει υπολογιστικά αναλυόμενα δεδομένα πρωτεϊνικών ακολουθιών που προέκυψαν από αυτόματη μετάφραση γονιδιωματικών αλληλουχιών [5].

3.2. Protein Information Resource

Η Protein Information Resource (PIR) αποτελεί ένα ολοκληρωμένο δημόσιο σύστημα παροχής πληροφοριών για πρωτεϊνικές ακολουθίες. Έχε έδρα το Πανεπιστήμιο του Georgetown και αποτελεί τμήμα του Εθνικού Ιδρύματος Βιοϊατρικής Έρευνας (NBRF) των Η.Π.Α. Η κυριότερη βάση που περιλαμβάνει η PIR είναι η PIR-International Protein Sequence Database (PSD) και τα δεδομένα της προκύπτουν από τη συνεργασία της PIR με το Munich Information Center for Protein Sequences (MIPS) και την Japanese International Protein Information Database (JIPID) [6].

Το 2002 η PIR, μαζί με το EBI (European Bioinformatics Institute) και το SIB ((Swiss Institute of Bioinformatics), δημιούργησαν το UniProt consortium, μια ενιαία παγκόσμια βάση δεδομένων με ακολουθίες πρωτεϊνών, που προέκυψε από την ενοποίηση των βάσεων δεδομένων PIR- PSD, Swiss-Prot και TrEMBL.

Σήμερα, η PIR διατηρεί προσωπικό σε UD και GUMC και συνεχίζει να προσφέρει πόρους για να βοηθήσει στην πρωτεομική και γονιδιωματική ανάλυση καθώς και στο σχολιασμό των πρωτεϊνών.

3.3. PROSITE

Η PROSITE αποτελεί μια βάση ταξινόμησης πρωτεϊνικών ακολουθιών και αυτοτελών περιοχών ακολουθιών (sequence domains) [7]. Βασίζεται στη γενικότερη παρατήρηση ότι ενώ υπάρχει ένας τεράστιος αριθμός διαφορετικών πρωτεϊνών στη φύση, αυτές μπορούν να ομαδοποιηθούν με βάση την ομοιότητα στην ακολουθία τους. Οι πρωτεΐνες ή οι αυτοτελείς δομικές περιοχές που ανήκουν στην ίδια οικογένεια έχουν την ίδια λειτουργία και προέρχονται από κοινό πρόγονο. Είναι φανερό ότι πρωτεΐνες που ανήκουν στην ίδια οικογένεια, έχουν τμήματα της ακολουθίας τους που είναι περισσότερο συντηρημένα στην πορεία της εξέλιξής τους. Αυτές οι περιοχές σχετίζονται άμεσα με τη λειτουργία τους και με τη δομή των πρωτεϊνών στον χώρο. Αναλύοντας τις ακολουθίες πρωτεϊνών που ανήκουν στην ίδια οικογένεια είναι δυνατό να προκύψει ένα 'αποτύπωμα' χαρακτηριστικό για κάθε ομάδα, ικανό ώστε να τη διαχωρίζει από τις άλλες πρωτεϊνικές αλληλουχίες που δεν ανήκουν στην οικογένεια αυτή. Η χρήση ενός αποτυπώματος μπορεί να χρησιμεύσει για να ταξινομηθεί μια άγνωστη πρωτεϊνική ακολουθία σε μια γνωστή οικογένεια πρωτεϊνών δίνοντας μας ενδείξεις για την πιθανή λειτουργία της. Αυτή τη στιγμή η PROSITE περιέχει 'αποτυπώματα' για περισσότερες από χίλιες οικογένειες. Για κάθε οικογένεια υπάρχει λεπτομερής ανάλυση για τη δομή και τη λειτουργία των πρωτεϊνών αυτών [8].

4. Βάσεις δεδομένων τρισδιάστατων δομών

Οι βάσεις δεδομένων τρισδιάστατων δομών περιέχουν δεδομένα που σχετίζονται με την τρισδιάστατη δομή βιολογικών μακρομορίων. Οι τρισδιάστατες δομές αποτελούν το τελικό στάδιο μιας διαδικασίας η οποία μετά τη χρήση μοριακών τεχνικών, οδηγεί τελικά στην υπολογιστική επίλυση της δομής μέσω της διαδικασίας της κρυσταλλογραφίας ακτίνων Χ, ή μέσω της φασματογραφίας NMR.

Μεγαλύτερο ενδιαφέρον έχουν οι δομές πρωτεϊνών, καθώς η μεγάλη ποικιλομορφία της δομής τους συνδέεται άμεσα με τη βιολογική δράση.

4.1. Protein Data Bank

Η Protein Data Bank (PDB) αποτελεί παγκοσμίως τη μοναδική βάση στην οποία περιέχονται τρισδιάστατες δομές βιολογικών μακρομορίων, συμπεριλαμβανομένων των πρωτεϊνών και των νουκλεϊνικών οξέων. Η βάση ανανεώνεται κάθε εβδομάδα και είναι ελεύθερα διαθέσιμη.

Ιδρύθηκε το 1971 στα εργαστήρια Brookhaven National Laboratories (BNL) των ΗΠΑ και περιελάμβανε 7 δομές μακρομορίων όπως αυτές προέκυψαν από κρυσταλλογραφικές μελέτες. Από το 1980 και μετά λόγω της τεχνολογικής εξέλιξης σε κάθε στάδιο του προσδιορισμού δομών ο ρυθμός προσθήκης δεδομένων στην PDB αυξήθηκε δραματικά. Οι εγγραφές στην PDB εκτός από τις συντεταγμένες των ατόμων που απαρτίζουν τη δομή περιλαμβάνουν και επιπρόσθετα βοηθητικά στοιχεία όπως βιβλιογραφικές αναφορές, λεπτομέρειες για τον προσδιορισμό της δομής καθώς και άλλα στοιχεία που προκύπτουν από τη συγκεκριμένη δομή. Κάθε δομή προτού διατεθεί στο κοινό υφίσταται έλεγχο για την ορθότητα της με τη χρήση ειδικού λογισμικού [9].

5. Βάσεις δεδομένων γονιδιακής έκφρασης

Εξαιτίας της εξέλιξης της τεχνολογίας τα πειράματα ανάλυσης γονιδιακής έκφρασης πραγματοποιούνται με μεγαλύτερο ρυθμό, με αποτέλεσμα να υπάρχει ανάγκη αποθήκευσης και ανάλυσης του όγκου δεδομένων που παρουσιάζεται από τα χιλιάδες πειράματα μικροσυστοιχιών. Η καταχώρηση των αποτελεσμάτων αυτών, καθώς και πληροφορίες σχετικά με το είδος των δεδομένων, τα γονίδια τα οποία μελετώνται και πληροφορίες σχετικά με τα είδη των δειγμάτων τα οποία χρησιμοποιήθηκαν γίνεται στις βάσεις δεδομένων γονιδιακής έκφρασης.

Τα δεδομένα που καταχωρούνται στις βάσεις αυτές έχουν σαν βασική δομή την μορφή πίνακα. Στον πίνακα αναγράφονται οι τιμές "έκφρασης" ενός γονιδίου για κάθε άτομο.

Η πολυπλοκότητα και ο όγκος των δεδομένων γονιδιακής έκφρασης καθιστούν την καταχώρηση των δεδομένων των μικροσυστοιχιών στις δημόσιες βάσεις πολύπλοκη, καθώς είναι απαραίτητο να ακολουθείται ένα συγκεκριμένο πρωτόκολλο με βάση το οποίο να καταχωρείται η ελάχιστη πληροφορία που περιγράφει ένα πείραμα μικροσυστοιχιών.

5.1. GeneExpression Omnibus

Η GeneExpression Omnibus (GEO) αποτελεί δημόσια βάση δεδομένων του NCBI που περιέχει δεδομένα μικροσυστοιχιών, αλληλούχισης (next generation sequenicng) και άλλες μορφές δεδομένων γονιδιακής έκφρασης που υποβλήθηκαν από την επιστημονική κοινότητα.

Η GEO παρέχει ένα ευέλικτο και ανοικτό σχεδιασμό που διευκολύνει την υποβολή, την αποθήκευση και την ανάκτηση των ετερογενών συνόλων δεδομένων γονιδιακής έκφρασης. Ο χρήστης έχει πρόσβαση σε εργαλεία που επιτρέπουν την ανάλυση των δεδομένων της βάσης, που είναι διαθέσιμη μέσω διαδικτύου [10].

5.2. Array Express

Η Array Express είναι δημόσια βάση δεδομένων που περιέχει δεδομένα γονιδιακής έκφρασης. Βρίσκεται υπό την αιγίδα του Ευρωπαϊκού Ινστιτούτου Βιοπληροφορικής (EBI) και είναι διαθέσιμη μέσω διαδικτύου, προσφέροντας τη δυνατότητα στην επιστημονική κοινότητα να ανακτήσει και να επεξεργαστεί τα δεδομένα κάνοντας χρήση των εργαλείων που παρέχονται.

Απαρτίζεται από δυο τμήματα, το ArrayExpress Repository που περιέχει δεδομένα μικροσυστοιχιών και το ArrayExpress Data Warehouse που περιέχει προφίλ (profile) γονιδιακής έκφρασης [11].

5.3. Stanford Microarray Database

Η Stanford Microarray Database (SMD) αποτελεί μια βάση δεδομένων που περιέχει δεδομένα που προκύπτουν από πειράματα μικροσυστοιχιών. Στην αρχή κατασκευάστηκε για να καλύπτει τις ανάγκες διαμοιρασμού αρχείων των ερευνητών του Stanford, αλλά στην πορεία εξελίχθηκε σε μια δημόσια βάση δεδομένων για μικροσυστοιχίες.

Μέσω των διεπαφών που παρέχει από τον διαδικτυακό τόπο που διατηρεί επιτρέπει στους ερευνητές να έχουν πρόσβαση στα δεδομένα της βάσης, παρέχοντας τους εργαλεία ανάκτησης και ανάλυσης δεδομένων [12].

6. Βάσεις δεδομένων γενετικής ποικιλομορφίας

Υπάρχουν παραλλαγές στην ακολουθία σε καθορισμένες θέσεις εντός των γονιδιωμάτων που είναι υπεύθυνες για τα ατομικά φαινοτυπικά χαρακτηριστικά, συμπεριλαμβανομένης και της τάσης ενός ατόμου προς πολύπλοκες διαταραχές, όπως η καρδιοπάθεια και ο καρκίνος [13]. Με τον εντοπισμό και την ταξινόμηση των γενετικών ομοιοτήτων και παραλλαγών στην αλληλουχία μπορεί να γίνει η χαρτογράφηση του γονιδίου και να οριστεί η δομή του πληθυσμού, δίνοντας στους ερευνητές τη δυνατότητα να εντοπίσουν γονίδια που επηρεάζουν την υγεία, την ασθένεια και μεμονωμένες απαντήσεις σε φάρμακα και περιβαλλοντικούς παράγοντες.

Στις βάσεις γενετικής ποικιλομορφίας υπάρχουν καταχωρημένοι οι πολυμορφισμοί και οι συχνότητες τους στους διάφορους πληθυσμούς καθώς και οι αλληλοσυσχετίσεις των πολυμορφισμών αυτών.

6.1. dbSNP

Η dbSNP έχει σχεδιαστεί για να υποστηρίξει την έρευνα σε ένα ευρύ φάσμα βιολογικών προβλημάτων. Σε αυτά περιλαμβάνεται η χαρτογράφηση, η λειτουργική ανάλυση, η φαρμακογονιδιωματική (pharmacogenomics), η πληθυσμιακή γενετική και η εξελικτική βιολογία.

Αποτελεί δημόσια βάση δεδομένων με ευρεία συλλογή από καταχωρήσεις γενετικών πολυμορφισμών και είναι αποτέλεσμα της συνεργασία του National Center for Biotechnology Information (NCBI) και του National Human Genome Research Institute (NHGRI). Στις καταχωρήσεις αυτές περιλαμβάνονται νουκλεοτιδικοί πολυμορφισμοί (single nucleotide polymorphisms), δεδομένα για πολυμορφικές θέσεις που αφορούν απαλοιφές ή εισαγωγές βάσεων (deletion insertion polymorphisms) και μικροδορυφορικές επαναλήψεις (short tandem repeats). Κάθε καταχώρηση περιέχει την αλληλουχία με την τοποθεσία του πολυμορφισμού, τη συχνότητα εμφάνισής του σε διάφορους πληθυσμούς ή άτομα, την πειραματική μεθοδολογία, τα πρωτόκολλα και τις συνθήκες κάτω από τις οποίες προσδιορίστηκε ο πολυμορφισμός.

Επιτρέπει στην ερευνητική κοινότητα την υποβολή για καταχωρήσεις πολυμορφισμών από κάθε είδος, αλλά και από διαφορετικά σημεία του γονιδιώματος. Μετά από κάθε νέα καταχώρηση γίνεται και ενημέρωση της βάσης [13].

6.2. HapMap

Το International HapMap Project αποτελεί μία διεθνή συνεργασία μεταξύ επιστημόνων και φορέων από τον Καναδά, την Κίνα, την Ιαπωνία, τη Νιγηρία, το Ηνωμένο Βασίλειο και τις Ηνωμένες Πολιτείες που έχει σκοπό τον εντοπισμό και την ταξινόμηση των γενετικών ομοιοτήτων και διαφορών στον άνθρωπο [14]. Ο στόχος του HapMap είναι η χαρτογράφηση και η κατανόηση των προτύπων της κοινής γενετικής ποικιλομορφίας του ανθρώπινου γονιδιώματος, προκειμένου να επιταχυνθεί η έρευνα για τις ασθένειες που προκαλούνται στον άνθρωπο από γενετικές αιτίες.

Η πρόσβαση στα δεδομένα της βάσης γίνεται μέσω ενός γραφικού προγράμματος περιήγησης που δίνει τη δυνατότητα στο ερευνητικό κοινό να κάνει αναζήτηση για ένα γονίδιο ή μια περιοχή ενδιαφέροντος και στη συνέχεια να απεικονίσει το νουκλεοτιδικό πολυμορφισμό. Επίσης παρέχει εργαλεία που διευκολύνουν την λήψη των δεδομένων και επιτρέπει στους χρήστες να κατεβάσουν τα δεδομένα σε μια μορφή που είναι κατάλληλη για ανάλυση [15].

7. Δευτερογενείς βάσεις δεδομένων

7.1. Βάσεις δεδομένων οικογενειών πρωτεϊνικών ακολουθιών

Οι πρωτεΐνες αποτελούνται από μια ή περισσότερες πρωτεϊνικές περιοχές (domains). Μια πρωτεϊνική περιοχή είναι η μικρότερη μονάδα που διαθέτει χαρακτηριστικά που συνδέονται με ολόκληρη την πρωτεΐνη. Είναι συμπαγής, έχει έναν υδρόφοβο πυρήνα και μερικές φορές λειτουργεί ανεξάρτητα από την υπόλοιπη δομή [16]. Η μεγάλη ποικιλία των πρωτεϊνών στη φύση οφείλεται συνήθως στο συνδυασμό πρωτεϊνικών περιοχών που αλληλεπιδρούν μεταξύ τους.

Οι βάσεις δεδομένων δομικών περιοχών ή οικογενειών πρωτεϊνών είναι χρήσιμες για την ανάλυση των πρωτεϊνών και κυρίως για τον χαρακτηρισμό των λειτουργιών τους. Οι βάσεις αυτές συνήθως αποκαλούνται «βάσεις δεδομένων πρωτεϊνικών υπογραφών» (protein signature database), καθώς περιέχουν συλλογές υπογραφών, που επιτρέπουν την αναγνώριση των δομικών περιοχών. Τα δεδομένα που είναι καταχωρημένα στις βάσεις αυτές έχουν προκύψει από την επεξεργασία πολλαπλών στοιχίσεων ενός συνόλου ομόλογων πρωτεϊνικών ακολουθιών [17].

Από τις πολλαπλές στοιχίσεις χωρίς κενά προκύπτουν τα μοτίβα που αντιστοιχούν στα συντηρημένα τμήματα των πρωτεϊνών. Η ομάδα από μοτίβα που παρατηρούνται σε μια οικογένεια πρωτεϊνών ονομάζεται αποτύπωμα (fingerprint). Όσο περισσότερα μοτίβα υπάρχουν σε ένα αποτύπωμα, τόσο καλύτερος είναι ο προσδιορισμός της

συσχέτισης. Αντιθέτως, λιγότερα μοτίβα, οδηγούν σε φτωχότερη διαγνωστική απόδοση [18]. Υπάρχουν δύο τρόποι για τη δημιουργία των αποτυπωμάτων. Ο ένας βασίζεται στη χρήση κανονικών εκφράσεων (regular expressions) που αντιπροσωπεύουν μια αλληλουχία που έχει προκύψει από ένα συντηρημένο τμήμα μιας πρωτεΐνης [19], ενώ ο άλλος βασίζεται στην κατασκευή πινάκων με ειδικές ανά θέση πιθανότητες εμφάνισης αμινοξέων (profiles), που αντιστοιχούν σε όλο το μήκος μιας πολλαπλής στοίχισης πρωτεϊνών, περιλαμβανομένων και των κενών [20].

Οι βάσεις αυτές διαθέτουν κατάλληλα υπολογιστικά εργαλεία για την ταυτοποίηση των δομικών περιοχών στην υπό εξέταση ακολουθία και για τον προσδιορισμό της οικογένειας πρωτεϊνών που ανήκει η ακολουθία.

7.1.1. PROSITE

Η PROSITE είναι μια βάση δεδομένων πρωτεϊνικών οικογενειών και πρωτεϊνικών περιοχών. Βασίζεται στην παρατήρηση ότι ενώ υπάρχει ένας τεράστιος αριθμός διαφορετικών πρωτεϊνών, οι περισσότερες μπορούν να ομαδοποιηθούν βάση των ομοιοτήτων στις αλληλουχίες τους, σε έναν περιορισμένο αριθμό οικογενειών. Οι πρωτεΐνες ή οι πρωτεϊνικές περιοχές που ανήκουν σε μια συγκεκριμένη οικογένεια, μοιράζονται λειτουργικά γνωρίσματα και προέρχονται από έναν κοινό πρόγονο.

Με την έρευνα των ακολουθιών των πρωτεϊνικών οικογενειών είναι εμφανές ότι κατά τη διάρκεια της εξέλιξης ορισμένες περιοχές είναι καλύτερα συντηρημένες από κάποιες άλλες. Αυτές οι περιοχές είναι γενικά σημαντικές για τη λειτουργία μιας πρωτεΐνης ή και για τη διατήρηση της τρισδιάστατης δομής της. Αναλύοντας τις σταθερές και μεταβλητές ιδιότητες τέτοιων ομάδων παρόμοιων ακολουθιών, είναι δυνατόν να παραχθεί μια υπογραφή για μια οικογένεια πρωτεϊνών ή μια πρωτεϊνική περιοχή, που διακρίνει τα μέλη της από όλες τις άλλες πρωτεΐνες. Η υπογραφή μπορεί να χρησιμοποιηθεί για να εκχωρηθεί μια πρωτεϊνική ακολουθία σε μία συγκεκριμένη οικογένεια πρωτεϊνών και έτσι να διατυπωθούν υποθέσεις σχετικά με τη λειτουργία της [8].

Κάθε καταχώρηση που υπάρχει στη βάση συνδέεται με σχολιασμό όπου ο χρήστης μπορεί να βρει πληροφορίες για την οικογένεια πρωτεϊνών, την πρωτεϊνική περιοχή ή τη λειτουργία που προσδιορίζεται από την υπογραφή, την προέλευση του ονόματός της, την ταξινομική εμφάνιση, την αρχιτεκτονική της πρωτεϊνικής

περιοχής, την τρισδιάστατη δομή της, τα κύρια χαρακτηριστικά της ακολουθίας, το μέγεθος της πρωτεϊνικής περιοχής και κάποιες αναφορές [21].

7.1.2. PFAM

Η Pfam αποτελεί μια συλλογή πρωτεϊνικών οικογενειών. Κάθε καταχώρηση ορίζεται από δυο πολλαπλές στοιχίσεις και ένα προφίλ Hidden Markov Model (HMM). Οι πρωτεϊνικές περιοχές που έχουν score πάνω από το όριο που έχει οριστεί για κάθε οικογένεια για την εξάλειψη των ψευδώς θετικών, στοιχίζεται με το HMM για να παραχθεί η πλήρης στοίχιση. Τα προφίλ HMM κατασκευάζονται χρησιμοποιώντας το λογισμικό HMMER Η PFAM αποτελείται από δύο υποσύνολα, την PFAM-A, και την PFAM-B.

Μερικές φορές, ένα προφίλ HMM δεν μπορεί να ανιχνεύσει όλα τα ομόλογα μιας διαφορετικής υπεροικογένειας, έτσι πολλαπλές καταχωρήσεις μπορεί να κατασκευαστούν για να αντιπροσωπεύσουν διαφορετικές οικογένειες ακολουθιών στην υπεροικογένεια. Αυτές οι καταχωρήσεις PFAM-A ομαδοποιούνται σε «φυλές» (clans) και είναι υψηλής «ποιότητας» δεδομένα, καθώς έχουν όλες υποστεί σχολιασμό από ειδικούς, ενώ υπάρχουν αναφορές σε άλλες βάσεις δεδομένων και σε βιβλιογραφία.

Η PFAM-B αποτελείται από αυτόματες καταχωρήσεις που προκύπτουν με τον εντοπισμό ομοιοτήτων ανάμεσα στις πρωτεϊνικές περιοχές που απομένουν όταν αφαιρεθούν οι περιοχές που αντιστοιχούν στις καταχωρήσεις της PFAM-A. Είναι αρκετά σημαντική, γιατί με στοχευμένη ανάλυση, μπορούν να προκύψουν οικογένειες που μετέπειτα θα καταχωρηθούν στην PFAM-A [22].

7.1.3. CATH

Η CATH είναι μια βάση ιεραρχικής ταξινόμησης πρωτεϊνικών δομών που είναι κατατεθειμένες στην PDB με βάση τις αυτοτελείς δομικές περιοχές (domains) που τις απαρτίζουν. Για τον καταρτισμό της CATH δεν λαμβάνονται υπόψη μη πρωτεϊνικές δομές, ενώ οι πρωτεϊνικές δομές που περιέχονται πρέπει να είναι προσδιορισμένες σε διακριτικότητα υψηλότερη των 3 Angstroms. Η CATH χρησιμοποιεί κυρίως αυτοματοποιημένες μεθόδους για την ταξινόμηση, αν και σε ειδικές περιπτώσεις τα ανθρώπινα κριτήρια είναι δυνατόν να δώσουν καλύτερα αποτελέσματα από τις αυτοματοποιημένες μεθόδους, όποτε και προτιμούνται.

Τα 4 κύρια επίπεδα της ιεραρχίας είναι η Τάξη (Class), η Αρχιτεκτονική (Architecture), η Τοπολογία (Topology fold family) και η Ομόλογη Οικογένεια (Homologous superfamily).

Οι πρωτεΐνες που αποτελούνται από περισσότερα του ενός domains αναλύονται στα επιμέρους στοιχεία αυτόματα, με βάση ειδικούς αλγόριθμους αναγνώρισης domains. Με την αυτόματη διαδικασία κατατάσσεται το 53% των δομών. Οι υπόλοιπες διαχωρίζονται στα επιμέρους domains με παρατηρήσεις που προκύπτουν είτε από τους αλγόριθμους αυτόματου διαχωρισμού, είτε από τη βιβλιογραφία. Η ταξινόμηση πραγματοποιείται μόνο στις αυτοτελείς δομικές περιοχές.

Ιεραρχία στην CATH:

- C: Τάξη (Class): Η κατάταξη σε τάξεις πραγματοποιείται λαμβάνοντας υπόψιν τα στοιχεία δευτεροταγούς δομής μιας αυτοτελούς δομικής περιοχής.
- A: Αρχιτεκτονική (Architecture): Η ταξινόμηση πραγματοποιείται με βάση τη γενικότερη δομή της αυτοτελούς δομικής περιοχής (domain), με βάση τον προσανατολισμό των στοιχείων δευτεροταγούς δομής χωρίς να λαμβάνει υπόψιν όμως τον τρόπο διασύνδεσής τους.
- T: Τοπολογία (Topology): Σε αυτό το επίπεδο οι δομές ταξινομούνται με βάση τον προσανατολισμό των στοιχείων δευτεροταγούς δομής, αλλά και με βάση τη σύνδεση αυτών των στοιχείων μεταξύ τους.
- H: Ομόλογη οικογένεια (Homology superfamily): Χαρακτηρίζεται από ομαδοποίηση των δομικών στοιχείων που εμφανίζουν 35% ομοιότητα μεταξύ τους στο επίπεδο της αλληλουχίας τους με αποτέλεσμα να θεωρείται ότι προέρχονται από ένα κοινό πρόγονο [23].

7.1.4. SCOP

Η βάση SCOP έχει σαν βασικό στόχο την ανάλυση των δομικών και εξελικτικών σχέσεων μεταξύ όλων των πρωτεϊνών γνωστής δομής που είναι κατατεθειμένες στην PDB. Για την αναγνώριση των σχέσεων αυτών και την ταξινόμηση των πρωτεϊνών η διαδικασία δεν είναι αυτοματοποιημένη αλλά πραγματοποιείται αποκλειστικά με βάση τον ανθρώπινο παράγοντα μετά από λεπτομερή μελέτη και σύγκριση των πρωτεϊνικών δομών. Αυτοματοποιημένες μέθοδοι χρησιμοποιούνται μόνο για την ομοιογένεια των δεδομένων που περιέχονται στη βάση.

Επίπεδα ταξινόμησης:

- Οικογένεια (Family): Ξεκάθαρη εξελικτική σχέση μεταξύ των μελών.
  Οι πρωτεΐνες που ταξινομούνται σε μια οικογένεια έχουν ξεκάθαρη εξελικτική σχέση μεταξύ τους. Η ομοιότητα σε επίπεδο ακολουθίας είναι της τάξης του 30% και άνω. Υπάρχουν όμως περιπτώσεις όπου οι δομές και η λειτουργία είναι παρόμοιες υποδηλώνοντας κοινό πρόγονο, ενώ η ομοιότητα σε επίπεδο ακολουθίας να είναι μικρότερη του 30% (σφαιρίνες, 15%).
- Υπεροικογένεια (Superfamily): Τα μέλη της έχουν πιθανά προέλθει από κοινό πρόγονο.
  Στο επίπεδο της υπεροικογένειας κατατάσσονται πρωτεΐνες που εμφανίζουν πολύ μικρή ομοιότητα στο επίπεδο της ακολουθίας, αλλά τα δομικά τους χαρακτηριστικά και η λειτουργία τους υποδηλώνουν πιθανή κοινή προέλευση.
- Δίπλωμα (Fold): Εμφάνιση ομοιότητας σε επίπεδο δομής.
  Οι πρωτεΐνες που εμφανίζουν το ίδιο δίπλωμα έχουν τα ίδια σε μεγάλο βαθμό χαρακτηριστικά δευτεροταγούς δομής, με κοινό προσανατολισμό και τις ίδιες τοπολογικές συνδέσεις μεταξύ τους. Πρωτεΐνες που έχουν το ίδιο δίπλωμα, αλλά δεν είναι όμοιες από άποψη αμινοξικής ακολουθίας, έχουν ορισμένα περιφερειακά στοιχεία της δευτεροταγούς τους δομής διαφορετικά και διαφορετικές στροφές, όσον αφορά το μέγεθος και τη διαμόρφωση. Πρωτεΐνες που εμφανίζουν κοινό δίπλωμα δεν είναι απαραίτητο να έχουν κοινή εξελικτική προέλευση.
- Τάξη (Class): Τέσσερις κύριες δομικές κατηγορίες πρωτεϊνών έχουν ταυτοποιηθεί με βάση το δίπλωμα των στοιχείων δευτεροταγούς δομής τους, οι all-α (η δομή σχηματίζεται από α-έλικες), all-β (η δομή αποτελείται από β-πτυχωτές επιφάνειες), α/β (α-έλικες και β-πτυχωτές επιφάνειες εναλλάσσονται στην δομή της πρωτεΐνης) και α+β (α-έλικες και β-πτυχωτές επιφάνειες βρίσκονται σε διακριτές περιοχές της δομής) [24].

8. Ολοκληρωμένα συστήματα ανάκτησης πληροφοριών

8.1. Entrez

Το Entrez αποτελεί ένα ολοκληρωμένο σύστημα ανάκτησης δεδομένων που παρέχει πρόσβαση σε όλες τις βάσεις δεδομένων που περιέχονται στο NCBI. Το Entrez υποστηρίζει την αναζήτηση κειμένου χρησιμοποιώντας απλά ερωτήματα, τη

λήψη δεδομένων σε διάφορες μορφές και τη διασύνδεση των αρχείων μεταξύ των βάσεων δεδομένων [25]. Δίνει τη δυνατότητα αναζήτησης σε βάσεις δεδομένων νουκλεοτιδικών και πρωτεϊνικών ακολουθιών, δομές βιομορίων, γονιδιωμάτων και στη βάση βιβλιογραφίας MEDLINE μέσω του ίδιου γραφικού περιβάλλοντος [26]. Μειονέκτημα αποτελεί το γεγονός ότι περιορίζεται μόνο στις βάσεις δεδομένων του NCBI και ότι δεν επιτρέπει πολύπλοκες αναζητήσεις.

8.2. Sequence Retrieval System (SRS)

Το SRS είναι ένα σύστημα ανάκτησης πληροφοριών που έχει σχεδιαστεί για βάσεις δεδομένων όπως η EMBL, η SwissProt ή η Prosite και διατίθεται από την εταιρία LION Bioscience [27].

Το SRS μέσω ενός γραφικού περιβάλλοντος παρέχει στον χρήστη τη δυνατότητα να αναζητήσει και ανακτήσει δεδομένα από περισσότερες από τετρακόσιες βάσεις δεδομένων, οι οποίες μπορεί να είναι αποθηκευμένες στον ίδιο κεντρικό υπολογιστή. Επίσης, ο χρήστης μπορεί να κάνει ταυτόχρονη αναζήτηση για ένα ζήτημα άμεσου ενδιαφέροντος σε παραπάνω από μια βάσεις δεδομένων που δεν περιέχουν ανάλογου είδους πληροφορία και η μορφοποίηση των δεδομένων σε καθεμιά να είναι διαφορετική. Παρά το γεγονός ότι διαχειρίζεται πραγματικά τεράστιο όγκο πληροφορίας, λόγω του μεγάλου αριθμού βάσεων που μπορεί να διαχειρίζεται ταυτόχρονα, είναι σε θέση να πραγματοποιεί τις αναζητήσεις με μεγάλη ταχύτητα. Επιπλέον, ο χρήστης του συστήματος μπορεί να ενσωματώνει σε αυτό και βάσεις που έχει δημιουργήσει ο ίδιος ή ακόμα και προγράμματα για κάθε είδος υπολογιστική ανάλυση χωρίς να επηρεάζεται η απόδοση του συστήματος [28].

ΜΕΡΟΣ 2ο
ΜΕΘΟΔΟΙ ΠΡΟΒΛΕΨΗΣ

1. Εισαγωγή

Οι μέθοδοι πρόβλεψης είναι τεχνικές που χρησιμοποιούν οι ερευνητές στην Βιοπληροφορική για να προσδιορίσουν τις βιολογικές δομές και λειτουργίες των μακρομορίων. Με τη χρήση των μεθόδων αυτών είναι δυνατό να γίνει προσδιορισμός μιας άγνωστης ακολουθίας νουκλεϊνικών οξέων ή αμινοξέων, ως προς τη δομή και λειτουργία της, με σκοπό τη δημιουργία ενός τρισδιάστατου μοντέλου της δομής της. Για να είναι εφικτός ο προσδιορισμός αυτός γίνεται χρήση των μεθόδων ομολογίας, έτσι ώστε να διασφαλιστεί η ομολογία μεταξύ της άγνωστης ακολουθίας και μιας άλλης ακολουθίας γνωστής δομής και λειτουργίας. Η ακρίβεια των μεθόδων αυτών καθορίζεται από το ποσοστό της ομολογίας. Όσο μεγαλύτερη η ομολογία, τόσο πιο μεγάλη η ακρίβεια στην πρόβλεψη της δομής και λειτουργίας της ακολουθίας. Στην περίπτωση όμως, που μία άγνωστη ακολουθία δεν έχει βρεθεί να έχει ομόλογη της με γνωστή δομή και λειτουργία, τότε δεν είναι δυνατός ο προσδιορισμός της και εξαιτίας του μεγάλου ρυθμού αύξησης των μοριακών δεδομένων, είναι αδύνατος και ο πειραματικός προσδιορισμός της, καθώς είναι δαπανηρή, επίπονη και χρονοβόρα διαδικασία.

Επομένως, προέκυψε η ανάγκη ανάπτυξης των μεθόδων πρόβλεψης, που έχουν ως σκοπό την πρόβλεψη των δομικών ή λειτουργικών χαρακτηριστικών μιας αλληλουχίας πρωτεΐνης ή μιας αλληλουχίας νουκλεϊνικών οξέων, χρησιμοποιώντας μόνο την ακολουθία. Στην παρούσα πτυχιακή εργασία θα αναφερθούμε στις μεθόδους πρόβλεψης της δευτεροταγούς δομής των πρωτεϊνών.

Οι πρωτεΐνες περιγράφονται από την ακολουθία των αμινοξέων τους. Ωστόσο, οι ειδικές λειτουργίες τους εξαρτώνται από την τρισδιάστατη δομή τους. Γνωρίζοντας την τρισδιάστατη δομή μιας πρωτεΐνης μας βοηθά να κατανοήσουμε τη λειτουργία της και μας παρέχει τα μέσα για τη διεξαγωγή πειραμάτων που οδηγούν στον σχεδιασμό φαρμάκων. Οι πειραματικές μέθοδοι με τις οποίες μπορεί να προσδιοριστεί μια πρωτεϊνική δομή είναι η κρυσταλλογραφία ακτίνων Χ και η φασματοσκοπία NMR. Όλες οι διαθέσιμες τρισδιάστατες δομές πρωτεϊνών, νουκλεϊνικών οξέων, υδατανθράκων και ποικίλων άλλων συμπλόκων, οι οποίες έχουν προσδιοριστεί πειραματικά από κρυσταλλογραφία ακτίνων Χ και φασματοσκοπία

NMR είναι καταχωρημένες στη βάση δεδομένων PDB και είναι διαθέσιμες στην επιστημονική κοινότητα μέσω του διαδικτύου. Εξαιτίας του ότι αυτές οι μέθοδοι είναι τεχνικά δύσκολες και δαπανηρές, η πρόβλεψη της δομής της πρωτεΐνης, η οποία βασίζεται κυρίως στην ακολουθία, έχει γίνει όλο και πιο σημαντική δραστηριότητα. Τα ομόλογα μοντέλα έχουν γίνει πιο ακριβή και το εύρος εφαρμογής τους έχει αυξηθεί. Η πρόοδος έχει έρθει, εν μέρει, από το πλήθος των ακολουθιών και των πληροφοριών για τη δομή, που έχει εμφανιστεί κατά τη διάρκεια των τελευταίων ετών, καθώς επίσης και από τις βελτιώσεις στα εργαλεία ανάλυσης. Επίσης σημαντική πρόοδος έγινε στην κατανόηση της φυσικής χημικής βάσης της σταθερότητας της πρωτεΐνης και στην αντίστοιχη χρήση των φυσικών χημικών δυνητικών λειτουργιών για τον προσδιορισμό του σωστού διπλώματος μιας πρωτεΐνης [29].

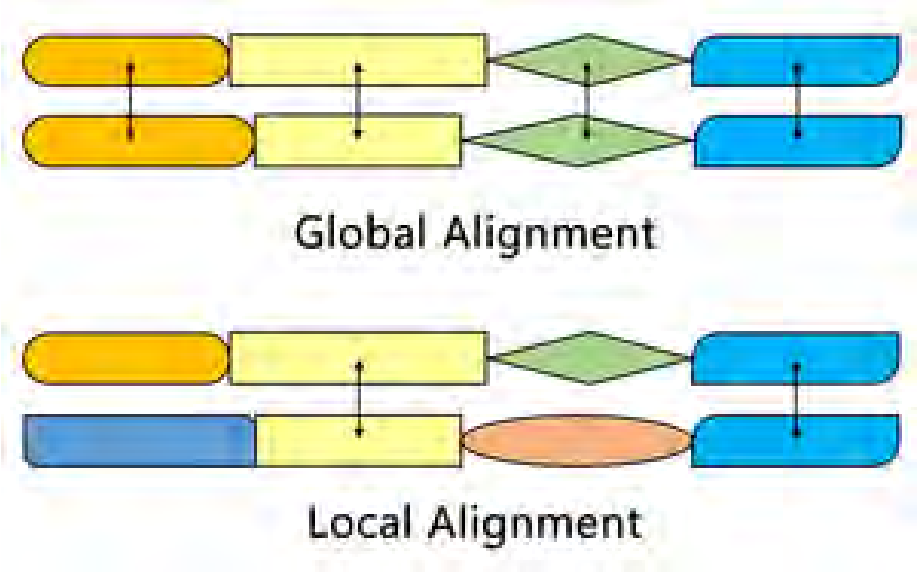
Το γεγονός ότι οι πληροφορίες σχετικά με το πώς οι αλληλουχίες αναδιπλώνονται στις δευτεροταγείς δομές μπορούν στη συνέχεια να χρησιμοποιηθούν για την πρόβλεψη της τριτοταγούς δομή με σταδιακό τρόπο (από την αλληλουχία των αμινοξέων στην δευτεροταγή δομή και από τη δευτεροταγή δομή στην τριτοταγή δομή) προωθούν την ανάπτυξη πολλών μεθόδων για την πρόβλεψη της δευτεροταγούς δομής [30]. Εκτός από την αναγνώριση του διπλώματος μια πρωτεΐνης, η πρόβλεψη της δευτεροταγούς δομής έχει επίσης ενσωματωθεί επιτυχώς σε μια σειρά από άλλα σημαντικά εργαλεία βιοπληροφορικής. Σε αυτά περιλαμβάνονται προγράμματα ανίχνευσης ομολογίας και πολλαπλής στοίχισης ακολουθιών. Σε όλες αυτές τις περιπτώσεις, το κοινό χαρακτηριστικό είναι ότι η δομή είναι περισσότερο διατηρημένη από την αλληλουχία. Αυτό ισχύει ιδιαίτερα σε πιο μακρινής συγγένειας πρωτεΐνες, όπου η εξελικτική συγγένεια μπορεί να μην είναι πια ευδιάκριτη σε επίπεδο αλληλουχίας, αλλά μπορεί ακόμα να ανιχνευθεί στη δομή [31]. Οι ακριβείς πολλαπλές στοιχίσεις των μελών των οικογενειών αυτών και η εξελικτική πληροφορία που παρουσιάζεται στην στοίχιση μπορεί να χρησιμοποιηθεί για να προσδιοριστούν ποια κατάλοιπα είναι καίριας σημασίας για το δίπλωμα και τη λειτουργία της πρωτεΐνης. Η αποτελεσματική εκμετάλλευση των εν λόγω πληροφοριών έχει οδηγήσει σε σημαντική αύξηση της ακρίβειας των μεθόδων πρόβλεψης της δευτεροταγούς δομής, έτσι ώστε σήμερα είναι συχνά δυνατό να προσδιοριστεί σωστά η πλειοψηφία των στοιχείων της δευτεροταγούς δομής σε μια πρωτεΐνη [32].

2. Στοίχιση Αλληλουχιών

Η πρόβλεψη της δευτεροταγούς δομής με τη χρήση υπολογιστικών τεχνικών βασισμένων στην αμινοξική ακολουθία εστιάζει στην τοπική πρόγνωση κατά μήκος της αλληλουχίας (Prediction) και στην ταξινόμηση των αλληλουχιών σε δύο ή περισσότερες πρωτεϊνικές οικογένειες ή κατηγορίες γενικότερα (Classification) [Εικόνα 1]. Εξαιτίας του γεγονότος ότι τα αμινοξέα έχουν πολύ μεγάλη ποικιλομορφία, μπορεί να επηρεάσουν τις φυσικές ιδιότητες των πρωτεϊνών, καθώς μπορεί να είναι υδρόφιλα ή υδρόφοβα, βασικά ή όξινα και έχουν ευθείες ή διακλαδισμένες αλυσίδες. Κατά συνέπεια κάθε αμινοξικό κατάλοιπο έχει προκαθορισμένες τάσεις να δημιουργεί δομές διαφορετικών τύπων.

Η σωστή χρήση αυτών των ιδιοτήτων, μπορεί να οδηγήσει στη δημιουργία μεθόδων πρόγνωσης διάφορων αλγορίθμων που στη συνέχεια θα δώσουν πολύτιμες πληροφορίες για τη δομή και λειτουργία μιας άγνωστης πρωτεϊνικής ακολουθίας και θα την ταξινομήσουν σε κάποια πρωτεϊνική κατηγορία.

Εικόνα 1: Τοπική και ολική στοίχιση αλληλουχιών



2.1. Τοπική πρόβλεψη της αλληλουχίας (Prediction)

Στην τοπική πρόβλεψη οι αλληλουχίες που θεωρείται ότι έχουν ομοιότητες ή ακόμα και διαφορές στην ακολουθία τους μπορούν να συγκριθούν μεταξύ τους, με σκοπό την εύρεση τοπικών περιοχών με υψηλά ποσοστά ομοιότητας. Η τοπική πρόβλεψη βασίζεται στην εύρεση παρόμοιων υπό-ακολουθιών ίδιου μήκους που βρίσκονται σε ένα σύνολο ακολουθιών. Αυτό επιτρέπει στους αλγόριθμους να

συγκρίνουν περιοχές στις ακολουθίες ξεχωριστά ανεξάρτητα από τη συνολική σειρά εντός της ακολουθίας με αποτέλεσμα να γίνεται σύγκριση μόνων των παρόμοιων περιοχών και παράλειψη των εξαιρετικά αποκλινόντων περιοχών. Οι αλγόριθμοι αυτοί ως επί το πλείστον χρησιμοποιούν δυναμικό προγραμματισμό, για τη βέλτιστη ευθυγράμμιση των ακολουθιών.

Στην υλοποίηση γίνεται χρήση ενός παραθύρου ορισμένου μήκους q που κινείται κατά μήκος της ακολουθίας και την χωρίζει σε «υπό-ακολουθίες». Αν η δοσμένη ακολουθία είναι μεγέθους n, τότε υπάρχουν n-q+1 «υπό-ακολουθίες». Στη συνέχεια χρησιμοποιώντας έναν πίνακα αντικατάστασης, που δημιουργείται παίρνοντας τα ζεύγη στοιχίσεων ομόλογων πρωτεϊνών και υπολογίζοντας τη συχνότητα αντικατάστασης όλων των χαρακτήρων, καθορίζονται οι εμφανίσεις κάθε «υπό-ακολουθίας» της δοσμένης ακολουθίας με υψηλό σκορ στο σύνολο των ακολουθιών σύγκρισης. Ανάλογα με το βιολογικό πρόβλημα επιλέγεται και το κατάλληλο σύστημα για τα σκορ. Επιπλέον, γίνεται δοκιμή διαφορετικού μήκους στο παράθυρο και διαφορετικών συνδυασμών πινάκων αντικατάστασης, με σκοπό την εύρεση του καλύτερου συνδυασμού Με βάση τον αριθμό των στοιχίσεων μπορεί να οριστεί η ομοιότητα μεταξύ των περιοχών των ακολουθιών και συνεπώς να καθοριστούν κάποιες από τις λειτουργίες που πιθανόν να έχει μια άγνωστη αλληλουχία.

2.2.Ταξινόμηση αλληλουχιών (Classification)

Οι πρωτεΐνες μπορούν να ταξινομηθούν σε ομάδες ανάλογα με την αλληλουχία ή τη δομική ομοιότητα τους. Αυτές οι ομάδες συχνά περιέχουν καλά χαρακτηρισμένες πρωτεΐνες των οποίων η λειτουργία είναι γνωστή. Έτσι, όταν γίνεται αναγνώριση μιας νέας πρωτεΐνης, οι λειτουργικές της ιδιότητες μπορούν να χαρακτηριστούν με βάση την ομάδα στην οποία προβλέπεται να ανήκει. Οι πρωτεΐνες μπορούν να ταξινομηθούν σε διαφορετικές ομάδες με βάση τις οικογένειες στις οποίες ανήκουν, τις περιοχές (domains) που βρίσκονται στην ακολουθία τους και κάποια άλλα χαρακτηριστικά που έχουν στην ακολουθία τους.

Μια οικογένεια πρωτεϊνών είναι μια ομάδα από πρωτεΐνες που μοιράζεται ένα κοινό εξελικτικό πρόγονο και οι πρωτεΐνες συνδέονται μεταξύ τους με βάση τις λειτουργίες και τις ομοιότητες που έχουν σε αλληλουχία ή δομή. Οι οικογένειες των πρωτεϊνών είναι ιεραρχικά δομημένες, με αποτέλεσμα οι πρωτεΐνες που μοιράζονται έναν κοινό πρόγονο να υποδιαιρούνται σε μικρότερες, πιο στενά συνδεδεμένες

ομάδες. Για να καθοριστεί η ιεραρχία χρησιμοποιούνται οι όροι υπεροικογένεια και υποοικογένεια. Ο όρος υπεροικογένεια αναφέρεται σε μια μεγάλη ομάδα πρωτεϊνών που έχουν μακρινή συγγένεια, ενώ ο όρος υποοικογένεια αναφέρεται σε μια μικρή ομάδα πρωτεϊνών που έχουν στενή συγγένεια μεταξύ τους.

Τα domains είναι ξεχωριστές λειτουργικές και δομικές περιοχές σε μια πρωτεΐνη. Συνήθως είναι υπεύθυνες για μια συγκεκριμένη λειτουργία ή αλληλεπίδραση, συμβάλλοντας έτσι στον συνολικό ρόλο μιας πρωτεΐνης. Οι περιοχές αυτές μπορεί να υπάρχουν σε μια ποικιλία βιολογικών χαρακτηριστικών και γι' αυτό τον λόγο οι ίδιες περιοχές μπορούν να βρεθούν σε πρωτεΐνες με διαφορετικές λειτουργίες. Επίσης μια πρωτεΐνη μπορεί να περιέχει πολλές διαφορετικές περιοχές. Συχνά οι επιμέρους περιοχές έχουν ειδικές λειτουργίες, όπως η δέσμευση ενός συγκεκριμένου μορίου ή η καταλυτική αντίδραση μιας δεδομένης αντίδρασης, που συμβάλλουν επικουρικά στον συνολικό ρόλο της πρωτεΐνης.

Οι ταξινομήσεις με βάση την οικογένεια και τις περιοχές δεν είναι πάντα απλές και μπορούν να αλληλεπικαλύπτονται, καθώς ορισμένες φορές οι πρωτεΐνες τοποθετούνται σε οικογένειες εξαιτίας των περιοχών που περιέχουν.

Επιπλέον υπάρχουν ομάδες αμινοξέων που προσδίδουν ορισμένα χαρακτηριστικά σε μια πρωτεΐνη και μπορεί να είναι σημαντικά για τη συνολική λειτουργία της, αλλά δεν είναι domains. Σε αυτά τα χαρακτηριστικά περιλαμβάνονται οι ενεργείς θέσεις, οι οποίες περιέχουν αμινοξέα που εμπλέκονται στην καταλυτική δραστηριότητα, οι θέσεις πρόσδεσης, που περιέχουν αμινοξέα τα οποία εμπλέκονται άμεσα στη δέσμευση μορίων ή ιόντων, οι θέσεις μετά μεταφραστικής τροποποίησης (PTM), οι οποίες περιέχουν κατάλοιπα (φωσφορυλιωμένα, παλμιτοϋλιωμένα, ακετυλιωμένα κλπ.) που τροποποιούνται χημικά, μετά τη διαδικασία της μετάφρασης και οι επαναλήψεις, οι οποίες είναι συνήθως μικρές αλληλουχίες αμινοξέων οι οποίες επαναλαμβάνονται εντός μίας πρωτεΐνης και μπορεί να δώσουν σε αυτές δεσμευτικές ή δομικές ιδιότητες.

Τα χαρακτηριστικά αυτά διαφέρουν από τις περιοχές, καθώς είναι αρκετά μικρά και αποτελούνται από μερικά μόνο αμινοξέα, ενώ οι περιοχές αντιπροσωπεύουν ολόκληρες δομικές ή λειτουργικές μονάδες της πρωτεΐνης. Επίσης κάποια από τα χαρακτηριστικά μπορεί να βρίσκονται εντός ενός domain.

Προκειμένου να ταξινομηθούν οι πρωτεΐνες σε οικογένειες και να προβλεφθεί η παρουσία σημαντικών περιοχών ή χαρακτηριστικών στις ακολουθίες, απαιτείται η χρήση υπολογιστικών εργαλείων. Τα εργαλεία αυτά κατασκευάζονται

χρησιμοποιώντας πολλαπλή στοίχιση ακολουθιών, που εστιάζουν σε διαφορετικές προσεγγίσεις και μπορεί να περιέχει patterns, προφίλ, αποτυπώματα (fingerprints) και κρυπτομαρκοβιανά μοντέλα (HMMs). Κάθε προσέγγιση ξεκινά με πολλαπλή στοίχιση ακολουθιών και μπορεί να επικεντρωθεί σε μια μόνο διατηρημένη περιοχή στην ακολουθία, σε πολλαπλά διατηρημένα μοτίβα ή στην πλήρη ευθυγράμμιση ολόκληρης της πρωτεΐνης ή μιας συγκεκριμένης περιοχής.

3. Πρόγνωση Δευτεροταγούς Δομής των Πρωτεϊνών

Η δευτεροταγής δομή μιας πρωτεΐνης αναφέρεται στην τοπική διαμόρφωση της πολυπεπτιδικής της αλυσίδας και έχει τρείς κατηγορίες, την α-έλικα, την β-πτυχωτή επιφάνεια και την τυχαία δομή.

Η πρόβλεψη της δευτεροταγούς δομής της πρωτεΐνης είναι ένα σημαντικό μέρος του γενικού προβλήματος αναδίπλωσης της πρωτεΐνης και είναι η πιο γενική μέθοδος απόκτησης ορισμένων δομικών πληροφοριών από οποιαδήποτε νέα αλληλουχία. Η πρόβλεψη της δευτεροταγούς δομής είναι χρήσιμη σε πολλά προβλήματα που αφορούν τις πρωτεΐνες.

- Στον σχεδιασμό νέων πρωτεϊνών, οι κανόνες που διέπουν τις δομές της α-έλικας και της β-πτυχωτής επιφάνειας παρέχουν τις κατευθυντήριες γραμμές στην επιλογή συγκεκριμένων μεταλλάξεων.
- Η δευτεροταγής δομή μπορεί να βοηθήσει στην επιβεβαίωση μιας δομικής και λειτουργικής σχέσης μεταξύ πρωτεϊνών που έχουν μια αδύναμη σχέση όσον αφορά την αλληλουχία τους.
- Η πρόβλεψη της δευτεροταγούς δομής είναι σημαντική για την καθιέρωση των ευθυγραμμίσεων κατά τη διάρκεια της κατασκευής ενός μοντέλου με ομολογία και για τον έλεγχο της εγκυρότητας των κρυσταλλογραφικών μοντέλων.
- Οι πληροφορίες σχετικά με το πώς οι αλληλουχίες αναδιπλώνονται στις δευτεροταγείς δομές μπορούν στη συνέχεια να χρησιμοποιηθούν για την πρόβλεψη της τριτοταγούς δομή με σταδιακό τρόπο (από την αλληλουχία των αμινοξέων στη δευτεροταγή δομή και από τη δευτεροταγή δομή στην τριτοταγή δομή).

Η πρόβλεψη της δευτεροταγούς δομής της πρωτεΐνης είναι η πιο γενική μέθοδος λήψης ορισμένων δομικών πληροφοριών για μια οποιαδήποτε νέα προσδιορισμένη

αλληλουχία και ένα κύριο βήμα στο γενικό πρόβλημα της αναδίπλωσης της πρωτεΐνης. Υπάρχουν πολλές μέθοδοι πρόβλεψης και αυτές μπορούν να χωριστούν σε τρεις κύριες ομάδες. Η πρώτη κατηγορία είναι οι στατιστικές / εμπειρικές μέθοδοι που προβλέπουν τη δομή με βάση στατιστικά δεδομένα που συγκεντρώνονται κατόπιν εξεύρεσης εμπειρικών σχέσεων μεταξύ των τύπων δομής. Οι στατιστικές μέθοδοι έχουν το πλεονέκτημα της χρήσης της βάσης δεδομένων γνωστών πρωτεϊνικών δομών, αλλά έχουν το μειονέκτημα ότι δεν λαμβάνουν πλήρως υπόψη τις φυσικοχημικές γνώσεις για τις πρωτεΐνες και γι' αυτό έχουν ανεπαρκή ερμηνευτική ισχύ. Η ποσότητα των διαθέσιμων δεδομένων είναι πολύ σημαντική για τις στατιστικές μεθόδους πρόβλεψης και γι' αυτό τον λόγο η έλλειψη δεδομένων αποτελεί σημαντικό θέμα στην ακρίβεια των στατιστικών μεθόδων. Η δεύτερη κατηγορία είναι οι φυσικοχημικές μέθοδοι που προβλέπουν τη δομή με βάση τη φυσική και χημική βάση της πρωτεϊνικής δομής και η τρίτη κατηγορία είναι οι αλγόριθμοι μηχανικής μάθησης που προσπαθούν να συνδυάσουν τα καλύτερα χαρακτηριστικά των στατιστικών και των φυσικοχημικών μεθόδων.

Παρόλο που η πρόγνωση της δευτεροταγούς δομής είναι πολύ σημαντική, δεν υπάρχει σαφής ενιαία μέθοδος καλύτερης πρόβλεψης και υπάρχουν αρκετά σημαντικά μεθοδολογικά προβλήματα που δυσχεραίνουν τη σωστή αξιολόγηση των μεθόδων αυτών.

4. Πρόγνωση Διαμεμβρανικών Τμημάτων

Οι βιολογικές μεμβράνες, είναι πολύπλοκα οργανωμένα συμπλέγματα, τα οποία αποτελούνται κυρίως από λιπίδια και πρωτεΐνες. Η ποικιλομορφία των μορίων αυτών παρέχουν σημαντικές λειτουργικές ιδιότητες στο κύτταρο, διαχωρίζοντας το εσωτερικό από το εξωκυττάριο περιβάλλον και ελέγχοντας τις ουσίες που εισέρχονται και εξέρχονται από αυτό.

Υπάρχουν αρκετές κατηγορίες μεμβρανικών λιπιδίων που διαθέτουν ως κοινό χαρακτηριστικό την υδρόφιλη κεφαλή και την υδρόφοβη ουρά.. Οι κυτταρικές μεμβράνες αποτελούνται από μια διπλοστιβάδα λιπιδίων, όπου η υδρόφιλη κεφαλή αλληλεπιδρά με το υδατικό εξωκυττάριο και ενδοκυττάριο περιβάλλον και η υδρόφοβη ουρά βρίσκεται στο εσωτερικό της διπλοστιβάδας. Ο σχηματισμός της λιπιδικής διπλοστιβάδας είναι αποτέλεσμα της τάσης των ουρών να αποστρέφονται το υδατικό περιβάλλον και γι' αυτό τον λόγο δημιουργείται μια σταθερή δομή. Η

ιδιότητα αυτή είναι σημαντική για τη συγκρότηση και τη λειτουργικότητα των μεμβρανών του κυττάρου, καθώς η διπλοστιβάδα αποτελεί φραγμό ανάμεσα στο εσωτερικό και εξωτερικό περιβάλλον του κυττάρου.

Οι πρωτεΐνες που βρίσκονται στην κυτταρική μεμβράνη και διασχίζουν την διπλοστιβάδα των λιπιδίων ονομάζονται μεμβρανικές πρωτεΐνες. Ανάλογα με την θέση τους στην μεμβράνη μπορεί να είναι περιφερειακές ή διαμεμβρανικές. Οι περιφερειακές πρωτεΐνες δεν διαπερνούν ολόκληρη τη μεμβράνη, αλλά συνδέονται χαλαρά με ηλεκτροστατικές δυνάμεις στην εσωτερική είτε στην εξωτερική της επιφάνεια, ενώ οι διαμεμβρανικές πρωτεΐνες είναι ολοκληρωτικά ενσωματωμένες στην μεμβράνη, καθώς διαπερνούν εξολοκλήρου τη μεμβράνη.

Γενικά οι μεμβρανικές πρωτεΐνες έχουν μεγάλη ποικιλία βιολογικών λειτουργιών. Είναι υπεύθυνες για τη δημιουργία διαφόρων τύπων υποδοχέων για νευροδιαβιβαστές, ορμόνες και άλλες ουσίες, σχηματίζουν μεγάλη ποικιλία ιοντικών καναλιών, αποτελούν βασικά δομικά στοιχεία των μεμβρανών και αναλαμβάνουν τη μεταφορά και αποθήκευση ουσιών ζωτικής σημασίας.

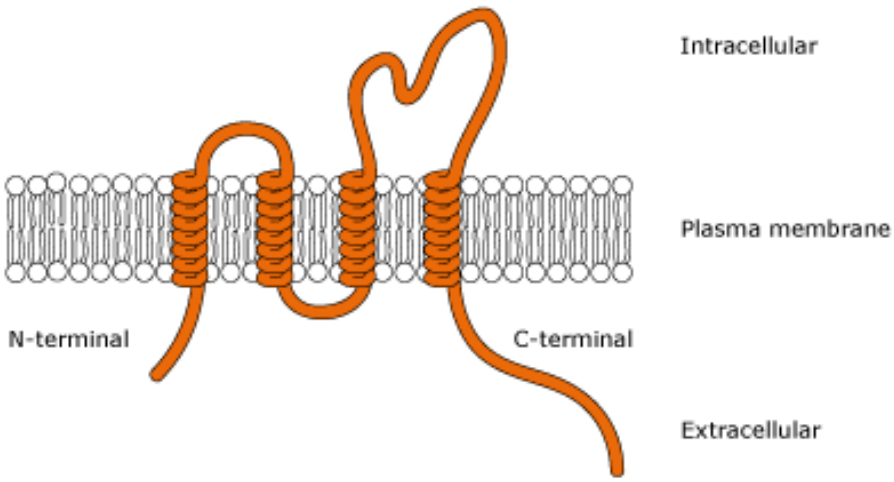
Η μεγαλύτερη κατηγορία μεμβρανικών πρωτεϊνών είναι οι διαμεμβρανικές πρωτεΐνες. Επειδή το εσωτερικό της κυτταρικής μεμβράνης είναι υδρόφοβο, είναι αναμενόμενο τα επιφανειακά αμινοξέα των διαμεμβρανικών πρωτεϊνών να είναι υδρόφοβα. Με αυτή την προϋπόθεση είναι δυνατή η πρόβλεψη των τμημάτων της πρωτεΐνης που βρίσκονται στο εσωτερικό, καθώς απαρτίζονται από υδρόφοβα αμινοξέα. Εξαιτίας αυτής της ιδιότητας έχουν οριστεί κλίμακες υδροφοβικότητας από μελέτες διαλυτότητας ή κρυσταλλογραφίας για κάθε αμινοξύ.

Τα περισσότερα διαμεμβρανικά τμήματα είναι α-έλικες [Εικόνα 2] και αυτό είναι γνωστό από τις πρωτεΐνες των οποίων η δομή έχει προσδιοριστεί. Οι περισσότερες πρωτεΐνες που διασχίζουν την κυτταρική μεμβράνη περισσότερο από μία φορά περιέχουν α-έλικες στο διαμεμβρανικό τους τμήμα, ενώ ελάχιστες τη διασχίζουν με β-πτυχωτές επιφάνειες. Το γεγονός ότι οι α-έλικες είναι περισσότερες οφείλεται στο ότι το μήκος των β-πτυχωτών επιφανειών σε αμινοξικά κατάλοιπα είναι σημαντικά μικρότερο από ότι αυτό των α-ελίκων. Επίσης μπορεί να οφείλεται στο ότι οι α-έλικες μπορούν να εισέλθουν στη μεμβράνη κάθε μία ανεξάρτητα κατά το δίπλωμα της πρωτεΐνης, ενώ οι β-κλώνοι πρέπει πρώτα να σχηματίσουν τις β-πτυχωτές επιφάνειες πριν την είσοδο τους στη μεμβράνη.

Η πρόβλεψη των διαμεμβρανικών τμημάτων των πρωτεϊνών αποτελεί κλασικό πρόβλημα στον τομέα της Βιοπληροφορικής, δεδομένου του ότι ο προσδιορισμός της

δομής μιας διαμεμβρανικής πρωτεΐνης με πειραματικά μέσα είναι δύσκολος, χρονοβόρος και δαπανηρός. Η χρήση υπολογιστικών εργαλείων αποδείχθηκε σημαντική στην αναγνώριση των διαμεμβρανικών τμημάτων αν και είναι πολύπλοκη, καθώς πολλές διαμεμβρανικές α-έλικες σε πρωτεΐνες πολλών στρώσεων είναι μερικώς ή πλήρως θωρακισμένες από άλλες διαμεμβρανικές α-έλικες και επίσης οι β-κλώνοι δεν είναι πλήρως εκτεθειμένοι στη λιπιδική διπλοστιβάδα, με αποτέλεσμα να είναι αμφιπαθικοί [33]. Η πρόγνωση των διαμεμβρανικών τμημάτων γίνεται με τη χρήση σε συνδυασμό της υδροφοβικότητας των αμινοξικών κατάλοιπων, των στατιστικών τεχνικών και των μεθόδων μηχανικής μάθησης.

Εικόνα 2: α-ελικοειδής μεμβρανική πρωτεΐνη



Ένας από τους σημαντικότερους και καλύτερους αλγόριθμους πρόβλεψης διαμεμβρανικών ελίκων στις πρωτεΐνες είναι ο TMHMM [34]. Επίσης γνωστές είναι και οι μέθοδοι PHOBIUS [35] και SPOCTOPUS [36], καθώς πραγματοποιούν ταυτόχρονη πρόγνωση των διαμεμβρανικών τμημάτων και των σηματοδοτικών αλληλουχιών. Για την πρόγνωση των διαμεμβρανικών β-βαρελιών το PRED-TMBB [37] αποτελεί μια πετυχημένη μέθοδο, που χρησιμοποιεί μόνο πληροφορία από την αμινοξική αλληλουχία, αλλά και διαφορετικούς αλγόριθμους για την εκπαίδευση και την αποκωδικοποίησή του.

5. Πρόγνωση Σηματοδοτικών Αλληλουχιών

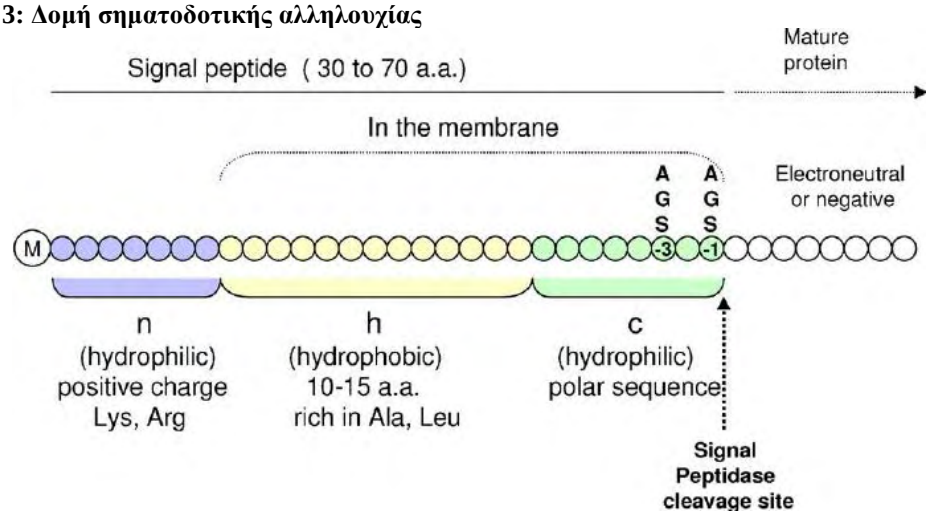
Στα βακτήρια, τους ευκαρυώτες και τα αρχαία, στην πλειοψηφία τους, οι πρωτεΐνες που εκκρίνονται από το κυτταρόπλασμα, συντίθεται σαν πρόδρομες πρωτεΐνες και φέρουν μια διασπάσιμη σηματοδοτική αλληλουχία στο αμινοτελικό τους άκρο [38]. Αυτή η αλληλουχία ονομάζεται σηματοδοτική αλληλουχία ή

πεπτίδιο οδηγητής (signal peptide) και συνήθως αποτελείται από δεκαέξι έως τριάντα αμινοξέα.

Η δομή της σηματοδοτικής αλληλουχίας [Εικόνα 3] χωρίζεται σε τρείς περιοχές, την αμινοτελική περιοχή (N-region), την υδρόφοβη περιοχή (H-region) και τη θέση διάσπασης (C-region). Η αμινοτελική περιοχή αποτελείται από ένα θετικά φορτισμένο τμήμα αμινοξέων μικρού μήκους, ενώ ο πυρήνας της αλληλουχίας περιέχει μια μεγάλη έκταση υδρόφοβων αμινοξέων που έχει την τάση να σχηματίζει μία άλφα έλικα. Στο άκρο του πεπτιδίου υπάρχει μία σειρά αμινοξέων που ονομάζεται θέση διάσπασης και αναγνωρίζεται και διασπάται από την πεπτιδάση σήματος, μία μεμβρανική πρωτεΐνη που αποκόπτει το πεπτίδιο από την πολυπεπτιδική αλυσίδα.

Η σημαντικότητα των σηματοδοτικών αλληλουχιών στη στόχευση των πρωτεϊνών, οδήγησε στη δημιουργία υπολογιστικών μεθόδων πρόβλεψης των αλληλουχιών αυτών. Η υλοποίηση τέτοιων αλγορίθμων είναι δύσκολη εξαιτίας της παρόμοιας υδρόφοβης σύνθεσής που παρουσιάζεται μεταξύ των σηματοδοτικών αλληλουχιών και των αμινοτελικών τμημάτων των διαμεμβρανικών περιοχών. Αυτή η ομοιότητα πολλές φορές μπορεί να προκαλέσει λανθασμένη πρόβλεψη, δηλαδή να προβλεφθεί ως σηματοδοτική αλληλουχία μια διαμεμβρανική περιοχή ή και αντίστροφα [39], παρόλο που οι διαμεμβρανικές έλικες τυπικά έχουν μακρύτερες υδρόφοβες περιοχές από τις σηματοδοτικές αλληλουχίες και επίσης δεν έχουν θέσεις διάσπασης [40]. Αν και οι διαφορές αυτές ίσως να μην είναι αρκετές για μια πρόβλεψη να διακρίνει τους δυο αυτούς τύπους, αποτελεί μια μεγάλη πρόκληση για τη βελτίωση των προβλέψεων αυτών.

Εικόνα 3: Δομή σηματοδοτικής αλληλουχίας



Η πιο καλή και πιο σύγχρονη μέθοδος πρόγνωσης σηματοδοτικών ακολουθιών είναι το SignalP [40], το οποίο διαθέτει ξεχωριστά εργαλεία για την κάθε ομάδα οργανισμών και βασίζεται στην εξαιρετική βιβλιογραφική αναζήτηση για την κατάρτιση του συνόλου εκπαίδευσης, περιλαμβάνοντας έτσι πολλές πρωτεΐνες, αλλά και απομακρύνοντας λάθος καταχωρίσεις [41]. Μια ακόμα σημαντική μέθοδος είναι το PRED-SIGNAL [38], καθώς αποτελεί τη μοναδική μέχρι στιγμής διαθέσιμη μέθοδο πρόγνωσης σηματοδοτικών αλληλουχιών για τα Αρχαία. Για τις σηματοδοτικές αλληλουχίες που κατευθύνουν τις πρωτεΐνες στους χλωροπλάστες, ο πιο γνωστός αλγόριθμος είναι το ChloroP [42], ενώ το TargetP [43] προβλέπει τις εκκριτικές πρωτεΐνες και των μιτοχονδρίων, αλλά και των χλωροπλαστών.

6. Νευρωνικά Δίκτυα

Τα τεχνητά νευρωνικά δίκτυα είναι υπολογιστικές μέθοδοι που μπορούν με διαδικασίες εκμάθησης, να ανιχνεύσουν συσχετισμούς στα δεδομένα και γι' αυτό έχουν μεγάλη επιτυχία και στις προβλέψεις. Τα νευρωνικά δίκτυα μπορούν να εξάγουν νέες πληροφορίες από ακατέργαστα δεδομένα και να δημιουργήσουν υπολογιστικά μοντέλα που είναι χρήσιμα για τη λήψη αποφάσεων. Είναι εμπνευσμένα από τα βιολογικά νευρωνικά δίκτυα και κυρίως τον εγκέφαλο. Όπως ο εγκέφαλος αποτελείται από νευρώνες, οι οποίοι λειτουργούν αυτόνομα αλλά παράλληλα συνεργάζονται και μεταξύ τους ανταλλάσοντας ηλεκτρικά σήματα, έτσι και τα τεχνητά νευρωνικά δίκτυα αποτελούνται από στρώσεις κόμβων που συνδέονται μεταξύ τους με τη χρήση του συντελεστή βάρους. Η επεξεργασία κάθε κόμβου καθορίζεται από τη συνάρτηση μεταφοράς, η οποία και καθορίζει την κάθε έξοδο σε σχέση με τις εισόδους και τους συντελεστές βάρους. Τα βάρη ρυθμίζονται όταν τα δεδομένα παρουσιάζονται στο δίκτυο κατά τη διάρκεια της «εκπαίδευσης» του δικτύου. Η επιτυχής εκπαίδευση καθορίζει και τα αποτελέσματα στην ακρίβεια των τεχνητών νευρωνικών δικτύων.

Η εκπαίδευση των δικτύων μπορεί να πραγματοποιηθεί είτε με επίβλεψη είτε χωρίς επίβλεψη. Στην εκπαίδευση με επίβλεψη (supervised learning) γίνεται χρήση ενός εκπαιδευτή απομνημόνευσης και της συνολικής γενικευμένης πληροφορίας. Με τη χρήση των αρχικών δεδομένων το δίκτυο μαθαίνει μια συνάρτηση που προσεγγίζει τη σχέση μεταξύ εισόδου και εξόδου. Η ικανότητα απομνημόνευσης της σχέσης των δεδομένων αυτών αποτελεί το μέτρο για την απόδοση του νευρωνικού δικτύου.

Κατόπιν με τη χρήση της συνολικής γενικευμένης πληροφορίας, γίνεται έλεγχος επιβεβαίωσης και αξιολόγησης του δικτύου. Όσο πιο μικρά είναι τα σφάλματα πρόβλεψης για τα δεδομένα εκπαίδευσης και μάθησης τόσο πιο σωστή θεωρείται η μάθηση του δικτύου. Στην εκπαίδευση χωρίς επίβλεψη γίνεται χρήση μόνο οργανωμένων δεδομένων εισόδου για την εκπαίδευση του νευρωνικού δικτύου που έχει ως σκοπό την εύρεση όμοιων χαρακτηριστικών στο δοσμένο σύνολο δεδομένων.

Με βάση τον τρόπο που συνδέονται οι κόμβοι μεταξύ τους καθορίζεται και η αρχιτεκτονική του δικτύου. Γενικά οι αρχιτεκτονικές των νευρωνικών δικτύων μπορούν να χωριστούν σε τρείς κατηγορίες. Η πρώτη κατηγορία είναι τα δίκτυα εμπρόσθιας τροφοδότησης (Feedforward Networks) και μπορεί να έχουν ένα επίπεδο ή πολλά επίπεδα (που ονομάζονται κρυφά επίπεδα). Στα δίκτυα αυτά η πληροφορία διαδίδεται μόνο προς τα εμπρός, από την είσοδο προς την έξοδο. Δεν υπάρχει ανάδραση, δηλαδή η έξοδος οποιουδήποτε επιπέδου δεν επηρεάζει το ίδιο το επίπεδο. Η δεύτερη κατηγορία είναι τα δίκτυα ανάδρασης (Feedback Networks), όπου η πληροφορία μπορεί να διαδίδεται και προς τις δύο κατευθύνσεις εισάγοντας κόμβους στο δίκτυο, ενώ ακόμα υπάρχει η δυνατότητα αυτοανάδρασης, δηλαδή ένας κόμβος μπορεί να δέχεται σαν είσοδο την ίδια του την έξοδο. Τέλος η τρίτη κατηγορία είναι τα δίκτυα με επαναλαμβανόμενες δομές στο χώρο και μπορεί να είναι μονοδιάστατα, δισδιάστατα ή μεγαλύτερης διάστασης και αποτελούνται από συστοιχίες κόμβων, οι οποίοι συνδέονται με τους κόμβους εισόδου.

Η πρόγνωση της δευτεροταγούς δομής των πρωτεϊνών γίνεται κυρίως με τη χρήση των νευρωνικών δικτύων εμπρόσθιας τροφοδότησης. Για την υλοποίηση ενός τέτοιου νευρωνικού δικτύου πρέπει καταρχάς να προσδιοριστεί η μέθοδος με την οποία θα γίνει η κωδικοποίηση της πρωτεϊνικής ακολουθίας. Ακολούθως ορίζεται ο αριθμός των κόμβων εισόδου και ο αριθμός των κρυφών επιπέδων που θα χρησιμοποιηθούν για να πραγματοποιηθούν οι προβλέψεις, ενώ παράλληλα ρυθμίζεται και ο συντελεστής βάρους με σκοπό να μειωθούν οι διαφορές που θα εμφανιστούν μεταξύ της προβλεπόμενης και δοσμένης ακολουθίας. Κατόπιν γίνεται επιλογή του αλγορίθμου εκπαίδευσης και των συνόλων δεδομένων που θα εκπαιδεύσουν το νευρωνικό δίκτυο. Μετά την ολοκλήρωση των βημάτων αυτών γίνεται έλεγχος του νευρωνικού δικτύου και περιορίζονται τα σφάλματα εξόδου με την αλλαγή του συντελεστή βάρους. Όσο πιο μικρά είναι τα σφάλματα, τόσο πιο επιτυχείς θεωρείται η πρόβλεψη.

Μερικές από τις μεθόδους πρόγνωσης που χρησιμοποιούν νευρωνικά δίκτυα είναι το PeroxiP Predictor [44], για τις πρωτεΐνες των υπεροξεισωμάτων, το Myristoylator [45] που προβλέπει την προσθήκη ενός λιπιδίου του μυριστικού οξέως στο αμινοτελικό άκρο και τα ANOLEA [46] και DNABIND [47] που είναι για τη δέσμευση των πρωτεϊνών στο DNA.

ΜΕΡΟΣ 3ο
ΣΥΝΟΛΑ ΔΕΔΟΜΕΝΩΝ

1. Εισαγωγή

Εξαιτίας του μεγάλου όγκου πληροφορίας που παρουσιάζεται στις βάσεις δεδομένων βιολογικών ακολουθιών, η ανάπτυξη αλγορίθμων που θα μπορούσαν να ταξινομήσουν και να προβλέψουν τις ιδιότητες μιας ακολουθίας είναι σημαντική. Ουσιαστικό εργαλείο για την υλοποίηση των αλγορίθμων αυτών αποτελεί η δημιουργία συνόλων δεδομένων (datasets). Προκειμένου να είναι αποτελεσματικός ο αλγόριθμος που θα αναπτυχθεί, είναι απαραίτητη η κατασκευή ενός μεγάλου και περιεκτικού συνόλου δεδομένων από ακολουθίες, με το οποίο θα γίνει η εκπαίδευση και η εκτέλεση του αντίστοιχου αλγόριθμου.

Με τον όρο σύνολο δεδομένων στην Βιοπληροφορική αναφερόμαστε σε μια συλλογή από αρχεία δεδομένων που αποτελούνται από ακολουθίες νουκλεϊνικών οξέων (Δεσοξυριβονουκλεϊκού οξέος (DNA) ή Ριβονουκλεϊκού οξέος (RNA)) ή αμινοξέων και προορίζονται για υπολογιστική επεξεργασία. Για τη δημιουργία των συνόλων δεδομένων ακολουθούνται κατά βάση κάποια βήματα:

- Βήμα 1^{ο}: Επιλογή ακολουθιών από βιολογική βάση δεδομένων.
- Βήμα 2^{ο}: Απόρριψη ακολουθιών με διφορούμενους όρους όπως «potential», «probable», «probably», «maybe».
- Βήμα 3^{ο}: Απόρριψη των ακολουθιών με τον όρο «fragment»
- Βήμα 4^{ο}: Αφαίρεση ακολουθιών με λιγότερο από ένα ποσοστό αμινοξέων που κρίνεται ότι μπορεί να θεωρηθεί η ακολουθία ως ατελής.
- Βήμα 5^{ο}: Αποκοπή των ακολουθιών που έχουν πάνω από ένα ποσοστό, που ορίζεται με βάση της προδιαγραφές της μεθόδου χρήσης, για γνωστά ζεύγη, όμοια με οποιαδήποτε άλλη ακολουθία της ίδιας κατηγορίας μιας οικογένειας, με σκοπό να μειωθεί η απόκλιση ομόλογων ακολουθιών.
- Βήμα 6^{ο}: Έλεγχος μη επανάληψης μιας ακολουθίας. Δηλαδή καμιά ακολουθία να μην περιλαμβάνεται στο σύνολο δεδομένων περισσότερο από μια φορά.

Ο τρόπος συλλογής και δημιουργίας ενός συνόλου δεδομένων μπορεί να διαφοροποιηθεί ανάλογα με τις απαιτήσεις και τις προδιαγραφές των μεθόδων που χρησιμοποιούνται για την υλοποίηση του αλγορίθμου πρόβλεψης ή και ταξινόμησης.

Αφού σχεδιαστεί και υλοποιηθεί ο αλγόριθμος πρόγνωσης είναι σημαντικό να γίνει η αξιολόγηση του. Για τη σωστή αξιολόγηση ενός αλγορίθμου γίνεται χρήση των ελέγχων αυτο-συνέπειας (self-consistency), cross-validation και της μεθόδου Jackknife. Στον έλεγχο αυτο-συνέπειας γίνεται έλεγχος με τα ίδια δεδομένα που έχει εκπαιδευτεί ο αλγόριθμος. Επειδή όμως υπάρχει περίπτωση προσαρμογής (over-fitting). του αλγορίθμου στο σύνολο εκπαίδευσης του, είναι καλύτερα να γίνεται έλεγχος με ένα ανεξάρτητο σύνολο δεδομένων (independent set). Στον έλεγχο cross-validation χωρίζεται το σύνολο εκπαίδευσης σε k υποσύνολα. Στη συνέχεια ένα υποσύνολο αφαιρείται από το σύνολο εκπαίδευσης και έτσι η εκπαίδευση πραγματοποιείται με τα υπόλοιπα υποσύνολα και ο αλγόριθμος ελέγχεται στο υποσύνολο το οποίο έχει αφαιρεθεί. Η διαδικασία αυτή επαναλαμβάνεται k φορές και δίνει μια αμερόληπτη εκτίμηση για την επιτυχία του αλγορίθμου. Στη μέθοδο Jackknife το σύνολο εκπαίδευσης χωρίζεται πάλι σε k υποσύνολα, με τη διαφορά ότι το k είναι ίσο με το μέγεθος του συνόλου εκπαίδευσης και συνεπώς κάθε υποσύνολο είναι ίσο με ένα. Η μέθοδος αυτή είναι καλύτερη σε σύνολα με μέτριο μέγεθος ή για αλγορίθμους που είναι γρήγοροι και δεν χρησιμοποιείται σε πολύ μεγάλα σύνολα, σε αργούς αλγορίθμους και αν δεν υπάρχει εύκολος τρόπος να εξασφαλιστούν οι συνθήκες ομοιότητας [41].

2. Σύνολα δεδομένων διαμεμβρανικών πρωτεϊνών (a helical transmembrane proteins datasets)

Ως «δομικό στοιχείο της ζωής», ένα κύτταρο θεωρείται η πιο βασική δομική και λειτουργική μονάδα όλων των ζωντανών οργανισμών. Είναι εξαιρετικά οργανωμένο με πολλές λειτουργικές μονάδες ή οργανίδια, σύμφωνα με την κυτταρική ανατομία. Οι περισσότερες από αυτές τις μονάδες «περιβάλλονται» από μία ή περισσότερες μεμβράνες, που είναι η δομική βάση για πολλές σημαντικές βιολογικές λειτουργίες. Παρόλο που η λιπιδική διπλοστοιβάδα είναι η βασική δομή των μεμβρανών, οι περισσότερες από τις ειδικές λειτουργίες της κυτταρικής μεμβράνης εκτελούνται από τις μεμβρανικές πρωτεΐνες [48]. Οι μεμβρανικές πρωτεΐνες αποτελούνται από διαμεμβρανικές πρωτεΐνες και από αγκυροβολημένες μεμβρανικές πρωτεΐνες. Η πρώτη κατηγορία περιέχει ένα ή περισσότερα υδρόφοβα τμήματα, και ως εκ τούτου είναι σχετικά εύκολη η διάκριση της από τις μη μεμβρανικές πρωτεΐνες. Η δεύτερη έχει ένα ομόφωνο μοτίβο αλληλουχίας στο N- ή C- τερματικό άκρο και ως εκ τούτου

μπορεί να αναγνωριστεί σε κάποιο βαθμό. Ο τρόπος που μια πρωτεΐνη δεμένη με την μεμβράνη συνδέεται με τη λιπιδική διπλοστοιβάδα συνήθως αντανακλά τη λειτουργία της πρωτεΐνης. Επίσης, η σύνδεση των μεμβρανικών πρωτεϊνών σε διαφορετικές τοποθεσίες, συνήθως υποδηλώνει και διαφορετικές βιολογικές λειτουργίες. Οι πρωτεΐνες που είναι συνδεδεμένες με την πλασματική μεμβράνη δρουν σαν αισθητήρες εξωκυττάριων σημάτων, μεταφέρουν πληροφορίες δια μέσου της μεμβράνης και επιτρέπουν στο κύτταρο να αλλάζει την συμπεριφορά του για να μπορεί να ανταποκριθεί σε διάφορα περιβαλλοντικά ερεθίσματα [49].

Οι μεμβρανικές πρωτεΐνες πιστεύεται ότι αποτελούν το 20-30% των πρωτεϊνών σε ένα γονιδίωμα και αντιπροσωπεύουν μια σημαντική αναλογία των θεραπευτικών στόχων των φαρμάκων. Ωστόσο, ως αποτέλεσμα των δυσκολιών στον πειραματικό προσδιορισμό της δομής, αυτές αποτελούν μόνο το 1% των δομών που διατίθενται στην PDB [50]. Η κατανόηση του μηχανισμού λειτουργίας των διαμεμβρανικών πρωτεϊνών απαιτεί να γνωρίζουμε την τρισδιάστατης δομής τους, αν και η επίλυση της τρισδιάστατης δομής αποτελεί πειραματική πρόκληση, καθώς οι διαμεμβρανικές πρωτεΐνες και η μεμβράνη συνδέονται με ισχυρές υδρόφοβες αλληλεπιδράσεις [51].

Οι ελικοειδείς μεμβρανικές πρωτεΐνες (Helical Membrane Proteins - HMPs) διαδραματίζουν κρίσιμο ρόλο σε ποικίλες φυσιολογικές διαδικασίες, συμπεριλαμβανομένης της παραγωγής ενέργειας, της μεταγωγή σήματος, της μεταφοράς των διαλυτών ουσιών κατά μήκος της μεμβράνης και της διατήρησης της βαθμωτής ανάλυσης των ιόντων και πρωτονίων [52]. Ως εκ τούτου, είναι επιθυμητό να αναπτυχθούν υπολογιστικές μέθοδοι που βασίζονται στην αλληλουχία για την πρόβλεψη των χαρακτηριστικών της δομής των HMPs, καθώς είναι πολύ δύσκολος ο προσδιορισμός των δομών των τους με πειραματικές τεχνικές [53].

Μια θεμελιώδης πτυχή της δομής των πρωτεϊνών είναι η τοπολογία τους, δηλαδή ο αριθμός των διαμεμβρανικών τμημάτων, η θέση τους στην αλληλουχία της πρωτεΐνης και ο προσανατολισμός τους στην μεμβράνη [39]. Η πρόβλεψη της τοπολογίας των διαμεμβρανικών τμημάτων είναι σημαντικό εργαλείο για την καλύτερη κατανόηση των διαμεμβρανικών πρωτεϊνών. Για τις περισσότερες μηχανές πρόβλεψης τοπολογίας και πρόβλεψης πεπτιδίου σήματος, μια N τερματική διαμεμβρανική περιοχή είναι συχνά δύσκολο να διακριθεί από ένα πεπτίδιο σήματος, καθώς και στα δύο ένα τμήμα τους αποτελείται από υδρόφοβα κατάλοιπα. Όταν εφαρμόζεται σε ολόκληρο το γονιδίωμα είναι ιδιαίτερα σημαντικό για την πρόγνωση της τοπολογίας να μην γίνεται ψευδής πρόβλεψη των πραγματικών πεπτιδίων

σήματος ως Ν τερματική περιοχή ή το αντίστροφο, καθώς αυτό μπορεί να οδηγήσει σε ανακριβή συμπεράσματα [36]. Μια σωστή πρόβλεψη της τοπολογίας παρέχει ένα εξαιρετικό πρότυπο για περεταίρω πειραματικές μελέτες και για τη δομική και λειτουργική κατάταξη της ακολουθίας σε γονιδιωματικό επίπεδο [54].

3. Σύνολα δεδομένων για GPCRs (GPCR datasets)

Οι υποδοχείς που είναι συζευγμένοι με G-πρωτεΐνες (G- Protein coupled receptors ή GPCRs) αποτελούν μια μεγάλη ομάδα της οικογένειας των πρωτεϊνών που βρίσκονται στην επιφάνεια του κυττάρου και διαδραματίζουν σημαντικό ρόλο σε πολλά φυσιολογικά συστήματα καθώς μετατρέπουν ένα εξωκυττάριο σήμα σε μια ενδοκυττάρια απόκριση. Είναι εκτιμημένο ότι το ανθρώπινο γονιδίωμα περιέχει περισσότερο από 1000 γονίδια, τα οποία κωδικοποιούν την δομή των πρωτεϊνών των υποδοχέων που συνδέονται με G-πρωτεΐνες. Πάνω από το 50% των φαρμάκων που διατίθενται στο εμπόριο στοχεύουν άμεσα ένα GPCR [55].

Επειδή έχουν χαρακτηριστική διαμεμβρανική τοπολογία, οι GPCRs είναι επίσης γνωστοί ως επτά ελικοειδείς (heptahelical) υποδοχείς, επτά διαμεμβρανικοί υποδοχείς, 7TM υποδοχείς, και ελισσόμενοι ελικοειδείς (serpentine) υποδοχείς, επειδή ελίσσονται στην κυτταρική μεμβράνη επτά φορές. Ένας GPCR συζευγμένος με πρωτεΐνες έχει σημαντικό ρόλο στην διαμεσολάβηση για τη σηματοδότηση υποδοχέων, στην ρύθμιση της σηματοδότησης των υποδοχέων μέσω ελέγχου της τοποθεσίας του υποδοχέα ή και της διακίνησης του (trafficking). Επίσης δρα σαν scaffold και ως αλλοστερικός ρυθμιστής της δομής του υποδοχέα [56].

Εξαιτίας κυρίως της σημαντικότητας των GPCRs στην φαρμακολογία είναι επιθυμητή η ανάπτυξη ενός αλγορίθμου που θα μπορούσε να προβλέψει αποτελεσματικά την λειτουργία ενός GPCR από την πρωτογενή ακολουθία του [57]. Με δεδομένη μια μη χαρακτηρισμένη πρωτεϊνική ακολουθία, να μπορεί να προσδιοριστεί αν είναι GPCR ή όχι και αν είναι, σε ποια λειτουργική κλάση οικογενειών ανήκει [56].

Για να μπορεί να αναπτυχθεί ένας αποτελεσματικός αλγόριθμος για την ταξινόμηση των ακολουθιών GPCRs, είναι απαραίτητη η κατασκευή ενός συνόλου δεδομένων από ακολουθίες GPCRs, με το οποίο θα γίνει η εκπαίδευση και ο έλεγχος του αλγορίθμου. Για τη δημιουργία του συνόλου δεδομένων γίνεται ανάκτηση των πρωτεϊνικών ακολουθιών από βάσεις δεδομένων πρωτεϊνών [57]. Στη συνέχεια με

βάση τις απαιτήσεις των μεθόδων πρόβλεψης και ταξινόμησης που χρησιμοποιούνται περιορίζονται οι ακολουθίες, δίνοντας το τελικό σύνολο δεδομένων.

4. Σύνολα δεδομένων για OMPs (OMPs datasets)

Οι ενσωματωμένες μεμβρανικές πρωτεΐνες (Integral Membrane Proteins or IMPs) διαιρούνται σε δύο διακριτές δομικές κατηγορίες, τις α-ελικοειδή διαμεμβρανικές πρωτεΐνες και τα διαμεμβρανικά β-βαρέλια. Η πρώτη κατηγορία είναι η πιο άφθονη και καλά μελετημένη, δεδομένου ότι οι πρωτεΐνες του εν λόγω είδους βρίσκονται ως επί το πλείστον στις κυτταρικές μεμβράνες τόσο των προκαρυωτικών όσο και των ευκαρυωτικών οργανισμών, εκτελώντας μία ποικιλία σημαντικών βιολογικών λειτουργιών. Οι πρωτεΐνες αυτής της κατηγορίας έχουν τα διαμεμβρανικά τους τμήματα να εκτείνονται στις περιφέρειες σχηματίζοντας α-έλικες, οι οποίες αποτελούνται κυρίως από υδρόφοβα κατάλοιπα. Μια ποικιλία αλγορίθμων και υπολογιστικών τεχνικών έχουν προταθεί για την πρόβλεψη των διαμεμβρανικών τμημάτων των α-ελικοειδή διαμεμβρανικών πρωτεϊνών, με υψηλή ορθότητα και ακρίβεια προσέγγισης. Τα μέλη της τελευταίας κατηγορίας βρίσκονται στην εξωτερική μεμβράνη των αρνητικών κατά Gram βακτηρίων, και πιθανώς στην εξωτερική μεμβράνη των χλωροπλαστών και των μιτοχονδρίων. Τα μέλη αυτής της κατηγορίας έχουν τα διαμεμβρανικά τους τμήματα να εκτείνονται σε τομείς που σχηματίζουν αντιπαράλληλους β-κλώνους, δημιουργώντας ένα κανάλι σε μορφή κυλίνδρου που καλύπτει την εξωτερική μεμβράνη [37].

Οι πρωτεΐνες της εξωτερικής μεμβράνης (Outer Membrane Proteins ή OMPs) εκτελούν μια ποικιλία λειτουργιών, επιτρέποντας στα κύτταρα να αντιδρούν σε εξωτερικά ερεθίσματα και να ρυθμίζουν την είσοδο θρεπτικών ουσιών καθώς και την παθητική μεταφορά ιόντων και μικρών μορίων, επιτρέποντας έτσι την επιλεκτική διέλευση μορίων όπως η μαλτόζη και η σακχαρόζη [58].

Εξαιτίας των σημαντικών βιολογικών λειτουργιών στις οποίες εμπλέκονται οι πρωτεΐνες της εξωτερικής μεμβράνης, προσελκύουν αυξημένο ιατρικό ενδιαφέρον. Αυτό επιβεβαιώνεται από τον συνεχώς αυξανόμενο αριθμό των πλήρως αλληλουχημένων γονιδιωμάτων αρνητικών κατά gram βακτηρίων που κατατίθεται στις δημόσιες βάσεις δεδομένων. Από την άλλη πλευρά, η εκτεταμένη μελέτη της δομής των διαμεμβρανικές β-βαρελιών, θα μπορούσε να αποκαλύψει ειδικές πτυχές της διαδικασίας της αναδίπλωσης των πρωτεϊνών, και να μας δώσει χρήσιμες

πληροφορίες σχετικά με τη δομή και τη λειτουργία των πρωτεϊνών. Για τους λόγους αυτούς, είναι σαφές ότι υπάρχει ανάγκη ανάπτυξης υπολογιστικών εργαλείων για την πρόβλεψη των εν λόγω πρωτεϊνών, καθώς επίσης και τη διάκριση τους από τις υδατοδιαλυτές πρωτεΐνες όταν γίνεται αναζήτηση σε πλήρη γονιδιώματα [37].

Μια συγκριτική ανάλυση για την κατανομή των αμινοξικών καταλοίπων σε α-ελικοειδείς διαμεμβρανικές πρωτεΐνες και σε διαμεμβρανικά β-βαρέλια, δείχνει ότι η μεμβρανική περιοχή στις OMPs είναι πιο περίπλοκη από εκείνη στις ελικοειδείς διαμεμβρανικές πρωτεΐνες εξαιτίας της παρεμβολής πολλών φορτισμένων και πολικών σημάτων στη μεμβράνη. Κατά συνέπεια, το ποσοστό επιτυχίας της διάκρισης μιας διαμεμβρανικά β-βαρέλι πρωτεΐνης από άλλες πρωτεΐνες είναι σημαντικά χαμηλότερη από εκείνη της α-ελικοειδούς διαμεμβρανικής πρωτεΐνης [58].

5. Πρόσβαση στα Σύνολα Δεδομένων

Σήμερα το διαδίκτυο αποτελεί πολύτιμο και απαραίτητο πόρο χρήσης στην πραγματοποίηση ερευνών, καθώς παρέχει πληθώρα πληροφοριών που μπορούν να ανακτηθούν εύκολα και γρήγορα, κάτι που ήταν ανέφικτο μέσα από τα έντυπα μέσα. Οι πόροι του διαδικτύου είναι μέρος μιας νέας επανάστασης στην επιστημονική έρευνα που αυξάνει την προσβασιμότητα στα δεδομένα και παρέχει πρόσβαση σε εργαλεία που βοηθούν στην επιστημονική έρευνα. Ωστόσο, οι πόροι αυτοί παρουσιάζουν επίσης νέες προκλήσεις στη διαδικασία αξιολόγησης των δεδομένων και επιπλέον, επειδή αυτοί οι ηλεκτρονικοί πόροι μπορούν να εξαφανιστούν εντελώς, παρέχουν μια μοναδική πρόκληση όσον αφορά τη διατήρηση και την αξιολόγηση τους. Οποιοσδήποτε χρήστης του Διαδικτύου θα έχει συναντήσει το παραδοσιακό μήνυμα «404 not found» που επιστρέφεται από έναν διακομιστή ιστού σαν απάντηση σε μια διεύθυνση URL που δεν είναι πλέον διαθέσιμη [59].

Οι λόγοι για τους οποίους μια διεύθυνση δεν είναι πλέον διαθέσιμη προκύπτουν από σφάλματα στη μορφοποίηση της διεύθυνσης URL, ορθογραφικά λάθη και μετακίνησης της ιστοσελίδας σε καινούργια διεύθυνση, χωρίς ανακατεύθυνση από την παλιά ιστοσελίδα στην καινούργια. Επιπλέον μπορεί να προκύπτουν και από τους ίδιους τους συγγραφείς και δημιουργούς των ιστοσελίδων, καθώς μπορεί να φύγουν από τα ιδρύματα στα οποία αναπτύσσουν τους διαδικτυακούς πόρους και ενώ το ενδιαφέρον τους για την συντήρηση του URL μπορεί να συνεχιστεί, για νομικούς λόγους, δεν είναι εφικτό.

Η πρόσβαση στα σύνολα δεδομένων βιολογικών ακολουθιών μπορεί να γίνει μέσα από την αναζήτηση στην βιβλιογραφία, και στη συνέχεια με επίσκεψη στις ιστοσελίδες που κατασκευάστηκαν από τους επιστήμονες που ασχολούνται με τις προγνωστικές μεθόδους.

Η αναζήτηση μέσα από τη βιβλιογραφία αποτελεί χρονοβόρα διαδικασία και δεν έχει πάντα θετικά αποτελέσματα, καθώς η ανακατεύθυνση που δίνει η βιβλιογραφία στις ιστοσελίδες που φιλοξενούνται τα σύνολα δεδομένων δεν είναι πάντα διαθέσιμη, για τους λόγους που αναφέρθηκαν πιο πάνω, με αποτέλεσμα οι δημιουργοί των ιστοσελίδων να σταματούν να συντηρούν τις ιστοσελίδες και να μην είναι πλέον δυνατή η πρόσβαση στα σύνολα δεδομένων.

Είναι σημαντικό να μπορεί κάποιος να αναζητήσει με βάση μια κατηγορία βιολογικού προβλήματος και να έχει στη διάθεση του εύκολα και γρήγορα τον μέγιστο αριθμό συνόλων δεδομένων που αναφέρονται στο πρόβλημα αυτό.

Εξαιτίας των προβλημάτων που παρουσιάζονται στην προσβασιμότητα των συνόλων δεδομένων στην παρούσα πτυχιακή εργασία δημιουργήθηκε μια βάση δεδομένων με σκοπό την καταχώρηση των συνόλων δεδομένων των πρωτεϊνικών ακολουθιών, που έχουν ήδη υλοποιηθεί και χρησιμοποιηθεί από τους επιστήμονες για την εκπαίδευση και τον έλεγχο αλγορίθμων πρόγνωσης και έχουν δημοσιευθεί στο παρελθόν στη βιβλιογραφία. Η συγκέντρωση τους σε μια βάση κάνει πιο εύκολη την αναζήτηση, την πρόσβαση και τέλος την ανάκτηση των συνόλων δεδομένων.

Με αυτό τον τρόπο όταν κάποιος επιστήμονας θελήσει να συγκρίνει τα αποτελέσματα του δικού του αλγορίθμου με ένα άλλο αλγόριθμο ή ακόμα και να χρησιμοποιήσει κάποια από τα σύνολα δεδομένων που έχουν ξανά χρησιμοποιηθεί για να εκπαιδεύσει και να ελέγξει τον αλγόριθμο πρόγνωσης που έχει υλοποιήσει, μπορεί να το κάνει μέσα από την ιστοσελίδα στην οποία και φιλοξενείται η βάση δεδομένων με τα σύνολα δεδομένων.

ΜΕΡΟΣ 4ο
ΥΛΟΠΟΙΗΣΗ

1. Εισαγωγή

Η χρήση των συνόλων δεδομένων στην Βιοπληροφορική αποτελεί σημαντικό εργαλείο πρόβλεψης και ταξινόμησης των βιολογικών ακολουθιών. Εξαιτίας όμως, του όγκου δεδομένων που υπάρχει διασκορπισμένος σε ψηφιακή μορφή, γίνεται ολοένα και πιο δύσκολη η αναζήτηση και εξαγωγή τέτοιων δεδομένων. Σκοπός της παρούσας πτυχιακής εργασίας είναι η δημιουργία μιας βάσης δεδομένων στην οποία θα καταχωρούνται τα σύνολα δεδομένων των βιολογικών ακολουθιών που έχουν εξαχθεί από επιστημονικές μελέτες.

2. Περιγραφή Δεδομένων

Για την μετατροπή όλων των συνόλων δεδομένων σε μια μορφή και συγκεκριμένα σε αρχεία Fasta, έγινε χρήση της γλώσσα προγραμματισμού Perl. Για την υλοποίηση της βάσης δεδομένων στην οποία καταχωρήθηκαν τα σύνολα δεδομένων έγινε χρήση της γλώσσας σήμανσης HTML, των γλωσσών προγραμματισμού PHP, sql και JavaScript και του εργαλείου XAMPP.

2.1.Γλώσσα προγραμματισμού Perl

Η Perl (Practical Extraction and Report Language) είναι μια γλώσσα προγραμματισμού γενικής χρήσης που αναπτύχθηκε και σχεδιάστηκε αρχικά για την επεξεργασία κειμένου, ενώ τώρα χρησιμοποιείται σε ένα ευρύ φάσμα εργασιών, συμπεριλαμβανομένων της διαχείρισης συστημάτων (system administration), της κατασκευής ιστοσελίδων (web development), του προγραμματισμού δικτύων (network programming), της ανάπτυξης γραφικού περιβάλλοντος χρήστη (GUI development) και πολλών άλλων.

Δημιουργήθηκε από τον Larry Wall και η πρώτη έκδοση κυκλοφόρησε το 1987. Αποτελεί μια διερμηνευόμενη γλώσσα (interpretive language) και διατίθεται δωρέαν. Τρέχει σε μια ποικιλία από πλατφόρμες λειτουργικών συστήματων, όπως τα Windows, τα Mac OS, και τις διάφορες εκδόσεις του UNIX [60].

2.2.Αρχεία Fasta

Στην Βιοπληροφορική, η μορφή Fasta είναι μια μορφή που βασίζεται σε κείμενο και περιέχει πρωτεϊνικές ή νουκλεοτιδικές ακολουθίες.. Η μορφή προέρχεται από το πακέτο λογισμικού FASTA, αλλά τώρα έχει γίνει πρότυπο στον τομέα της βιοπληροφορικής. Η απλότητα της μορφής FASTA καθιστά εύκολο τον χηρισμό και την ανάλυση των ακολουθιών χρησιμοποιώντας εργαλεία επεξεργασίας κειμένου και γλώσσες scripting, όπως το Python, η Ruby, και η Perl.

Η γραμμή περιγραφής ή γραμμή κεφαλίδας, που αρχίζει με «>», δίνει ένα όνομα ή και ένα μοναδικό αναγνωριστικό για την ακολουθία, και μπορεί επίσης να περιέχει πρόσθετες πληροφορίες. Μετά τη γραμμή κεφαλίδας και σχολίων, σε μία ή περισσότερες γραμμές ακολουθεί η ακολουθία με κάθε γραμμή να έχει λιγότερο από 80 χαρακτήρες. Το NCBI ορίζει ένα πρότυπο για το μοναδικό αναγνωριστικό που χρησιμοποιείται για την ακολουθία (SeqID) στη γραμμή κεφαλίδας [Πίνακας 4] [61].

2.3.Γλώσσα σήμανσης HTML

Η HTML (HyperText Markup Language) είναι η κύρια γλώσσα σήμανσης υπερκειμένου για την δημιουργία ιστοσελίδων. Είναι μια περιγραφική γλώσσα, δηλαδή ένας ειδικός τρόπος γραφής κειμένου και κλήσης άλλων αρχείων ή εφαρμογών βασισμένος σε ετικέτες (tags). Ο Web client αναγνωρίζει αυτόν τον ειδικό τρόπο γραφής και εκτελεί τις εντολές που περιέχονται σε αυτόν. Οι ετικέτες περικλείονται μέσα σε σύμβολα «μεγαλύτερο από» και «μικρότερο από» και συνήθως λειτουργούν ανά ζεύγη με την πρώτη να ονομάζεται ετικέτα έναρξης και τη δεύτερη ετικέτα λήξης ή σε κάποιες περιπτώσεις ετικέτα ανοίγματος και ετικέτα κλεισίματος αντίστοιχα. Οι ετικέτες δεν επηρεάζονται από το αν έχουν γραφτεί με πεζά (μικρά) ή κεφαλαία γράμματα. Ανάμεσα στις ετικέτες, οι σχεδιαστές ιστοσελίδων μπορούν να τοποθετήσουν κείμενο, πίνακες, εικόνες κλπ.

Η HTML μπορεί να αντιπροσωπεύσει υπερκείμενα ειδήσεων, αλληλογραφία, κείμενα και υπερμέσα, μενού επιλογών, αποτελέσματα ερωτημάτων βάσεων δεδομένων και απλά δομημένα έγγραφα με γραφικά. Η χρήση της στον παγκόσμιο ιστό (World Wide Web) γίνεται από το 1990 [62].

2.4.Γλώσσα προγραμματισμού PHP

Η PHP (Hypertext Preprocessor) είναι μια server side scripting γλώσσα προγραμματισμού που χρησιμοποιείται συνήθως για τη δημιουργία ιστοσελίδων

δυναμικού περιεχομένου. Η PHP μπορεί να χρησιμοποιηθεί για τη δημιουργία εφαρμογών γραμμής εντολών και επιφάνειας εργασίας GUI μαζί με τις δυναμικές ιστοσελίδες. Είναι επίσης πολύ ευέλικτη και μπορεί να εγκατασταθεί και να ρυθμιστεί, σε σχεδόν, οποιοδήποτε λειτουργικό σύστημα. Επιτρέπει στους χρήστες να αλληλεπιδρούν περισσότερο με το διαδίκτυο και να δημιουργούν σελίδες με λιγότερη προσπάθεια [63].

2.5.Γλώσσα στυλ CSS

Η γλώσσα CSS (Cascading Style Sheets) είναι μια γλώσσα στυλ και χρησιμοποιείται για να διαμορφωθεί η εμφάνιση ενός αρχείου που έχει γραφτεί σε γλώσσα σήμανσης. Έχει αποδειχθεί ότι είναι ένα από τα πιο ισχυρά εργαλεία που είναι διαθέσιμα για τον σχεδιασμό ιστοσελίδων. Είναι υπεύθυνη για τη διαμόρφωση των χρωμάτων, τη στοίχιση, το μεγέθος της γραμματοσειράς, το στυλ της γραμματοσειράς και γενικά τη στυλιστική εμφάνιση μιας ιστοσελίδας [64].

2.6.Γλώσσα προγραμματισμού SQL

Η SQL (Structured Query Language) είναι μια γλώσσα προγραμματισμού που σχεδιάστηκε για τη διαχείριση σχεσιακών βάσεων δεδομένων και η οποία βασίστηκε στη σχεσιακή άλγεβρα και στο σχεσιακό λογισμό πλειάδων. Χρησιμοποιείται για τη δημιουργία, τροποποίηση και ενημέρωση δεδομένων και για τη διατύπωση ερωτημάτων. Έχει απλή σύνταξη και αποδεσμεύει το χρήστη από λεπτομέρειες υλοποίησης. Χρησιμοποιεί τους όρους πίνακας, γραμμή και στήλη, οι οποίοι αντιπροσωπεύουν τις έννοιες σχέση, πλειάδα και χαρακτηριστικό αντιστοίχως [65].

2.7.Γλώσσα προγραμματισμού JavaScript

Η JavaScript είναι μια διερμηνευμένη γλώσσα προγραμματισμού, με την οποία μπορεί να δημιουργηθούν σενάρια που να εκτελούν αυτόματες εργασίες και ενέργειες ανταποκρινόμενα σε ένα συγκεκριμένο γεγονός. Γενικά, τα σενάρια εκτελούνται αμέσως μόλις μια ιστοσελίδα φορτωθεί σε ένα φυλλομετρητή. Σε αρκετές περιπτώσεις είναι επιθυμητό ένα σενάριο να εκτελείται όταν φορτωθεί η ιστοσελίδα, ενώ σε άλλες περιπτώσεις να εκτελείται όταν ο χρήστης κάνει κάποια ενέργεια [66].

2.8.XAMPP

Το XAMPP (X (cross-platform), Apache HTTP εξυπηρετητής, MySQL, PHP, Perl) αποτελεί ένα πακέτο προγραμμάτων ελεύθερου λογισμικού, λογισμικού ανοικτού κώδικα και ανεξάρτητης πλατφόρμας που περιέχει Apache HTTP Server, PHP, MySQL, phpMyAdmin, Openssl, και SQLite. Το XAMPP είναι ένα ελεύθερο λογισμικό ανεξάρτητης πλατφόρμας και τρέχει σε όλα τα διαθέσιμα λειτουργικά συστήματα. Αποτελεί εργαλείο ανάπτυξης και δοκιμής ιστοσελίδων τοπικά (localhost) στον υπολογιστή χωρίς να είναι απαραίτητη η σύνδεση στο διαδίκτυο. Ορισμένες φορές όμως χρησιμοποιείται και για την φιλοξενία ιστοσελίδων [67].

3. Μεθοδολογία

Χρησιμοποιώντας ως λέξεις κλειδιά τους όρους «*dataset*», «*protein prediction*», «*protein classification*», αλλά και ειδικές λέξεις κλειδιά που αναφέρονται στο βιολογικό πρόβλημα όπως «*membrane proteins*», «*signal peptides*», «*GPCRs*», «*OMPs*», «*Post-translational modifications*, «*non-classical protein secretion*», «*DNA/RNA binding proteins*», «*subcellular localization*» και «*secondary structure prediction or classification*» πραγματοποιήθηκε αναζήτηση στις βιολογικές βάσεις δεδομένων και κυρίως στην βιβλιογραφία για την ανάκτηση των συνόλων δεδομένων πρωτεϊνικών ακολουθιών. Στην συνέχεια δημιουργήθηκε ένας πίνακας στον οποίο καταχωρήθηκαν τα δεδομένα που συλλέχθηκαν από την αναζήτηση [Πίνακας 5]. Ο πίνακας αποτελείτε από τα εξής πεδία:

- Το όνομα του συνόλου δεδομένων, Dataset Name, το οποίο αποτελεί πληροφορία από την βιβλιογραφία του συνόλου δεδομένων και μπορεί να είναι είτε το όνομα που δόθηκε από τους συγγραφείς είτε το όνομα που δόθηκε στον αλγόριθμο πρόβλεψης ή ταξινόμησης. Στην περίπτωση που δεν υπήρχε κάποιο στοιχείο για το όνομα του συνόλου δεδομένων το πεδίο αυτό παρέμεινε κενό.
- Τον κωδικό του συνόλου δεδομένων, Dataset ID που αποτελεί μοναδικό αλφαριθμητικό κωδικό για το κάθε σύνολο ξεχωριστά και έχει δοθεί με σκοπό την καλύτερη αναζήτηση πληροφοριών για κάθε σύνολο δεδομένων στην βάση.
- Το είδος της βιολογικής ακολουθίας, Sequence Type που μπορεί να είναι μια ακολουθία νουκλεϊνικών οξέων, (DNA ή RNA), ή μια ακολουθία αμινοξέων

(πρωτεΐνη). Στην παρούσα εργασία το πεδίο αυτό αποτελείτε μόνο από πρωτεϊνικές ακολουθίες.

- Το πεδίο τύπος βιολογικού προβλήματος, Problem Type, το οποίο αποτελεί την μέθοδο πρόβλεψης (prediction) ή ταξινόμησης (classification) που έχει χρησιμοποιηθεί στην συγκεκριμένη μελέτη και είχε ως αποτέλεσμα την δημιουργία του συνόλου δεδομένων.
- Το βιολογικό πρόβλημα, Biological Problem, στο οποίο γίνεται μια σύντομη αναφορά του βιολογικού προβλήματος το οποίο αναφέρεται το συγκεκριμένο άρθρο.
- Τη μορφή δεδομένων, Data, που μπορεί να είναι fasta format, IDs που έχουν εξαχθεί από τις βιολογικές βάσεις δεδομένων και άλλες μορφές βιολογικών δεδομένων.
- Την περιγραφή του συνόλου δεδομένων, Description, στην οποία γίνεται μια σύντομη περιγραφή του συνόλου δεδομένων. Από πόσες πρωτεϊνικές ακολουθίες και υποσύνολα δεδομένων αποτελείτε και σε πόσες κατηγορίες – οικογένειες υποδιαιρείτε ο αριθμός των ακολουθιών αυτών.
- Τη βιβλιογραφία, References, η οποία αναφέρεται στα άρθρα που σχετίζονται με το συγκεκριμένο σύνολο δεδομένων και αποτελεί το PubMed ID των αντίστοιχων άρθρων.
- Την ηλεκτρονική σελίδα από την οποία μπορείς να κατεβάσεις το σύνολο δεδομένων, URL και
- Την ηλεκτρονική σελίδα στην οποία υπάρχει ο αλγόριθμος πρόβλεψης ή ταξινόμησης, web server.

Εξαιτίας του ότι τα σύνολα δεδομένων δεν είχαν όλα την ίδια μορφή δεδομένων, χρειάστηκαν περαιτέρω επεξεργασία, με σκοπό την δημιουργία μιας κοινής μορφής. Όσα σύνολα δεδομένων ήταν σε απλή μορφή fasta δεν έτυχαν επεξεργασίας, ενώ αυτά που ήταν σε οποιαδήποτε μορφή fasta label μετατράπηκαν σε απλή fasta. Για τα υπόλοιπα αρχεία χρησιμοποιήθηκε η γλώσσα προγραμματισμού Perl για την δημιουργία script αρχείων [Script αρχεία], καθώς είχαν μόνο τους κωδικούς των πρωτεϊνικών ακολουθιών και για να μετατραπούν σε αρχεία fasta, χρειαζόταν να γίνει η εξαγωγή των ακολουθιών από τις βιολογικές βάσεις που αντιστοιχούσαν οι συγκεκριμένοι κωδικοί [Εικόνα 4]. Μετά το τέλος της διαδικασίας αυτής παρουσιάζονται όλα τα σύνολα δεδομένων σε απλή μορφή fasta [Εικόνα 5].

Εικόνα 4: Κωδικοί πρωτεϊνικών ακολουθιών

UniProt IDs:			PDB IDs:		
Q99VY4	O34385	O32167	2ADF:A	1JRH:I	1SY6:A
P29230	O34335	P54941	2ADF:A	1LK3:A	1TZI:V
Q8VQS9	P24141	Q600S6	1AFV:A	1MHP:B	1WEJ:F
Q5HET4	P36949	Q5ZZQ6	1BGX:T	1NL0:G	1YJD:C
Q2FI86	O34966	Q8KVR9	1E6J:P	1NSN:S	1YNT:F
Q7A603	O05497	O05121	1EGJ:A	1OAZ:A	1YY9:A
Q99U04	O31567	Q50327	1FSK:A	1ORQ:C	1ZA3:R
Q9ZEP5	O34348	P55801	1H0D:C	1ORS:C	1ZTX:E
P40409	P54535	P0A671	1I9R:A	1PKQ:E	2JEL:P
P37580	O05410	P21625	1IQD:C	1RJL:C	1A14:N

Εικόνα 5: Μορφή αρχείου FASTA

```
>gi|476885|pir||A46390 cAMP receptor subtype 2, CAR2 - slime mold (Dictyostelium discoideum)
MTIMSDIIAQRTILLIADFSSIIGCSLVLIGFWRLKLLRNHITKIISLFCATSLFKDVISTIITLLYKPDQTESGFPCYLHAIVI
TFGSLACWLWTLMLSFSIYNLIVRREPEPERFEKFYFCLCYGLPLISTIVMLSTHIIQPVGGWCWIGDNYDGYRFGLF
YGPFFFIWGTSAILVGLTSKYTYSVIRSSVSDNKDKHMTYQFKLINYIVVFLVCWVFAIVNRILNGLNQFPTVPNVLH
TYFSVSHGFYASITFIYNNPLMWRYFGAKFLLIFSKFGLFVQAQQRLELNKNNNNPSPIMRSKNALDNGADSSVVE
LPCLSKADSLSLDAENNIETPKENENQNHHHHHHHHHHHNHYNNNNNNNNNINNKNDMI
>gi|60470107|gb|EAL68087.1| cAMP receptor [Dictyostelium discoideum AX4]
MENLNTTSTAALTGMTKQENDASYAVLLIADFTSIIGCTLVLLGFWRLKLLRNHITKIITFFCSTSLAKDLISTILTLIEKK
QSNGSFQCYLYATVITYGSLACWLWTLCLSFSIYNLIVKREPEPEKFEKYYHVFCWVVPFIMSVIMLSKGVIEVTGN
WCWIGNTYVGYRFGLFYGPFLAIWFLAAVLVGLTSRYTYKVIRSSVSDNKDRHMTYQFKLINYIIVFLLCWVFAVIN
RIVNGLNMFPAWVSILHTYLSVSHGFYASVTFIYNNPLMWRYLASIILIPFTKFGYFVETQQRLEKNKNNNNHSPV
GLSNNAQNNNHHHNHNNNHNNNHNNHNNNNNNNNNSDFVNNDSSNYYTASMIESFSVQNENSKSINGADNF
KQNGASQQDDKDSPNSNNNNNNNNNNNNNNNNNNNNNNNNNNNNYNNKDIEPIDNCNTNSIPMDNIATRIEI
PPQHPTLTPQQSLQEINLNDDDNKINTHQSNKKKDSNV
```

Στην συνέχεια δημιουργήθηκε η βάση δεδομένων χρησιμοποιώντας το εργαλείο XAMPP και η γλώσσα προγραμματισμού sql, ενώ η παρουσίαση της στην ιστοσελίδα έγινε με την χρήση html στην οποία και έχουν πρόσβαση οι χρήστες που ενδιαφέρονται να κατεβάσουν τα σύνολα δεδομένων. Κατόπιν με την χρήση php και javascript έγινε η σύνδεση μεταξύ της βάσης δεδομένων και της ιστοσελίδας.

4. Ηλεκτρονική Σελίδα

Για τη δημιουργία της ιστοσελίδας χρησιμοποιήθηκε η γλώσσα προγραμματισμού html, με σκοπό την υλοποίηση της πλευράς του χρήστη και η PHP για την υλοποίηση της πλευράς του server. Κατόπιν δημιουργήθηκαν τα απαραίτητα ερωτήματα σε γλώσσα sql, με σκοπό την εξαγωγή πληροφοριών από τη βάση δεδομένων dataset που δημιουργήθηκε με τη χρήση των εργαλείων XAMPP και phpMyAdmin.

Η πλευρά του χρήστη αποτελείται από τη σελίδα υποδοχής (Home) [Εικόνα 6], τη σελίδα στην οποία παρουσιάζονται σε μορφή πίνακα τα σύνολα δεδομένων (DataSets) [Εικόνα 7], τη σελίδα στην οποία μπορεί κάποιος να κάνει αναζήτηση στη βάση δεδομένων (Search) [Εικόνα 8] χρησιμοποιώντας λέξεις κλειδιά, με σκοπό την εξαγωγή περισσότερων πληροφοριών, σχετικών με τα σύνολα δεδομένων, τη σελίδα στην οποία ο χρήστης μπορεί να κατεβάσει (Downloads) [Εικόνα 9] τα σύνολα δεδομένων στην αρχική τους μορφή και σε μορφή fasta και τέλος τη σελίδα με τη φόρμα επικοινωνίας (Contact) [Εικόνα 10] στην οποία μπορεί να επικοινωνήσει ο χρήστης με τον διαχειριστή για απορίες σχετικές με τα σύνολα δεδομένων.

Για την παρουσίαση των συνόλων δεδομένων στη σελίδα DataSets έγινε χρήση της γλώσσας προγραμματισμού javaScript με σκοπό την εμφάνιση των δεδομένων κατά την φόρτωση της σελίδας. Επιπλέον χρήση της javaScript έγινε και στην σελίδα Search όπου ο χρήστης κάνει αναζήτηση στην βάση για περισσότερες πληροφορίες όσον αφορά το σύνολο δεδομένων που τον ενδιαφέρει.

Η εμφάνιση και το στυλ της ιστοσελίδας σχεδιάστηκε με τη χρήση της γλώσσας στυλ css. Το κυρίως μενού τοποθετήθηκε στο πάνω μέρος της ιστοσελίδας σε χρώμα πλαισίου μαύρο και χρώμα γραμματοσειράς λευκό, ενώ η σελίδα στην οποία βρίσκεται ο χρήστης την συγκεκριμένη χρονική στιγμή περιήγησης περικλείεται από κόκκινο πλαίσιο. Κάτω ακριβώς από το μενού τοποθετήθηκε μια εικόνα, κοινή σε όλες τις σελίδες του ιστότοπου και για την υπόδειξη του σώματος της κάθε σελίδας ο τίτλος χρωματίστηκε με κόκκινο. Με τη χρήση της css δημιουργήθηκε πιο εύκολα το σώμα της ιστοσελίδας, στο οποίο και παρουσιάζονται τα σύνολα δεδομένων σε απλό και εύχρηστο περιβάλλον.

Η ιστοσελίδα που δημιουργήθηκε αποτελεί εύχρηστο εργαλείο, αναζήτησης και εξαγωγής των συνόλων δεδομένων των πρωτεϊνικών ακολουθιών.

5. Αποτελέσματα

Με την αναζήτηση που πραγματοποιήθηκε έγινε ανάκτηση εκατόν ογδόντα συνόλων δεδομένων πρωτεϊνικών ακολουθιών. Με βάση τα δεδομένα που περιέχει κάθε σύνολο δεδομένων και την περιγραφή του από τις ερευνητικές εργασίες μπορεί να ταξινομηθεί σε μια ευρύτερη ομάδα, όπως φαίνεται και στον πιο κάτω πίνακα.

Πίνακας 1: Ομαδοποίηση συνόλων δεδομένων

Ομάδα Πρωτεϊνών	Κωδικοί Συνόλων Δεδομένων
Υποδοχείς που συνδέονται με G-πρωτεΐνες (GPCRs)	DSDB0001, DSDB0002, DSDB0004, DSDB0005, DSDB0006, DSDB0007, DSDB0008
Σηματοδοτικές ακολουθίες (signal peptides)	DSDB0003, DSDB0017, DSDB0023, DSDB0041, DSDB0069, DSDB0070, DSDB0071, DSDB0072, DSDB0073, DSDB0104, DSDB0108
Πρωτεΐνες της εξωτερικής μεμβράνης (OMPs)	DSDB0009, DSDB0010
Διαμεμβρανικές πρωτεΐνες (transmembrane proteins)	DSDB0011, DSDB0012, DSDB0013, DSDB0014, DSDB0015, DSDB0016, DSDB0018, DSDB0019, DSDB0020, DSDB0021, DSDB0022, DSDB0024, DSDB0025, DSDB0104
Πυρηνικές πρωτεΐνες (NLS, Nuclear Localization Sequence)	DSDB0027, DSDB0028
Υπεροξυσωμικό σήμα στόχευσης (PTS, peroxisomal targeting signal)	DSDB0029, DSDB0030, DSDB0031, DSDB0032, DSDB0033
Δευτεροταγής δομή πρωτεΐνης (Protein secondary structure)	DSDB0034
Μετα Μεταφραστική τροποποίηση (PTMs, Post-translational modifications) και γλυκοζυλίωση (Glycosylation)	DSDB0035, DSDB0036, DSDB0037, DSDB0038, DSDB0039, DSDB0040, DSDB0044, DSDB0045, DSDB0046, DSDB0048, DSDB0049, DSDB0050, DSDB0051, DSDB0052, DSDB0053, DSDB0054, DSDB0055, DSDB0056, DSDB0058, DSDB0059, DSDB0060, DSDB0061, DSDB0062, DSDB0063, DSDB0064, DSDB0067
Επαφή μεταξύ των κατάλοιπων (protein residue-residue contact)	DSDB0042, DSDB0115
Φυλογενετική (Phylogenetic)	DSDB0043
Πρόβλεψη των κρυμμένων και μη κατάλοιπων (prediction of buried vs exposed residues)	DSDB0047
Επιτόπιο των β κυττάρων (epitope prediction)	DSDB0057, DSDB0065
Πρωτεΐνες που δεσμεύονται με ATP (ATP binding proteins)	DSDB0068
Μη κλασσικής έκκρισης πρωτεΐνες (non-classical protein secretion)	DSDB0074, DSDB0075, DSDB0076, DSDB0077, DSDB0078, DSDB0079, DSDB0080, DSDB0081, DSDB0082, DSDB0083, DSDB0084, DSDB0085, DSDB0086, DSDB0087, DSDB0088, DSDB0089
Αναδίπλωση των πρωτεϊνών (protein fold)	DSDB0090, DSDB0091, DSDB0092, DSDB0106, DSDB0114
Αλλαγή στην σταθερότητα των πρωτεϊνών κατά τις μεταλλάξεις (protein stability changes upon mutations)	DSDB0093, DSDB0094, DSDB0095, DSDB0096, DSDB0097, DSDB0098, DSDB0099, DSDB0100, DSDB0101, DSDB0102, DSDB0109, DSDB0111, DSDB0112
Πρόβλεψη στοχευμένων πεπτιδίων (targeting peptides prediction)	DSDB0113
Ανακατασκευή της τρισδιάστατης δομής της πρωτεΐνης (three dimensional structure reconstruction)	DSDB0105, DSDB0106
Πρόβλεψη πρωτεϊνών τεταρτοταγούς δομής (protein quaternary structure prediction)	DSDB0116
Δέσμευση των πρωτεϊνών στο DNA (DNA binding proteins)	DSDB0117, DSDB0118, DSDB0119, DSDB0120, DSDB0122, DSDB0123, DSDB0124, DSDB0125, DSDB0126, DSDB0128, DSDB0129, DSDB0130, DSDB0131
Προσβασιμότητα του Διαλύτη (solvent accessibility)	DSDB0127, DSDB0132, DSDB0133

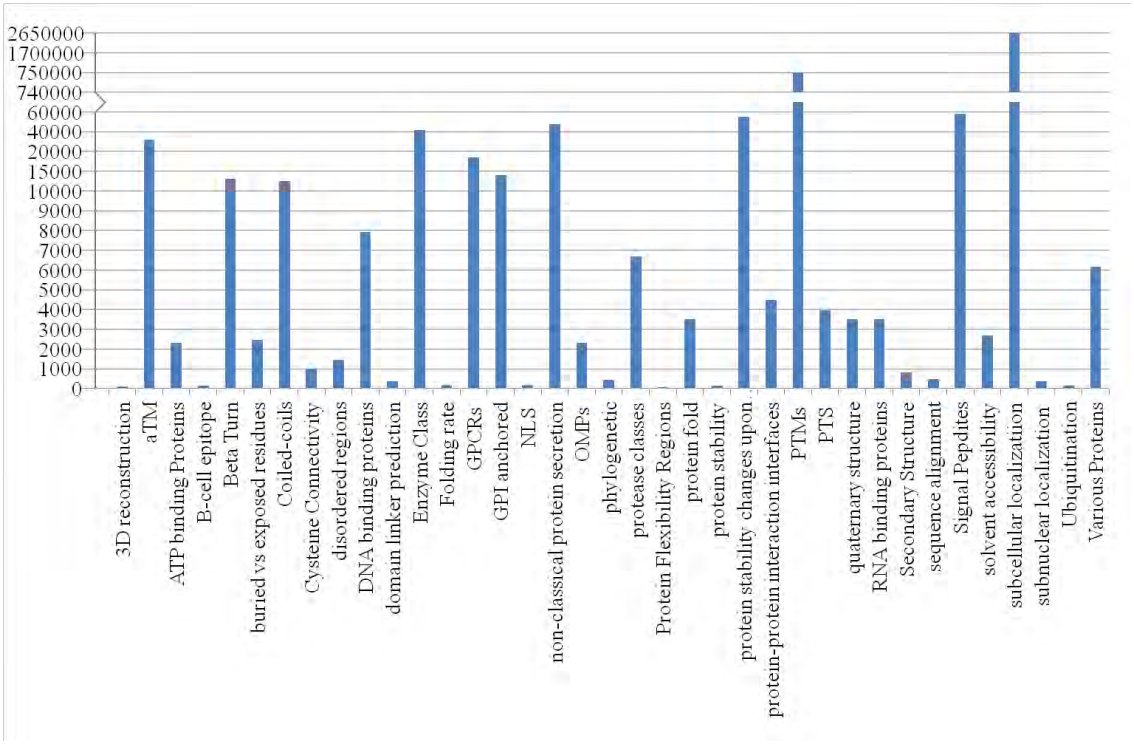
Δέσμευση πρωτεϊνών στο RNA (RNA binding proteins)	DSDB0134, DSDB0135, DSDB0136, DSDB0137, DSDB0138
Αλληλεπιδράσεις μεταξύ των πρωτεϊνών (Protein-protein interaction interfaces)	DSDB0139
Πρόβλεψη υποκκυταρικού εντοπισμού της πρωτεΐνης (subcellular localization prediction)	DSDB0026, DSDB0103, DSDB0107, DSDB0108, DSDB0140, DSDB0141, DSDB0142, DSDB0143, DSDB0144, DSDB0145, DSDB0146, DSDB0147, DSDB0148, DSDB0149, DSDB0150, DSDB0151, DSDB0152, DSDB0153
Υποπυρηνικός εντοπισμός (Subnuclear localization)	DSDB0154
Κατηγορίες πρωτεάσης (protease classes)	DSDB0155, DSDB0156
Ύπαρξη υπερελίκων (Coiled-coils)	DSDB0157, DSDB0158, DSDB0159
Πρόβλεψη των βήτα στροφών (Beta Turn prediction)	DSDB0161, DSDB0162, DSDB0163, DSDB0164, DSDB0165
GPI αγκυροβολημένες πρωτεΐνες (GPI-anchored)	DSDB0110, DSDB0166, DSDB0167
Πρόβλεψη συνδετικών τμημάτων (domain linker prediction)	DSDB0168, DSDB0169
Πρόβλεψη των τμημάτων των δομικά διαταραγμένων πρωτεϊνών (Prediction of Protein Binding Regions in Disordered Proteins)	DSDB0170, DSDB0171
Δομική στοίχιση (Protein Structure Alignments)	DSDB0172, DSDB0173, DSDB0174
Ευελιξία των πρωτεϊνών σε μη δομημένες περιοχές (Protein Flexibility Regions)	DSDB0175
Ένζυμα (Enzyme class)	DSDB0176, DSDB0177
Συνδεσιμότητα της κυστεΐνης (Cysteine connectivity)	DSDB0178
Ουμπικουιτίνη (Ubiquitination)	DSDB0066
Διάφορες πρωτεΐνες (Various Proteins)	DSDB0121, DSDB0160, DSDB0179, DSDB0180

Με τη χρήση περιγραφικών στατιστικών έγινε περαιτέρω ομαδοποίηση των δεδομένων με σκοπό την ανάλυση μέσω της γραφικής παρουσίασης τους. Για το λόγο αυτό έγινε ομαδοποίηση των δεδομένων ως προς τον αριθμό των πρωτεϊνών για κάθε κατηγορία, τον αριθμό των συνόλων ανά την ευρύτερη ομάδα στην οποία ανήκουν, πόσα σύνολά δεδομένων δημοσιεύθηκαν ανά έτος και τον αριθμό των συνόλων ανά κατηγορία του προβλήματος πρόβλεψης, αν είναι δηλαδή Prediction ή Classification. Κατόπιν με την χρήση ραβδογράμματων, ιστογραμμάτων και κυκλικών διαγραμμάτων έγινε η γραφική παρουσίαση των δεδομένων.

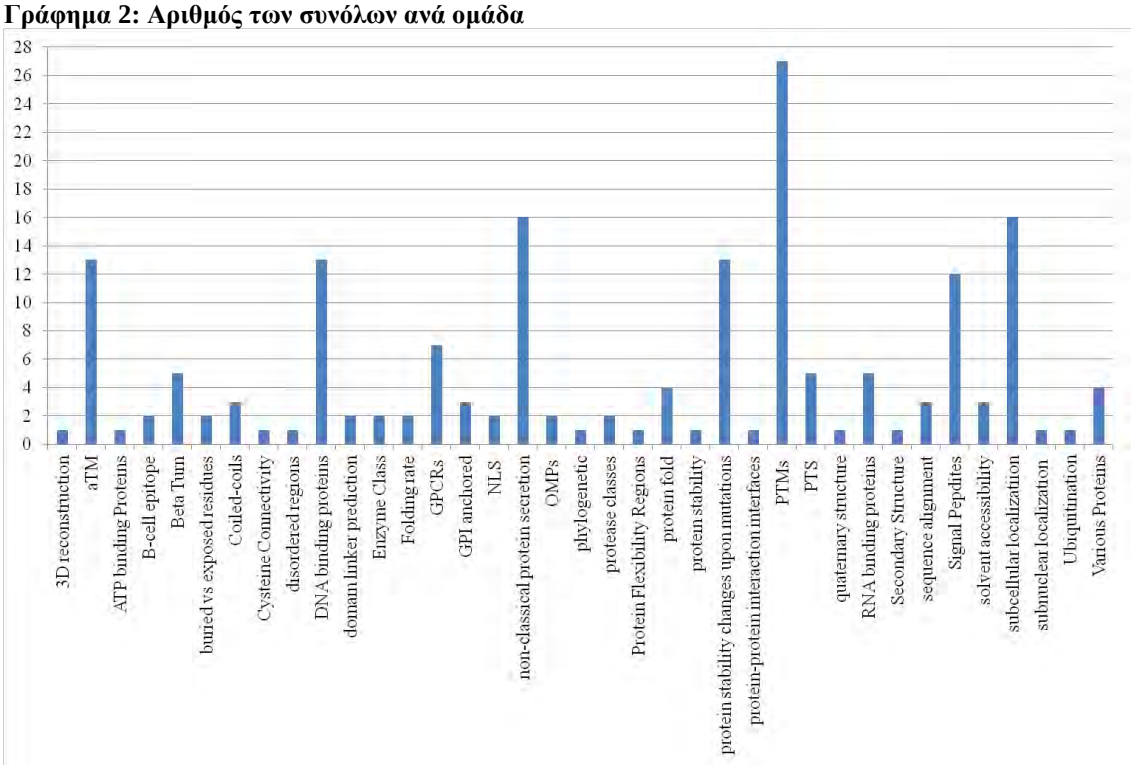
Παρατηρώντας το γράφημα 1 και το γράφημα 2 βλέπουμε ότι οι περισσότερες πρωτεΐνες που χρησιμοποιήθηκαν είναι στις κλάσεις subcellular localization και PTMs, με τον αριθμό των πρωτεϊνών να ξεπερνά τις εφτακόσες πενήντα χιλιάδες,

καθώς από τα εκατόν ογδόντα σύνολα δεδομένων, είκοσι εφτά είναι για PTMs και δεκαέξι για subcellular localization.

Γράφημα 1: Αριθμός των πρωτεϊνών ανά Ομάδα



Γράφημα 2: Αριθμός των συνόλων ανά ομάδα



Για την δημιουργία του ιστογράμματος που φαίνεται στο γραφήματος 3 χρειάστηκε να γίνει ομαδοποίηση των δεδομένων, καθώς τα σύνολα δεδομένων ανά

ομάδα έχουν μεγάλη διαφορά στον αριθμό των πρωτεϊνών τους. Για αυτό τον λόγο κατασκευάστηκε ο πιο κάτω πίνακας κλάσεων.

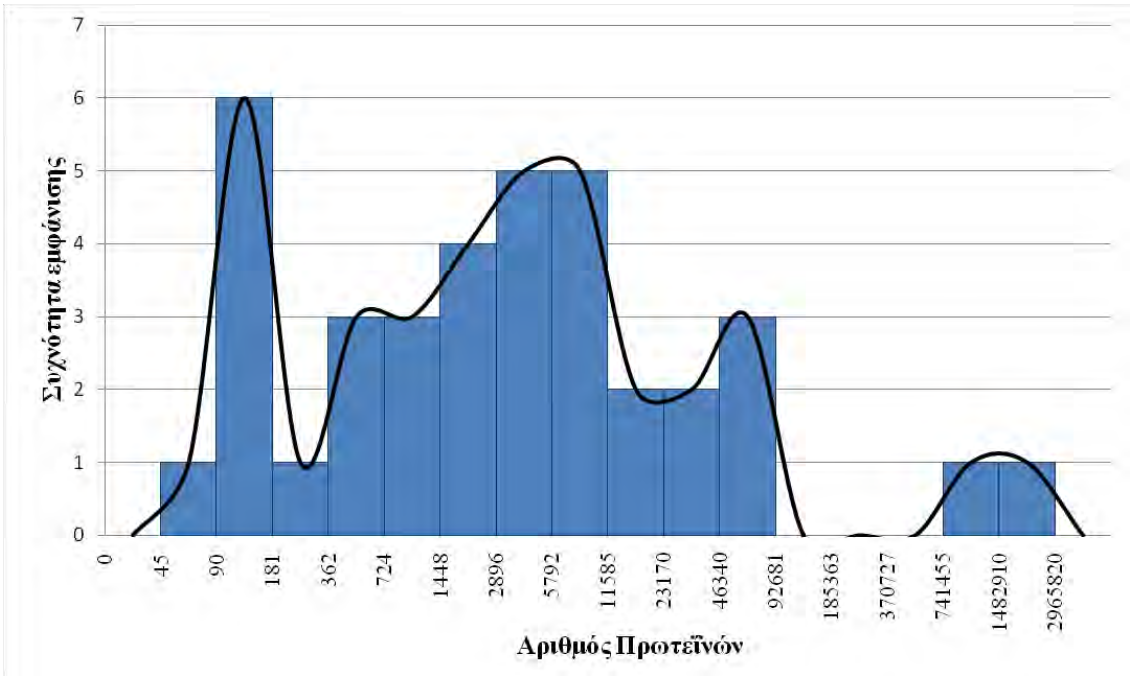
Πίνακας 2: Πίνακας συχνοτήτων για τις Πρωτεΐνες

Κλάσεις	Κέντρο Κλάσης (mi)	Κάτω όριο	Πάνω όριο	Συχνότητα εμφάνισης (fi)	fi*mi
0	1	0,7	1,4	0	0
1	2	1,4	2,8	0	0
2	4	2,8	5,7	0	0
3	8	5,7	11,3	0	0
4	16	11,3	22,6	0	0
5	32	22,6	45,3	0	0
6	64	45,3	90,5	1	64
7	128	90,5	181,0	6	768
8	256	181,0	362,0	1	256
9	512	362,0	724,1	3	1536
10	1024	724,1	1448,2	3	3072
11	2048	1448,2	2896,3	4	8192
12	4096	2896,3	5792,6	5	20480
13	8192	5792,6	11585,2	5	40960
14	16384	11585,2	23170,5	2	32768
15	32768	23170,5	46341,0	2	65536
16	65536	46341,0	92681,9	3	196608
17	131072	92681,9	185363,8	0	0
18	262144	185363,8	370727,6	0	0
19	524288	370727,6	741455,2	0	0
20	1048576	741455,2	1482910,4	1	1048576
21	2097152	1482910,4	2965820,8	1	2097152

Παρατηρώντας το γράφημα 3 φαίνεται ότι ο αριθμός των πρωτεϊνών στις περισσότερες ομάδες βρίσκεται στο εύρος τιμών [1448, 11585]. Για να γίνει καλύτερη περιγραφή του γραφήματος υπολογίστηκαν η μέση τιμή, η διάμεσος και η κορυφή. Η μέση τιμή για ομαδοποιημένα δεδομένα δίνεται από την σχέση $\tilde{x} = \frac{1}{n}\sum_{i=1}^{k} f_i m_i$ και με βάση τα δεδομένα είναι ίση με 95026. Εξαιτίας των ακραίων τιμών παρατηρούμε ότι η μέση τιμή είναι παραπλανητική. Για αυτό τον λόγο για τον υπολογισμό της γίνεται χρήση της σχέσης $\tilde{x} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i}$ με σκοπό την εύρεση της μέσης τιμής με βάση τις κλάσεις. Χρησιμοποιώντας την σχέση αυτή η μέση τιμή είναι ίση με 11,59 και συνεπώς οι περισσότερες παρατηρήσεις έχουν σαν κέντρο τις κλάσεις 11 και 12, δηλαδή το κέντρο ισορροπίας είναι το εύρος τιμών [1448, 5792]. Η διάμεσος ορίζεται ως η κεντρική θέση της κατανομής των παρατηρήσεων και για τον υπολογισμό της χρησιμοποιείται το πλήθος των παρατηρήσεων. Αν το πλήθος είναι περιττός αριθμός τότε η διάμεσος δ είναι η τιμή στη θέση (n+1)/2, ενώ αν είναι

άρτιος η διάμεσος είναι το ημιάθροισμα των τιμών στις θέσεις n/2 και n/2 + 1. Επειδή τα δεδομένα είναι ομαδοποιημένα η διάμεσος θα υπολογιστεί με βάση τις κλάσεις. Το πλήθος των κλάσεων είναι άρτιος αριθμός και συνεπώς η διάμεσος είναι ίση με (10+11)/2 και με εύρος τιμών [1448, 2896]. Η κορυφή Μ ορίζεται ως η τιμή της κλάσης με την μεγαλύτερη συχνότητα και είναι ίση με 7. Από τα αποτελέσματα των μέτρων θέσης ισχύει $\tilde{x}$> δ> Μ και προκύπτει ότι η κατανομή είναι ασύμμετρη.

Γράφημα 3: Ιστόγραμμα συχνοτήτων για τον αριθμό των πρωτεϊνών ανά ομάδα



Από τις 37 πρωτεϊνικές ομάδες στις οποίες ανήκουν τα σύνολα δεδομένων προκύπτει ο πιο κάτω πίνακας συχνοτήτων για τα σύνολα δεδομένων. Με την χρήση του πίνακα αυτού δημιουργήθηκε το ιστόγραμμα που φαίνεται στο γράφημα 4.

Πίνακας 3: Πίνακας συχνοτήτων για τα Σύνολα δεδομένων

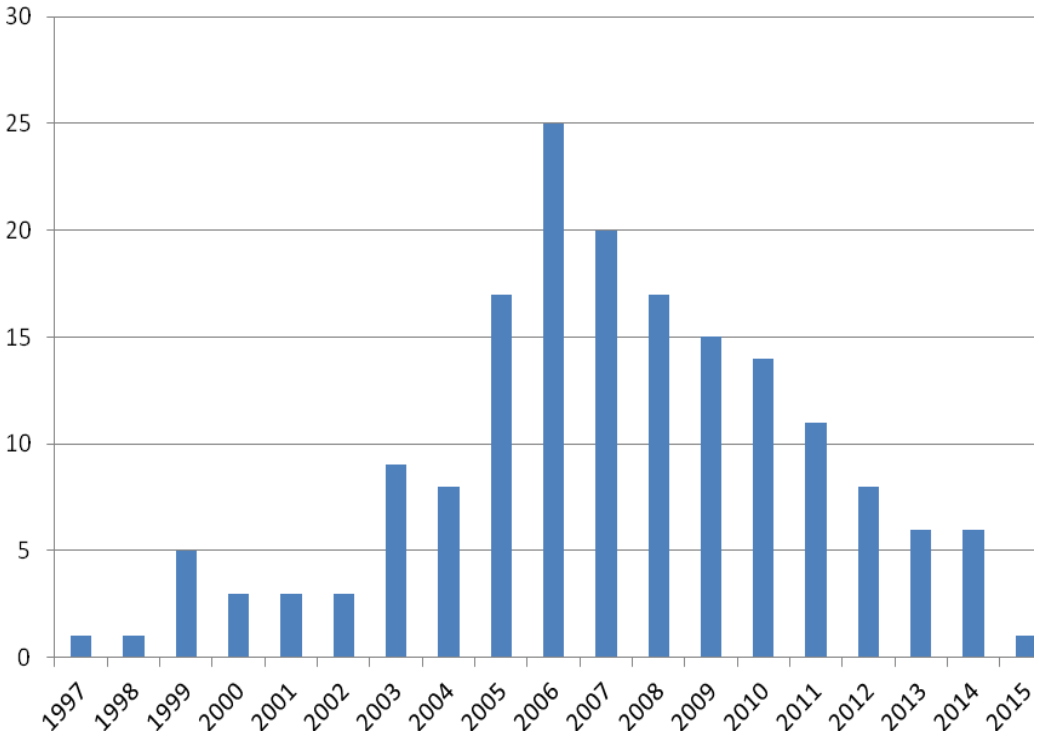
Αριθμός Συνόλων Δεδομένων	Συχνότητα Εμφάνισης	Σχετική συχνότητα εμφάνισης (%)
1	12	32,43243243
2	8	21,62162162
3	4	10,81081081
4	2	5,405405405
5	3	8,108108108
7	1	2,702702703
12	1	2,702702703
13	3	8,108108108
16	2	5,405405405
27	1	2,702702703
Σύνολο	**37**	**100**

Παρατηρώντας το γράφημα 4 φαίνεται ότι η καμπύλη συχνοτήτων παρουσιάζει θετική ασυμμετρία, καθώς δεξιά του κατακόρυφου άξονα βρίσκεται το μεγαλύτερο ποσοστό των παρατηρήσεων. Επίσης φαίνεται ότι 12 από τις 37 ομάδες έχουν ένα σύνολο δεδομένων που ανήκει στην ομάδα τους. Για να γίνει καλύτερη περιγραφή του γραφήματος χρησιμοποιήθηκαν τα μέτρα θέσης, που προσδιορίζουν την θέση της κατανομής των παρατηρήσεων και έτσι υπολογίστηκαν η μέση τιμή, η διάμεσος και η κορυφή. Η μέση τιμή δίνεται από την σχέση $\tilde{x} = \frac{1}{n}\sum_{i=1}^{n} f_i x_i$ και με βάση τα δεδομένα είναι ίση με 4,86. Η διάμεσος ορίζεται ως η κεντρική θέση της κατανομής των παρατηρήσεων και για τον υπολογισμό της χρησιμοποιείται το πλήθος των παρατηρήσεων. Αν το πλήθος είναι περιττός αριθμός τότε η διάμεσος δ είναι η τιμή στη θέση (n+1)/2, ενώ αν είναι άρτιος η διάμεσος είναι το ημιάθροισμα των τιμών στις θέσεις n/2 και n/2 + 1. Με βάση τα δεδομένα η διάμεσος είναι ίση με 2. Η κορυφή Μ ορίζεται ως η τιμή με την μεγαλύτερη συχνότητα και στην προκειμένη είναι ο αριθμός 1. Με βάση τα αποτελέσματα ισχύει η σχέση $\tilde{x}$> δ> Μ και συνεπώς επαληθεύεται ότι η κατανομή παρουσιάζει θετική ασυμμετρία.

Γράφημα 4: Ιστόγραμμα συχνοτήτων για τον αριθμό των συνόλων δεδομένων ανά ομάδα

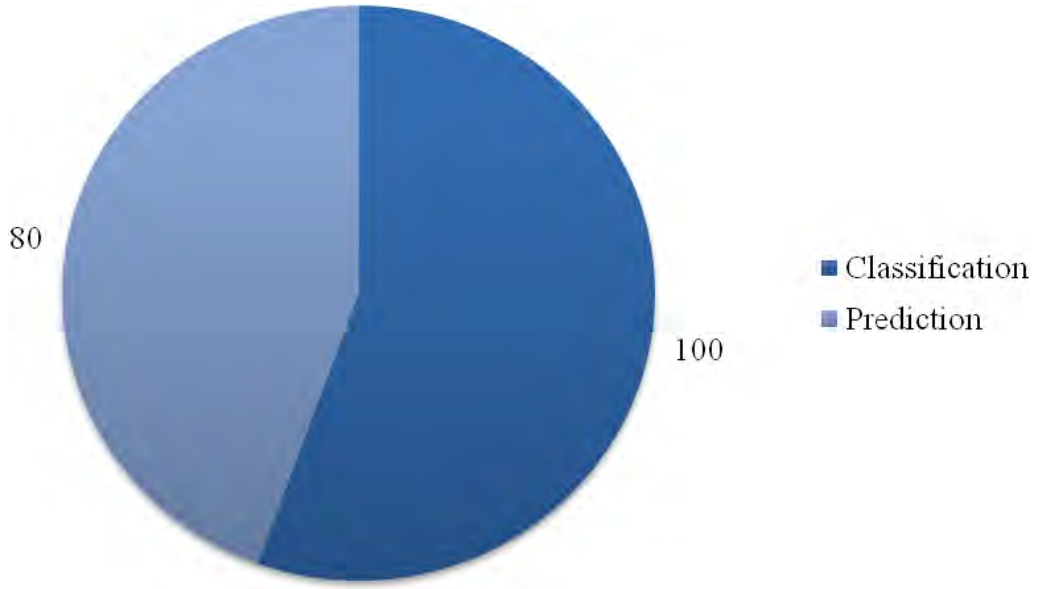
Στο γράφημα 5 παρατηρούμε ότι οι περισσότερες δημοσιεύσεις από τα σύνολα δεδομένων που ανακτήθηκαν στην παρούσα εργασία έγιναν μεταξύ των ετών 2005 και 2011.

Γράφημα 5: Αριθμός των συνόλων που δημοσιεύθηκαν ανά έτος



Τέλος στο γράφημα 6 φαίνεται ότι από τα εκατόν ογδόντα σύνολα δεδομένων τα εκατόν είναι για την κατηγοριοποίηση (Classification) των πρωτεϊνών στις ομάδες που ανήκουν και τα υπόλοιπα ογδόντα είναι για την πρόβλεψη (Prediction) των πρωτεϊνών.

Γράφημα 6: Αριθμός των συνόλων ανά κατηγορία του προβλήματος πρόβλεψης



6. Συμπεράσματα

Στην παρούσα πτυχιακή εργασία συλλέχθηκαν σύνολα δεδομένων πρωτεϊνικών ακολουθιών και μετέπειτα τοποθετήθηκαν σε μια βάση δεδομένων. Κατόπιν δημιουργήθηκε ένας ιστότοπος στον οποίο και αναρτήθηκαν οι πληροφορίες από την βάση, δίνοντας στους χρήστες την ευκαιρία να αναζητήσουν και να εξάγουν πληροφορίες για τα σύνολα δεδομένων.

Η ιστοσελίδα που δημιουργήθηκε είναι λειτουργική και εύκολη στην χρήση της. Ξεκινώντας την πλοήγηση του ο χρήστης έρχεται σε επαφή με την αρχική σελίδα, όπου και ενημερώνεται για το περιεχόμενο και τις λειτουργίες που παρέχει ο ιστότοπος [Εικόνα 6]. Η πλοήγηση μπορεί να συνεχιστεί στη σελίδα με τα σύνολα δεδομένων [Εικόνα 7], τα οποία και εμφανίζονται σε ένα πίνακα με βάση το όνομα, τον μοναδικό κωδικό τους και μια σύντομη περιγραφή για το περιεχόμενο τους. Αν ο χρήστης ενδιαφέρεται να αποκτήσει περισσότερες πληροφορίες για κάποιο από τα σύνολα δεδομένων, μπορεί να επιλέξει τον αριθμό του συνόλου δεδομένων που θέλει και έτσι μεταφέρεται σε μια καινούργια σελίδα στην οποία εμφανίζονται επιπλέον πληροφορίες για το συγκεκριμένο σύνολο δεδομένων. Οι επιπρόσθετες πληροφορίες αφορούν την ηλεκτρονική διεύθυνση από την οποία και ανακτήθηκε το σύνολο, η μορφή στην οποία είναι τα δεδομένα του, ποιο είναι το πρόβλημα πρόβλεψης, η βιβλιογραφία, η οποία είναι σε μορφή pubmed id και είναι άμεσα συνδεδεμένη με link στο NCBI, κατευθύνοντας τον χρήστη στο συγκεκριμένο άρθρο και το σύνολο δεδομένων διαθέσιμο για κατέβασμα [Εικόνα 9]. Επιπλέον έχει τη δυνατότητα να κάνει αναζήτηση στη βάση δεδομένων χρησιμοποιώντας λέξεις κλειδιά [Εικόνα 8]. Καταχωρώντας τη λέξη κλειδί και πατώντας το κουμπί της αναζήτησης, γίνεται έλεγχος εμφάνισης της λέξης ή φράσης στα πεδία: όνομα συνόλου δεδομένων, κωδικός, περιγραφή, βιολογικό πρόβλημα, είδος ακολουθίας και στην κατηγορία του προβλήματος πρόβλεψης. Αν υπάρχει σε ένα ή περισσότερα πεδία, τότε εμφανίζεται ένας πίνακας με όλα τα πεδία των συνόλων δεδομένων στα οποία εμφανίζεται η λέξη κλειδί [Εικόνα 12 , Εικόνα 13]. Με την επιλογή της αναζήτησης ο χρήστης μπορεί να περιορίσει την έρευνα του στο πεδίο που τον ενδιαφέρει. Στην συνέχεια αν επιθυμεί να αποκτήσει τα σύνολα δεδομένων, μπορεί να τα κατεβάσει επιλέγοντας ποια θέλει ή και όλα μαζί από την σελίδα Downloads [Εικόνα 9]. Για τυχόν απορίες και διευκρινήσεις ο χρήστης μπορεί να επικοινωνήσει μέσω της φόρμας επικοινωνίας με

τον διαχειριστή του ιστότοπου [Εικόνα 10]. Καθ' όλη τη διάρκεια της πλοήγησης ο χρήστης είναι σε επαφή με την γραμμή μενού που βρίσκεται στο πάνω μέρος της σελίδας, έχοντας την ευκαιρία οποιαδήποτε στιγμή θελήσει να αλλάξει την περιήγηση του μεταπηδώντας από τη μια σελίδα στην άλλη [Εικόνα 14].

Επειδή η πρόσβαση στα σύνολα δεδομένων βιολογικών ακολουθιών δεν είναι εύκολη, καθώς τα δεδομένα είναι διάσπαρτα στην βιβλιογραφία και οι ιστοσελίδες που τα φιλοξενούν πολλές φορές δεν συντηρούνται από τους δημιουργούς τους, με αποτέλεσμα τα σύνολα δεδομένων να «χάνονται», είναι σημαντική η καταχώρηση τους σε μια βάση δεδομένων.

Η συγκέντρωση των συνόλων δεδομένων των πρωτεϊνικών ακολουθιών σε μια βάση δεδομένων, δίνει την ευκαιρία στην επιστημονική κοινότητα να πραγματοποιήσει συγκριτικές αναλύσεις και δοκιμές σε αλγόριθμους πρόβλεψης και ταξινόμησης ακολουθιών με τη χρήση όσο το δυνατόν περισσότερων δεδομένων, καθώς δίνεται η δυνατότητα στον χρήστη να έχει πρόσβαση σε όλα τα δεδομένα της βάσης. Έτσι όταν κάποιος θελήσει να συγκρίνει τον αλγόριθμο του χρησιμοποιώντας κάποιο σύνολο δεδομένων μπορεί να το πραγματοποιήσει εύκολα και γρήγορα κάνοντας μια αναζήτηση στη βάση δεδομένων μέσα από την ιστοσελίδα.

7. Μελλοντική Εργασία

Η παρούσα πτυχιακή εργασία έχει παρουσιάσει μια βάση δεδομένων αποτελούμενη από σύνολα δεδομένων πρωτεϊνικών ακολουθιών. Οι χρήστες είναι σε θέση να έχουν πρόσβαση στη βάση και τα σύνολα δεδομένων μέσω εύχρηστης διαδικτυακής σελίδας.

Η μελλοντική εργασία περιλαμβάνει τον εμπλουτισμό της βάσης δεδομένων, με την καταχώρηση περισσοτέρων συνόλων δεδομένων από πρωτεϊνικές ακολουθίες, καθώς και νέων συνόλων δεδομένων από νουκλεοτιδικές ακολουθίες. Η καταχώρηση συνόλων δεδομένων από όλα τα είδη των βιολογικών ακολουθιών αποτελεί σημαντικό βήμα στη μετέπειτα έρευνα.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Nakamura, Y., G. Cochrane, and I. Karsch-Mizrachi, *The International Nucleotide Sequence Database Collaboration.* Nucleic Acids Res, 2013. **41**(Database issue): p. D21-4.
2. Benson, D.A., et al., *GenBank.* Nucleic Acids Res, 2013. **41**(Database issue): p. D36-42.
3. Kanz, C., et al., *The EMBL Nucleotide Sequence Database.* Nucleic Acids Res, 2005. **33**(Database issue): p. D29-33.
4. Miyazaki, S., et al., *DDBJ in the stream of various biological data.* Nucleic Acids Res, 2004. **32**(Database issue): p. D31-4.
5. *The universal protein resource (UniProt).* Nucleic Acids Res, 2008. **36**(Database issue): p. D190-5.
6. Wu, C.H., et al., *The Protein Information Resource.* Nucleic Acids Res, 2003. **31**(1): p. 345-7.
7. Sigrist, C.J., et al., *PROSITE, a protein domain database for functional characterization and annotation.* Nucleic Acids Res, 2010. **38**(Database issue): p. D161-6.
8. PROSITE. Available from: http://prosite.expasy.org/prosite_details.html.
9. PDB. Available from: http://www.rcsb.org/pdb/home/home.do.
10. Edgar, R., M. Domrachev, and A.E. Lash, *Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.* Nucleic Acids Res, 2002. **30**(1): p. 207-10.
11. Parkinson, H., et al., *ArrayExpress--a public database of microarray experiments and gene expression profiles.* Nucleic Acids Res, 2007. **35**(Database issue): p. D747-50.
12. Sherlock, G., et al., *The Stanford Microarray Database.* Nucleic Acids Res, 2001. **29**(1): p. 152-5.
13. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation.* Nucleic Acids Res, 2001. **29**(1): p. 308-11.
14. *Integrating ethics and science in the International HapMap Project.* Nat Rev Genet, 2004. **5**(6): p. 467-75.
15. Thorisson, G.A., et al., *The International HapMap Project Web site.* Genome Res, 2005. **15**(11): p. 1592-3.
16. Holland, T.A., et al., *Partitioning protein structures into domains: why is it so difficult?* J Mol Biol, 2006. **361**(3): p. 562-90.
17. Mulder, N.J. and R. Apweiler, *Tools and resources for identifying protein families, domains and motifs.* Genome Biol, 2002. **3**(1): p. REVIEWS2001.
18. Attwood, T.K., *The role of pattern databases in sequence analysis.* Brief Bioinform, 2000. **1**(1): p. 45-59.
19. Hofmann, K., *Sensitive protein comparisons with profiles and hidden Markov models.* Brief Bioinform, 2000. **1**(2): p. 167-78.
20. Gribskov, M., A.D. McLachlan, and D. Eisenberg, *Profile analysis: detection of distantly related proteins.* Proc Natl Acad Sci U S A, 1987. **84**(13): p. 4355-8.
21. Hulo, N., et al., *The PROSITE database.* Nucleic Acids Res, 2006. **34**(Database issue): p. D227-30.
22. Finn, R.D., et al., *Pfam: the protein families database.* Nucleic Acids Res, 2014. **42**(Database issue): p. D222-30.

23. Knudsen, M. and C. Wiuf, *The CATH database.* Hum Genomics, 2010. **4**(3): p. 207-12.
24. Murzin, A.G., et al., *SCOP: a structural classification of proteins database for the investigation of sequences and structures.* J Mol Biol, 1995. **247**(4): p. 536-40.
25. *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2013. **41**(Database issue): p. D8-D20.
26. NCBI. Available from: http://www.ncbi.nlm.nih.gov/.
27. Etzold, T. and P. Argos, *SRS--an indexing and retrieval tool for flat file data libraries.* Comput Appl Biosci, 1993. **9**(1): p. 49-57.
28. Zdobnov, E.M., et al., *The EBI SRS server--recent developments.* Bioinformatics, 2002. **18**(2): p. 368-73.
29. Al-Lazikani, B., E.E. Hill, and V. Morea, *Protein structure prediction.* Methods Mol Biol, 2008. **453**: p. 33-85.
30. Meng, F. and L. Kurgan, *Computational Prediction of Protein Secondary Structure from Sequence.* Curr Protoc Protein Sci, 2016. **86**: p. 2 3 1-2 3 10.
31. Pirovano, W. and J. Heringa, *Protein secondary structure prediction.* Methods Mol Biol, 2010. **609**: p. 327-48.
32. Barton, G.J., *Protein secondary structure prediction.* Curr Opin Struct Biol, 1995. **5**(3): p. 372-6.
33. Kall, L., A. Krogh, and E.L. Sonnhammer, *A combined transmembrane topology and signal peptide prediction method.* J Mol Biol, 2004. **338**(5): p. 1027-36.
34. Sonnhammer, E.L., G. von Heijne, and A. Krogh, *A hidden Markov model for predicting transmembrane helices in protein sequences.* Proc Int Conf Intell Syst Mol Biol, 1998. **6**: p. 175-82.
35. Kall, L., A. Krogh, and E.L. Sonnhammer, *An HMM posterior decoder for sequence feature prediction that includes homology information.* Bioinformatics, 2005. **21 Suppl 1**: p. i251-7.
36. Viklund, H., et al., *SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology.* Bioinformatics, 2008. **24**(24): p. 2928-9.
37. Bagos, P.G., et al., *A Hidden Markov Model method, capable of predicting and discriminating beta-barrel outer membrane proteins.* BMC Bioinformatics, 2004. **5**: p. 29.
38. Bagos, P.G., et al., *Prediction of signal peptides in archaea.* Protein Eng Des Sel, 2009. **22**(1): p. 27-35.
39. Tsirigos, K.D., et al., *The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides.* Nucleic Acids Res, 2015. **43**(W1): p. W401-7.
40. Petersen, T.N., et al., *SignalP 4.0: discriminating signal peptides from transmembrane regions.* Nat Methods, 2011. **8**(10): p. 785-6.
41. Μπάγκος, Π.Γ., *Βιοπληροφορική*. 2015: Πανεπιστήμιο Θεσσαλίας. 419.
42. Emanuelsson, O., H. Nielsen, and G. von Heijne, *ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites.* Protein Sci, 1999. **8**(5): p. 978-84.
43. Emanuelsson, O., et al., *Predicting subcellular localization of proteins based on their N-terminal amino acid sequence.* J Mol Biol, 2000. **300**(4): p. 1005-16.
44. Emanuelsson, O., et al., *In silico prediction of the peroxisomal proteome in fungi, plants and animals.* J Mol Biol, 2003. **330**(2): p. 443-56.

45. Bologna, G., et al., *N-Terminal myristoylation predictions by ensembles of neural networks.* Proteomics, 2004. **4**(6): p. 1626-32.
46. Melo, F., et al., *ANOLEA: a www server to assess protein structures.* Proc Int Conf Intell Syst Mol Biol, 1997. **5**: p. 187-90.
47. Szilagyi, A. and J. Skolnick, *Efficient prediction of nucleic acid binding function from low-resolution protein structures.* J Mol Biol, 2006. **358**(3): p. 922-33.
48. Chou, K.C. and H.B. Shen, *MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM.* Biochem Biophys Res Commun, 2007. **360**(2): p. 339-45.
49. Chou, K.C. and D.W. Elrod, *Prediction of membrane protein types and subcellular locations.* Proteins, 1999. **34**(1): p. 137-53.
50. Forrest, L.R., C.L. Tang, and B. Honig, *On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins.* Biophys J, 2006. **91**(2): p. 508-17.
51. Zhou, H. and Y. Zhou, *Predicting the topology of transmembrane helical proteins using mean burial propensity and a hidden-Markov-model-based method.* Protein Sci, 2003. **12**(7): p. 1547-55.
52. Park, Y. and V. Helms, *On the derivation of propensity scales for predicting exposed transmembrane residues of helical membrane proteins.* Bioinformatics, 2007. **23**(6): p. 701-8.
53. Park, Y., S. Hayat, and V. Helms, *Prediction of the burial status of transmembrane residues of helical membrane proteins.* BMC Bioinformatics, 2007. **8**: p. 302.
54. Viklund, H. and A. Elofsson, *OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar.* Bioinformatics, 2008. **24**(15): p. 1662-8.
55. Tilakaratne, N. and P.M. Sexton, *G-Protein-coupled receptor-protein interactions: basis for new concepts on receptor structure and function.* Clin Exp Pharmacol Physiol, 2005. **32**(11): p. 979-87.
56. Xiao, X., P. Wang, and K.C. Chou, *GPCR-CA: A cellular automaton image approach for predicting G-protein-coupled receptor functional classes.* J Comput Chem, 2009. **30**(9): p. 1414-23.
57. Davies, M.N., et al., *On the hierarchical classification of G protein-coupled receptors.* Bioinformatics, 2007. **23**(23): p. 3113-8.
58. Park, K.J., et al., *Discrimination of outer membrane proteins using support vector machines.* Bioinformatics, 2005. **21**(23): p. 4223-9.
59. Wren, J.D., *404 not found: the stability and persistence of URLs published in MEDLINE.* Bioinformatics, 2004. **20**(5): p. 668-72.
60. Wall, L., *The PERL Programming Language.*
61. Madden, T., *The NCBI Handbook.*
62. T. Berners-Lee, D.C., *Hypertext Markup Language - 2.0.* 1995.
63. Pleva, J.T., *PHP: Hypertext Preprocessor.*
64. Duncan Reed, P.J.T., *Cascading Style Sheets.*
65. Ιωάννης Μανωλόπουλος, Α.Ν.Π., *Συστήματα Βάσεων Δεδομένων Θεωρία και πρακτική εφαρμογή*.
66. Flanagan, D., *JavaScript: The Definitive Guide*.
67. Dvorski, D.D., *INSTALLING, CONFIGURING, AND DEVELOPING WITH XAMPP.* 2007.

68. Adzhubei, I.A., et al., *A method and server for predicting damaging missense mutations.* Nat Methods, 2010. **7**(4): p. 248-9.
69. Ahmad, S., M.M. Gromiha, and A. Sarai, *Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information.* Bioinformatics, 2004. **20**(4): p. 477-86.
70. Ahmad, S., et al., *ReadOut: structure-based calculation of direct and indirect readout energies and specificities for protein-DNA recognition.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W124-7.
71. Ahmad, S. and A. Sarai, *Moment-based prediction of DNA-binding proteins.* J Mol Biol, 2004. **341**(1): p. 65-71.
72. Andrabi, M., et al., *Prediction of mono- and di-nucleotide-specific DNA-binding sites in proteins using neural networks.* BMC Struct Biol, 2009. **9**: p. 30.
73. Bacardit, J., et al., *Automated alphabet reduction for protein datasets.* BMC Bioinformatics, 2009. **10**: p. 6.
74. Basu, S. and D. Plewczynski, *AMS 3.0: prediction of post-translational modifications.* BMC Bioinformatics, 2010. **11**: p. 210.
75. Bendtsen, J.D., et al., *Feature-based prediction of non-classical and leaderless protein secretion.* Protein Eng Des Sel, 2004. **17**(4): p. 349-56.
76. Bendtsen, J.D., et al., *Non-classical protein secretion in bacteria.* BMC Microbiol, 2005. **5**: p. 58.
77. Bernsel, A., et al., *Prediction of membrane-protein topology from first principles.* Proc Natl Acad Sci U S A, 2008. **105**(20): p. 7177-81.
78. Bernsel, A., et al., *TOPCONS: consensus prediction of membrane protein topology.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W465-8.
79. Bhasin, M., A. Garg, and G.P. Raghava, *PSLpred: prediction of subcellular localization of bacterial proteins.* Bioinformatics, 2005. **21**(10): p. 2522-4.
80. Blom, N., S. Gammeltoft, and S. Brunak, *Sequence and structure-based prediction of eukaryotic protein phosphorylation sites.* J Mol Biol, 1999. **294**(5): p. 1351-62.
81. Blum, T., S. Briesemeister, and O. Kohlbacher, *MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction.* BMC Bioinformatics, 2009. **10**: p. 274.
82. Briesemeister, S., J. Rahnenfuhrer, and O. Kohlbacher, *YLoc--an interpretable web server for predicting subcellular localization.* Nucleic Acids Res, 2010. **38**(Web Server issue): p. W497-502.
83. Cai, R., et al., *GPS-MBA: computational analysis of MHC class II epitopes in type 1 diabetes.* PLoS One, 2012. **7**(3): p. e33884.
84. Cai, Y.D. and K.C. Chou, *Predicting enzyme subclass by functional domain composition and pseudo amino acid composition.* J Proteome Res, 2005. **4**(3): p. 967-71.
85. Capriotti, E., R. Calabrese, and R. Casadio, *Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information.* Bioinformatics, 2006. **22**(22): p. 2729-34.
86. Capriotti, E., P. Fariselli, and R. Casadio, *I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W306-10.
87. Carugo, O., *Predicting residue solvent accessibility from protein sequence by considering the sequence environment.* Protein Eng, 2000. **13**(9): p. 607-9.

88. Chen, H., et al., *MeMo: a web tool for prediction of protein methylation modifications.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W249-53.
89. Chen, H. and H.X. Zhou, *Prediction of interface residues in protein-protein complexes by a consensus neural network method: test against NMR data.* Proteins, 2005. **61**(1): p. 21-35.
90. Chen, H. and H.X. Zhou, *Prediction of solvent accessibility and sites of deleterious mutations from protein sequence.* Nucleic Acids Res, 2005. **33**(10): p. 3193-9.
91. Chen, K., L.A. Kurgan, and J. Ruan, *Prediction of flexible/rigid regions from protein sequences using k-spaced amino acid pairs.* BMC Struct Biol, 2007. **7**: p. 25.
92. Chen, K., M.J. Mizianty, and L. Kurgan, *Prediction and analysis of nucleotide-binding residues using sequence and sequence-derived structural descriptors.* Bioinformatics, 2012. **28**(3): p. 331-41.
93. Chen, Y.L. and Q.Z. Li, *Prediction of the subcellular location of apoptosis proteins.* J Theor Biol, 2007. **245**(4): p. 775-83.
94. Cheng, C.W., et al., *Predicting RNA-binding sites of proteins using support vector machines and evolutionary information.* BMC Bioinformatics, 2008. **9 Suppl 12**: p. S6.
95. Cheng, J. and P. Baldi, *Three-stage prediction of protein beta-sheets by neural networks, alignments and graph algorithms.* Bioinformatics, 2005. **21 Suppl 1**: p. i75-84.
96. Cheng, J. and P. Baldi, *A machine learning information retrieval approach to protein fold recognition.* Bioinformatics, 2006. **22**(12): p. 1456-63.
97. Cheng, J. and P. Baldi, *Improved residue contact prediction using support vector machines and a large feature set.* BMC Bioinformatics, 2007. **8**: p. 113.
98. Cheng, J., A. Randall, and P. Baldi, *Prediction of protein stability changes for single-site mutations using support vector machines.* Proteins, 2006. **62**(4): p. 1125-32.
99. Cheng, J., et al., *SCRATCH: a protein structure and structural feature prediction server.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W72-6.
100. Cheng, J., H. Saigo, and P. Baldi, *Large-scale prediction of disulphide bridges using kernel methods, two-dimensional recursive neural networks, and weighted graph matching.* Proteins, 2006. **62**(3): p. 617-29.
101. Chou, K.C., *A key driving force in determination of protein structural classes.* Biochem Biophys Res Commun, 1999. **264**(1): p. 216-24.
102. Chou, K.C. and Y.D. Cai, *Predicting protein quaternary structure by pseudo amino acid composition.* Proteins, 2003. **53**(2): p. 282-9.
103. Chou, K.C. and H.B. Shen, *ProtIdent: a web server for identifying proteases and their types by fusing functional domain and sequential evolution information.* Biochem Biophys Res Commun, 2008. **376**(2): p. 321-5.
104. Dang, T.H., et al., *Prediction of kinase-specific phosphorylation sites using conditional random fields.* Bioinformatics, 2008. **24**(24): p. 2857-64.
105. de Castro, E., et al., *ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W362-5.
106. Diella, F., et al., *Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins.* BMC Bioinformatics, 2004. **5**: p. 79.

107. Diella, F., et al., *Phospho.ELM: a database of phosphorylation sites--update 2008.* Nucleic Acids Res, 2008. **36**(Database issue): p. D240-4.
108. Ding, C.H. and I. Dubchak, *Multi-class protein fold recognition using support vector machines and neural networks.* Bioinformatics, 2001. **17**(4): p. 349-58.
109. Dinkel, H., et al., *Phospho.ELM: a database of phosphorylation sites--update 2011.* Nucleic Acids Res, 2011. **39**(Database issue): p. D261-7.
110. Dinkel, H., et al., *The eukaryotic linear motif resource ELM: 10 years and counting.* Nucleic Acids Res, 2014. **42**(Database issue): p. D259-66.
111. Dou, Y., B. Yao, and C. Zhang, *PhosphoSVM: prediction of phosphorylation sites by integrating various protein sequence attributes with a support vector machine.* Amino Acids, 2014. **46**(6): p. 1459-69.
112. Dunker, A.K., et al., *Intrinsic disorder and protein function.* Biochemistry, 2002. **41**(21): p. 6573-82.
113. Dunker, A.K., et al., *Intrinsically disordered protein.* J Mol Graph Model, 2001. **19**(1): p. 26-59.
114. Durek, P., et al., *PhosPhAt: the Arabidopsis thaliana phosphorylation site database. An update.* Nucleic Acids Res, 2010. **38**(Database issue): p. D828-34.
115. Ebina, T., H. Toh, and Y. Kuroda, *Loop-length-dependent SVM prediction of domain linkers for high-throughput structural proteomics.* Biopolymers, 2009. **92**(1): p. 1-8.
116. Ebina, T., H. Toh, and Y. Kuroda, *DROP: an SVM domain linker predictor trained with optimal features selected by random forest.* Bioinformatics, 2011. **27**(4): p. 487-94.
117. Eickholt, J. and J. Cheng, *Predicting protein residue-residue contacts using deep networks and boosting.* Bioinformatics, 2012. **28**(23): p. 3066-72.
118. Eisenhaber, B., et al., *Glycosylphosphatidylinositol lipid anchoring of plant proteins. Sensitive prediction from sequence- and genome-wide studies for Arabidopsis and rice.* Plant Physiol, 2003. **133**(4): p. 1691-701.
119. Fang, Y., et al., *Predicting DNA-binding proteins: approached from Chou's pseudo amino acid composition and other specific sequence features.* Amino Acids, 2008. **34**(1): p. 103-9.
120. Fankhauser, N. and P. Maser, *Identification of GPI anchor attachment signals by a Kohonen self-organizing map.* Bioinformatics, 2005. **21**(9): p. 1846-52.
121. Fariselli, P., et al., *Grammatical-Restrained Hidden Conditional Random Fields for Bioinformatics applications.* Algorithms Mol Biol, 2009. **4**: p. 13.
122. Fayech, S., N. Essoussi, and M. Limam, *Partitioning clustering algorithms for protein sequence data sets.* BioData Min, 2009. **2**(1): p. 3.
123. Fischer, J.D., C.E. Mayer, and J. Soding, *Prediction of protein functional residues from sequence by probability density estimation.* Bioinformatics, 2008. **24**(5): p. 613-20.
124. Fu, S.C., K. Imai, and P. Horton, *Prediction of leucine-rich nuclear export signal containing proteins with NESsential.* Nucleic Acids Res, 2011. **39**(16): p. e111.
125. Gao, J., et al., *Musite, a tool for global prediction of general and kinase-specific phosphorylation sites.* Mol Cell Proteomics, 2010. **9**(12): p. 2586-600.
126. Gardy, J.L. and F.S. Brinkman, *Methods for predicting bacterial protein subcellular localization.* Nat Rev Microbiol, 2006. **4**(10): p. 741-51.

127. Gardy, J.L., et al., *PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis.* Bioinformatics, 2005. **21**(5): p. 617-23.
128. Gardy, J.L., et al., *PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria.* Nucleic Acids Res, 2003. **31**(13): p. 3613-7.
129. Gnad, F., J. Gunawardena, and M. Mann, *PHOSIDA 2011: the posttranslational modification database.* Nucleic Acids Res, 2011. **39**(Database issue): p. D253-60.
130. Gorodkin, J., et al., *Using sequence motifs for enhanced neural network prediction of protein distance constraints.* Proc Int Conf Intell Syst Mol Biol, 1999: p. 95-105.
131. Goudenege, D., et al., *CoBaltDB: Complete bacterial and archaeal orfeomes subcellular localization database and associated resources.* BMC Microbiol, 2010. **10**: p. 88.
132. Gould, C.M., et al., *ELM: the status of the 2010 eukaryotic linear motif resource.* Nucleic Acids Res, 2010. **38**(Database issue): p. D167-80.
133. Granseth, E., H. Viklund, and A. Elofsson, *ZPRED: predicting the distance to the membrane center for residues in alpha-helical membrane proteins.* Bioinformatics, 2006. **22**(14): p. e191-6.
134. Guindon, S., et al., *New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0.* Syst Biol, 2010. **59**(3): p. 307-21.
135. Guo, J., et al., *Predicting protein folding rates using the concept of Chou's pseudo amino acid composition.* J Comput Chem, 2011. **32**(8): p. 1612-7.
136. Gupta, R., et al., *Scanning the available Dictyostelium discoideum proteome for O-linked GlcNAc glycosylation sites using neural networks.* Glycobiology, 1999. **9**(10): p. 1009-22.
137. Hamby, S.E. and J.D. Hirst, *Prediction of glycosylation sites using random forests.* BMC Bioinformatics, 2008. **9**: p. 500.
138. Heazlewood, J.L., et al., *PhosPhAt: a database of phosphorylation sites in Arabidopsis thaliana and a plant-specific phosphorylation site predictor.* Nucleic Acids Res, 2008. **36**(Database issue): p. D1015-21.
139. Hennerdal, A. and A. Elofsson, *Rapid membrane protein topology prediction.* Bioinformatics, 2011. **27**(9): p. 1322-3.
140. Hessa, T., et al., *Molecular code for transmembrane-helix recognition by the Sec61 translocon.* Nature, 2007. **450**(7172): p. 1026-30.
141. Ho, S.Y., et al., *Design of accurate predictors for DNA-binding sites in proteins using hybrid SVM-PSSM method.* Biosystems, 2007. **90**(1): p. 234-41.
142. Hoglund, A., et al., *MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition.* Bioinformatics, 2006. **22**(10): p. 1158-65.
143. Hoof, I., et al., *NetMHCpan, a method for MHC class I binding prediction beyond humans.* Immunogenetics, 2009. **61**(1): p. 1-13.
144. Huang, H.D., et al., *KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W226-9.
145. Huang, L.T., M.M. Gromiha, and S.Y. Ho, *iPTREE-STAB: interpretable decision tree based method for predicting protein stability changes upon mutations.* Bioinformatics, 2007. **23**(10): p. 1292-3.

146. Hwang, S., Z. Gou, and I.B. Kuznetsov, *DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins.* Bioinformatics, 2007. **23**(5): p. 634-6.
147. Iakoucheva, L.M., et al., *Intrinsic disorder in cell-signaling and cancer-associated proteins.* J Mol Biol, 2002. **323**(3): p. 573-84.
148. Iakoucheva, L.M., et al., *The importance of intrinsic disorder for protein phosphorylation.* Nucleic Acids Res, 2004. **32**(3): p. 1037-49.
149. Indio, V., et al., *The prediction of organelle-targeting peptides in eukaryotic proteins with Grammatical-Restrained Hidden Conditional Random Fields.* Bioinformatics, 2013. **29**(8): p. 981-8.
150. Jessen, L.E., et al., *SigniSite: Identification of residue-level genotype-phenotype correlations in protein multiple sequence alignments.* Nucleic Acids Res, 2013. **41**(Web Server issue): p. W286-91.
151. Johansen, M.B., et al., *Prediction of disease causing non-synonymous SNPs by the Artificial Neural Network Predictor NetDiseaseSNP*. PLoS One, 2013. **8**(7): p. e68370.
152. Johansen, M.B., L. Kiemer, and S. Brunak, *Analysis and prediction of mammalian protein glycation.* Glycobiology, 2006. **16**(9): p. 844-53.
153. Julenius, K., *NetCGlyc 1.0: prediction of mammalian C-mannosylation sites.* Glycobiology, 2007. **17**(8): p. 868-76.
154. Juncker, A.S., et al., *Prediction of lipoprotein signal peptides in Gram-negative bacteria.* Protein Sci, 2003. **12**(8): p. 1652-62.
155. Jung, I., et al., *PostMod: sequence based prediction of kinase-specific phosphorylation sites with indirect relationship.* BMC Bioinformatics, 2010. **11 Suppl 1**: p. S10.
156. Karchin, R., K. Karplus, and D. Haussler, *Classifying G-protein coupled receptors with support vector machines.* Bioinformatics, 2002. **18**(1): p. 147-59.
157. Karypis, G., *YASSPP: better kernels and coding schemes lead to improvements in protein secondary structure prediction.* Proteins, 2006. **64**(3): p. 575-86.
158. Kaur, H. and G.P. Raghava, *An evaluation of beta-turn prediction methods.* Bioinformatics, 2002. **18**(11): p. 1508-14.
159. Kiemer, L., J.D. Bendtsen, and N. Blom, *NetAcet: prediction of N-terminal acetylation sites.* Bioinformatics, 2005. **21**(7): p. 1269-70.
160. Kountouris, P. and J.D. Hirst, *Prediction of backbone dihedral angles and protein secondary structure using support vector machines.* BMC Bioinformatics, 2009. **10**: p. 437.
161. Kountouris, P. and J.D. Hirst, *Predicting beta-turns and their types using predicted backbone dihedral angles and secondary structures.* BMC Bioinformatics, 2010. **11**: p. 407.
162. Kringelum, J.V., et al., *Reliable B cell epitope predictions: impacts of method development and improved benchmarking.* PLoS Comput Biol, 2012. **8**(12): p. e1002829.
163. Kumar, M., M.M. Gromiha, and G.P. Raghava, *Identification of DNA-binding proteins using support vector machines and evolutionary profiles.* BMC Bioinformatics, 2007. **8**: p. 463.
164. Kumar, M., M.M. Gromiha, and G.P. Raghava, *SVM based prediction of RNA-binding proteins using binding residues and evolutionary information.* J Mol Recognit, 2011. **24**(2): p. 303-13.

165. Kuznetsov, I.B., *Ordered conformational change in the protein backbone: prediction of conformationally variable positions from sequence and low-resolution structural data.* Proteins, 2008. **72**(1): p. 74-87.
166. Kuznetsov, I.B., et al., *Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins.* Proteins, 2006. **64**(1): p. 19-27.
167. Larsen, J.E., O. Lund, and M. Nielsen, *Improved method for predicting linear B-cell epitopes.* Immunome Res, 2006. **2**: p. 2.
168. Larsen, M.V., et al., *Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction.* BMC Bioinformatics, 2007. **8**: p. 424.
169. Li, A., et al., *Prediction of Nepsilon-acetylation on internal lysines implemented in Bayesian Discriminant Method.* Biochem Biophys Res Commun, 2006. **350**(4): p. 818-24.
170. Lin, W.Z., X. Xiao, and K.C. Chou, *GPCR-GIA: a web-server for identifying G-protein coupled receptors and their families with grey incidence analysis.* Protein Eng Des Sel, 2009. **22**(11): p. 699-705.
171. Lingner, T., et al., *Identification of novel plant peroxisomal targeting signals by a combination of machine learning methods and in vivo subcellular targeting analyses.* Plant Cell, 2011. **23**(4): p. 1556-72.
172. Litou, Z.I., et al., *Prediction of cell wall sorting signals in gram-positive bacteria with a hidden markov model: application to complete genomes.* J Bioinform Comput Biol, 2008. **6**(2): p. 387-401.
173. Liu, B., et al., *Prediction of protein binding sites in protein structures using hidden Markov support vector machine.* BMC Bioinformatics, 2009. **10**: p. 381.
174. Liu, Z., et al., *GPS-CCD: a novel computational program for the prediction of calpain cleavage sites.* PLoS One, 2011. **6**(4): p. e19001.
175. Liu, Z., et al., *GPS-PUP: computational prediction of pupylation sites in prokaryotic proteins.* Mol Biosyst, 2011. **7**(10): p. 2737-40.
176. Liu, Z., et al., *Systematic analysis of the Plk-mediated phosphoregulation in eukaryotes.* Brief Bioinform, 2013. **14**(3): p. 344-60.
177. Liu, Z., et al., *GPS-ARM: computational analysis of the APC/C recognition motif by predicting D-boxes and KEN-boxes.* PLoS One, 2012. **7**(3): p. e34370.
178. Lomize, M.A., et al., *OPM: orientations of proteins in membranes database.* Bioinformatics, 2006. **22**(5): p. 623-5.
179. Magnan, C.N. and P. Baldi, *SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity.* Bioinformatics, 2014. **30**(18): p. 2592-7.
180. Magnan, C.N., A. Randall, and P. Baldi, *SOLpro: accurate sequence-based prediction of protein solubility.* Bioinformatics, 2009. **25**(17): p. 2200-7.
181. Maurer-Stroh, S., B. Eisenhaber, and F. Eisenhaber, *N-terminal N-myristoylation of proteins: refinement of the sequence motif and its taxon-specific differences.* J Mol Biol, 2002. **317**(4): p. 523-40.
182. Maurer-Stroh, S., B. Eisenhaber, and F. Eisenhaber, *N-terminal N-myristoylation of proteins: prediction of substrate proteins from amino acid sequence.* J Mol Biol, 2002. **317**(4): p. 541-57.
183. Mayr, G., F.S. Domingues, and P. Lackner, *Comparative analysis of protein structure alignments.* BMC Struct Biol, 2007. **7**: p. 50.

184. McDonnell, A.V., et al., *Paircoil2: improved prediction of coiled coils from sequence.* Bioinformatics, 2006. **22**(3): p. 356-8.
185. Meszaros, B., I. Simon, and Z. Dosztanyi, *Prediction of protein binding regions in disordered proteins.* PLoS Comput Biol, 2009. **5**(5): p. e1000376.
186. Miller, M.L., et al., *NetPhosBac - a predictor for Ser/Thr phosphorylation sites in bacterial proteins.* Proteomics, 2009. **9**(1): p. 116-25.
187. Muppirala, U.K., V.G. Honavar, and D. Dobbs, *Predicting RNA-protein interactions using only sequence information.* BMC Bioinformatics, 2011. **12**: p. 489.
188. Nagata, K., A. Randall, and P. Baldi, *SIDEpro: a novel machine learning approach for the fast and accurate prediction of side-chain conformations.* Proteins, 2012. **80**(1): p. 142-53.
189. Neuberger, G., et al., *Prediction of peroxisomal targeting signal 1 containing proteins from amino acid sequence.* J Mol Biol, 2003. **328**(3): p. 581-92.
190. Neuberger, G., et al., *Motif refinement of the peroxisomal targeting signal 1 and evaluation of taxon-specific differences.* J Mol Biol, 2003. **328**(3): p. 567-79.
191. Neuberger, G., G. Schneider, and F. Eisenhaber, *pkaPS: prediction of protein kinase A phosphorylation sites with the simplified kinase-substrate binding model.* Biol Direct, 2007. **2**: p. 1.
192. Ng, P.C. and S. Henikoff, *Predicting deleterious amino acid substitutions.* Genome Res, 2001. **11**(5): p. 863-74.
193. Nguyen Ba, A.N., et al., *NLStradamus: a simple Hidden Markov Model for nuclear localization signal prediction.* BMC Bioinformatics, 2009. **10**: p. 202.
194. Nielsen, M. and O. Lund, *NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction.* BMC Bioinformatics, 2009. **10**: p. 296.
195. Niu, B., et al., *HIV-1 protease cleavage site prediction based on amino acid property.* J Comput Chem, 2009. **30**(1): p. 33-9.
196. Pan, Z., et al., *Systematic analysis of the in situ crosstalk of tyrosine modifications reveals no additional natural selection on multiply modified residues.* Sci Rep, 2014. **4**: p. 7331.
197. Park, Y. and V. Helms, *How strongly do sequence conservation patterns and empirical scales correlate with exposure patterns of transmembrane helices of membrane proteins?* Biopolymers, 2006. **83**(4): p. 389-99.
198. Petsalaki, E.I., et al., *PredSL: a tool for the N-terminal sequence-based prediction of protein subcellular localization.* Genomics Proteomics Bioinformatics, 2006. **4**(1): p. 48-55.
199. Pierleoni, A., et al., *MemPype: a pipeline for the annotation of eukaryotic membrane proteins.* Nucleic Acids Res, 2011. **39**(Web Server issue): p. W375-80.
200. Pierleoni, A., P.L. Martelli, and R. Casadio, *PredGPI: a GPI-anchor predictor.* BMC Bioinformatics, 2008. **9**: p. 392.
201. Pierleoni, A., P.L. Martelli, and R. Casadio, *MemLoci: predicting subcellular localization of membrane proteins in eukaryotes.* Bioinformatics, 2011. **27**(9): p. 1224-30.
202. Pierleoni, A., et al., *BaCelLo: a balanced subcellular localization predictor.* Bioinformatics, 2006. **22**(14): p. e408-16.

203. Plewczynski, D., S. Basu, and I. Saha, *AMS 4.0: consensus prediction of post-translational modifications in protein sequences.* Amino Acids, 2012. **43**(2): p. 573-82.
204. Plewczynski, D., et al., *AutoMotif server: prediction of single residue post-translational modifications in proteins.* Bioinformatics, 2005. **21**(10): p. 2525-7.
205. Plewczynski, D., et al., *AutoMotif Server for prediction of phosphorylation sites in proteins using support vector machine: 2007 update.* J Mol Model, 2008. **14**(1): p. 69-76.
206. Poisson, G., et al., *FragAnchor: a large-scale predictor of glycosylphosphatidylinositol anchors in eukaryote protein sequences by qualitative scoring.* Genomics Proteomics Bioinformatics, 2007. **5**(2): p. 121-30.
207. Ponomarenko, J.V. and P.E. Bourne, *Antibody-protein interactions: benchmark datasets and prediction tools evaluation.* BMC Struct Biol, 2007. **7**: p. 64.
208. Puntervoll, P., et al., *ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins.* Nucleic Acids Res, 2003. **31**(13): p. 3625-30.
209. Qin, S. and H.X. Zhou, *meta-PPISP: a meta web server for protein-protein interaction site prediction.* Bioinformatics, 2007. **23**(24): p. 3386-7.
210. Radivojac, P., et al., *Identification, analysis, and prediction of protein ubiquitination sites.* Proteins, 2010. **78**(2): p. 365-80.
211. Rask, T.S., et al., *Plasmodium falciparum erythrocyte membrane protein 1 diversity in seven genomes--divide and conquer.* PLoS Comput Biol, 2010. **6**(9).
212. Rose, A., et al., *RHYTHM--a server to predict the orientation of transmembrane helices in channels and membrane-coils.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W575-80.
213. Scott, M.S., D.Y. Thomas, and M.T. Hallett, *Predicting subcellular localization via protein motif co-occurrence.* Genome Res, 2004. **14**(10A): p. 1957-66.
214. Sgourakis, N.G., et al., *A method for the prediction of GPCRs coupling specificity to G-proteins using refined profile Hidden Markov Models.* BMC Bioinformatics, 2005. **6**: p. 104.
215. Shatkay, H., et al., *SherLoc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data.* Bioinformatics, 2007. **23**(11): p. 1410-7.
216. Shen, H.B. and K.C. Chou, *Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition.* Biochem Biophys Res Commun, 2005. **337**(3): p. 752-6.
217. Shen, H.B. and K.C. Chou, *EzyPred: a top-down approach for predicting enzyme functional classes and subclasses.* Biochem Biophys Res Commun, 2007. **364**(1): p. 53-9.
218. Soding, J., M. Remmert, and A. Biegert, *HHrep: de novo protein repeat detection and the origin of TIM barrels.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W137-42.
219. Song, C., et al., *Systematic analysis of protein phosphorylation networks from phosphoproteomic data.* Mol Cell Proteomics, 2012. **11**(10): p. 1070-83.

220. Sreekumar, K.R., et al., *Predicting GPCR-G-protein coupling using hidden Markov models.* Bioinformatics, 2004. **20**(18): p. 3490-9.
221. Steentoft, C., et al., *Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology.* EMBO J, 2013. **32**(10): p. 1478-88.
222. Stranzl, T., et al., *NetCTLpan: pan-specific MHC class I pathway epitope predictions.* Immunogenetics, 2010. **62**(6): p. 357-68.
223. Suo, S.B., et al., *PSEA: Kinase-specific prediction and analysis of human phosphorylation substrates.* Sci Rep, 2014. **4**: p. 4524.
224. Terribilini, M., et al., *Prediction of RNA binding sites in proteins from amino acid sequence.* RNA, 2006. **12**(8): p. 1450-62.
225. Tjong, H. and H.X. Zhou, *DISPLAR: an accurate method for predicting DNA-binding sites on protein surfaces.* Nucleic Acids Res, 2007. **35**(5): p. 1465-77.
226. Trolle, T. and M. Nielsen, *NetTepi: an integrated method for the prediction of T cell epitopes.* Immunogenetics, 2014. **66**(7-8): p. 449-56.
227. Vassura, M., et al., *FT-COMAR: fault tolerant three-dimensional structure reconstruction from protein contact maps.* Bioinformatics, 2008. **24**(10): p. 1313-5.
228. Viklund, H. and A. Elofsson, *Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information.* Protein Sci, 2004. **13**(7): p. 1908-17.
229. Walia, R.R., et al., *RNABindRPlus: a predictor that combines machine learning and sequence homology-based methods to improve the reliability of predicted RNA-binding residues in proteins.* PLoS One, 2014. **9**(5): p. e97725.
230. Wan, J., et al., *Meta-prediction of phosphorylation sites with weighted voting and restricted grid search parameter selection.* Nucleic Acids Res, 2008. **36**(4): p. e22.
231. Wang, L. and S.J. Brown, *BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W243-8.
232. Wang, L., M.Q. Yang, and J.Y. Yang, *Prediction of DNA-binding residues from protein sequence information using random forests.* BMC Genomics, 2009. **10 Suppl 1**: p. S1.
233. Wong, Y.H., et al., *KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns.* Nucleic Acids Res, 2007. **35**(Web Server issue): p. W588-94.
234. Wu, J., et al., *Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature.* Bioinformatics, 2009. **25**(1): p. 30-5.
235. Xiao, X., P. Wang, and K.C. Chou, *GPCR-2L: predicting G protein-coupled receptors and their types by hybridizing two different modes of pseudo amino acid compositions.* Mol Biosyst, 2011. **7**(3): p. 911-9.
236. Xu, Y., et al., *Prediction of posttranslational modification sites from amino acid sequences with kernel methods.* J Theor Biol, 2014. **344**: p. 78-87.
237. Xue, Y., et al., *GPS 2.1: enhanced prediction of kinase-specific phosphorylation sites with an algorithm of motif length selection.* Protein Eng Des Sel, 2011. **24**(3): p. 255-60.
238. Xue, Y., et al., *GPS-SNO: computational prediction of protein S-nitrosylation sites with a modified GPS algorithm.* PLoS One, 2010. **5**(6): p. e11290.

239. Xue, Y., et al., *GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy.* Mol Cell Proteomics, 2008. **7**(9): p. 1598-608.
240. Yabuki, Y., et al., *GRIFFIN: a system for predicting GPCR-G-protein coupling selectivity using a support vector machine and a hidden Markov model.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W148-53.
241. Yao, Q., et al., *Predicting and analyzing protein phosphorylation sites in plants using musite.* Front Plant Sci, 2012. **3**: p. 186.
242. Yu, N.Y., et al., *PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes.* Bioinformatics, 2010. **26**(13): p. 1608-15.
243. Yuan, Z., J.S. Mattick, and R.D. Teasdale, *SVMtm: support vector machines to predict transmembrane segments.* J Comput Chem, 2004. **25**(5): p. 632-6.
244. Yuan, Z. and R.D. Teasdale, *Prediction of Golgi Type II membrane proteins based on their transmembrane domains.* Bioinformatics, 2002. **18**(8): p. 1109-15.
245. Zhao, Q., et al., *GPS-SUMO: a tool for the prediction of sumoylation sites and SUMO-interaction motifs.* Nucleic Acids Res, 2014. **42**(Web Server issue): p. W325-30.
246. Zheng, C. and L. Kurgan, *Prediction of beta-turns at over 80% accuracy based on an ensemble of predicted secondary structures and multiple alignments.* BMC Bioinformatics, 2008. **9**: p. 430.
247. Zhou, G.P. and K. Doctor, *Subcellular location prediction of apoptosis proteins.* Proteins, 2003. **50**(1): p. 44-8.
248. Zou, L., et al., *PKIS: computational identification of protein kinases for experimentally discovered protein phosphorylation sites.* BMC Bioinformatics, 2013. **14**: p. 247.
249. Zulawski, M., R. Braginets, and W.X. Schulze, *PhosPhAt goes kinases--searchable protein kinase target information in the plant phosphorylation site database PhosPhAt.* Nucleic Acids Res, 2013. **41**(Database issue): p. D1176-84.

ΠΑΡΑΡΤΗΜΑ

1. Πίνακες

Πίνακας 4: Αναγνωριστικά Βάσεων δεδομένων στην γραμμή κεφαλίδας του FASTA

Όνομα Βάσης Δεδομένων	**Σύνταξη γραμμής κεφαλίδας**
GenBank	gb|accession|locus
EMBL Data Library	emb|accession|locus
DDBJ, DNA Database of Japan	dbj|accession|locus
NBRF PIR	pir||entry
Protein Research Foundation	prf||name
SWISS-PROT	sp|accession|entry name
Brookhaven Protein Data Bank	pdb|entry|chain
Patents	pat|country|number
GenInfo Backbone Id	bbs|number
General database identifier	gnl|database|identifier
NCBI Reference Sequence	ref|accession|locus
Local Sequence identifier	lcl|identifier

Πίνακας 5: Σύνολα Δεδομένων

Dataset Name	Dataset ID	Sequence Type	Problem Type	Biological Problem	Data	Description	References	URL	Web Server
GDS	DSDB0001	Protein	Classification	Hierarchical classification of GPCRs	Fasta	The datasets contain 8354 protein sequences in 5 classes at the family level (A-E), 40 classes at the sub-family level, and 108 classes at the sub-subfamily level. The dataset contain only human protein sequences, with the exception of Class D proteins, which are found only in fungi and Class E, which are found in Dictyostelium.	17956878	http://111.68.99.218/gpcr-mpredictor/ http://www.cs.kent.ac.uk/projects/biasprofs/downloads.html	http://igrid-ext.cryst.bbk.ac.uk/gpcrtree/
GPCR-CA	DSDB0002	Protein	Classification	Preticting GPCR with Celullar Automaton Image Approach	Fasta	The DataSet contains 365 GPCRs, of which (1) 232 are of rhodopsin-like, (2) 39 of secretin-like, (3) 44 of metabotrophic/ glutamate/ pheromone, (4) 23 of fungal pheromone, (5) 10 of cAMP, and (6) 17 of frizzled/smoothened family.	19037861	http://onlinelibrary.wiley.com/doi/10.1002/jcc.21163/suppinfo	
PRED-LIPO	DSDB0003	Protein	Classification	Prediction of Lipoprotein and Secretory Signal Peptides in Gram-positive Bacteria	UniProt ID	The training set contained 67 lipoproteins from Gram-positive bacteria, 127 secreted proteins containing a signal peptide cleaved by SPase I from Gram-positive bacteria, 111 cytoplasmic proteins from Gram-positive bacteria and 58 Gram-positive bacterial sequences with an N-terminal TM segment. The test set contains 66 TM proteins, 117 Lipoproteins, 109 secreted proteins and 713 cytoplasmic proteins.	18975931	http://bioinformatics.biol.uoa.gr/PRED-LIPO/datasets.html	http://bioinformatics.biol.uoa.gr/PRED-LIPO/input.jsp
PRED-COUPLE	DSDB0004	Protein	Classification	Predicting GPCRs using refined profile Hidden Markov Models	Fasta	The validation set, consist of 479 GPCR species homologues of the receptor subtypes with known coupling specificity (256 Gi/o, 102 Gq/11 and 121 Gs).	15847681	http://athina.biol.uoa.gr/bioinformatics/PRED-COUPLE2/dataset.fas	http://athina.biol.uoa.gr/bioinformatics/PRED-COUPLE/
S	DSDB0005	Protein	Classification	GPCR	Fasta	The original dataset S contains 780 GPCRs, of which are, 540 of Rhodopsin-like receptor, 75 of Peptide hormones receptor, 25 of Glutamate and calcium receptor, 12 of	19776029	http://peds.oxfordjournals.org/content/suppl/2009/09/23/gzp057.DC1	http://218.65.61.89:8080/bioinfo/GPCR-GIA

						Fungal mating pheromone receptor, 4 of Cyclic AMP Receptor, 56 of Odorant receptors in Drosophila, 21 of Gustatory receptor of Drosophila, 16 of Frizzled/Smoothened Family and 31 of T2R family in Mammals.		/gzp057_supp_A.pdf	
SGPCR	DSDB0006	Protein	Classification	Identifying GPCR and their families with grey incidence analysis	Fasta	The dataset contains 1,478 protein sequences, of which 367 are of GPCR protein and 1,101 are of non-GPCR (training test). The 367 GPCRs are further classified into the following six main families according to their binding with different ligand types: (1) Class A (rhodopsin-like), (2) Class B (secretin-like), (3) Class C (metabotropic glutamate/pheromone), (4) Class D (fungal pheromone), (5) Class E (cAMP receptors), and (6) Class F (frizzled/smoothened family).	21180772	http://www.rsc.org/suppdata/MB/c0/c0mb00170h/c0mb00170h-S1.pdf	http://icpr.jci.edu.cn/bioinfo/GPCR-2L
GRIFFIN	DSDB0007	Protein	Classification	Predicting GPCR with G-protein coupling selectivity using SVM and HMM	Fasta	The number of Class A sequences (training data) for SVM classification is 132 (Gi/o: 61 sequences; Gq/11: 47 sequences; Gs: 24 sequences). Class C sequences is classified into two types, Gi/o and Gq/11; The numbers of GPCRs are 170, 394, 34, 20, 9, 40 and 5 for opsins, olfactory receptors, Class B, Class C for Gi/o, Class C for Gq/11, frizzled and smoothened families, respectively.	15980445	http://griffin.cbrc.jp/training.html	http://griffin.cbrc.jp/
gpcr	DSDB0008	Protein	Classification	Classifying GPCRs with SVMs	Fasta	The GPCR superfamily set contains 692 sequences from Class A, 56 from Class B, 16 from Class C, 11 from Class D and 3 from Class E. Also included 99 decoy negative examples and 2425 additional negative examples from SCOP. The GPCR subfamily set contains in Level 1 1267 sequences and in Level 2 1171 sequences.	11836223	http://www.soe.ucsc.edu/research/compbio/gpcr/ http://compbio.soe.ucsc.edu/gpcr/	http://www.soe.ucsc.edu/research/compbio/
TMBETA-SVM	DSDB0009	Protein	Classification	Discrimination of OMPs using	Fasta	The total number of proteins in the dataset are 208 OMPs, 673 globular proteins and	16204348	http://www.cbrc.jp/~gromiha/omp/dataset2.ht	

				SVMs		206 a-helical membrane proteins.		ml	
S1	DSDB0010	Protein	Classification	Discrimination of β-barrel OMPs using HMMs	IDs	The training set consist of 14 OMPs, the validation set 119 OMPs and the test set 1100 sequences of globular proteins.	15070403	http://www.biomedcentral.com/1471-2105/5/29/additional/	http://bioinformatics.biol.uoa.gr/PRED-TMBB/input.jsp
PHOBIUS	DSDB0011	Protein	Prediction	Decoding a seguence with Homology information using HMMs	Labeled Fasta	Four different test sets: 292 sequences from transmembrane proteins in a 'TM' set. 2362 sequences from soluble proteins in a 'non-TM' set. 1320 sequences with signal peptides in a 'SP' set. 1334 sequences without signal peptide in a 'non-SP' set.	15961464#15111065	http://phobius.binf.ku.dk/data.html	http://phobius.sbc.su.se/ http://phobius.binf.ku.dk/
TMHMM	DSDB0012	Protein	Prediction	Predicting transmembrane helices in protein sequences using HMM	Labeled Fasta	set 1: It consists of 38 multi-spanning and 45 single-spanning proteins whose topologies have been experimentally determined. set 2: It contains 108 multi-spanning and 52 single-spanning proteins.	9783223	http://people.binf.ku.dk/krogh/TMHMM/index.html	http://www.cbs.dtu.dk/services/TMHMM-2.0/
HOMEP	DSDB0013	Protein	Classification	Applied Homology Modeling and Sequence Alignment Methods in Membrane Proteins	Fasta	The HOMEP data set contains 94 query-template pairs, from which 94 alignments and homology models can be onstructed	16648166	http://bhapp.c2b2.columbia.edu/software/Data_sets/homep/	
TMX	DSDB0014	Protein	Prediction	Predict the Burial status of TransMembrane residues	PDB IDs	The final data set comprise 41 protein chains of 2901 TM residues.	16397007#17237049#16838301	http://service.bioinformatik.uni-saarland.de/tmx/dataSet.html	http://service.bioinformatik.uni-saarland.de/tmx/index.html
MemType-2L	DSDB0015	Protein	Classification	Predicting membrane protein types	Fasta	(Supp-A) training dataset: 3,249 membrane proteins classified into 8 subsets according to their experimental annotations. (Supp-B) testing dataset: 4,333 membrane proteins classified into 8 subsets according to their experimental annotations. (Supp-C) non-membrane dataset: 7,965 non-membrane	17586467	http://www.csbio.sjtu.edu.cn/bioinf/MemType/Supp-A.pdf http://www.csbio.sjtu.edu.cn/bioinf/MemType/Supp-B.pdf	http://www.csbio.sjtu.edu.cn/bioinf/MemType/

						proteins.		http://www.csbio.sjtu.edu.cn/bioinf/MemType/Supp-C.pdf	
OCTOPUS	DSDB0016	Protein	Prediction	Predicting transmembrane protein topology using a combination of HMM and artificial neural networks.	Fasta	The dataset contains 124 protein chains with known three-dimensional structures.	18474507	http://topcons.net/index.php?about=octopus	http://octopus.cbr.su.se/
SPOCTOPUS	DSDB0017	Protein	Prediction	Predict a signal peptide using a neural network	Fasta	Two datasets: 1. ECOLI.TOP : contain 613 sequences with experimentally determined C-terminal locations for membrane proteins in Escherichia coli. 2. YEAST.TOP : contain 546 sequences with experimentally determined C-terminal locations for membrane proteins in Saccharomyces cerevisiae	18945683	http://octopus.cbr.su.se/index.php?about=download	http://octopus.cbr.su.se/
PRO/PRODIV-TMHMM	DSDB0018	Protein	Prediction	Predicting the topology of transmembrane helical proteins	Fasta	The data set contains 73 3D helix proteins	12824500	http://www.smbs.buffalo.edu/phys_bio/Softwares-Services_files/database.htm#tmp	http://www.smbs.buffalo.edu/phys_bio/Softwares-Services_files/service_0.htm
SCAMPI	DSDB0019	Protein	Prediction	Prediction of membrane-protein topology	Fasta	The high-resolution benchmark set consist of 123 membrane protein chains. The low-resolution set consist of 146 membrane protein chains.	18477697	http://scampi.cbr.su.se/index.php?about=SCAMPI	http://scampi.cbr.su.se/index.php
TOPCONS	DSDB0020	Protein	Prediction	Prediction of membrane protein topology and signal peptides	Fasta	Four Datasets: 313 proteins in the 'TM-set', 752 in the 'SP+TM', 3597 in the 'Globular' and 2194 in the 'Globular+SP' set	25969446	http://topcons.cbr.su.se/pred/download/	http://topcons.cbr.su.se/
TOPCONS-single	DSDB0021	Protein	Prediction	Rabid prediction of	Fasta	The benchmarked set is a modified version of the dataset used in SCAMPI (Bernsel et	21493661	http://single.topcons.net/index.php?about=TO	http://single.topcons.net/

				membrane protein topology		al., 2008). The original set consisted of two subsets stemming from the high-resolution structures (123 sequences) and from structures of lower resolution (146 sequences). The reduced set contain 101 sequences and was further divided into multi-spanning (79 sequences) and single-spanning (22 sequences) proteins resulting in three sets labeled ‘all’, ‘multi’ and ‘single’.		PCONS-single	
ZPRED	DSDB0022	Protein	Prediction	Predicting the distance to the membrane center for residues in a-helical membrane proteins	IDs	The dataset consisted of 101 non-homologous protein chains from 46 PDB structures obtained by X-ray diffraction. 147 membrane protein sequences with experimentally verified topologies were used for evaluating the topology prediction.	16873471	http://www.sbc.su.se/~erikgr/	
CW-PRED	DSDB0023	Protein	Classification	Prediction of cell wall sorting signals in gram-positive bacteria with a HMM	Fasta	The final set consisted of 55 sequences	18464329	http://bioinformatics.biol.uoa.gr/CW-PRED-results/	http://bioinformatics.biol.uoa.gr/CW-PRED/input.jsp
mem-type	DSDB0024	Protein	Classification	Prediction of Membrane Protein Types and Subcellular Locations	IDs	The 1st dataset contain 2,059 protein sequences of which 435 are type I transmembrane proteins, 152 type II transmembrane proteins, 1,311 multipass transmembrane proteins, 51 lipid chain-anchored membrane proteins, and 110 GPI-anchored membrane proteins. This dataset was used as a training dataset for predicting the membrane protein types. The 2nd dataset contain 2,105 protein sequences, of which 55 are chloroplast membrane proteins, 64 endoplasmic reticulum membrane proteins, 44 Golgi membrane proteins, 21 lysosome membrane proteins, 154 mitochondria membrane proteins, 26 nucleus membrane proteins, 37 peroxisome membrane proteins, 1,680 plasma membrane proteins, and 24 vacuole membrane proteins. This dataset	10336379	http://www.scirp.org/kcchou/papers/Protein_1999_mem-type.pdf	

						was used as a training dataset for predicting the cellular locations of membrane proteins.			
Golgi	DSDB0025	Protein	Classification	Prediction of Golgi Type II membrane proteins	IDs	The training dataset contain 129 non-redundant protein sequences. (Golgi non-redundant dataset, Post-Golgi non-redundant dataset)	12176834	http://ccb.imb.uq.edu.au/golgi/documents/Training_Set.html	http://ccb.imb.uq.edu.au/golgi/golgi_predictor.shtml
PredSL	DSDB0026	Protein	Classification	Prediction of Subcellular Location from the N-terminal Sequence	Fasta	The plant set consisting of 1,309 sequences (249 chloroplast,62 mitochondrial, 422 secreted, 171 cytoplasmic, and 405 nuclear sequences) and the non-plant set consisting of 10,559 sequences (366 mitochondrial, 5,247 secreted, 1,458 cytoplasmic, and 3,488 nuclear sequences)	16689702	http://aias.biol.uoa.gr/PredSL/datasets_down.html	http://aias.biol.uoa.gr/PredSL/input.html
NLStradamus	DSDB0027	Protein	Prediction	Using hidden Markov models (HMMs) to predict novel NLSs	IDs	The cNLS and non-cNLS set contain 26 sequences and the unknown set contain 22 sequences.	19563654		http://www.moseslab.csb.utoronto.ca/NLStradamus/
NESsential	DSDB0028	Protein	Prediction	Predict leucine-rich NESs from amino acid sequence	Fasta	The test data set conatins 70 sequences and the training data set contains 60 sequences.	21705415	http://seq.cbrc.jp/NESsential/DATA/	
PTS1 motifs	DSDB0029	Protein	Prediction	Prediction of Peroxisomal Proteome in Fungi, Plants and Animals	Fasta	152 peroxisomal proteins with PTS1 motif and 308 non-peroxisomal proteins with PTS1-like motif set used for the training and testing of neural networks and SVMs	12823981	http://www.sbc.su.se/~olofe/peroxi/	
NMT	DSDB0030	Protein	Prediction	Prediction of Substrate Proteins from Amino Acid Sequence	SWISSPROT ID	learning set of 390 proteins	11955008#11955007	http://mendel.imp.ac.at/myristate/SUPLsubstrates.htm	http://mendel.imp.ac.at/myristate/SUPLpredictor.htm
pkaPS	DSDB0031	Protein	Classification	Prediction of protein kinase A phosphorylation	UniProt ID	The final learning set consist of 143 sequences with 239 phosphorylated sites and 28 sequences with 1026 nonphosphorylated	17222345	http://mendel.imp.ac.at/sat/pkaPS/dataset.html	http://mendel.imp.ac.at/sat/pkaPS/

				sites using the simplified kinase binding model		serines and threonines			
plant	DSDB0032	Protein	Prediction	Prediction of plant proteins carrying PTS1	Fasta	The data set contain 2,562 peroxisomal proteins from known proteins and ESTs	21487095	http://www.plantcell.org/content/suppl/2011/04/06/tpc.111.084095.DC1.html	
PTS1 Predictor	DSDB0033	Protein	Prediction	Prediction of PTS1 containing proteins from amino acid sequence	UniProt ID	"LH set" contain 150 sequences, "SW set" contain 205 sequences. The learning set has a total size of 355 sequences, with 211 and 72 entries belonging to the metazoan and fungal taxonomic groups, in addition to 72 sequences from plants and protozoa.	12706717#12706718	http://mendel.imp.ac.at/pts1/learningSet.jsp	http://mendel.imp.ac.at/pts1/PTS1predictor.jsp
YASSPP	DSDB0034	Protein	Classification	A protein secondary structure prediction algorithm YASSPP use SVM-based models to compute a three-state prediction	Fasta	RS126 (contains 126 sequences), CB513 (contains 513 non-homologous sequences), EVAc4 (contains 165 sequences) datasets	16763996	http://glaros.dtc.umn.edu/yasspp/supplement/	http://www-users.cs.umn.edu/~karypis/servers/yasspp
CRPhos	DSDB0035	Protein	Prediction	Prediction of kinase-specific phosphorylation sites using conditional random fields	other	The dataset contains 5362 sequences.	18940828	http://www.ptools.ua.ac.be/CRPhos	http://www.ptools.ua.ac.be/CRPhos
ELM instance	DSDB0036	Protein	Prediction	Investigating candidate functional sites in eukaryotic proteins	Fasta	The dataset contains 2585 proteins.	12824381#19920119#24214962	http://elm.eu.org/downloads.html	http://elm.eu.org/

GPS 2.0	DSDB0037	Protein	Classification	Prediction of Kinase-Specific Phosphorylation Sites	UniProt ID	The training data set containing 3,161 verified phosphorylation sites with respective kinase information.	18463090#21062758	http://gps.biocuckoo.org/download/Large-scale_Prediction.txt.gz	http://gps.biocuckoo.org/online.php
AutoMotif 2.0	DSDB0038	Protein	Classification	Prediction of phosphorylation sites in proteins using SVM	Protein Sequences	Use of Phospho.ELM dataset version 8.2 which contains 4,687 substrate proteins covering 2,217 tyrosine, 14,518 serine and 2,914 Threonine instances	20423529#17994256#22555647#15728119	https://code.google.com/p/automotifserver/downloads/list	
KinasePhos 1.0	DSDB0039	Protein	Prediction	A tool for identifying protein kinase-specific phosphorylation sites	UniProt ID	PhosphoBase consists of 1,083 experimentally verified phosphorylation sites and the Swiss-Prot with 3,614 entries.	15980458	http://kinasephos.mbc.nctu.edu.tw/download.html	http://kinasephos.mbc.nctu.edu.tw/
MetaPS06	DSDB0040	Protein	Classification	Meta-prediction of phosphorylation sites with weighted voting and restricted grid search parameter selection	Sequences	The data set contains 3,252 sites. (CDK, CK2, PKA, PKC)	18234718	http://nar.oxfordjournals.org/content/36/4/e22/suppl/DC1	
PRED-SIGNAL	DSDB0041	Protein	Prediction	Prediction of signal peptides in archaea	Fasta	three sets: the SP contains 69 archaeal proteins with a verified SP, the TM contains 69 proteins with an N-terminal TM segment and the Globular contains 183 archaeal cytoplasmic proteins	18988691	http://bioinformatics.biol.uoa.gr/PRED-SIGNAL/input.jsp	http://bioinformatics.biol.uoa.gr/PRED-SIGNAL/input.jsp
DNcon	DSDB0042	Protein	Prediction	Predicting protein residue-residue contacts using deep networks and boosting	Fasta	The primary dataset, DNCON, consisting of 1426 proteins. This dataset was randomly split into two sets: DNCON_TRAIN consisting of 1230 proteins and DNCON_TEST consisting of 196 proteins. The evaluation datasets used included D329, a set of 329 proteins; SVMCON_TEST, a set of 48 short to medium length proteins; CASP9, a set of 111 targets;	23047561	http://iris.rnet.missouri.edu/dncon/datasets/	http://iris.rnet.missouri.edu/dncon/

						CASP9_HARD, a subset of 16 targets.			
PhyML	DSDB0043	Protein	Prediction	Phylogenetic prediction based on Likelihood methods	Sequences	Real datasets: RDPII data set (218 taxa, 4182 bp), rbcL data set (500 taxa, 1428 bp). Benchmark datasets: Medium-size data set contains 50 protein alignments and Large-size data sets contains 10 protein alignments. The benchmark contains 100 simulated data sets of 40 sequences and 500 sites.	20525638	http://www.atgc-montpellier.fr/phyml/benchmarks/	http://www.atgc-montpellier.fr/phyml/
NetPhosBac1.0	DSDB0044	Protein	Prediction	A predictor for Ser/Thr phosphorylation sites in bacterial proteins	UniProt ID	The training data set contains 14 phosphorylation sites from PSD , 71 sites from B. subtilis and 102 sites from E. coli.	19053140	http://onlinelibrary.wiley.com/store/10.1002/pmic.200800285/asset/supinfo/pmic_200800285_sm_miscellaneous_information.pdf?v=1&s=b386b4997eed11bd6e7b9fa026351447f91f5b31	http://www.cbs.dtu.dk/services/NetPhosBac-1.0/
PostMod	DSDB0045	Protein	Prediction	An effective method to recognize phosphorylation sites by combining sequence patterns and evolutionary information	UniProt ID	The test set contains 48 different kinase groups	20122181	http://pbil.kaist.ac.kr/PostMod/	http://pbil.kaist.ac.kr/PostMod/
ScanProsite	DSDB0046	Protein	Prediction	A web interface to identify protein matches against signatures	UniProt ID	2.371 subsets	16845026	ftp://ftp.expasy.org/databases/prosite/	http://prosite.expasy.org/scanprosite/
RHYTHM	DSDB0047	Protein	Prediction	Predict the orientation of transmembrane helices in channels and	PDB ID	Two different sets of propensity matrices derived from representative and non-redundant datasets of 21 channels and 14 membrane-coils	19465378	http://proteinformatics.charite.de/rhythm/index.php?site=methods&sub=learning	

				membrane-coils					
PHOSIDA	DSDB0048	Protein	Prediction	Identifying de novo phosphorylation motifs in datasets	Sequences	Four datasets: EGF_Stimulation, Cell Cycle, HEPA1-6, Melanoma	18039369#21081558	http://www.phosida.com/	http://www.phosida.com/
PhosPhAt 4.0	DSDB0049	Protein	Prediction	The PhosPhAt service has a built-in plant specific phosphorylation site predictor trained on the experimental dataset for Serine, threonine and tyrosine phosphorylation (pSer, pThr, pTyr)	TAIR ID	The Dataset contains 8627 proteins of Arabidopsis thaliana	23172287#17984086#19880383	http://phosphat.uni-hohenheim.de/download.html	http://phosphat.uni-hohenheim.de/index.html
PhosphoSVM	DSDB0050	Protein	Prediction	Prediction of phosphorylation sites with SVMs	Fasta	Two datasets: PELM (S- type 6,635 Sequences,T- type 3,227 Sequences, Y- type 1,392 Sequences), PPA (S- type 3,037 Sequences, T- type 1,359 Sequences, Y- type 617 Sequences)	24623121	http://sysbio.unl.edu/PhosphoSVM/download.php	http://sysbio.unl.edu/PhosphoSVM/prediction.php
PKIS	DSDB0051	Protein	Classification	Computational Identification of Protein Kinases for Experimentally Discovered Protein Phosphorylation Sites	Fasta	The dataset contains 5,374 proteins.	23941207	http://bioinformatics.ustc.edu.cn/pkis/download.html	http://bioinformatics.ustc.edu.cn/pkis/tool.html
PTMPred	DSDB0052	Protein	Prediction	Predicting post-transcriptional modification	Sequences	Six training datasets: CDK_Neg, CDK_Pos, OGly_Neg, OGly_Pos, PKA_Neg,	24291233	http://doc.aporc.org/wiki/PTMPred	

				sites based on protein sequence		PKA_Pos			
Musite	DSDB0053	Protein	Classification	Prediction of General and Kinase-specific Phosphorylation Sites	Sequences	Six training datasets: A.thaliana (33410 proteins), C.elegans (23102 proteins), D.melanogaster (16206 proteins), H.sapiens (20595 proteins), M.musculus (18732 proteins), S.cerevisiae (19012 proteins)	20702892#22934099	http://musite.sourceforge.net/datasets.php	
PSEA	DSDB0054	Protein	Prediction	Predict the kinase types and kinase-specific phosphorylation site for a protein	Fasta	The dataset contains 128122 phosphorylation sites within 32148 proteins, where the number of serine (S), threonine (T) and tyrosine (Y) are 69315, 30398 and 28409	24681538	http://bioinfo.ncu.edu.cn/PKPred_Download.aspx	http://bioinfo.ncu.edu.cn/PKPred_Prediction.aspx
iGPS	DSDB0055	Protein	Prediction	Prediction of in vivo site-specific kinase-substrate relations mainly from the phosphoproteomic data	Fasta	The phosphorylation data set contains 145,646 p-sites in 28,457 substrates, with 14,534, 5555, 15,622, 49,119, and 60,816 p-sites in S. cerevisiae, C.elegans, D. melanogaster, M. musculus, and H. sapiens.	22798277	http://igps.biocuckoo.org/faq.php	
Myristoylator	DSDB0056	Protein	Prediction	N-Terminal myristoylation predictions by ensembles of neural networks	UniProt ID	Two datasets, Positive (327 proteins) and Negative (390 proteins)	15174132	http://web.expasy.org/myristoylator/myristoylator-data.html	http://web.expasy.org/myristoylator/
Dataset #1	DSDB0057	Protein	Prediction	3D structure-based epitope prediction methods	PDB ID	The dataset #1 contains 82 3D structures of antibody-protein complexes.	17910770	http://www.biomedcentral.com/1472-6807/7/64/additional	
GPS-ARM	DSDB0058	Protein	Prediction	Computational Analysis of the APC/C Recognition Motif by Predicting D-	UniProt ID	The dataset contains 74 D-boxes (from 68 unique proteins) and 44 KEN-boxes (from 42 unique proteins)	22479614	http://arm.biocuckoo.org/faq.php	http://arm.biocuckoo.org/online.php

				Boxes and KEN-Boxes					
GPS-MBA	DSDB0059	Protein	Prediction	Prediction of I-Ag7 and HLA-DQ8	UniProt ID	The dataset contains 318 mouse I-Ag7 binding peptides (from 177 proteins), and 134 human HLA-DQ8 epitopes (from 85 proteins)	22479466	http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0033884#s5	http://mba.biocuckoo.org/online.php
GPS-PUP	DSDB0060	Protein	Prediction	Computational prediction of pupylation sites in prokaryotic proteins	UniProt ID	Tha dataset contains 127 pupylation sites from 109 unique proteins	21850344	http://www.rsc.org/suppdata/mb/c1/c1mb05217a/c1mb05217a.pdf	http://pup.biocuckoo.org/online.php
GPS-Polo	DSDB0061	Protein	Prediction	Analysis of Plk-specific phospho-binding and phosphorylation sites (p-sites) in proteins	UniProt ID	The dataset contains 56 phospho-binding sites (from 47 distinct substrates) and 275 phosphoylation sites (from 124 unique proteins)	22851512	http://polo.biocuckoo.org/faq.php	http://polo.biocuckoo.org/online.php
GPS-CCD	DSDB0062	Protein	Prediction	Prediction of Calpain Cleavage Sites	UniProt ID	The dataset contains 368 calpain cleavage sites from 130 proteins	21533053	http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0019001#s5	http://ccd.biocuckoo.org/online.php
GPS-SNO	DSDB0063	Protein	Prediction	Prediction of Protein S-Nitrosylation Sites with a Modified GPS Algorithm	UniProt ID	The dataset contains 504 S-nitrosylation sites from 327 unique proteins	20585580	http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0011290#s5	http://sno.biocuckoo.org/online.php
GPS-SUMO	DSDB0064	Protein	Prediction	Prediction of sumoylation sites and SUMO-interaction motifs	UniProt ID	The training data set contains 912 sumoylation sites in 510 protein and 137 SIMs in 80 proteins.	24880689	http://nar.oxfordjournals.org/content/42/W1/W325/suppl/DC1	http://sumosp.biocuckoo.org/online.php

Dataset #2	DSDB0065	Protein	Prediction	3D structure-based epitope prediction methods	PDB ID	The dataset #2 contains 59 structures of one-chain (monomer) antigens in complexes with two-chain antibody fragments.	17910770	http://www.biomedcentral.com/1472-6807/7/64/additional	
UbPred	DSDB0066	Protein	Prediction	Random forest-based predictor of potential ubiquitination sites in proteins	Fasta	The dataset of 4,651 non-ubiquitinated (negative) fragments were extracted from 124 mitochondrial matrix proteins.	19722269	http://ubpred.org/help.html#datasets	http://ubpred.org/index.html
KohGPI	DSDB0067	Protein	Classification	Identification of GPI anchor attachment signals by a Kohonen self-organizing map	Fasta	The negative training and evaluation sets consisted of 256 known cytosolic and 128 transmembrane proteins of all eukaryote kingdoms. The negative test and evaluation sets consisted of 229 and 129 transmembrane proteins	15691858	http://gpi.unibe.ch/	http://gpi.unibe.ch/
NsitePred	DSDB0068	Protein	Prediction	Prediction and Analysis of Nucleotide Binding Residues Using Sequence and Sequence-derived Structural Descriptors	other	Dataset 1 includes 227, 321, 140, 56 and 105 chains that bind to ATP, ADP, AMP, GTP and GDP. Dataset 2 includes 17, 25, 18, 6, and 9 chains that bind to ATP, ADP, AMP, GTP, and GDP. Dataset 3 consists1372 chains that do not interact with the nucleotides	22130595	http://biomine-ws.ece.ualberta.ca/NsitePred.html	http://biomine-ws.ece.ualberta.ca/NsitePred.html
ChloroP	DSDB0069	Protein	Classification	Prediction of chloroplast transit peptides and their cleavage sites	Fasta	The training set contains 150 sequences, whereof 75 were chloroplast transit peptide (cTP) containing	10338008	http://www.cbs.dtu.dk/services/ChloroP/pages/datasets.php	http://www.cbs.dtu.dk/services/ChloroP/
LipoP	DSDB0070	Protein	Classification	Prediction of lipoprotein signal peptides in Gram-negative bacteria	UniProt ID	The data set consisted of 63 nonhomologous lipoproteins, 328 SPaseI-cleaved proteins and 388 cytoplasmic proteins.	12876315		http://www.cbs.dtu.dk/services/LipoP/

Secretome P	DSDB0071	Protein	Classification	Prediction of non-classical protein secretion	Fasta	The positive data sets consists 152 and 350 sequences for Firmicutes and Proteobacteria. The negative data sets, consists 140 and 334 sequences for Firmicutes and Proteobacteria	15115854#16212653	http://www.cbs.dtu.dk/services/SecretomeP/datasets.php	http://www.cbs.dtu.dk/services/SecretomeP/
SignalP	DSDB0072	Protein	Prediction	Predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms: Gram-positive prokaryotes, Gram-negative prokaryotes, and eukaryotes	Fasta label	Three datasets: eukaryotes, Gram-positive bacteria and Gram-negative bacteria	21959131	http://www.cbs.dtu.dk/services/SignalP/data.php	http://www.cbs.dtu.dk/services/SignalP/
TargetP	DSDB0073	Protein	Classification	Predicting subcellular localization of proteins based on their N-terminal amino acid sequence	Fasta	Two Datasets, Plant (940 proteins) and Non-plant (2738 proteins)	10891285	http://www.cbs.dtu.dk/services/TargetP/datasets/datasets.php	http://www.cbs.dtu.dk/services/TargetP/
BepiPred	DSDB0074	Protein	Prediction	Predicts the location of linear B-cell epitopes using HMM	Fasta label	Three data sets, the Pellequer data set that contains 14 protein sequences, the AntiJen data set that contains 127 protein sequences and the HIV data set that contains 10 protein sequences	16635264	http://www.cbs.dtu.dk/suppl/immunology/Bepipred.php	http://www.cbs.dtu.dk/services/BepiPred/
DiscoTope	DSDB0075	Protein	Prediction	Predicts discontinuous B cell epitopes from protein 3D structures	PDB ID	The Training dataset consist of 75 x-ray crystal structures of antigen-antibody complexes. The evaluation dataset contains 52 structures identified as antigen-antibody complexes	23300419	http://www.cbs.dtu.dk/suppl/immunology/DiscoTope-2.0/	http://www.cbs.dtu.dk/services/DiscoTope/
NetCTL	DSDB0076	Protein	Prediction	Predicts CTL epitopes in	Fasta	The HIV dataset contains 216 epitope-protein pairs restricted to all 12 recognized	17973982	http://www.cbs.dtu.dk/suppl/immunology/CT	http://www.cbs.dtu.d

				protein sequences		supertypes. The other dataset is called HIVEpiJen and contains 87 epitopes restricted to the A1, A2, or A3 supertypes. The SYFPETHI dataset contained a total of 863 epitope-protein pairs		L-1.2.php	k/services/NetCTL/
NetCTLpan	DSDB0077	Protein	Prediction	Predicts CTL epitopes in protein sequences	Fasta	The SYF data set consisting of 2,267 HLA class I ligand pairs with corresponding source proteins, where 226 ligands are 8-mers, 1,443 are 9-mers, 430 are 10-mers, and 168 ligands belong to the group of 11-mers. The HIV dataset contains 216 epitope-protein pairs	20379710	http://www.cbs.dtu.dk/suppl/immunology/NetCTLpan.php	http://www.cbs.dtu.dk/services/NetCTLpan/
NetMHCII	DSDB0078	Protein	Prediction	Predicts binding of peptides to HLA-DR, HLA-DQ, HLA-DP and mouse MHC class II allele	other	The data set comprises 14 HLA-DR alleles each characterized by at least 420 and up to 5166 peptide binding data points. The binding data for each HLADR allele was partitioned into 5 data sets	19765293	http://www.cbs.dtu.dk/suppl/immunology/NetMHCII-2.0.php	http://www.cbs.dtu.dk/services/NetMHCII/
NetMHCpan	DSDB0079	Protein	Prediction	Predicts binding of peptides to any known MHC molecule using ANNs	Fasta	The data set consisted of 79,137 unique peptide–MHC class I	19002680	http://www.cbs.dtu.dk/suppl/immunology/NetMHCpan-2.0.php	http://www.cbs.dtu.dk/services/NetMHCpan/
distanceP	DSDB0080	Protein	Prediction	Predicts distance constraints between amino acids in proteins from the amino acid sequence	PDB ID	The training data set contains 650 proteins	10786291	http://www.cbs.dtu.dk/services/distanceP/intro.html	http://www.cbs.dtu.dk/services/distanceP/
NetDiseaseSNP	DSDB0081	Protein	Prediction	Predicts whether a single non-synonymous SNP causes a disease or is	Fasta	The training data set contains 10003 proteins	23935863	http://www.cbs.dtu.dk/services/NetDiseaseSNP/trainingset.php	http://www.cbs.dtu.dk/services/NetDiseaseSNP/

				invariant.					
SigniSite	DSDB0082	Protein	Prediction	Identifying amino acid residues significantly associated with the phenotype of the data set	Fasta	18 different benchmark data sets	23761454	http://www.cbs.dtu.dk/services/SigniSite/	http://www.cbs.dtu.dk/services/SigniSite/
VarDom	DSDB0083	Protein	Classification	Classification analysis of the malaria antigen family PfEMP1	Fasta	The DataSet contains 403 proteins	20862303	http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000933#pcbi.1000933.s001	http://www.cbs.dtu.dk/services/VarDom/
NetTepi	DSDB0084	Protein	Classification	Predicts T-cell epitopes from protein sequences	Fasta	The training data sets contained of 1201 epitopes and the evaluation data sets contained of 557 epitopes	24863339	http://www.cbs.dtu.dk/suppl/immunology/NetTepi-1.0/	http://www.cbs.dtu.dk/services/NetTepi-1.0/
DictyOGlyc	DSDB0085	Protein	Prediction	Prediction for GlcNAc O-glycosylation sites in Dictyostelium discoideum proteins	UniProt ID	The Data Set contains 652 Dictyostelium discoideum proteins	10521537	http://www.cbs.dtu.dk/services/DictyOGlyc/STATS/AllNames.html	http://www.cbs.dtu.dk/services/DictyOGlyc/
NetACet	DSDB0086	Protein	Prediction	Predicts substrates of N-acetyltransferase A	Fasta Label	Two training DataSets: the positive set contains 61 sequences and the negative set contains 76 sequences	15539450	http://www.cbs.dtu.dk/services/NetAcet/background/trainingset.php#trainingset	http://www.cbs.dtu.dk/services/NetAcet/
NetCGlyc	DSDB0087	Protein	Prediction	Prediction of C-mannosylation sites in mammalian proteins	Fasta Label	The training Data set contains 53 mammalian mucin type glycoproteins	17494086	http://www.cbs.dtu.dk/databases/OGLYCBASE/	http://www.cbs.dtu.dk/services/NetCGlyc/
NetGlycate	DSDB0088	Protein	Prediction	Predicts glycation of ε	Fasta	The data set consists of 20 proteins with 89 glycated lysines and 126 non-glycated	16762979	http://www.cbs.dtu.dk/databases/GlycateBase-	http://www.cbs.dtu.dk/services/NetGlycat

				amino groups of lysines in mammalian proteins	Label	lysines		1.0/	e/
NetOGlyc	DSDB0089	Protein	Prediction	Predictions of mucin type GalNAc O-glycosylation sites in mammalian proteins	UniProt ID	The Dataset contains 662 Glycoproteins	23584533	http://emboj.embopress.org/content/32/10/1478.long	http://www.cbs.dtu.dk/services/NetOGlyc/
	DSDB0090	Protein	Classification	Multi-class protein fold pretiction using SVMs and NN	Fasta	The training set contains 311 proteins and testing set contains 383 proteins	11301304	http://ranger.uta.edu/~chqding/protein/	
	DSDB0091	Protein	Classification	Determination of Protein Structural Classes	PDB ID	The Data Set contains 204 Proteins	10527868	http://www.sciencedirect.com/science/article/pii/S0006291X99913256	
25PDB	DSDB0092	Protein	Classification	Prediction of structural classes for protein sequences and domains	PDB ID	The 25PDB dataset contains 1673 proteins and domains. The 1189 datasets contains 223 all-a, 294 all-b, 334 a/b, and 241 a+b domains and sequences.		http://biomine.ece.ualberta.ca/papers/PR-ProteinStructuralClasses2006.pdf	
S1859	DSDB0093	Protein	Classification	Predicting protein stability changes upon mutations	PDB ID	The dataset consists of 1859 different single point mutations and is comprised of 64 protein sequences.	17379687	http://210.60.98.19/IPTREEr/iptree.htm	http://210.60.98.19/IPTREEr/iptree.htm
I-Mutant	DSDB0094	Protein	Classification	Prediction of protein stability changes upon single-site mutations	PDB ID	The dataset contains 2087 different single mutations in 65 different proteins. The subset of structures known with atomic resolution contains 1948 different single mutations	15980478	http://folding.biofold.org/i-mutant//pages/dbMut.html	http://folding.biofold.org/i-mutant/i-mutant2.0.html

MUpro	DSDB0095	Protein	Classification	Prediction of protein stability changes for single-site Mutations using SVMs	Fasta Label	The dataset S1615 contains s 1615 single site mutations obtained from 42 different proteins and the subset S388 contains 388 unique mutations	16372356	http://download.igb.uci.edu/	http://www.ics.uci.edu/~baldig/mutation.html
SCRATCH	DSDB0096	Protein	Prediction	Prediction of protein tertiary structure and structural features	Fasta	The training dataset contains 5772 proteins	15980571#24860169	http://download.igb.uci.edu/	
DOMpro	DSDB0097	Protein	Prediction	Protein Domain Prediction	Fasta Label	The dataset contains 354 multi-domain chains and 963 single-domain chains		http://download.igb.uci.edu/	http://scratch.proteomics.ics.uci.edu/
DIpro	DSDB0098	Protein	Classification	Large-Scale Prediction of Disulphide Bridges	Fasta	The dataset (SPX) contains 1018 proteins	16320312	http://download.igb.uci.edu/intro.html#5	http://download.igb.uci.edu/bridge.html
BETApro	DSDB0099	Protein	Prediction	Prediction of beta residue pairs, stand pairs, strand alignments, and beta sheets	Fasta Label	The dataset (BetaSheet916) contains 916 proteins	15961501	http://www.ics.uci.edu/~baldig/betasheet_data.html	http://www.ics.uci.edu/~baldig/betasheet.html
DISpro	DSDB0100	Protein	Prediction	Prediction of disordered regions from protein sequences	Fasta Label	The dataset (Disorder723) contains 723 non-redundant protein chains		http://download.igb.uci.edu/disorder.dataset	http://scratch.proteomics.ics.uci.edu/
SOLpro	DSDB0101	Protein	Classification	Prediction of protein solubility upon overexpression	Fasta	The dataset contains (SOLP) 17408 proteins (8704 soluble, 8704 insoluble)	19549632	http://download.igb.uci.edu/SOLP.fa	http://scratch.proteomics.ics.uci.edu/
SIDEpro	DSDB0102	Protein	Classification	Prediction of side-chain	PDB ID	Three DataSets: (1) a benchmark dataset of 379 proteins (SCWRL4 dataset), (2) 94	22072531	http://download.igb.uci.edu/	

				conformations for protein backbones		proteins determined by X-ray crystallography (CASP9 dataset) and (3) a small set of seven large protein complexes ranging in size from 2760 to 8767 residues (COMPLEXES dataset)			
BaCello	DSDB0103	Protein	Classification	Prediction of the subcellular localization for proteins in eukaryotes	Fasta	Three datasets: 2597 proteins from animals, 1198 proteins from fungi and 491 proteins from plants, distributed in five locations: nucleus, cytoplasm, secretory pathway, mitochondrion and chloroplast	16873501	http://gpcr.biocomp.unibo.it/bacello/dataset.htm	http://gpcr.biocomp.unibo.it/bacello/pred.htm
DisLocate	DSDB0104	Protein	Prediction	Prediction of cysteine connectivity patterns in a protein chain	Fasta	The training set consists of 38 high-resolution experimentally determined outer-membrane proteins of Prokaryotes	19849839	http://dislocate.biocomp.unibo.it/dislocate/default/method	http://dislocate.biocomp.unibo.it/dislocate/default/index
FT-COMAR	DSDB0105	Protein	Prediction	Fault Tolerance Reconstruction of 3D Structure from Protein Contact Maps	PDB ID	The dataset contains 100 mono domain proteins	18381401	http://bioinformatics.cs.unibo.it/FT-COMAR/index.html#data	http://bioinformatics.cs.unibo.it/FT-COMAR/page0/page0.html
K-Fold	DSDB0106	Protein	Prediction	Prediction of the Protein Folding Kinetic Order and Rate	PDB ID	The Dataset contains 63 proteins		http://folding.biofold.org/k-fold/pages/dbFold.html	http://folding.biofold.org/k-fold/K-Fold.html
MemLoci	DSDB0107	Protein	Classification	Membrane Protein Subcellular Localization Predicto	Fasta	The training and evaluating dataset consists of 10 634 sequences: 4016 from plasma membrane, 2308 from organelle membranes and 4310 from the internal membranes	21367869	http://mu2py.biocomp.unibo.it/memloci/default/dataset	http://mu2py.biocomp.unibo.it/memloci/default/predict
MemPype	DSDB0108	Protein	Classification	Prediction of signal peptides	UniProt ID	Four Datasets: Spep (contains 1287 eukaryotic proteins), PredGPI (contains 340 and 10 630 GPI- and non-GPI-anchored proteins), ENSEMBLE 3.0 (train and test), MemLoci (68 proteins)	21543452	http://mu2py.biocomp.unibo.it/mempype/default/dataset	http://mu2py.biocomp.unibo.it/mempype/default/predict

PhD-SNP	DSDB0109	Protein	Classification	Prediction of human Deleterious Single Nucleotide Polymorphisms	UniProt ID	The dataset consists of 21 185 different single point mutations (12 944 of which are disease-related and 8241 are described as neutral polymorphisms), obtained from 3587 protein sequences. The subset HumVarProf contains 8718 mutations (3852 of which disease-related and 4866 neutral polymorphisms. The third set HumVar contains 935 single point protein mutations (149 of which are disease-related and 786 are described as neutral polymorphisms) from a total of 469 different proteins	16895930	http://snps.biofold.org/phd-snp/pages/PhD-SNP_HelpOld.html	http://snps.biofold.org/phd-snp/PhD-SNP.html
PredGPI	DSDB0110	Protein	Classification	Prediction of GPI-anchored proteins	Fasta	Four datasets: GPIω-Set, which contains 26 proteins whose ω-sites are known, GPI-Set, which contains 145 proteins GPI-anchored, All-GPI-Set, which contains 340 proteins GPI-anchored, Non-GPI-Set, comprising 10,630 proteins chains not-GPI-anchored	18811934	http://gpcr.biocomp.unibo.it/predgpi/dataset.htm	http://gpcr.biocomp.unibo.it/predgpi/pred.htm
PolyPhen	DSDB0111	Protein	Classification	Prediction of the possible impact of an amino acid substitution on the structure and function of a human protein	Fasta	Two Datasets: HumDiv: 5564 deleterious and 7539 neutral mutations from 978 human proteins.HumVar: 22196 deleterious and 21119 neutral mutations in 9679 human proteins, no restriction on deleterious and neutral mutations	20354512	http://genetics.bwh.harvard.edu/pph2/dokuwiki/downloads	http://genetics.bwh.harvard.edu/pph2/index.shtml
SIFT	DSDB0112	Protein	Prediction	Prediction whether an amino acid substitution affects protein function	Gene Name	Three test Datasets: LacI (4004 substitutions), HIV-1 protease (336 substitutions), and bacteriophage T4 lysozyme (2015 substitutions)	11337480	http://sift.bii.a-star.edu.sg/index.html	
TPpred	DSDB0113	Protein	Prediction	Prediction of organelle-targeting peptides in eukaryotic proteins	Fasta	The dataset consists of 297 sequences with targeting peptide (DBþ) and 8010 without targeting peptide (DB)	23428638	http://biocomp.unibo.it/~valentina/TPpred/	http://tppred.biocomp.unibo.it/tppred/default/index

FOLDpro	DSDB0114	Protein	Classification	Prediction of protein 3D structure using a machine learning fold recognition approach	Query Name	The Dataset (Lindahl's) contains 976 proteins	16547073	http://mine5.ics.uci.edu:1026/fold_help.html	http://mine10.ics.uci.edu/
SVMcon	DSDB0115	Protein	Classification	Prediction of medium- to long-range residue-residue contacts using SVMs	Label	The training dataset contains 485 proteins and the test dataset contains 48 proteins	17407573	http://scratch.proteomics.ics.uci.edu/explanation.html#SVMcon	http://scratch.proteomics.ics.uci.edu/
	DSDB0116	Protein	Classification	Predicting Protein Quaternary Structure by Pseudo Amino Acid Composition	UniProt ID	The training data set consists of 3174 protein sequences, of which 382 are with annotation of monomer, 817 of dimer, 593 of trimer, 884 of tetramer, 54 of pentamer, 287 of hexamer, and 157 of octamer. The testing data set contains 332 protein sequences, of which 50 are with annotation of monomer, 102 of dimer, 56 of trimer, 80 of tetramer, 6 of pentamer, 28 of hexamer, and 10 of octamer.	14517979	http://onlinelibrary.wiley.com/doi/10.1002/prot.10500/suppinfo	
DNA-BP	DSDB0117	Protein	Classification	Predicting DNA-binding proteins	PDB ID	The positive data set contains 118 DNA-BPs. The negative dataset contains 231 non-DNA-BPs	17624492	http://link.springer.com/article/10.1007%2Fs00726-007-0568-2	http://www.csbio.sjtu.edu.cn/bioinf/PseAA/
DNAbinder	DSDB0118	Protein	Classification	Predicting DNA-binding proteins	Sequences	The DNAset or main dataset, consists of 146 DNA-binding and 250 non-binding protein chains. The DNAaset or alternate dataset consists of 1153 DNA-BPs and 1153 NBPs. The DNAiset or independent dataset has 92 DNA-binding protein chains. The DNArset or realistic dataset has 146 DNA-BPs used in DNAset and 1500 NBPs	18042272	http://www.imtech.res.in/raghava/dnabinder/download.html	http://www.imtech.res.in/raghava/dnabinder/submit.html
SVM–PSSM	DSDB0119	Protein	Classification	Predicting DNA-binding proteins using	PDB ID	Three datasets: The PDNA-62 Dataset contains 62 proteins, the PDNA-48 dataset contains 48 protein chains and the PDC-59	17275170	http://ir.lib.nctu.edu.tw/bitstream/987654321/3	

				hybrid SVM–PSSM		contains 59 proteinc hains		0454/1/050305008.pdf	
DBS-PRED	DSDB0120	Protein	Classification	Prediction of DNA-binding in proteins using Neural Networks	PDB ID	Three data sets: PDNA-62, NRTF-915, CNTR-3332	14990443	http://bioinformatics.oxfordjournals.org/content/20/4/477.full.pdf	http://dbs-pred.netasa.org/
ANOLEA	DSDB0121	Protein	Prediction	Prediction of a protein structure in the last stages of the model building process	PDB ID	The data set contains 147 proteins	9322034	http://melolab.org/anolea/protein_database.html	http://melolab.org/anolea/
ReadOut	DSDB0122	Protein	Prediction	Calculation of direct and indirect readout energies and specificities for protein–DNA recognition	PDB ID	The Dataset contains 62 proteins	16844974	http://readout.netasa.org/	http://readout.netasa.org/
SDCPred	DSDB0123	Protein	Prediction	Prediction of mono- and di-nucleotide specific DNA-binding sites in proteins using neural networks	PDB ID	The training dataset PDNA159 contains 159 protein chains	19439068	http://sdcpred.netasa.org/	http://sdcpred.netasa.org/
DNABIND	DSDB0124	Protein	Classification	DNA Binding Protein prediction	PDB ID	Two Datasets: The PD138 data set consists of 138 DNA-binding protein chain structures. The PD54 data set (unbound set UD54 and the bound set BD54) contains 54 proteins each.	16551468	http://www.szialab.org/data/uploads/ownpdf/nucpred.pdf	http://dnabind.szialab.org/
DP-Bind	DSDB0125	Protein	Classification	Sequence-based prediction of DNA-binding residues in DNA-binding	PDB ID	The dataset is a non-redundant set of 62 experimentally solved protein–DNA complexes	17237068#16568445	http://lcg.rit.albany.edu/dp-bind/dpbind_supplement.html	http://lcg.rit.albany.edu/dp-bind/

				proteins					
DISPLAR	DSDB0126	Protein	Classification	Prediction of DNA-binding sites on protein surfaces	PDB ID	The Dataset contains 264 DNA-binding proteins	17284455	http://pipe.scs.fsu.edu/displar_data.pdf	http://pipe.scs.fsu.edu/displar.html
WESA	DSDB0127	Protein	Classification	Prediction of solvent accessibility and sites of deleterious mutations from protein sequence	PDB ID	The Dataset contains 2,148 nonhomologous protein chains	15937195	http://pipe.scs.fsu.edu/wesa_data.pdf	http://pipe.scs.fsu.edu/wesa.html
cons-PPISP	DSDB0128	Protein	Classification	Prediction of interface residues in protein-protein complexes by a consensus neural network method: test against NMR data	PDB ID	The Dataset consists of 1,256 protein chains	16080151	http://pipe.scs.fsu.edu/ppisp_data.pdf	http://pipe.scs.fsu.edu/ppisp.html
meta-PPISP	DSDB0129	Protein	Classification	Protein-protein interaction site prediction	PDB ID	The Dataset contains 35 proteins (Enz35)	17895276	http://pipe.scs.fsu.edu/meta-ppisp-SI.pdf	http://pipe.scs.fsu.edu/meta-ppisp.html
BindN-RF	DSDB0130	Protein	Classification	Prediction of DNA-binding residues in proteins	PDB ID	Four Datasets.The PRINR25 dataset contains 107 protein sequences. The PDNA-62 dataset contains 62 non-redundant sequences. The test set TestPDB contains 92 protein sequences. The test set TestSP contains 100 protein sequences.	19594868#16845003	http://bioinfo.ggc.org/bindn-rf/	http://bioinfo.ggc.org/bindn-rf/
DBindR	DSDB0131	Protein	Classification	Prediction of DNA-binding residues in proteins from	PDB ID	Three Datasets. The DBP-374 contains 374 structures of representative protein–DNA complexes. The PDNA-62 contains 62 proteins. The test Dataset TS75 contains 75	19008251	http://www.cbi.seu.edu.cn/DBindR/supplementary.htm	http://www.cbi.seu.edu.cn/DBindR/DBindR.htm

				amino acid sequences		proteis.			
	DSDB0132	Protein	Classification	Predicting residue solvent accessibility from protein sequence	PDB ID	The learning set contains 338 monomeric, non-homologous and high-resolution protein crystal structures	11054454	http://peds.oxfordjournals.org/content/13/9/607.full	
Manesh	DSDB0133	Protein	Classification	Prediction of relative solvent accessibility	PDB ID	The Dataset consists of 215 low-similarity proteins		http://people.eng.unimelb.edu.au/jgl/Manesh.txt	
RNABindR	DSDB0134	Protein	Classification	Prediction of RNA Binding Residues in Proteins	PDB ID	The dataset consists of 109 nonredundant protein chains containing a total of 25,118 amino acids	16790841	http://rnajournal.cshlp.org/content/12/8/1450.full.pdf	http://einstein.cs.iastate.edu/RNABindR/
	DSDB0135	Protein	Classification	Prediction of RNA-binding sites of proteins using SVMs	Fasta Label	Three Datasets: The RBP86 data set consists of 86 protein chains extracted from RNA-protein complexes with X-ray crystallography resolution. The RBP109 data set contains 109 protein sequences obtained from 56 RNA-protein complexes with X-ray crystallography resolution. The RBP107 data set is comprised of 107 protein chains with X-ray crystallography resolutio	19091029	http://www.biomedcentral.com/1471-2105/9/S12/S6	
RNABindRPlus	DSDB0136	Protein	Classification	Prediction of RNA-binding residues from protein sequences using SVMs	Fasta Label	The RB198 contains 198 unique RNA-binding protein chains. The RB44 contains 44 protein chains. The RB111 contains 111 non-redundant RNA-binding protein chains.	24846307	http://einstein.cs.iastate.edu/RNABindRPlus/datasets.html	http://einstein.cs.iastate.edu/RNABindRPlus/index.html
RPISeq	DSDB0137	Protein	Classification	RNA-Protein Interaction Prediction	PDB ID	Two Datasets: The RPI2241, which contains a total of 952 protein chains and 443 RNA chains and the RPI369 which contains 369 RNA-protein partners	22192482	http://pridb.gdcb.iastate.edu/RPISeq/download.php	http://pridb.gdcb.iastate.edu/RPISeq/index.html
RNApred	DSDB0138	Protein	Classification	Prediction of RNA-binding	Fasta	Main Dataset: 377 RNA-binding proteins, 377 non RNA-binding proteins Independent	20677174	http://www.imtech.res.in/raghava/rnapred/dow	http://www.imtech.res.in/raghava/rnapred/

				proteins		Dataset: 69 RNA-binding protein chains, 100 non RNA-binding proteins		nload.html	submit.html
	DSDB0139	Protein	Classification	Prediction of protein binding sites in protein structures	Fasta Label	Six datasets: Hetero-complex I contains 504 proteins, Homo-complex I contains 620 proteins, Mix I contains 1124 proteins, Hetero-complex II contains 504 proteins, Homo-complex II contains 620 proteins and Mix II contains 1124 proteins	19925685	http://www.biomedcentral.com/1471-2105/10/381	
PSLpred	DSDB0140	Protein	Classification	Prediction of subcellular localization of bacterial proteins	other	The Dataset contains 1302 proteins	15699023	http://www.imtech.res.in/raghava/pslpred/data/	http://www.imtech.res.in/raghava/pslpred/submit.html
PSORTb v.1.0	DSDB0141	Protein	Classification	Protein subcellular localization prediction for Gram-negative bacteria	Fasta	The dataset comprises 1302 proteins resident at a single localization site: 248 cytoplasmic, 268 inner membrane, 244 periplasmic, 352 outer membrane and 190 extracellular; and contains a further 141 proteins resident at multiple localization sites: 14 cytoplasmic/inner membrane, 50 inner membrane/periplasmic and 77 outer membrane/extracellular	12824378	http://www.psort.org/dataset/datasetv1.html	http://www.psort.org/psortb/index.html
PSORTb v.2.0	DSDB0142	Protein	Classification	Prediction of bacterial protein subcellular localization	Fasta	The dataset contains 1591 Gram-negative and 576 Gram-positive proteins	15501914	http://www.psort.org/dataset/datasetv2.html	
PSORTb v.3.0	DSDB0143	Protein	Classification	Prediction of bacterial protein subcellular localization	Fasta	The Gram-negative training dataset contains 8230 proteins. The Gram-positive dataset contains 2652 proteins, and the archaeal contains 810 proteins	20472543	http://www.psort.org/dataset/datasetv3.html	
NRM2006	DSDB0144	Protein	Classification	Prediction of bacterial protein subcellular localization	Fasta	The dataset contains 299 Gram-negative proteins	16964270	http://www.psort.org/dataset/datasetnrm2006.html	

CoBaltDB	DSDB0145	Protein	Classification	Prediction of prokaryotic protein localization	RefSeq ID	The dataset contains 2,548,292 predicted non-redundant proteins	20331850	http://www.biomedcentral.com/1471-2180/10/88	
PSLT	DSDB0146	Protein	Classification	Predicting subcellular localization via protein motif co-occurrence	Fasta	The Hera human data set contains 2216 proteins. The Yeast data set contains 1612 proteins. The GFP data set contains 392 proteins. The Mouse data set contains 2095 proteins.	15466294	http://genome.cshlp.org/content/14/10a/1957/suppl/DC2	
MultiLoc	DSDB0147	Protein	Classification	Prediction of subcellular locations	Fasta	The Dataset contains 9761 sequences	16428265	http://abi.inf.uni-tuebingen.de/Services/MultiLoc/multiloc_dataset	http://abi.inf.uni-tuebingen.de/Services/MultiLoc
Höglund	DSDB0148	Protein	Classification	Prediction of subcellular locations	Fasta	The Dataset contains 5959 eukaryotic proteins	19723330	http://abi.inf.uni-tuebingen.de/Services/MultiLoc2/multiloc2_information	http://abi.inf.uni-tuebingen.de/Services/MultiLoc2
DBMLoc	DSDB0149	Protein	Classification	Prediction of subcellular locations	Fasta	The Dataset contains 3054 proteins	20507917	http://abi.inf.uni-tuebingen.de/Services/YLoc/webloc.cgi?page=info	http://abi.inf.uni-tuebingen.de/Services/YLoc/webloc.cgi
Unknown	DSDB0150	Protein	Classification	Prediction of subcellular locations	UniProt ID	The data set contains about 19000 proteins with unkown or uncertain localization	17392328	http://abi.inf.uni-tuebingen.de/Services/SherLoc/sherloc_information	http://abi.inf.uni-tuebingen.de/Services/SherLoc
	DSDB0151	Protein	Classification	Prediction of the subcellular location of apoptosis proteins	PDB ID	The Dataset contains 317 apoptosis proteins (112 cytoplasmic proteins, 55 membrane proteins, 34 mitochondrial proteins, 17 secreted proteins, 52 nuclear proteins and 47 endoplasmic reticulum proteins)	17189644	http://www.sciencedirect.com/science/article/pii/S0022519306005522	
	DSDB0152	Protein	Classification	Subcellular location prediction of apoptosis	UniProt ID	The training dataset contains 98 apoptosis proteins were classified into the following four subcellular locations: (1) cytoplasmic, (2) plasma membrane-bound, (3)	12471598	http://onlinelibrary.wiley.com/doi/10.1002/prot.10251/pdf	

				proteins		mitochondrial, and (4) other			
	DSDB0153	Protein	Prediction	Predicting protein folding rates using the concept of Chou's pseudo amino acid composition	PDB ID	The Dataset contains 99 proteins	21328402	http://onlinelibrary.wiley.com/doi/10.1002/jcc.21740/suppinfo	
	DSDB0154	Protein	Classification	Predicting protein subnuclear location	UniProt ID	The Dataset contains 370 nuclear proteins classified into 9 subnuclear locations	16213466	http://www.sciencedirect.com/science/article/pii/S0006291X05021388#	
ProtIdent	DSDB0155	Protein	Classification	Identification of proteases and their types	Fasta	the benchmark dataset S- contains 3,278 proteins and the benchmark dataset S+ contains 3,051 proteins	18774775	http://www.sciencedirect.com/science/article/pii/S0006291X08016902	http://www.csbio.sjtu.edu.cn/bioinf/Protease/
	DSDB0156	Protein	Classification	HIV-1 protease cleavage site prediction based on amino acid property	PubChem ID	The training dataset contains 299 octapeptides, which 60 HIV protease substrates assigned as positive samples and 239 non HIV protease substrates assigned as negative samples. The test dataset contains 63 octapeptides, which 54 HIV protease substrate as positive samples and 9 non-HIV protease substrate as negative samples are selected for the test set	18496789	http://onlinelibrary.wiley.com/doi/10.1002/jcc.21024/suppinfo	
Paircoil2	DSDB0157	Protein	Prediction	Prediction of coiled coils	PDB ID	The PDB-minus dataset consists of 6397 sequences that comprises 1 486 055 residues. The NEW-PDB21 dataset consists of 216 sequences that comprises 6288 residues. The NEW-PDB28 dataset consists of 85 sequences that comprises 3261 residues. The paircoil2_train dataset contains 824 proteins	16317077	http://groups.csail.mit.edu/cb/paircoil2/supplementary.html	http://groups.csail.mit.edu/cb/paircoil2/paircoil2.html
MARCOIL	DSDB0158	Protein	Classification	Prediction of coiled-coil domains in protein	Label	The posotive dataset contains 420 proteins and 820 CCDs. The negative dataset contains 1531 proteins	120160059	http://bcf.isb-sib.ch/Delorenzi/Marcoil/index.html	http://bcf.isb-sib.ch/webmarcoil/webmarcoilC1.html

				sequences Posterior probabilities generated by a HMM					
FRpred	DSDB0159	Protein	Prediction	Prediction of protein functional residues from sequence by probability density estimation	Fasta	The SITE set contains 726 alignments and the CSA set contains 428 alignments	18174181	ftp://ftp.tuebingen.mpg.de/pub/protevo/FRpred/	http://frpred.tuebingen.mpg.de/
HHrep	DSDB0160	Protein	Prediction	De novo protein repeat detection and the origin of TIM barrels	other	The benchmark dataset consists of the 50 most populated folds in the SCOP 1.69 database	16844977	http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1538828/bin/nar_34_suppl-2_W137__index.html	http://toolkit.tuebingen.mpg.de/hhrep
DEBT	DSDB0161	Protein	Classification	Prediction of beta turns and their type	PDB ID	The GR426 dataset consists of 426 protein chains. The FA547 dataset consist of 547 protein chains and the FA823 dataset consist of 823 protein chains. The PDB1296 dataset contains 1296 protein chains	20673368	http://comp.chem.nottingham.ac.uk/debt/index.html	http://comp.chem.nottingham.ac.uk/cgi-bin/debt/bin/getparams.cgi
DISSPred	DSDB0162	Protein	Classification	Prediction of dihedral angles and secondary structure	PDB ID	The dataset CB513 is a non-redundant non-homologous set of 513 protein sequences. The PDB-Select25 dataset was divided into two subsets. The subset one contains 280 128 residues from 1989 chains and the subset two contains 279 945 residues from 1988 chains. The four EVA subsets set1, set2, set4 and set6	20025785	http://comp.chem.nottingham.ac.uk/disspred/	http://comp.chem.nottingham.ac.uk/cgi-bin/disspred/bin/getparams.cgi
GPP	DSDB0163	Protein	Classification	Prediction of glycosylation sites from amino acid sequence	Fasta	The training dataset contains 242 protein sequences and 2413 verified glycosylation sites	19038042	http://comp.chem.nottingham.ac.uk/glyco/	http://comp.chem.nottingham.ac.uk/cgi-bin/glyco/bin/getparams.cgi

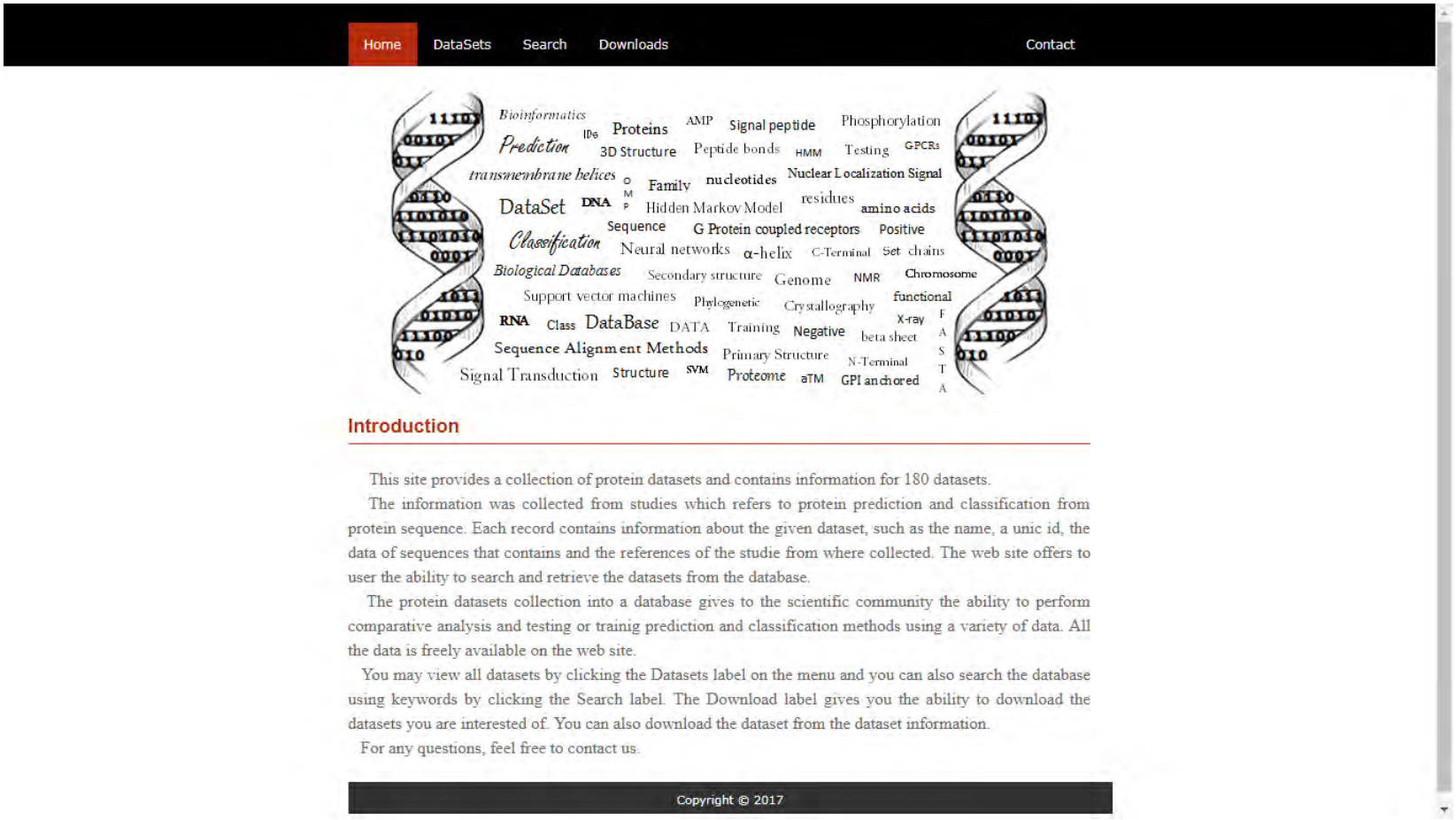
BTEVAL	DSDB0164	Protein	Classification	Beta turn prediction	Fasta	The data set consists of 426 proteins. The dataset divided into seven different subsets (setsI-VII), each containing equal number of proteins (61 proteins)	12424123	http://imtech.res.in/raghava/bteval/dataset.html	http://imtech.res.in/raghava/bteval/evaluate.html
BTNpred	DSDB0165	Protein	Classification	Beta turn prediction	Fasta Label	The BT426 dataset contains 426 proteins. The BT547 dataset contains 547 proteins and the BT823 dataset contains 823 proteins.	18847492	http://biomine.ece.ualberta.ca/BTNpred/BTNpred.html	
FragAnchor	DSDB0166	Protein	Classification	Prediction of GPI-anchor in proteins	Fasta	The testHMM dataset contains 66 proteins. The testNN dataset contains 268 poteins. The trainHMM dataset contains 87 proteins and the trainNN dataset contains 158 proteins	17893077	http://navet.ics.hawaii.edu/~fraganchor/NNHMM/NNHMM.html	http://navet.ics.hawaii.edu/~fraganchor/NNHMM/NNHMM.html
big-PI	DSDB0167	Protein	Classification	GPI Modification Site Prediction in Plants	Gene ID	The dataset contains 219 proteins	14681532	http://mendel.imp.ac.at/gpi/plants/l_set/plants.learn.html#down_12	http://mendel.imp.ac.at/gpi/plant_server.html
DLP-SVM	DSDB0168	Protein	Prediction	Prediction of Domain Linkers	PDB ID	The dataset (DS-All) contains 182 proteins	18844295	http://domserv.lab.tuat.ac.jp/LinkerList.txt	http://domserv.lab.tuat.ac.jp/dlpsvm.html
DROP	DSDB0169	Protein	Prediction	SVM domain linker prediction	PDB ID	The dataset (DS-All) contains 169 proteins	21169376	http://domserv.lab.tuat.ac.jp/DROP_LinkerList.txt	http://domserv.lab.tuat.ac.jp/dropform.html
ANCHOR	DSDB0170	Protein	Prediction	Prediction of Protein Binding Regions in Disordered Proteins	SWISSPROT ID	Seven Datasets: S1 (46 complexes of short disordered and long globular proteins), S2 (28 complexes of long disordered and long globular proteins), S3 (553 monomeric globular proteins that were used as a negative dataset), S4 (72 complexes of ordered proteins), S5 (53 complete archaea proteomes), S6 (639 complete bacteria proteomes), S7 (44 complete eukaryota proteomes)	19412530	http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000376#pcbi.1000376.s001	
FlexProt	DSDB0171	Protein	Prediction	Prediction of conformationally variable	morph ID	The dataset contains 137 proteins	18186479	http://lcg.rit.albany.edu/flexprot/flexprot_nr20.txt	

				positions from sequence and low-resolution structural data					
ASTRAL40	DSDB0172	Protein	Classification	Comparative Analysis of Protein Structure Alignments	SCOP ID	The dataset contains 355 pairs of remote homologous proteins	17672887	http://www.biomedcentral.com/1472-6807/7/50/additional	
SISY	DSDB0173	Protein	Classification	Comparative Analysis of Protein Structure Alignments	SCOP ID	The dataset contains 69 protein pairs	17672887	http://www.biomedcentral.com/1472-6807/7/50/additional	
RIPC	DSDB0174	Protein	Classification	Comparative Analysis of Protein Structure Alignments	SCOP ID	The datatset contains 40 protein pairs	17672887	http://www.biomedcentral.com/1472-6807/7/50/additional	
FlexRP	DSDB0175	Protein	Classification	Prediction of flexible/rigid regions from protein sequences using k-spaced amino acid pairs	Fasta Label	The Dataset contains 66 proteins	17437643	http://biomine.ece.ualberta.ca/FlexRP/FlexRPdataset.txt	
EzyPred	DSDB0176	Protein	Classification	Prediction of enzyme functional classes and sub-classes	Fasta	Sezy dataset contains 9,832 enzyme protein. Snon-ezy dataset contains 9,850 non-enzyme protein. Six subsets contains 10,442 enzyme proteins	17931599	http://www.csbio.sjtu.edu.cn/bioinf/EzyPred/Data.htm	http://www.csbio.sjtu.edu.cn/bioinf/EzyPred/
	DSDB0177	Protein	Classification	Prediction of Enzyme Subclass by Functional Domain	UniProt ID	The dataset, SA, consists of 1,697 oxidoreductases. The dataset, SB, contains 8 subclasses and 3582 transferases. The dataset, SC, contains 8 subclasses and 2902 hydrolases. The dataset, SD, contains 6	15952744	http://pubs.acs.org/doi/suppl/10.1021/pr0500399	

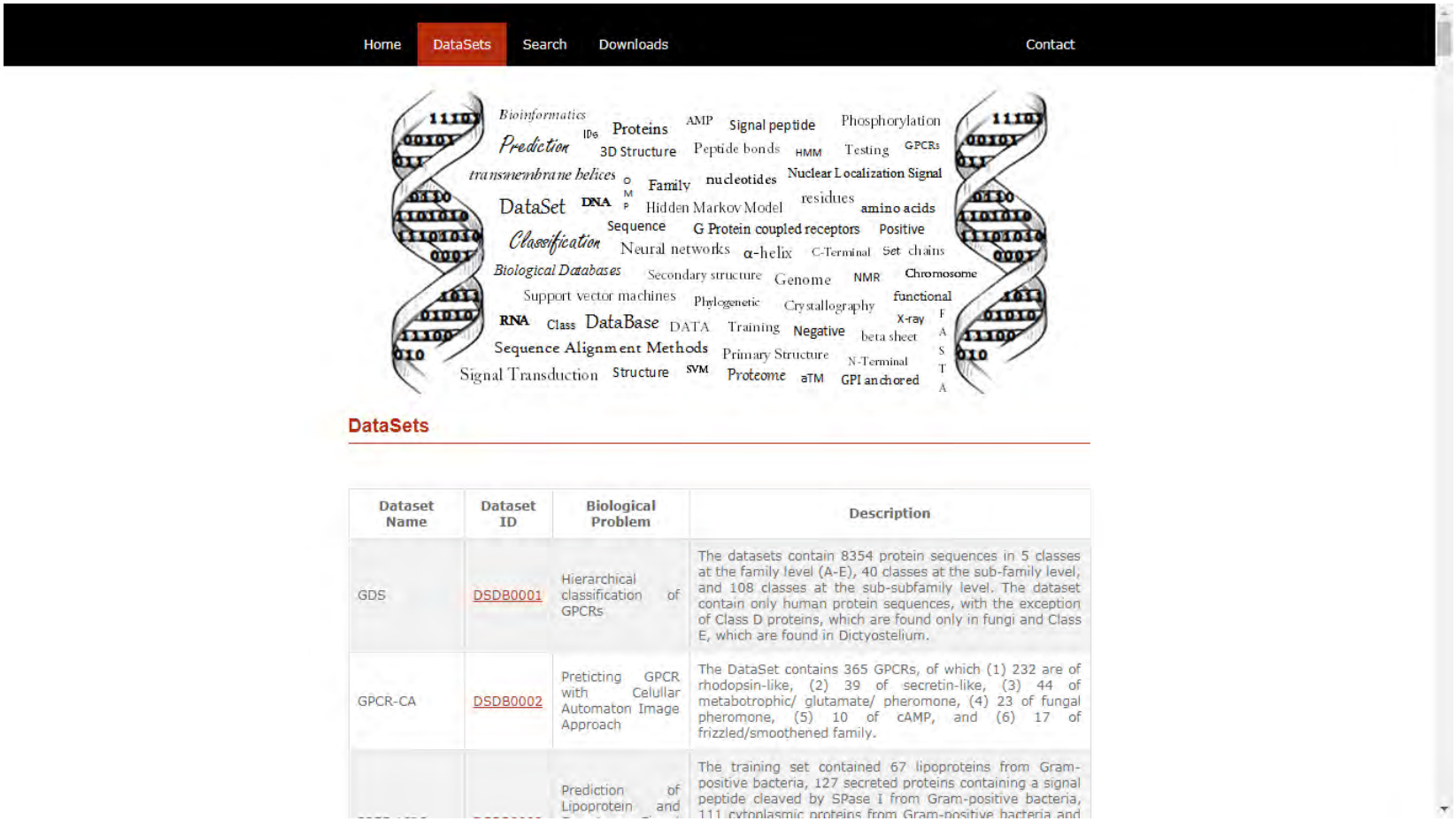
				Composition and Pseudo Amino Acid Composition		subclasses and 939 lyases. The dataset, SE, contains 6 subclasses and 503 isomerases. The dataset, SF, contains 6 subclasses and 840 ligases.			
SPX	DSDB0178	Protein	Classification	Prediction of the disulfide bonding connectivity pattern	Label	The dataset contains 1018 proteins	16320312	http://biomedical.ctust.edu.tw/edbcp/dataset.htm	http://biomedical.ctust.edu.tw/edbcp/
DS4	DSDB0179	Protein	Classification	Partitioning clustering algorithms for protein sequence data sets	Fasta	The data set has a total of 4922 sequences, out of which 3500 sequences are randomly for training, and 1422 for testing	19341454	http://www.biodatamining.org/content/2/1/3/additional	
PSP	DSDB0180	Protein	Prediction	Automated Alphabet Reduction for Protein Datasets	PDB ID	The dataset contains 1050 protein chains and 257560 residues	19126227	http://www.infobiotic.net/alphabet_reduction/	

2. Ιστοσελίδα

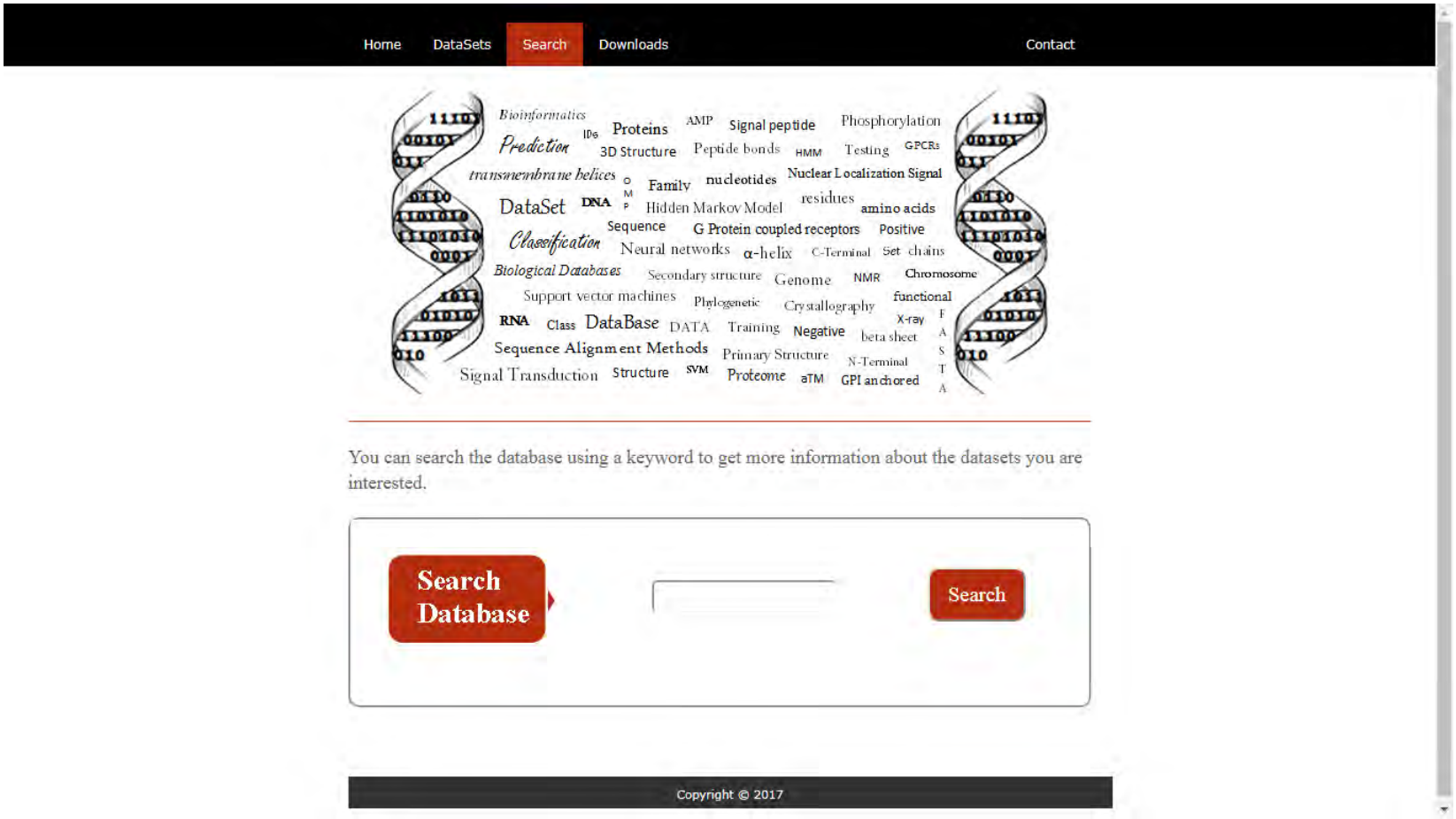
Εικόνα 6: Εισαγωγή



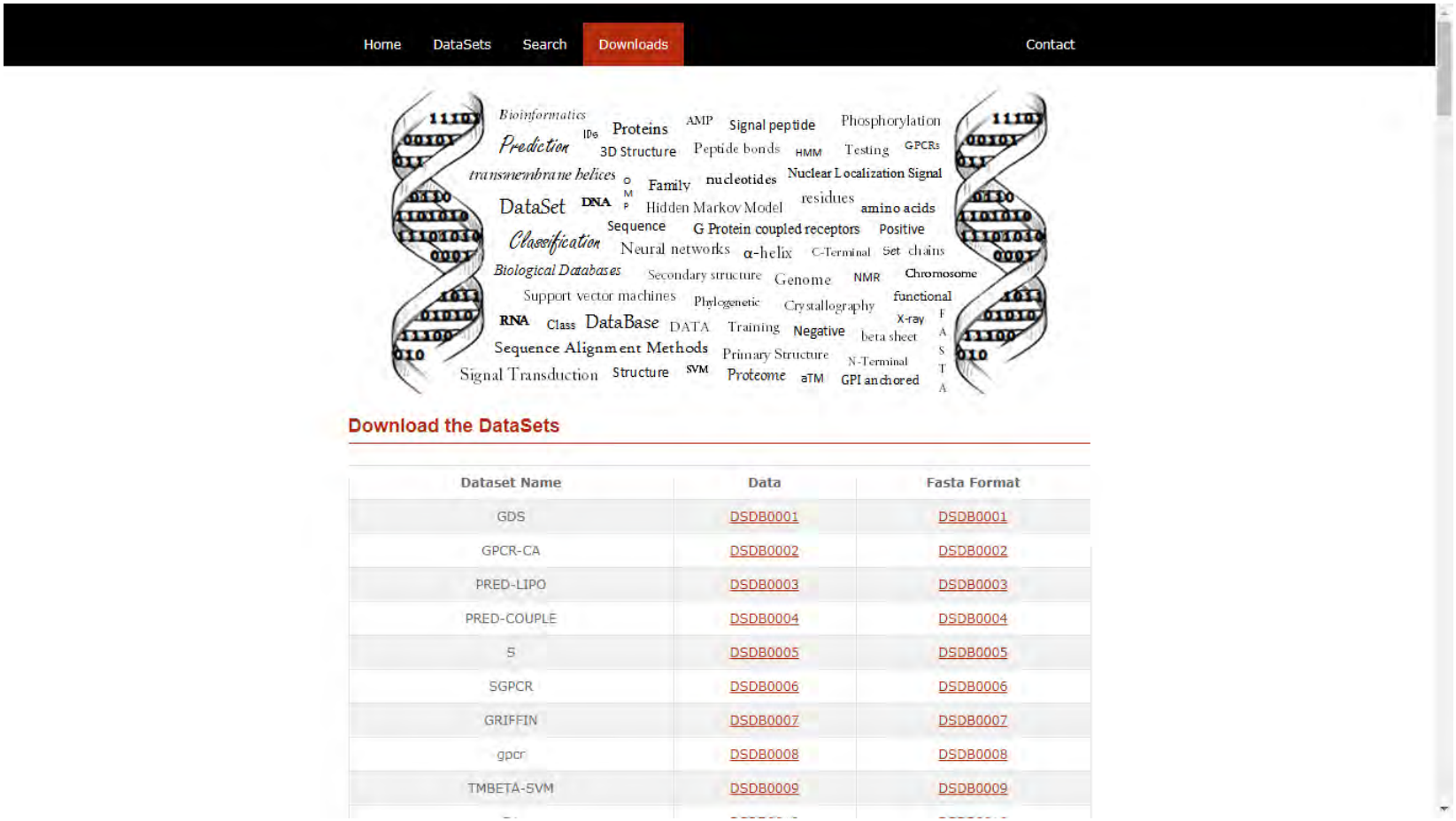
Εικόνα 7: Σύνολα Δεδομένων



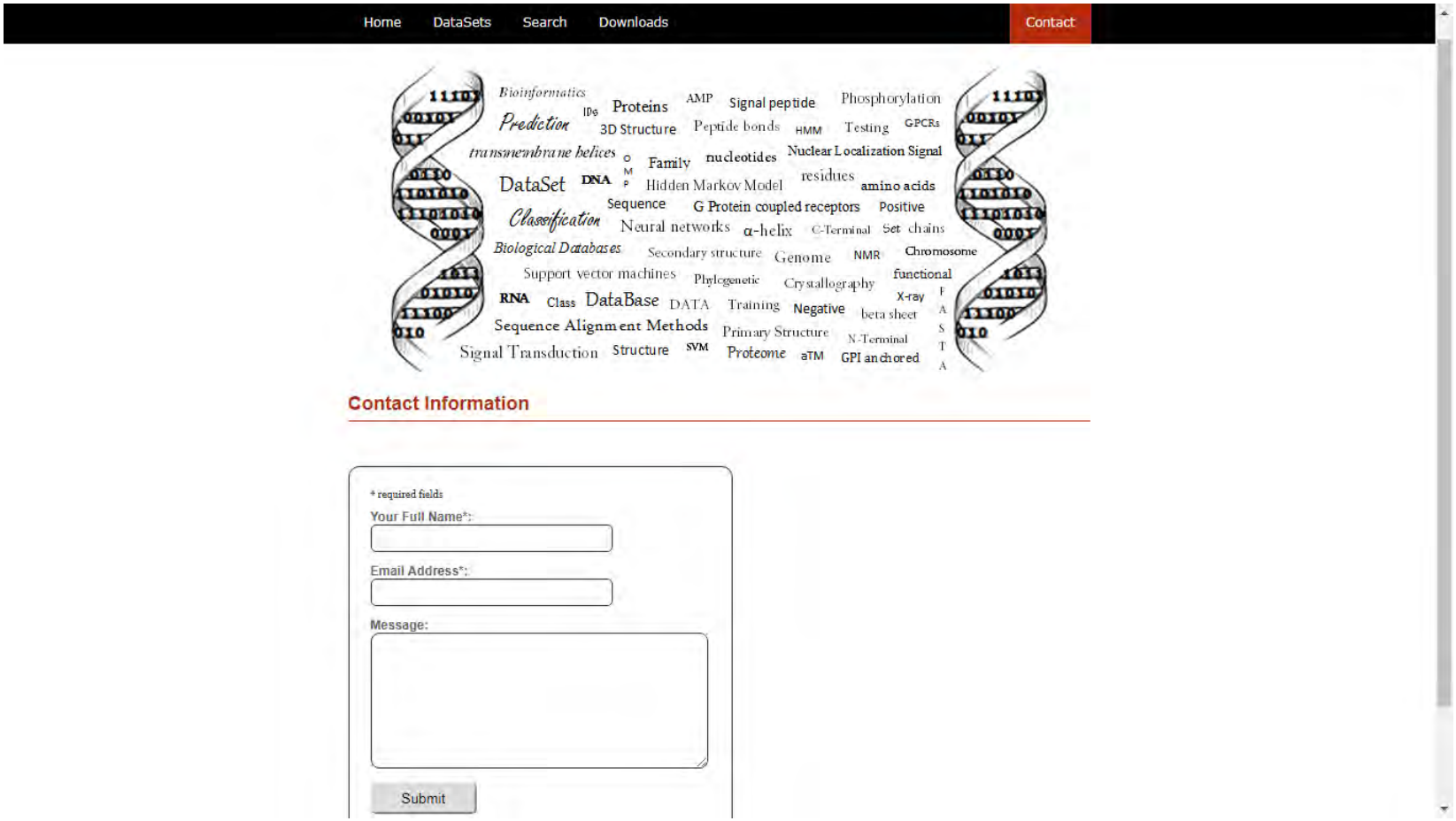
Εικόνα 8: Αναζήτηση



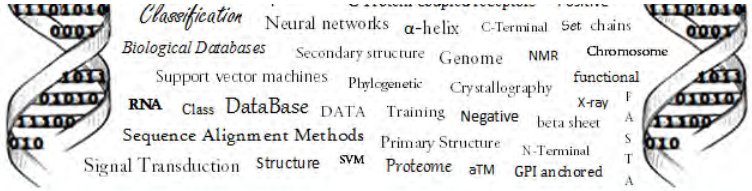
Εικόνα 9: Ανάκτηση Αρχείων



Εικόνα 10: Φόρμα Επικοινωνίας



Εικόνα 11: Επιπρόσθετες πληροφορίες Συνόλου Δεδομένων



DataSets

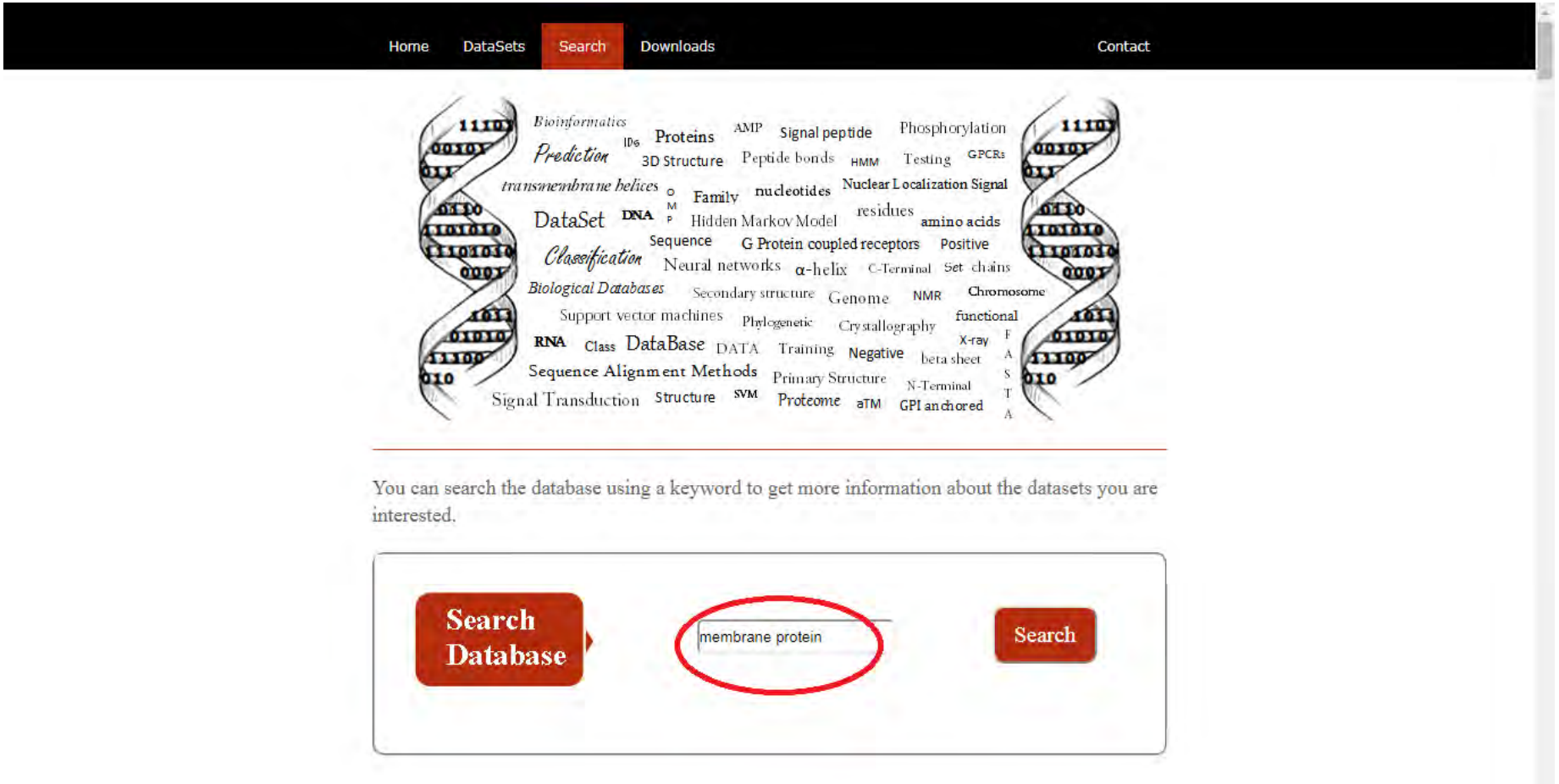
Dataset Name	Dataset ID	Biological Problem	Description	Sequence Type
PRED-LIPO	DSDB0003	Prediction of Lipoprotein and Secretory Signal Peptides in Gram-positive Bacteria	The training set contained 67 lipoproteins from Gram-positive bacteria, 127 secreted proteins containing a signal peptide cleaved by SPase I from Gram-positive bacteria, 111 cytoplasmic proteins from Gram-positive bacteria and 58 Gram-positive bacterial sequences with an N-terminal TM segment. The test set contains 66 TM proteins, 117 Lipoproteins, 109 secreted proteins and 713 cytoplasmic proteins.	Protein

Data	URL	Web Server	Problem Type
UniProt ID	http://bioinformatics.biol.uoa.gr/PRED-LIPO/datasets.html	http://bioinformatics.biol.uoa.gr/PRED-LIPO/input.jsp	Classification

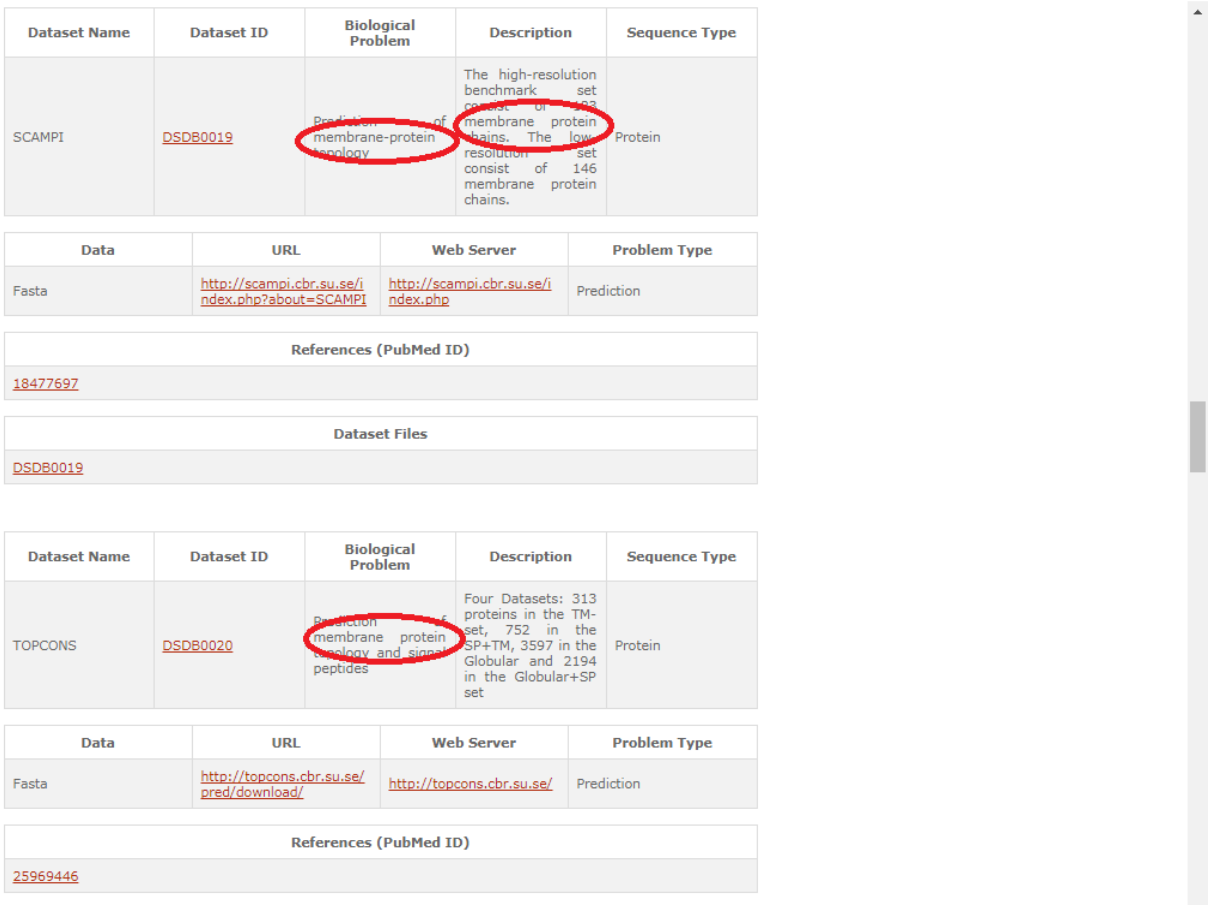
References (PubMed ID)
18975931

Dataset Files
DSDB0003

Εικόνα 12: Παράδειγμα αναζήτησης (i)



Εικόνα 13: Παράδειγμα αναζήτησης (ii)



Dataset Name	Dataset ID	Biological Problem	Description	Sequence Type
SCAMPI	DSDB0019	Prediction of membrane-protein topology	The high-resolution benchmark set consist of 103 membrane protein chains. The low-resolution set consist of 146 membrane protein chains.	Protein

Data	URL	Web Server	Problem Type
Fasta	http://scampi.cbr.su.se/index.php?about=SCAMPI	http://scampi.cbr.su.se/index.php	Prediction

References (PubMed ID)
18477697

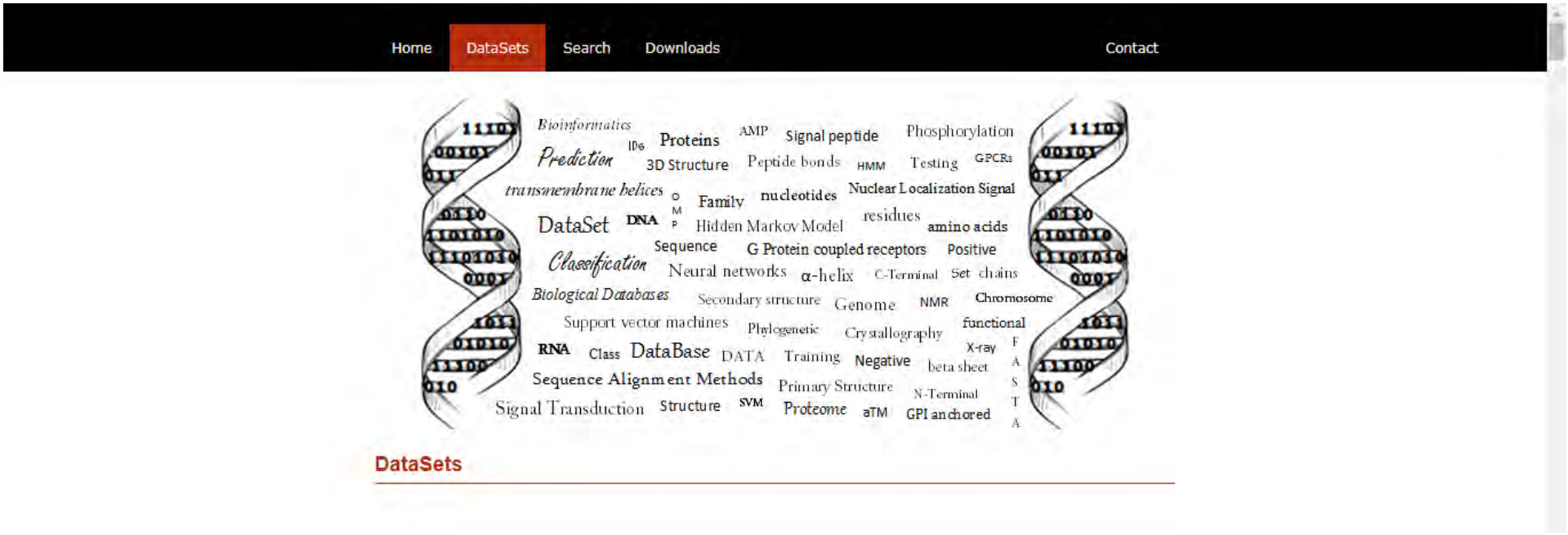
Dataset Files
DSDB0019

Dataset Name	Dataset ID	Biological Problem	Description	Sequence Type
TOPCONS	DSDB0020	Prediction of membrane protein topology and signal peptides	Four Datasets: 313 proteins in the TM-set, 752 in the SP+TM, 3597 in the Globular and 2194 in the Globular+SP set	Protein

Data	URL	Web Server	Problem Type
Fasta	http://topcons.cbr.su.se/pred/download/	http://topcons.cbr.su.se/	Prediction

References (PubMed ID)
25969446

Εικόνα 14: Γραμμή Μενού



3. Script αρχεία

Script1: DSDB0004.pl

```
while(<>)
{
	if($_=~/^(\w{6})(\s)/)
	{
		print "$1,";
	}
}
```

Script2: DSDB0022.pl

```
#function for removing duplicate items from array
sub uniq {
  my %seen;
  grep !$seen{$_}++, @_;
}
while(<>)
{
	if($_=~/^(\w{5})/)
	{
		push (@array, $1);
	}
}
my @filtered = uniq (@array);
print map { "$_\n" } @filtered;		#print elements on separate lines
```

Script3: DSDB0045.pl

```
sub uniq {
  my %seen;
  grep !$seen{$_}++, @_;
}

$dirname = "C:/Users/user/Desktop/PERL/IDs/DSDB0045";
my $existingdir = 'C:/Users/user/Desktop/PERL/IDs/DSDB0045/IDs';

opendir ( DIR, $dirname ) || die "Error in opening dir $dirname\n";
while( ($filename = readdir(DIR)))
{
   push(@names,$filename);
}
```

```
closedir(DIR);

for ( $i=4; $i <= scalar (@names); $i++)
{
	push(@new_names,$names[$i]);
}

print $_,"\n" foreach @new_names;
print scalar(@new_names);
undef(@array);
$j=0;
for (my $i=0; $i <=scalar(@new_names); $i++)
{
$filename = "$new_names[$i]";
if (-e $filename)
{
	open(DATA, "<$new_names[$i]");

  open my $out, '>>', "$existingdir/$new_names[$i]" or die "Can't
write new file: $!";

	while(<DATA>)
	{
		if($_=~/^(\w{6,})\s/)
		{
			push (@array, $1);
		}
	}
	my @filtered = uniq (@array);
	print $out map { "$_\n" } @filtered ;    #print elements on separate
lines
	undef(@array);
}
else
{
	$j++;
}
}
```

Script4: DSDB0045_1.pl

```
use warnings;
use LWP::UserAgent;
use LWP::Simple;

$dirname = "C:/Users/user/Desktop/PERL/IDs/DSDB0045/IDs";
my $existingdir = 'C:/Users/user/Desktop/PERL/IDs/DSDB0045/Fasta'
;
```

```
opendir ( DIR, $dirname ) || die "Error in opening dir $dirname\n";
while( ($filename = readdir(DIR)))
{
    push(@names,$filename);
}
closedir(DIR);

for ( $i=2; $i <= scalar (@names); $i++)
{
     push(@new_names,$names[$i]);
}

print $_,"\n" foreach @new_names;
print scalar(@new_names);

undef(@array);

for (my $i=0; $i <=scalar(@new_names); $i++)
{
$filename = "$new_names[$i]";
if (-e $filename)
{
     open(DATA, "<$new_names[$i]");

  open my $out, '>>', "$existingdir/$new_names[$i]" or die "Can't
write new file: $!";

     while(<DATA>)
     {
          if($_=~/^(\w{6,})/)
          {
               push (@array, $1);
          }
     }

     $f= scalar @array ;
     for (my $j=0; $j <=$f ; $j++)
     {
  my $content = get("http://www.uniprot.org/uniprot/$array[$j].fa
sta");
          print $out $content;

     }
     undef(@array);
}
```

Script5: DSDB0046.pl

```
#function for removing duplicate items from array
sub uniq {
   my %seen;
   grep !$seen{$_}++, @_;
}

my $existingdir = 'C:/Users/user/Desktop/PERL/IDs/DSDB0046/prosit
e_alignments/DSDB0046_IDs';

for (my $i=10; $i <=99; $i++)
{
  #print "PS000$i\n";
$filename = "PS000$i.msa";
if (-e $filename)
{
      #print "$i File Exists!\n";

      open(DATA, "<PS000$i.msa");
      open my $out, '>>', "$existingdir/PS000$i.txt" or die "Can't
write new file: $!";

      while(<DATA>)
      {
            if($_=~/^>(\w*\|)(\w{6,})/)
            {
                  push (@array, $2);
            }
      }

      my @filtered = uniq (@array);
      print $out map { "$_\n" } @filtered ;      #print elements on separate
lines
      undef(@array);
}
```

Script6: DSDB0046_1.pl

```
use warnings;
use LWP::UserAgent;
use LWP::Simple;

my $existingdir = 'C:/Users/user/Desktop/PERL/IDs/DSDB0046/Fasta'
;
undef(@array);
for (my $i=10; $i <=99; $i++)
{
$filename = "PS000$i.txt";
if (-e $filename)
{
```

```
        open(DATA, "<PS000$i.txt");

  open my $out, '>>', "$existingdir/PS000$i.fasta" or die "Can't
write new file: $!";

        while(<DATA>)
        {
                if($_=~/^(\w{6,})/)
                {
                        push (@array, $1);
                }
        }
        $f= scalar @array ;
        for (my $j=0; $j <=$f ; $j++)
        {
  my $content = get("http://www.uniprot.org/uniprot/$array[$j].fa
sta");
                print $out $content;

        }
        undef(@array);
}
```

Script7: DSDB0091.pl

```
use warnings;
use LWP::UserAgent;
use LWP::Simple;

        while(<>)
        {
                if($_=~/^(\w{4})(\w)/)
                {
                        push (@id, $1);
                        $k=$2;
                        if($k eq "_")
                        {
                                $k="A";
                                push(@chain,$k);
                        }
                        else
                        {
                                push(@chain,$2);
                        }
                }
        }
for (my $i=0; $i <=scalar(@id); $i++)
{
```

```
my $content = get("http://www.rcsb.org/pdb/download/downloadFile
.do?fileFormat=fastachain&compression=NO&structureId=$id[$i]&ch
ainId=$chain[$i]");
      print $content;

}
```

Script8: DSDB0126.pl

```
use warnings;
use LWP::UserAgent;
use LWP::Simple;

while(<>)
{
      if($_=~/^(\w{4})\s(.*\D)\s(.*)/)
      {
            my @new = split /,/,  $2 ;
            for(my $i=0; $i< scalar(@new);$i++)
            {
                  push (@id, $1);
                  push(@chain, $new[$i]);
            }
            undef(@new);
      }
}

for (my $i=0; $i <=scalar(@id); $i++)
{
my $content = get("http://www.rcsb.org/pdb/download/downloadFile
.do?fileFormat=fastachain&compression=NO&structureId=$id[$i]&ch
ainId=$chain[$i]");
      print $content;
}
```

Script9: DSDB0130.pl

```
use warnings;
use LWP::UserAgent;
use LWP::Simple;

while(<>)
      {
            if($_=~/^(\w{6})/)
            {
                  push (@id, substr($1, 0, 4));
                  push (@chain, substr($1, 5, 1));
            }
      }
```

```
for (my $i=0; $i <=scalar(@id); $i++)
{
my $content = get("http://www.rcsb.org/pdb/download/downloadFile
.do?fileFormat=fastachain&compression=NO&structureId=$id[$i]&ch
ainId=$chain[$i]");
      print $content;
}
```

Script10: DSDB0167.pl

```
while(<>)
{
      if($_=~/^ID\s{3}(.+?)\s.+\s(\d+\sAA)/)
      {
            print ">$1\t$2\n";
      }
      if($_=~/^\s{5}(.*)/)
      {
            $x=$1;
            $x=~s/\s//g;
            print "$x\n";
      }
}
```

Script11: duplicate.pl

```
use strict;
my %lines;

while (<>)
{
   print if not $lines{$_}++;
}
```

Script12: duplicate_1.pl

```
use strict;

my $file = "1.txt";
my $thiscount;
my $fullcount;
my $max = 10;
my %lines;

open(DATA, $file);
while(<DATA>)
{
      chomp;
      if(exists $lines{$_})
      {
```

```
            print "duplicate line (on read):$_\n";
        }
        else
        {
            $lines{$_} = 1;
        }
        $thiscount ++;
        $fullcount ++;

        if($thiscount >= $max)
        {
            my $checkcount=0;
            open(CHECK, $file);
            while(<CHECK>)
            {
                $checkcount ++;
                if($checkcount > $fullcount)
                {
                    chomp;
                    if(exists $lines{$_})
                    {
                        print "duplicate line (on check):$_\n";
                    }
                }
            }
            undef %lines;
            $thiscount = 0;
        }
}
```