ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΪΑΤΡΙΚΗ

ΠΡΟΒΛΕΨΗ ΣΥΝΔΕΤΙΚΩΝ ΠΕΡΙΟΧΩΝ ΤΩΝ ΠΡΩΤΕΪΝΩΝ ΜΕ ΜΕΘΟΔΟΥΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

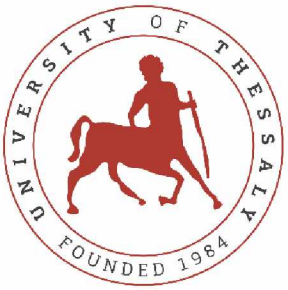
ΚΡΑΠΗ ΑΝΑΣΤΑΣΙΑ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ :

ΠΑΝΤΕΛΗΣ ΜΠΑΓΚΟΣ

Λαμία, 2021



UNIVERSITY OF THESSALY

COMPUTER SCIENCE AND BIOMEDICAL INFORMATICS

PREDICTION OF PROTEIN LINKERS WITH MACHINE LEARNING TECHNIQUES

KRAPI ANASTASIA

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

SUPERVISOR PROFESSOR:

PANDELIS BAGOS

Lamia, 2021

«Υπεύθυνη Δήλωση μη λογοκλοπής και ανάληψης προσωπικής ευθύνης»

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, και γνωρίζοντας τις συνέπειες της λογοκλοπής, δηλώνω υπεύθυνα και ενυπογράφως ότι η παρούσα εργασία με τίτλο «ΠΡΟΒΛΕΨΗ ΣΥΝΔΕΤΙΚΩΝ ΠΕΡΙΟΧΩΝ ΤΩΝ ΠΡΩΤΕΪΝΩΝ ΜΕ ΜΕΘΟΔΟΥΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ» αποτελεί προϊόν αυστηρά προσωπικής εργασίας και όλες οι πηγές από τις οποίες χρησιμοποίησα δεδομένα, ιδέες, φράσεις, προτάσεις ή λέξεις, είτε επακριβώς (όπως υπάρχουν στο πρωτότυπο ή μεταφρασμένες) είτε με παράφραση, έχουν δηλωθεί κατάλληλα και ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή. Αναλαμβάνω πλήρως, ατομικά και προσωπικά, όλες τις νομικές και διοικητικές συνέπειες που δύναται να προκύψουν στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής.

Ο/Η ΔΗΛΩΝ/-ΟΥΣΑ, ΚΡΑΠΗ ΑΝΑΣΤΑΣΙΑ

Ημερομηνία: 1/11/2021

Υπογραφή

HIDDEN MARKOV MODELS ΣΤΗ ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ

ΚΡΑΠΗ ΑΝΑΣΤΑΣΙΑ

Τριμελής Επιτροπή:

ΠΑΝΤΕΛΗΣ ΜΠΑΓΚΟΣ, ΚΑΘΗΓΗΤΗΣ

ΓΕΩΡΓΙΑ ΜΠΡΑΛΙΟΥ, ΕΠΙΚΟΥΡΟΣ ΚΑΘΗΓΗΤΡΙΑ

ΑΡΤΕΜΙΣ ΧΑΤΖΗΓΕΩΡΓΙΟΥ, ΚΑΘΗΓΗΤΡΙΑ

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΕΡΙΛΗΨΗ

Είναι κοινώς αποδεκτό πως ο προσδιορισμός της δομής των πρωτεϊνών είναι μείζονος σημασίας και ποικίλες μέθοδοι έχουν αναπτυχθεί σε βάθος χρόνου. Σημαντική πρόκληση στο εγχείρημα αυτό, εμφανίζεται ο εντοπισμός των περιοχών domain στις multidomain πρωτεΐνες. Μέσω της αναγνώρισης των διαφορετικών τμημάτων της πρωτεΐνης, επιτυγχάνεται η απλοποίηση περίπλοκων πρωτεϊνικών δομών, γεγονός που θα ευεργετήσει πολυάριθμες genomics εφαρμογές. Στην συγκεκριμένη πτυχιακή εργασία, παρουσιάζεται ένα μοντέλο που κατασκευάστηκε για τον σκοπό αυτό. Το μοντέλο δοκιμάζεται με τη χρήση ενός εργαλείου λογισμικού, που είναι εξειδικευμένο στην ανάλυση βιολογικών ακολουθιών και προσφέρει ποικίλες εναλλακτικές αλγορίθμων για Hidden Markov Models και επεκτάσεις αυτών. Για τον έλεγχο της απόδοσης του μοντέλου χρησιμοποιούνται οι μετρικές Q2, SOV και Correct Topolgy.

Το μοντέλο δοκιμάζεται για διαφορετικά σύνολα παραμέτρων, με σκοπό την εύρεση αυτών, που βελτιστοποιούν τα παραγόμενα αποτελέσματα. Στόχος της εργασίας λοιπόν αποτελεί η δημιουργία μιας ανταγωνιστικής μεθόδου που θα επιτύχει να διαχωρίζει της δομικές (και λειτουργικές) αυτές περιοχές. υπολογιστικά, χωρίς την ανάγκη ανθρώπινης παρέμβασης .

Τέλος, παρατίθενται τα κορυφαία αποτελέσματα και σχολιάζονται. Δηλαδή, αναφέρονται οι παράμετροι που παράγουν τα υψηλότερα ποσοστά, και πώς αυτοί επηρεάζουν την έκβαση του αποτελέσματος, ενώ παράλληλα οι αποδόσεις τους συγκρίνονται με αποδόσεις αντίστοιχων εφαρμογών που εμφανίζονται στην βιβλιογραφία.

ΚΕΦΑΛΑΙΟ 1^{Ο}: ΕΙΣΑΓΩΓΗ

1.1. ΠΡΩΤΕΪΝΕΣ

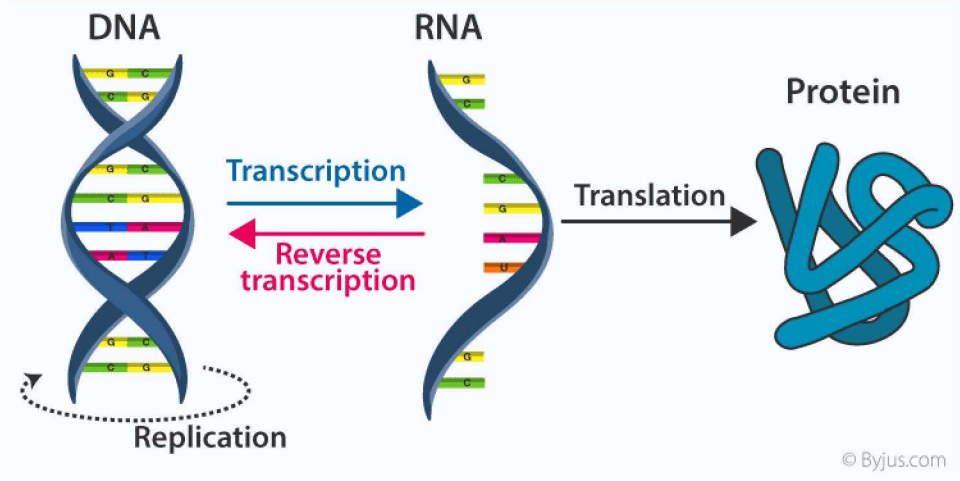
Οι πρωτεΐνες εμφανίζονται ως βασικό συστατικό των κυττάρων όλων των ζωντανών οργανισμών, από τους απλούστερους βακτηριακούς μέχρι τους πολύπλοκους ευκαρυωτικούς. Αποτελούν περισσότερο από το 50% του στέρεου βάρους των κυττάρων, και υπερνικούν σε ποσότητα όλα τα άλλα βιομόρια που συναντάται σε αυτά. Είναι εξέχουσας σημασίας για τη διατήρηση και τη διαιώνιση της ζωής, καθώς είναι υπεύθυνα για την διεκπεραίωση όλων των αντιδράσεων που λαμβάνουν χώρα στα βιολογικά συστήματα.

Εφόσον συμμετέχουν σχεδόν σε κάθε πτυχή της φυσιολογίας και βιοχημείας των ζωντανών οργανισμών, παρουσιάζουν μεγάλη ποικιλομορφία στη δομή και τη λειτουργία τους. Στις λειτουργίες τους συγκαταλέγονται η κατάλυση βιολογικών διεργασιών από τα ένζυμα και τις ορμόνες, η διαβίβαση ζωτικών μορίων όπως το οξυγόνο, διεργασίες που σχετίζονται με τον μεταβολισμό, την κίνηση, την άμυνα του κυττάρου, την αναγνώριση μορίων και τη διακυτταρική επικοινωνία ενώ ρυθμίζουν την γονιδιακή έκφραση και εξυπηρετούν σαν δομικά συστατικά. [1, 2]

Με έναυσμα το πρόγραμμα χαρτογράφησης του ανθρώπινου γονιδιώματος, το ενδιαφέρον για την αναγνώριση των διαφορετικών ρόλων των πρωτεϊνών στα βιολογικά συστήματα γνώρισε σημαντική αύξηση. Υπολογίζεται ότι εντοπίστηκαν περίπου 25.000 γονίδια, που όμως μέσω του εναλλακτικού ματίσματος και διαφοροποιήσεων των υπομονάδων τους, ο αριθμός των πρωτεϊνών που παράγεται από αυτά είναι ακόμα μεγαλύτερος. Έτσι, παρά την χαρτογράφηση του γονιδιώματος η λειτουργία της πλειοψηφίας των πρωτεϊνών παραμένει άγνωστος και παρατηρείται χάσμα μεταξύ της άφθονης πληροφορίας των ακολουθιών και της περιορισμένης γνώσης της δομής/λειτουργίας τους. [3]

1.2. ΣΥΝΘΕΣΗ ΠΡΩΤΕΪΝΩΝ

Η ροή της γενετικής πληροφορίας απεικονίζεται στην παρακάτω εικόνα που αναπαριστά το κεντρικό δόγμα της μοριακής βιολογίας [4].



Εικόνα 1Το δόγμα της μοριακής βιολογίας,[5]

Σύμφωνα με αυτό, το DNA μπορεί να μεταφέρει πληροφορία σε άλλα μόρια DNA, μέσω της αντιγραφής ή σε μόρια RNA, μέσω της μεταγραφής. Τα μόρια RNA μπορούν να μεταφέρουν πληροφορία σε μόρια DNA μέσω της αντίστροφης μεταγραφής ή να μεταφράζονται σε πρωτεΐνες. Προς το παρόν, δεν έχει βρεθεί μεταφορά πληροφορίας από τις πρωτεΐνες σε μόρια RNA.

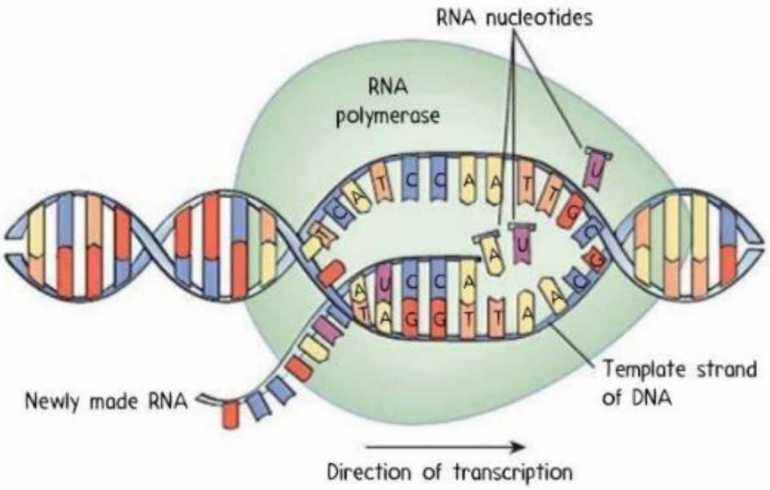
Το γενετικό υλικό (DNA: DeoxyriboNucleic Acid) είναι ένα μόριο που εμπεριέχει όλες τις γενετικές πληροφορίες στα περισσότερα είδη κυττάρων. Εξαίρεση αποτελούν κάποιοι ιοί, στους οποίους η αποθήκευση της γενετικής πληροφορίας πραγματοποιείται στο RNA. Αποτελείται από δύο συμπληρωματικές νουκλεοτιδικές αλυσίδες που συνδέονται μεταξύ τους με δεσμούς υδρογόνου και περιελίσσονται η μία γύρω από την άλλη σε μια δομή που καλείται διπλή έλικα.[6]

Το RNA είναι ένα μόριο παρόμοιο με το DNA. Απαρτίζεται ωστόσο από μία νουκλεοτιδική αλυσίδα ενώ διαφορές παρατηρούνται τόσο στο μέγεθος όσο και στη σύσταση αυτών των μορίων. Υπάρχουν πολλές κατηγορίες RNA που επιτελούν διαφορετικές λειτουργίες. Οι πιο γνωστές λειτουργίες τους είναι ως αγγελιοφόροι (mRNA), υπεύθυνοι για την τήρηση του γενετικού κώδικα (tRNA) ή ως δομικά συστατικά των ριβοσωμάτων (rRNA). Όμως επιτελούν σημαντικό ρόλο και στη ρύθμιση κυτταρικών διεργασιών όπως η διαίρεση του κυττάρου, η διαφοροποίηση, η ανάπτυξη, η γήρανση και ο θάνατός του.

Η διαδικασία που ακολουθεί η πληροφορία για να μεταβιβαστεί από το DNA σε πρωτεΐνη παρατίθεται στη συνέχεια.

1.2.1. ΜΕΤΑΓΡΑΦΗ

Παρά το μεγάλο μέγεθος του DNA ($3*10^9$ ζεύγη βάσεων), μόνο ένα μικρό κομμάτι αυτού χρησιμοποιείται για την παραγωγή των πρωτεϊνών ή και μορίων RNA (1-2%). Οι περιοχές αυτές καλούνται γονίδια. [7]



Εικόνα 2. Μεταγραφή RNA από το δίκλωνο μόριο DNA[8]

Το DNA στην περιοχή των γονιδίων, ξετυλίγεται τοπικά και ακολουθεί η διαδικασία της μεταγραφής. Ως μεταγραφή ονομάζουμε την παραγωγή μορίων RNA, που προκύπτουν έχοντας τη μία μόνο αλυσίδα DNA ως εκμαγείο. Στην διαδικασία συμμετέχουν ρυθμιστικά και καταλυτικά ένζυμα, τριφοσφωρικά ριβονουκλεοτίδια και άλλοι παράγοντες απαραίτητοι για τη βέλτιστη λειτουργία της αντίδρασης και για την δράση των καταλυτικών ενζύμων.

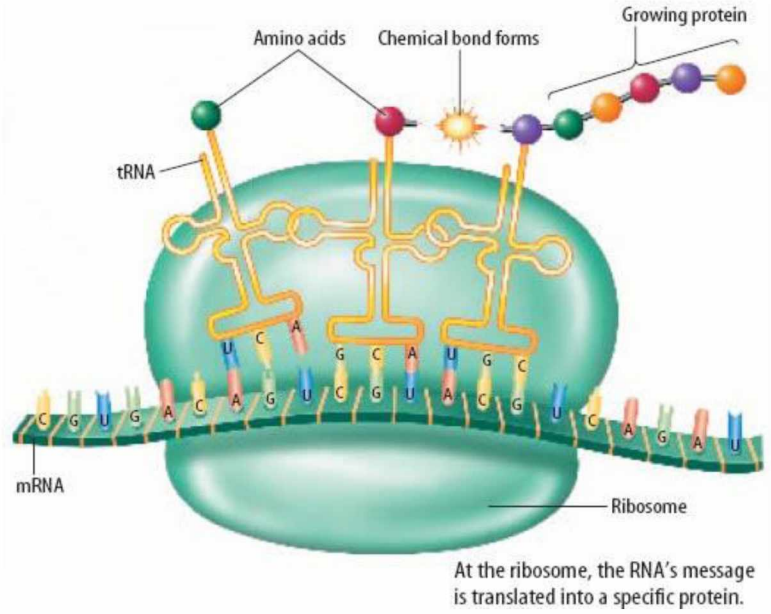
1.2.2. ΡΥΘΜΙΣΗ ΤΗΣ ΓΟΝΙΔΙΑΚΗΣ ΕΚΦΡΑΣΗΣ

Όμως, δεν εκφράζονται όλα τα γονίδια του κυττάρου, κάθε χρονική στιγμή. Για τον λόγο αυτό χρειάζεται ένα σύστημα ελέγχου του είδους και του πλήθους των μετάγραφων που παράγονται το οποίο να ανταποκρίνεται άμεσα στις αλλαγές του περιβάλλοντος και τις ανάγκες του κυττάρου. Η ρύθμιση επομένως της μεταγραφής πραγματοποιείται από πρωτεΐνες γνωστές ως μεταγραφικοί παράγοντες. Η ρύθμιση μπορεί να είναι θετική ή αρνητική, ποσοτική ή ποιοτική και συνήθως γίνεται με την αλληλεπίδραση των μεταγραφικών παραγόντων με το μόριο DNA (άμεση ή έμμεση πρόσδεση).

Το RNA που χρησιμοποιείται για την παραγωγή των πρωτεϊνών καλείται mRNA (messenger RNA) καθώς μεταφέρει την γενετική πληροφορία από τον πυρήνα, που βρίσκεται το DNA, στα ριβοσώματα, όπου πραγματοποιείται η πρωτεϊνοσύνθεση (μετάφραση). Όμως, η αλληλουχία του mRNA στα περισσότερα ευκαρυωτικά γονίδια εμπεριέχει αλληλουχίες που δεν μεταφράζονται και απομακρύνονται με τον μηχανισμό του ματίσματος. Έτσι δημιουργείται το ώριμο mRNA.

1.2.3. ΜΕΤΑΦΡΑΣΗ

Τα μόρια mRNA ύστερα από τη σύνθεση τους, εξέρχονται στο κυτταρόπλασμα. Εκεί αλληλεπιδρούν με σύμπλοκα πρωτεϊνών και ριβοσωμικού RNA, τα ριβοσωμάτια, και ξεκινάει η διαδικασία της μετάφρασης. Συγκεκριμένα, το ριβόσωμα προσδένεται στο κωδικόνιο έναρξης του mRNA, και μετακινούμενο κάθε φορά κατά μήκος του μορίου, επεκτείνει την πολυπεπτιδική αλυσίδα μέχρι να αντικρίσει το κωδικόνιο λήξης. Όταν επέλθει ο τερματισμός της μετάφρασης, το πολυπεπτίδιο απελευθερώνεται και οι υπομονάδες του ριβοσώματος αποσπώνται από το mRNA.



Εικόνα 3. Μετάφραση mRNA στο ριβόσωμα με τη βοήθεια tRNA[9].

Η μετάφραση πραγματοποιείται με βάση τον γενετικό κώδικα, δηλαδή μία τριάδα ριβονουκλεοτιδίων αντιστοιχίζεται σε ένα αμινοξύ. Η αντιστοίχιση και η μεταφορά των αμινοξέων στο ριβοσωμάτιο εκτελείται από το μεταφορικό tRNA. Όταν το σωστό πεπτίδιο τοποθετηθεί στη νεοσυντιθέμενη πεπτιδική αλυσίδα, συνδέεται με το προηγούμενό του με πεπτιδικό δεσμό και το tRNA που το μεταφέρει απελευθερώνεται από το σύμπλοκο της μετάφρασης. Ο μηχανισμός της μετάφρασης διαφέρει στα ευκαρυωτικά από τα προκαρυωτικά κύτταρα, καθώς στα τελευταία συναντάται το φαινόμενο όπου ένα mRNA μπορεί να κωδικοποιεί μια ομάδα γονιδίων, γνωστά ως πολυσιστρονικά.

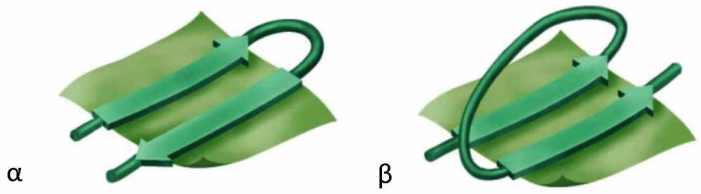
1.3.1. ΑΝΑΔΙΠΛΩΣΗ ΤΩΝ ΠΡΩΤΕΪΝΩΝ ΣΤΟΝ ΧΩΡΟ

1.3.1. ΠΡΩΤΟΤΑΓΗΣ ΔΟΜΗ

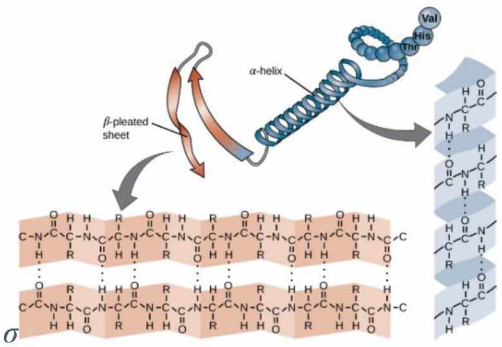
Οι πρωτεΐνες εμφανίζουν πολλά επίπεδα που καθορίζουν την τελική δομή, και κατ' επέκταση την λειτουργία τους. Η αλληλουχία των αμινοξέων που συνδέονται με πεπτιδικούς δεσμούς ορίζεται ως πρωτοταγής δομή. Οι μετέπειτα δομές που σχηματίζονται και κατ' επέκταση η λειτουργία της πρωτεΐνης καθορίζονται από την πρωτοταγή δομή της. [3]

1.3.2. ΔΕΥΤΕΡΟΤΑΓΗΣ ΔΟΜΗ

Τα αμινοξέα, παρουσιάζουν διαφορετικά μεγέθη, βιοχημικές ιδιότητες και φορτίο. Το γεγονός αυτό ωθεί την πολυπεπτιδική αλυσίδα, να αναδιπλωθεί σε δομές όπως η α-έλικα, η β-πτυχωτή επιφάνεια, οι στροφές και οι θηλές. Στην α-έλικα η πεπτιδική αλυσίδα περιστρέφεται γύρω από έναν άξονα δεξιόστροφα, ενώ κάθε στροφή της έλικας περιλαμβάνει 3,6 αμινοξέα. Η β-πτυχωτή δομή σχηματίζεται από τμήματα πεπτιδικής αλυσίδας που διατάσσονται παράλληλα. Οι αλυσίδες αποκτούν πτυχωτή διαμόρφωση, καθώς οι υποκαταστάτες διατάσσονται ενταλλάξ πάνω και κάτω από το επίπεδο. Οι στροφές και οι θηλές είναι μη κανονικές δομές που παρατηρούνται μεταξύ δύο άλλων στοιχείων κανονικής δομής (α-έλικα, β πτυχωτή επιφάνεια). Οι αναδιπλώσεις αυτές των αμινοξέων με τα γειτονικά τους, καλούνται δευτεροταγείς δομές και σταθεροποιούνται με δεσμούς υδρογόνου μεταξύ των ομάδων C=O και N-H δύο πεπτιδικών δεσμών. Σε μία πρωτεΐνη μπορεί να συνυπάρχουν διαφορετικές δευτεροταγείς δομές.



Εικόνα 4. Δευτεροταγείς δομές στροφών (α) και θηλών (β)[10]



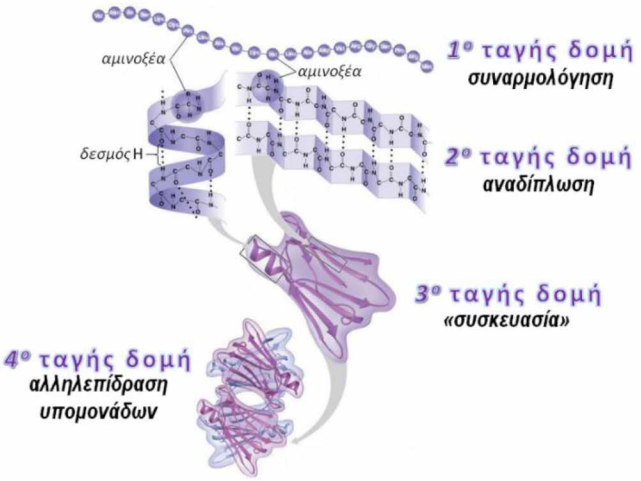
Εικόνα 5. Απεικόνιση δευτεροταγών δομών α-έλικας και β πτυχωτής επιφάνειας[11]

1.3.3. ΤΡΙΤΟΤΑΓΗΣ ΔΟΜΗ

Ακολουθεί η τριτοταγής δομή, δηλαδή η διαμόρφωση της πρωτεΐνης στον τρισδιάστατο χώρο. Η αναδίπλωση δεν περιορίζεται μόνο σε αμινοξέα κοντινής απόστασης μεταξύ τους, αλλά επεκτείνεται σε όλο το μήκος της πολυπεπτιδικής αλυσίδας. Η σταθεροποίηση της τριτοταγούς δομής γίνεται με αλληλεπιδράσεις μεταξύ των πλευρικών ομάδων R των αμινοξέων. Οι αλληλεπιδράσεις αυτές μπορεί να είναι ομοιοπολικοί δισουλφιδικοί δεσμοί, ηλεκτροστατικές δυνάμεις που αναπτύσσονται μεταξύ αντίθετα φορτισμένων αμινοξέων, δεσμοί υδρογόνου και υδρόφοβες αλληλεπιδράσεις, κατά τις οποίες τα υδρόφοβα αμινοξέα όταν βρεθούν σε υδατικό περιβάλλον τείνουν να συσσωματωθούν ώστε να μειώσουν την έκθεσή τους στο νερό.

1.3.4. ΤΕΤΑΡΤΟΤΑΓΗΣ ΔΟΜΗ

Η τεταρτοταγής δομή, αποτελεί το ύστατο επίπεδο αναδίπλωσης των πρωτεϊνών. Σε αυτή, διαφορετικές πολυπεπτιδικές αλυσίδες συνδυάζονται για την συγκρότηση ενός ενιαίου συμπλόκου. Η σταθεροποίηση του συμπλόκου υλοποιείται με δεσμούς υδρογόνου, ηλεκτροστατικές ή υδρόφοβες αλληλεπιδράσεις ανάμεσα στις πεπτιδικές αλυσίδες. Όλες οι πρωτεΐνες έχουν τριτοταγή δομή, όμως τεταρτοταγή δομή παρουσιάζουν μόνο όσες αποτελούνται από δύο ή περισσότερες πολυπεπτιδικές αλυσίδες. [12]



Εικόνα 6. Δομές πρωτεϊνών[13]

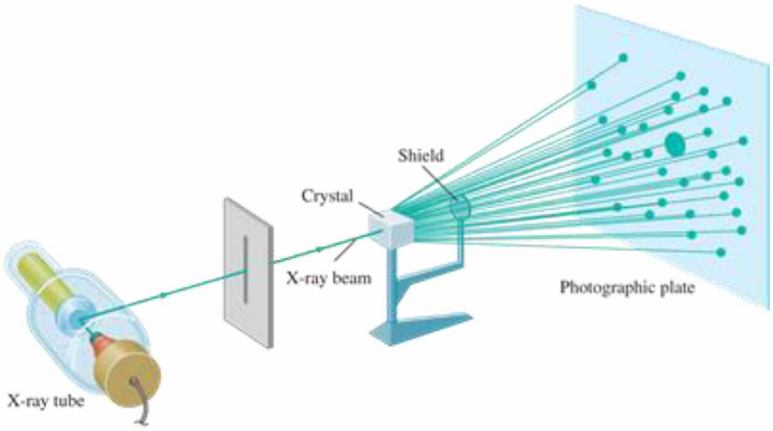
Μία από τις μεγαλύτερες προκλήσεις στη μελέτη μεγάλων πρωτεϊνών είναι η ανάκτηση σημαντικών πληροφοριών τριτοταγούς και τεταρτοταγούς δομής. Εφόσον η δομή είναι άρρηκτα συνδεδεμένη με τη λειτουργία τους, η ανακάλυψή τους, παρέχει πληροφορίες απαραίτητες για μία σε βάθος κατανόηση των συνθετικών ενζυμικών και όχι μόνο αλληλεπιδράσεων.

1.4. ΠΑΡΑΔΟΣΙΑΚΕΣ ΜΕΘΟΔΟΙ ΠΡΟΣΔΙΟΡΙΣΜΟΥ ΠΡΩΤΕΪΝΙΚΗΣ ΔΟΜΗΣ

1.4.1. ΚΡΥΣΤΑΛΛΟΓΡΑΦΙΑ ΑΚΤΙΝΩΝ Χ

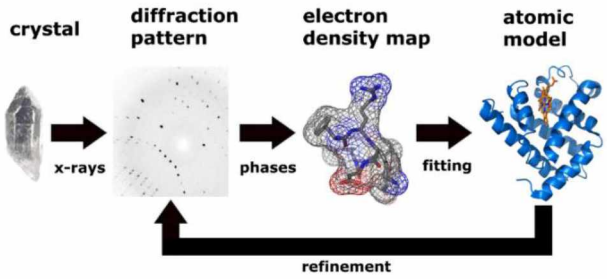
Η κρυσταλλογραφία με ακτίνες Χ έχει συνεισφέρει στο μεγαλύτερο πλήθος υψηλής ανάλυσης δομών μακρομορίων στην πρωτεϊνική βάση PDB. Με τη βοήθεια της μεθόδου αυτής, ανακτώνται ακριβείς πληροφορίες για τη θέση και την φύση των ατόμων από τα οποία συνίσταται η πρωτεΐνη.

Τα αποτελέσματα προκύπτουν από την αλληλεπίδραση του πρωτεϊνικού κρυστάλλου με τις ακτίνες-Χ, καθώς αυτές σκεδάζονται προς διάφορες διευθύνσεις. Οι σκεδαζόμενες ακτίνες-Χ συλλέγονται από έναν ανιχνευτή και από τη θέση πρόσπτωσης των κυμάτων στην επιφάνεια του ανιχνευτή καθίσταται δυνατός ο προσδιορισμός της διεύθυνσής τους.



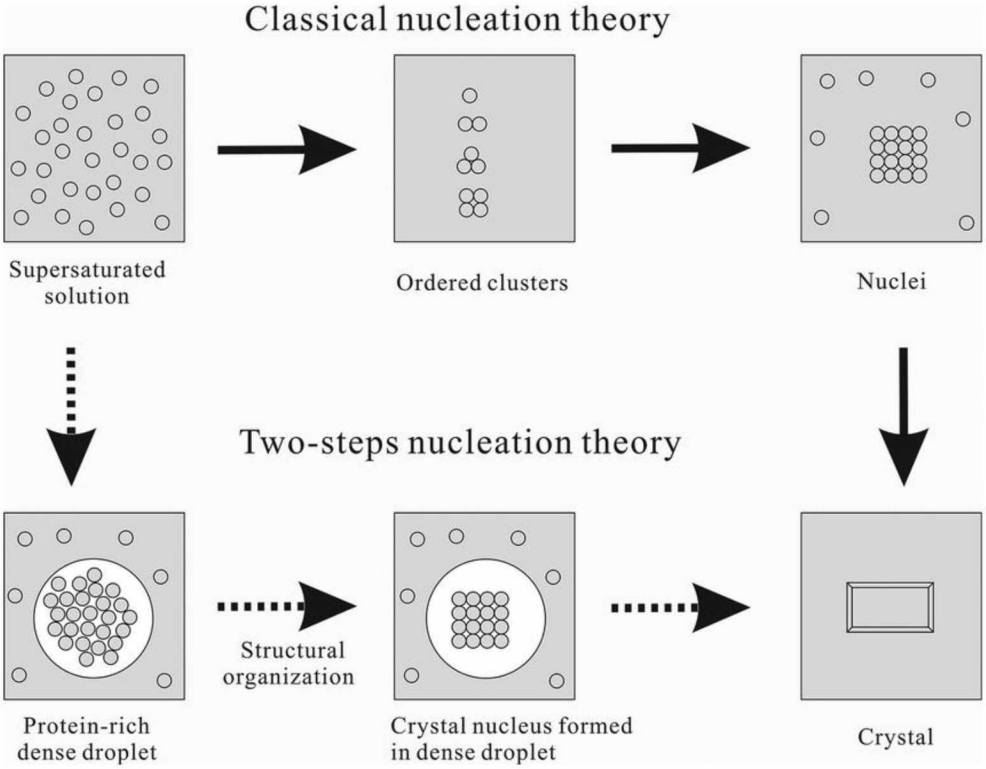
Εικόνα 7 x-ray crystallography [14]

Η ένταση του μαύρου χρώματος στον ανιχνευτή υποδηλώνει την ενέργεια που φέρει το κάθε κύμα. Κατ' αυτόν τον τρόπο μπορεί να προσδιοριστεί το πλάτος του, ενώ το μήκος κύματος των σκεδαζόμενων από τον κρύσταλλο κυμάτων, αντιστοιχεί στην προσπίπτουσα ακτινοβολία. Η διαδικασία αυτή επαναλαμβάνεται, με τον κρύσταλλο προσανατολισμένο υπό διαφορετική γωνία κάθε φορά, ώστε να ληφθούν πληροφορίες για την τρισδιάστατη δομή του. [15]



Εικόνα 8 Από τα αποτελέσματα της κρυσταλλογραφίας, στην αναγνώριση της δομής των πρωτεϊνών[16]

Όμως, η δημιουργία πρωτεϊνικών κρυστάλλων υψηλής ποιότητας, είναι μία χρονοβόρος, απαιτητική διαδικασία, που δεν προσφέρει εγγυημένα επιτυχή αποτελέσματα. Γι' αυτό, η επιτυχία κρυστάλλωσης είναι καθοριστική του προσδιορισμού της δομής του βιομακρομορίου. Αρχικά, πραγματοποιείται απομόνωση της «καθαρής» επιθυμητής πρωτεΐνης, από ένα περίπλοκο μείγμα, και στόχος είναι η δημιουργία ενός υπέρκορου διαλύματος αυτής. Με τον όρο υπέρκορο, αναφερόμαστε στο διάλυμα που η συγκέντρωση της διαλυμένης ουσίας είναι μεγαλύτερη από τη διαλυτότητα ισορροπίας για τις συγκεκριμένες συνθήκες περιβάλλοντος. Κατόπιν, εισάγεται ο παράγοντας κατακρήμνισης που πυροδοτεί την πυρηνοποίηση (nucleation) των πρωτεϊνικών κρυστάλλων στο διάλυμα, και τον σχηματισμό μεγάλων τρισδιάστατων κρυστάλλων. [17]



Εικόνα 9 Nucleation in protein crystallization [18]

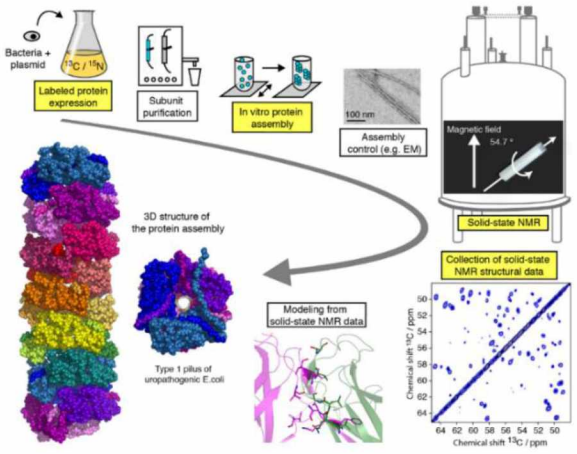
1.4.2. NUCLEAR MAGNETIC RESONANCE (NMR)

Η φασματοσκοπία πυρηνικού μαγνητικού συντονισμού είναι η διαδικασία διέγερσης πυρήνων, μέσω ηλεκτρομαγνητικής ακτινοβολίας καθώς αυτοί βρίσκονται σε μαγνητικό πεδίο. Όταν το δείγμα των πυρήνων αποδιεγερθεί, εκπέμπει την ενέργεια που δώσαμε.

Εν συντομία στο βασικό πείραμα NMR, αφήνεται ένα διάστημα χαλάρωσης μερικών δευτερολέπτων, που επιτρέπει στα spins να έρθουν σε ισορροπία. Στη συνέχεια, εφαρμόζεται η αποστολή του παλμού της ραδιοσυχνότητας που επιλέγεται και καταγράφεται η αποδιέγερση των μαγνητικών πυρήνων με το χρόνο για διάστημα tacq (50ms – μερικά sec). Η μαγνήτιση μειώνεται εκθετικά με το χρόνο και το αρχικά λαμβανόμενο σήμα, ως συνάρτηση του χρόνου ονομάζεται FID (Free Induction Decay). Τέλος, εφαρμόζεται μετασχηματισμός fourier και προκύπτει το επιθυμητό φάσμα.

Καθώς το NMR σήμα είναι ασθενές, το φάσμα από ένα FID δεν αρκεί. Για να βελτιωθεί το signal-to-noise ratio χρησιμοποιείται time averaging. Δηλαδή το πείραμα επαναλαμβάνεται πολλές φορές και τα FIDs προστίθενται. Αξίζει να αναφέρουμε πως καθώς αυξάνεται η ισχύς του σήματος, αυξάνεται και η ποσότητα του θορύβου, και για N επαναλήψεις αυτός ισούται περίπου με √N. [19]

Μέσω των 2D NMR μεθόδων (COSY, NOESY) δίνεται η δυνατότητα να καταγραφούν ιδιότητες των αμινοξέων, αλληλεπιδράσεις μεταξύ των ατόμων, και να παρατηρηθούν παράμετροι άμεσοι συνδεδεμένοι με την τρισδιάστατη μοριακή δομή τους. [20]



Εικόνα 10. Διαδικασία NMR για προσδιορισμό δομής[21]

Η φασματοσκοπία NMR μπορεί να εφαρμοστεί σε υδατικά και μη διαλύματα, εμφανίζει μεγάλη κλίμακα πειραματικών συνθηκών και μπορεί να εκτιμήσει την δυναμική συμπεριφορά των βιομορίων. Το πιο σημαντικό πλεονέκτημα της μεθόδου αυτής, συγκριτικά με την κρυσταλλογραφία ακτίνων-Χ είναι ότι δεν απαιτείται κρυστάλλωση των μορίων. Ένα εμπόδιο

όμως που προς το παρόν δεν έχει ξεπεραστεί είναι πως το μέγεθος μορίων που αναλύονται με αυτή είναι 35-50 kDa, που αντιστοιχεί σε πρωτεΐνες περίπου 315 – 450 αμινοξέων.[22]

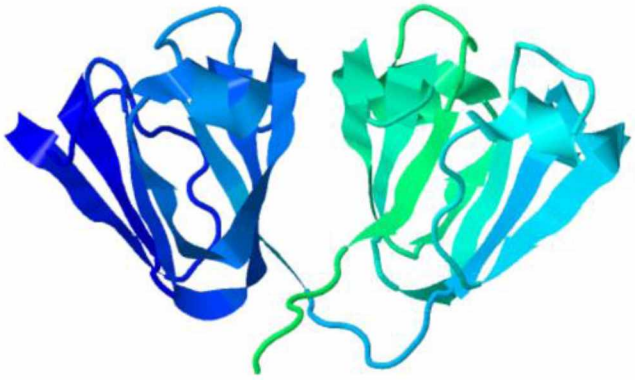
1.4.3. ΜΕΙΟΝΕΚΤΗΜΑΤΑ ΚΑΙ ΑΔΥΝΑΜΙΕΣ ΠΑΡΑΔΟΣΙΑΚΩΝ ΜΕΘΟΔΩΝ

Όμως, ο προσδιορισμός της δομής των μεγάλων πρωτεϊνών είναι μια σημαντική πρόκληση, ειδικά για τη μέθοδο NMR, που απαιτεί πολύπλοκες τεχνικές και χρονοβόρες αναλύσεις. Όμως, ακόμη και με την εφαρμογή της κρυσταλλογραφίας ακτίνων-Χ, ευδοκιμεί η ανάλυση πρωτεϊνικών μονάδων που επιτελούν μία λειτουργία, χωρίς την εναλλαγή των χαρακτηριστικών τους. Συγκεκριμένα, η κρυσταλλοποίηση πολύπλοκων, πολυλειτουργικών πρωτεϊνών, θεωρείται ως ισοδύναμη της ταυτόχρονης κρυσταλλοποίησης πολλών διαφορετικών πρωτεϊνών, κάτι που ίσως με την ισχύουσα τεχνολογία να μην είναι εφικτό. Γι' αυτό το λόγο, παρατηρείται πως το μέσο μήκος πρωτεϊνών που έχουν προσδιοριστεί και καταχωρηθεί στην PDB είναι περίπου 230 κατάλοιπα. Επίσης, η πρωτεΐνη ενδέχεται να χρειάζεται περεταίρω τροποποιήσεις, όπως γλυκοζηλίωση ή φωσφορυλίωση, ή την παρουσία συμπαραγόντων που να μην είναι διαθέσιμοι στον οργανισμό ξενιστή. [23] [24, 25]

Το γεγονός αυτό αντικατοπτρίζει τις δυσκολίες στον προσδιορισμό της δομής μεγάλων πρωτεϊνών αλλά και στις μεθόδους άμεσα συνυφασμένες με αυτόν, όπως η έκφραση και η απομόνωσή τους. Μεγάλης πρακτικής σημασίας, λοιπόν, παρουσιάζεται η τομή των πρωτεϊνών με τέτοιο τρόπο, που να μην επηρεάζει την αναδίπλωση, και να μην οδηγεί σε απώλεια της λειτουργικότητάς τους. Πειραματικές τεχνικές ταυτοποίησης τέτοιων υπομονάδων, βασίζονται σε πρωτεόλυση, όμως δεν επιλύονται έτσι τα προβλήματα που σχετίζονται με την έκφραση και την απομόνωση τους. Χρησιμοποιούνται επίσης μέθοδοι screening οι οποίες απομονώνουν τμήματα αναδιπλωμένα, από μια βιβλιοθήκη που δημιουργείται από τυχαία θραύσματα της πρωτεΐνης, αποφεύγοντας έτσι την διαδικασία απομόνωσης ολόκληρης της έκτασης της πρωτεΐνης. Ωστόσο, οι πειραματικές μέθοδοι, είναι χρονοβόρες, απαιτούν προσπάθεια και εμφανίζουν υψηλό κόστος. Έτσι, γίνεται αντιληπτό, γιατί είναι σημαντικό να αναπτυχθούν υπολογιστικές μέθοδοι, που θα υποβοηθήσουν το εγχείρημα αυτό. [26]

1.5.1. DOMAIN

Με τον όρο domain αναφερόμαστε στη στοιχειώδη μονάδα πρωτεϊνικής δομής. Δηλαδή, πρόκειται για μια μονάδα που μπορεί να διπλωθεί ανεξάρτητα σε σταθερή τριτοταγή δομή και εμφανίζεται ως συμπαγής, ημιαυτόνομη, με υδρόφοβο πυρήνα. Συνήθως εξελίσσεται ανεξάρτητα και για το λόγο αυτό χαρακτηρίζεται και ως εξελικτική μονάδα. [27] Κάθε domain σε μια multidomain πρωτεΐνη έχει τη δική της λειτουργία (και δομή) και δρα συνεργικά με τους γείτονές της [28] καθορίζοντας τα βιολογικά μονοπάτια στα οποία συμμετέχει η πρωτεΐνη, αλλά και τα μόρια με τα οποία αλληλοεπιδρά. [29] Τα domain συνδυάζονται σε διαφορετικές πρωτεΐνες και αυτός είναι ο λόγος που η οικογένεια στην οποία αυτά ανήκουν, ενδεχομένως να διαφέρει από την οικογένεια που κατατάσσεται η συνολική πρωτεΐνη. [30, 31]



Εικόνα 11. Δομή της 1HKO όπου διαφαίνονται τα δύο domain και η συνδετική περιοχή που τις ενώνει.[32, 33]

Η προέλευση των multidomain πρωτεϊνών αποδίδεται στην επιλεκτική πίεση με απώτερο σκοπό την δημιουργία νέων λειτουργιών, στα πλαίσια της εξέλιξης. Παρατηρείται λοιπόν ότι τα 2/3 των πρωτεϊνών σε μονοκύτταρους οργανισμούς και περισσότερο από το 80% των πρωτεϊνών των μεταζώων είναι multidomain πρωτεΐνες. Όπως είναι αναμενόμενο, όσο αυξάνεται η πολυπλοκότητα του οργανισμού, τόσο αυξάνεται και ο αριθμός των domain στις πρωτεΐνες του. [34] Επομένως, τα πρωτεϊνικά domain παίζουν καθοριστικό ρόλο στην κατανόηση της δομικής αρχιτεκτονικής μεγάλων πρωτεϊνών, αλλά και αποκαλύπτουν στοιχεία για την εξελικτική τους πορεία εφόσον πολλές πρωτεΐνες έχουν διαφοροποιηθεί από τους κοινούς τους πρόγονους παρουσιάζοντας διαφορετικούς συνδυασμούς και συσχετισμούς domain. [34]

Τέλος, έχει παρατηρηθεί ότι είναι genetically mobile και μπορούν να μετακινούνται μεταξύ βιολογικών συστημάτων με τον μηχανισμό του gene- exon shuffling. Αυτή η γενετική κινητικότητα, λοιπόν, θεωρείται υπεύθυνη για πολλές οικογένειες modular πρωτεϊνών που διαφέρουν ως προς την διάταξη και το πλήθος των domain.

1.6. ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ ΓΙΑ DOMAINS

Ποικίλες μέθοδοι ανάθεσης και βάσεις δεδομένων έχουν αναπτυχθεί για πρωτεΐνες γνωστών δομών. Η ανάθεση αυτή, επιτυγχάνεται με οπτική παρατήρηση, αυτοματοποιημένο προσδιορισμό ή και των συνδυασμό αυτών. Η βάση Structural Classification of Proteins (SCOP)[35] βασίζεται εξ ολοκλήρου σε οπτική ανάθεση από ειδικούς, λαμβάνοντας υπόψιν την εξελικτική κληρονομιά των περιοχών.

Οι πρώιμες αυτοματοποιημένες μέθοδοι δημιουργούσαν γεωμετρικά μοντέλα και ενίσχυαν την πρόβλεψη με πληροφορίες δευτεροταγούς δομής (DOMAK [36]) ή με εντοπισμό υδρόφοβων πυρήνων (DETECTIVE [37]). Μετέπειτα έρευνες στράφηκαν προς συστάδες δευτεροταγούς δομής [38], τη συμπαγότητα (PUU[39]), εφαρμοσμένες προσεγγίσεις σε θεωρίες γράφων[40] ή στη χρήση consensus ακολουθιών[41, 42]. Η βάση δεδομένων Class Architecture Topology Homology (CATH)[42] συνδυάζει διαφορετικές μεθόδους αποτίμησης και συμπεριλαμβάνει αυτοματοποιημένο και χειροκίνητο έλεγχο των αποτελεσμάτων. Ακόμα κάποιες ημι-αυτοματοποιημένες βάσεις, είναι η Dali Domain Dictionary[43] που χρησιμοποιεί τον αλγόριθμο PUU[39] και η MMDB που εφαρμόζει τον VAST[44]. Οι βάσεις αυτές, ορίζουν κατά κύριο λόγο, τα domain ως την δομικά συμπαγέστερη περιοχή.

Μία πλήρως αυτοματοποιημένη μέθοδος για πρόβλεψη domain είναι άκρως επιθυμητή, εφόσον θα εξαλείψει την εξάρτηση από την υποκειμενική γνώμη ειδημόνων. Η μόνη μέθοδος που προσδιορίζει συνεχόμενης αλληλουχίας domain από τρισδιάστατες δομές ,αυτοματοποιημένα, είναι η Protein Informatics System for Modeling (PrISM [45]). Η σπουδαιότητα του εγχειρήματος αυτού, διαφαίνεται και από το γεγονός ότι η ταύτιση των CATH και SCOP στην ανάθεση domain και όριων αυτών για λυμένες πρωτεϊνικές δομές εκτιμάται γύρω στο 68%. Η ασυμφωνία αυτή, οφείλεται όχι μόνο στην ύπαρξη ενδεχομένων ατελειών στις μεθόδους, αλλά και στην ασάφεια που υπάρχει σχετικά με τον προσδιορισμό του όρου domain, εφόσον δεν υπάρχουν σαφή διατηρημένα τμήματα ακολουθιών που να σηματοδοτούν τα σύνορά τους. [28, 46]

1.7. LINKER

Ισοδύναμος με την αναζήτηση domain στις πρωτεϊνικές ακολουθίες είναι ο εντοπισμός των linker. Ως linker ορίζεται η περιοχή που μεσολαβεί ανάμεσα σε δύο διαδοχικά domain. Οι συνδέτες στερούνται κανονικής δευτεροταγούς δομής και εμφανίζουν διαφορετικούς βαθμούς ευκαμψίας (flexibility) ώστε να ανταποκριθούν στην εκάστοτε βιολογική διαδικασία. Το μεγαλύτερο ποσοστό των linker καταλοίπων, υιοθετούν δευτεροταγή δομή α-έλικας (38.3%), ενώ παρατηρούνται και οι δομές β-strands και turns σε μικρότερα ποσοστά(13.6% και 8.4% αντίστοιχα). Σημαντικό είναι εξίσου και το ποσοστό των αμινοξέων που βρίσκονται σε coil ή bend δευτεροταγείς δομές (37.6%). Έρευνες έχουν δείξει ότι οι linkers παίζουν σημαντικό ρόλο στη διατήρηση συνεργατικών interdomain αλληλεπιδράσεων. Επίσης, η σύσταση και το μήκος τους επηρεάζουν την ευστάθεια/ σταθερότητα της πρωτεΐνης αλλά και την αναδίπλωση και τον προσανατολισμό των domains. [27]

Εκτός των διαφορετικών δομικών χαρακτηριστικών με τα domain, διαφοροποιείται και η σύστασή τους σε αμινοξέα. Ακριβέστερα, τα κατάλοιπα στις περιοχές των linker εμφανίζονται κυρίως υδρόφιλα, συχνά πολικά και μικρά σε μέγεθος[26]. Ειδικότερα, έχει παρατηρηθεί προτίμηση σε 5 αμινοξέα, τη Γλυκίνη, το Ασπαρτικό οξύ, την Ασπαραγίνη, τη Λυσίνη και την Προλίνη. Η τελευταία, είναι και αυτή που εμφανίζει μεγαλύτερη διαφορά εμφάνισης μεταξύ συνδετικών και περιοχών domain, ενώ έχει παρατηρηθεί ακόμα πως κατάλοιπα Προλίνης συνωστίζονται κυρίως στο κέντρο των συνδετικών περιοχών. Τα ποσοστά των αμινοξέων αυτών διαφέρουν τόσο στις linker και non-linker περιοχές, όσο και ανάμεσα στις linker και intra-domain loop ακολουθίες. Αντιθέτως, στις περιοχές συμπαγών domain, παρατηρούνται πιο υδρόφοβα κατάλοιπα όπως η Λευκίνη, η Βαλίνη, η Ισολευκίνη, η Αλανίνη, και σε μικρότερη έκταση η Αργινίνη, η Μεθειονίνη, η Τυροσίνη και το Γλουταμινικό οξύ [47].

Συνεπώς, η κατανόηση των ιδιοτήτων των linker είναι σημαντική γιατί θα διευκολυνθεί ο κατακερματισμός των πολύπλοκων multidomain πρωτεϊνών σε απλούστερες δομές(single domain) γεγονός που θα μπορούσε να ευεργετήσει μια πληθώρα επιστημονικών ερευνών και πειραματικών μεθόδων. Αρχικά, θα βελτιστοποιήσει τις μεθόδους πρόβλεψης δομής πρωτεϊνών αλλά και θα συνδράμει στα structural genomics, μέσω της καλύτερης αντίληψης του διπλώματός τους και διευκόλυνση του χαρακτηρισμού της τρισδιάστατης δομής τους. Κατ' επέκταση, αναγνωρίζοντας τη δομή, θα παραχθούν περισσότερες πληροφορίες και για τη λειτουργία άγνωστων πρωτεϊνών (functional genomics). Η συνεισφορά θα είναι επίσης σημαντική για καινοτόμα επιστημονικά εγχειρήματα όπως αυτό της μηχανικής of fusion proteins (protein

engineering) αλλά και την ερμηνεία των προϊόντων τροποποιημένων ακολουθιών DNA στα πειράματα που πραγματοποιούνται στα πλαίσια του site-directed mutagenesis.

1.8. ΥΠΟΛΟΓΙΣΤΙΚΕΣ ΜΕΘΟΔΟΙ ΠΡΟΒΛΕΨΗΣ LINKER

Καθώς πρόκειται για ένα θέμα που απασχολεί εκτενώς την επιστημονική κοινότητα, ποικίλες μέθοδοι έχουν αναπτυχθεί για τον σκοπό αυτό. Οι τρεις βασικές κατηγορίες στις οποίες συνοψίζονται οι περισσότερες προσεγγίσεις είναι μέθοδοι σύγκρισης, μέθοδοι στατιστικής ή τεχνητής νοημοσύνης και ab initio.

1.8.1. ΜΕΘΟΔΟΙ ΣΥΓΚΡΙΣΗΣ (ΒΑΣΙΣΜΕΝΕΣ ΣΤΗΝ ΟΜΟΛΟΓΙΑ)

Όλες οι μέθοδοι που συγκαταλέγονται σε αυτή την κατηγορία, (SBASE[48], SUPERFAMILY[49] and Domain Fishing[50] χρησιμοποιούν εξαντλητική αναζήτηση έναντι γνωστών πρωτεϊνικών δομών στην αντίστοιχη βάση δεδομένων για domain. Η αναζήτηση δημιουργεί αξιόπιστα αποτελέσματα όταν υπάρχουν ομόλογες περιοχές domain στη βάση δεδομένων που χρησιμοποιείται για τη σύγκριση. Όπως γίνεται αντιληπτό, καθώς οι μέθοδοι σύγκρισης απαιτούν πρότερη γνώση, δυσκολεύονται στην πρόβλεψη domain που ανήκουν σε μικρές οικογένειες που δεν έχουν προσδιοριστεί και σε καινούργιες ακολουθίες που δεν εμφανίζουν ομολογία με τις ήδη υπάρχουσες. Γενικά τα πρωτόκολλα αναζήτησης στις βάσεις για την αναγνώριση domain είναι standard σε όλες τις μεθόδους όπως τα PSI-BLAST[51] and HMM.

1.8.2. ΜΕΘΟΔΟΙ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΤΕΧΝΗΤΗΣ ΝΟΗΜΟΣΥΝΗΣ

Πρόκειται αναντίρρητα, για τις πιο διαδεδομένες μεθόδους, στις οποίες περιλαμβάνονται οι DGS [52], DomCut [23], Armadillo [27], PPRODO [53], DOMPro [54], DomNet [55], DROP [56], DOBO [57], PRODOM [58], ADDA [59] and EVEREST [60].

Στους αλγορίθμους DGS[52], DomCut[23] και Armadillo[46] χρησιμοποιούνται στατιστικά στοιχεία που παρατηρούνται με την πληροφορία να αντλείται από την ακολουθία και μόνο, όπως η κατανομή μεγέθους των domain και η διαφορετική σύσταση αμινοξέων μεταξύ των περιοχών domain και linker.

Στις μεθόδους PRODOM, ADDA και EVEREST, τα όρια των domain προκύπτουν από μεγάλης κλίμακας συγκρίσεις ακολουθιών και κατόπιν ανάλυση ομαδοποίησης.

Στους υπόλοιπους αλγορίθμους, PPRODO, DOMPro, DomNet, DROP και DOBO, τα στατιστικά χαρακτηριστικά των αμινοξέων, συνδυασμένα με position-specific scoring matrix από αναζήτηση PSI-BLAST, εκπαιδεύονται με τεχνικές μηχανικής μάθησης. Στις τεχνικές αυτές,

συγκαταλέγονται τα νευρωνικά δίκτυα, τα support vector machines (SVM) και οι ταξινομητές random forest. Η συνολική ακρίβεια παρατηρείται χαμηλότερη από τις μεθόδους που βασίζονται στην ομολογία, όμως μπορούν να αντλήσουν πληροφορία για τα όρια των ακολουθιών αποκλειστικά και μόνο από την ακολουθία.

1.8.3. ΜΕΘΟΔΟΙ ΤΡΙΣΔΙΑΣΤΑΤΩΝ ΜΟΝΤΕΛΩΝ Ab initio μέθοδοι

Οι μέθοδοι αυτές όπως οι SnapDRAGON [61], RosettaDom[62] και OPUS-DOM[63] χρησιμοποιούν μοντέλα τρισδιάστατων δομών της ακολουθίας-στόχου με τη χρήση ab initio folding ή μέσω της προσομοίωσης μοντέλων βασισμένες σε προηγούμενη γνώση. Μετά τη δημιουργία των μοντέλων, εφαρμόζονται εργαλεία για την ανάθεση των domain. Η ακρίβεια των αποτελεσμάτων εξαρτάται από την ποιότητα των 3D μοντέλων, που συνήθως μειώνεται με την αύξηση του μεγέθους των πρωτεϊνών, εξ αιτίας των περιορισμών δημιουργίας ab initio προσομοίωσης διπλώματος. Γενικά, δημιουργούν αξιόπιστα αποτελέσματα και παρέχουν τη δυνατότητα πρόβλεψης ακόμη και μη συνεχών domain, όμως είναι ιδιαίτερα απαιτητικά υπολογιστικά και δεν ενδείκνυνται σε αναλύσεις μεγάλης κλίμακας.

ΚΕΦΑΛΑΙΟ 2^{Ο}: ΥΛΙΚΑ ΚΑΙ ΜΕΘΟΔΟΙ

2.1 DATASET

Το dataset που χρησιμοποιήθηκε για την μελέτη προήλθε από το dataset του DOMpro 1.0.

Το αρχικό dataset περιλαμβάνει 354 πρωτεϊνικές ακολουθίες, καθεμιά από τις οποίες εμπεριέχει δύο ή περισσότερες domain περιοχές.

Το τελικό dataset που αξιοποιήθηκε για την εργασία, απαρτίζεται από 347 ακολουθίες. Συγκεκριμένα, αφαιρέθηκαν οι εξής 8 εγγραφές 1k7tA, 1meyC, 1tf3A, 1rmd0, 1dx5I, 2adr0, 2drpA.

Οι ακολουθίες αυτές κρίθηκε σκόπιμο να απομακρυνθούν καθώς, αφενός σύμφωνα με τη βιβλιογραφία ήταν πολύ σύντομες για να σχηματίζουν domains που διαχωρίζονται με linker (αλλά και γιατί παρήγαγαν ασταθή αποτελέσματα).

Το τελικό σύνολο πρωτεϊνών περιλαμβάνει 279 πρωτεΐνες με 1 linker, 47 πρωτεΐνες με 2 linker και 21 πρωτεΐνες με 3 ή περισσότερους linker (14 με 3, 2 με 4, 3 με 5, 1 με 6 και 1 με 7)

Το ελάχιστο μήκος linker στο σύνολο αυτό είναι 20 αμινοξέα και το μέγιστο είναι 68 ενώ το μέσο μήκος υπολογίστηκε 21,9. Επίσης το ελάχιστο μήκος domain στο σύνολο αυτό είναι 20 κατάλοιπα, το μέγιστο είναι 558 ενώ το μέσο μήκος υπολογίστηκε 124,1.

2.2 ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ ΣΤΗ ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ

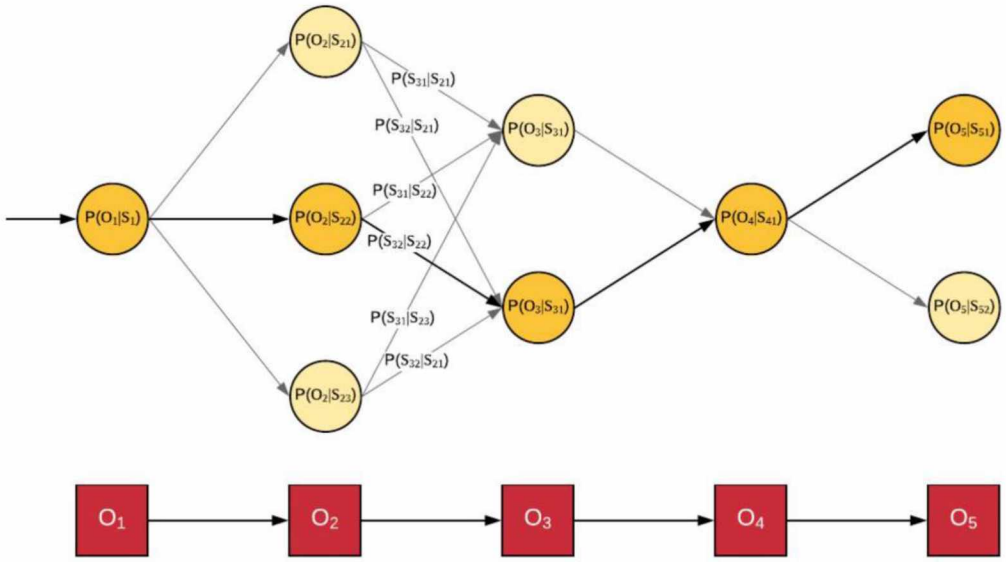
2.2.1.HIDDEN MARKOV MODELS (HMM)

Ποικίλοι αλγόριθμοι και μοντέλα για επεξεργασία σημάτων έχουν εφαρμοστεί για την ανάλυση βιολογικών ακολουθιών. Μια ιδιαίτερα δημοφιλής μέθοδος με ευρείς τομείς εφαρμογής είναι τα HMM. Είναι γνωστό ότι μπορούν να μοντελοποιούν τις συσχετίσεις μεταξύ γειτονικών συμβόλων, domains ή γεγονότων ενώ έχουν χρησιμοποιηθεί εκτενώς στην αναγνώριση φωνής και στην ψηφιακή επικοινωνία. Όσων αφορά την συνεισφορά τους στη Βιοπληροφορική, έχουν χρησιμοποιηθεί σε θεμελιώδη ζητήματα στα οποία συγκαταλέγονται η πρόβλεψη γονιδίων, η στοίχιση βιολογικών αλληλουχιών, η πρόβλεψη δευτεροταγούς δομής των πρωτεϊνών αλλά και δομικές στοιχίσεις σε μόρια RNA.

Πρόκειται για στατιστικά μοντέλα που χρησιμοποιούνται για την περιγραφή της εξέλιξης των παρατηρούμενων γεγονότων, τα οποία βασίζονται σε εσωτερικούς παράγοντες που δεν είναι δυνατή η παρατήρησή τους. Τα παρατηρούμενα γεγονότα καλούνται σύμβολα, ενώ οι λανθάνοντες παράγοντες, καταστάσεις. Ένα HMM αποτελείται από δύο στοχαστικές διαδικασίες, μία αόρατη διεργασία των κρυφών καταστάσεων, και μια ορατή επεξεργασία των παρατηρούμενων συμβόλων. Οι κρυμμένες καταστάσεις σχηματίζουν μια αλυσίδα Markov και η πιθανότητα κατανομής των παρατηρούμενων συμβόλων εξαρτάται από τις υποβόσκουσες καταστάσεις.

Καθώς σύμφωνα με τη βιβλιογραφία οι multidomain πρωτεΐνες αποτελούνται από διαφορετικές δομικές και λειτουργικές μονάδες, αναμένεται να χαρακτηρίζονται και από διαφορετικές στατιστικές ιδιότητες. Για τον λόγο αυτό, η μέθοδος των HMM είναι η κατάλληλη για το εγχείρημα αυτό.

Συγκεκριμένα, ορίζουμε ως $x=x_1x_2...x_L$ το διάνυσμα των παρατηρούμενων συμβόλων. Καθώς η πρόβλεψη αναφέρεται σε πρωτεΐνες, L θα είναι το μήκος της ακολουθίας και με x_i συμβολίζονται οι παρατηρήσεις αποτελούμενες από ένα εκ των 20 αμινοξέων. Οι υποβόσκουσες καταστάσεις συμβολίζονται ως $y=y_1y_2...y_L$, όπου y_n είναι η κρυφή κατάσταση της n-οστής παρατήρησης x_n. Στη συγκεκριμένη περίπτωση, οι κρυφές καταστάσεις υποδηλώνουν την αλληλουχία καταστάσεων έως μια θέση i στην αλληλουχία, ή αλλιώς το «μονοπάτι». Κάθε σύμβολο x_n παίρνει ένα πεπερασμένο πλήθος πιθανών τιμών από το σετ των παρατηρήσεων $O=O_1O_2,..,O_N$ και κάθε κατάσταση y_n λαμβάνει μία από τις τιμές από το σύνολο των καταστάσεων $S=1,2,..,M$, όπου N είναι το πλήθος των διακριτών παρατηρήσεων και M το πλήθος των διακριτών καταστάσεων του μοντέλου.



Εικόνα 12 Απεικόνιση ενός HMM[64].

Υποθέτουμε ότι η αλληλουχία των κρυφών καταστάσεων είναι μία ομογενής ως προς το χρόνο πρώτης τάξεως αλυσίδα Markov. Δηλαδή, η πιθανότητα μετάβασης στην κατάσταση j την επόμενη χρονική στιγμή εξαρτάται μόνο από τη δεδομένη κατάσταση i, και αυτή η πιθανότητα δεν μεταβάλλεται με την πάροδο του χρόνου.

$$P\{y_{n+1} = j | y_n = i, y_{n-1} = i_{n-1}, \ldots, y_1 = i_1\} = P\{y_{n+1} = j | y_n = i\} = t(i, j)$$

Για όλες τις καταστάσεις $i, j \in S$ όπου $n \geq 1$.

Η σταθερή πιθανότητα για τη μεταφορά από την κατάσταση i στην κατάσταση j ονομάζεται πιθανότητα μετάβασης και συμβολίζεται ως $t(i,j)$. Για την αρχική κατάσταση y_1 θεωρούμε την αρχική πιθανότητα καταστάσεων ως $\pi(i) = P\, y_1 = i\}$ ενώ η πιθανότητα η n-οστή παρατήρηση να είναι $x_n = x$ εξαρτάται μόνο από την κρυφή κατάσταση y_n. Έτσι:

$$P\{x_n = x | y_n = i, y_{n-1}, x_{n-1}, \ldots\} = P\{x_n = x | y_n = i\} = e(x|i)$$

Για όλες τις πιθανές παρατηρήσεις $x \in O$, κάθε κατάσταση $i \in S$, όπου $n \geq 1$. Αυτή καλείται η πιθανότητα εκπομπής του x στην κατάσταση i, και συμβολίζεται ως $e(x|i)$. Τα τρία αυτά μέτρα $t(i,j)$, $\pi(i)$ και $e(x \mid i)$ χαρακτηρίζουν πλήρως ένα HMM και θα τα συμβολίζουμε ως Θ.

Γνωρίζοντας αυτές τις παραμέτρους, είναι δυνατός ο υπολογισμός της πιθανότητας ένα HMM να παράξει την αλληλουχία παρατηρήσεων $x = x_1 x_2 \ldots x_L$ με την αλληλουχία κρυφών καταστάσεων $y = y_1 y_2 \ldots y_L$. Η εκ των προτέρων πιθανότητα $P\{x, y \mid \Theta\}$ υπολογίζεται ως

$$P\{x, y\,|\Theta\} = P\{x|y, \Theta\}P\{y|\Theta\}$$

Όπου

$$P\{x|y, \Theta\} = e(x_1|y_1)e(x_2|y_2)e(x_3|y_3)\ldots e(x_L|y_L)$$

$$P\{y|\Theta\} = \pi(y_1)t(y_1, y_2)t(y_2, y_3)\ldots t(y_{L-1}, y_L).$$

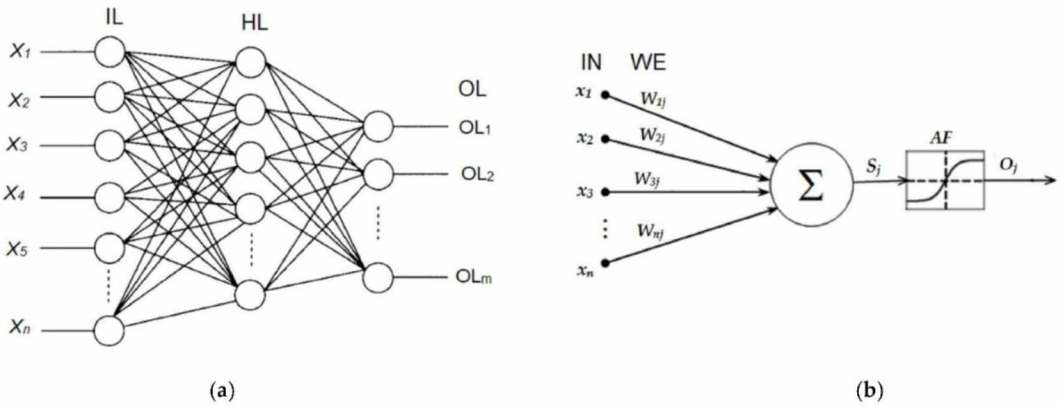
Όπως γίνεται αντιληπτό, έχοντας τη γνώση της αλληλουχίας των κρυφών καταστάσεων ο υπολογισμός των παρατηρούμενων πιθανοτήτων είναι απλός. [65]

2.2.2. ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ (NN)

Παρότι τα Νευρωνικά δίκτυα αποτελούν ξεχωριστή οντότητα, συνήθως αποτελούν τμήμα μια μεγαλύτερης εφαρμογής και δεν χρησιμοποιούνται ανεξάρτητα. Αποτελούν σημαντικό κομμάτι της Βιοπληροφορικής ενώ, κάποια από τα βιολογικά προβλήματα για την επίλυση των οποίων έχουν αξιοποιηθεί περιλαμβάνουν την αναγνώριση γονιδίων και κωδικών περιοχών, την αναγνώριση μεταγραφικών και μεταφραστικών σημάτων και την πρόβλεψη πρωτεϊνικής δομής.

Τα νευρωνικά δίκτυα σχεδιάστηκαν ως ένα υπολογιστικό μοντέλο που μιμείται τον τρόπο λειτουργίας του εγκεφάλου. Ο εγκέφαλος αποτελείται από τους νευρώνες. Κάθε νευρώνας αποτελείται από το κυτταρικό σώμα, αρκετούς δενδρίτες και έναν μοναδικό και εκτενή νευράξονα. Κάθε νευρώνας συνδέεται με άλλους, μέσω των δενδριτών και του νευράξονα. Συγκεκριμένα, οι δενδρίτες λαμβάνουν σήματα από άλλους νευρώνες και τα σήματα αυτά λειτουργούν ως είσοδος. Το ηλεκτρικό δυναμικό του κυτταρικού σώματος, μεταβάλλεται σύμφωνα με τις εισόδους που αυτό λαμβάνει αν αυτές ξεπεράσουν ένα κατώφλι και στη συνέχεια ένας ηλεκτρικός παλμός στέλνεται κατά μήκος του άξονα. Η έξοδος αυτή, παρομοίως αποτελεί είσοδο για πολλούς άλλους νευρώνες.

Αντιστοίχως, ένα τεχνητό νευρωνικό δίκτυο απαρτίζεται από πολλές υπολογιστικές μονάδες που συνδέονται μεταξύ τους με ακμές, καθεμία από τις οποίες έχει ένα βάρος που σχετίζεται με αυτή. Τα βάρη αντιστοιχούν στην μακροπρόθεσμη μνήμη, και είναι ενδεικτικά της σημαντικότητας της εκάστοτε εισόδου. Όπως και στα βιολογικά δίκτυα, οι είσοδοι λαμβάνονται από τις ακμές εισόδου. Κατόπιν, κάθε μονάδα υπολογίζει το σταθμισμένο άθροισμα των τιμών εισόδου και περνάει από συνάρτηση μεταφοράς απ' όπου εξάγεται μία και μοναδική τιμή (εικόνα 7(β)).



Εικόνα 13 (α) παράδειγμα Νευρωνικού δικτύου με κρυφούς νευρώνες
(β) η δομή ενός απλού τεχνητού νευρώνα[66]

Σε ένα νευρωνικό δίκτυο μπορούν να υπάρχουν ενδιάμεσα στρώματα νευρώνων (μεταξύ της εισόδου και της εξόδου) που αναφέρονται ως κρυφά στρώματα νευρώνων. Κάθε νευρώνας (κόμβος) του ενός στρώματος, συνδέεται με κάθε νευρώνα του επόμενου στρώματος.

Όμοια με τον εγκέφαλο, τα νευρωνικά μπορούν να μάθουν από παραδείγματα και να εφαρμόσουν την γνώση αυτή σε καινούργιες καταστάσεις. Η διαδικασία αυτή καλείται εκπαίδευση του δικτύου. [67] Πρόκειται για μια διαδικασία εκτίμησης των παραμέτρων και υπάρχουν πολυάριθμοι αλγόριθμοι που χρησιμοποιούνται για αυτό το σκοπό. Ένας από αυτούς είναι και ο «back-propagation». Η μεθοδολογία του οποίου περιλαμβάνει:

1. Αρχικοποίηση των βαρών με ανάθεση τυχαίων τιμών, που είθισται να ανήκει σε συγκεκριμένο εύρος τιμών που προκύπτει από τον αριθμό των νευρώνων.
2. Υπολογισμός του αποτελέσματος που προκύπτει με τα δεδομένα βάρη, για όλες τις παρατηρήσεις.
3. Υπολογισμός του σφάλματος, δηλαδή της απόκλισης του αποτελέσματος από τις παρατηρηθείσες τιμές.
4. Επαναυπολογισμός των βαρών με τη μέθοδο gradient descent.
5. Μεταβίβαση του σήματος προς τα πίσω, και διαδοχική τροποποίηση των βαρών.
6. Οι αντίστοιχες συναρτήσεις ενεργοποίησης καθορίζουν τους υπολογισμούς σε κάθε βήμα προς τα πίσω.
7. Όταν το σήμα φτάσει στους νευρώνες εισόδου, ολοκληρώνεται η εποχή και τα βάρη τροποποιούνται ώστε να μειωθεί το συνολικό σφάλμα.
8. Η παραπάνω διαδικασία (2-7) επαναλαμβάνεται, μέχρι να προκύψει το επιθυμητό συνολικό σφάλμα ή μέχρι να ολοκληρωθεί ο αριθμός επαναλήψεων που έχει οριστεί.

Η εκπαίδευση νευρωνικών αποτελεί μια διαδικασία που απαιτεί παρακολούθηση, εξαιτίας προβλημάτων που προκύπτουν από τα πολυάριθμα συνοπτικά βάρη της κωδικοποίησης των αλληλουχιών και του αριθμού των κρυφών νευρώνων. Ένας τρόπος αποφυγής τους, είναι η χρήση της τεχνικής cross validation, που θα αναλυθεί παρακάτω.

2.3. JUCHMME

Το λογισμικό που χρησιμοποιήθηκε για αυτή την εργασία ονομάζεται JUCHMME[68].

Πρόκειται για ένα εργαλείο σε Java που αναπτύχθηκε για την ανάλυση βιολογικών ακολουθιών με την εφαρμογή Class Hidden Markov Models, όπως υποδηλώνουν και τα αρχικά του (Java Utility for Class Hidden Markov Models and Extensions). Κάποια από τα βασικά πλεονεκτήματα του είναι:

- ✓ Η ευελιξία και η προσαρμοστικότητα σε ποικίλα προβλήματα, συναντούν την ευχρηστία καθώς μπορεί να χρησιμοποιηθεί και από άτομα που δεν έχουν καμία γνώση προγραμματισμού για τη δημιουργία μοντέλων οποιασδήποτε βιολογικής ακολουθίας (DNA, πρωτεϊνών κ.α.).
- ✓ Διαθέτει μεγάλο εύρος σε μεθόδους εκπαίδευσης HMMs για επισημασμένες ακολουθίες.
- ✓ Δίνεται η επιλογή δύο κριτηρίων εκπαίδευσης. Όταν πρόκειται για ακολουθίες με διαθέσιμες ετικέτες, συνήθης, είναι η μοντελοποίηση με Maximum Likelihood (ML), που υποστηρίζει τους αλγορίθμους του Baum-Welch καθώς και επεκτάσεις αυτού, gradient-descent και Viterbi. Επίσης, ως εναλλακτικό κριτήριο εκπαίδευσης παρουσιάζεται το Conditional Maximum Likelihood (CML) που προτείνεται για διακεκριμένη εκπαίδευση, υποστηρίζει μόνο gradient based αλγορίθμους και μπορεί να εφαρμοστεί και με κρυφά νευρωνικά δίκτυα.
- ✓ Η αποκωδικοποίηση του μοντέλου μπορεί να γίνει με έναν εκ τους αλγορίθμους που υποστηρίζονται, όπως οι Viterbi, N–Best , posterior–Viterbi και Optimal Accuracy Posterior Decoder .
- ✓ Επιπρόσθετα, δίνεται η ευχέρεια να διαλεγεί διαδικασία αξιολόγησης ανάμεσα από τις independent tests, self-consistency tests, jacknife tests, k–fold cross validation και early stopping. Σημαντικό είναι και ότι όλες οι προαναφερόμενες μέθοδοι παράγουν ευρέως γνωστά μέτρα αξιοπιστίας για την εκτίμηση των αποτελεσμάτων.
- ✓ Τέλος, αξίζει να αναφερθούν και οι επεκτάσεις στα HMM με σκοπό να ξεπεράσουν τους περιορισμούς που συνεπάγονται οι κλασσικές μέθοδοι HMM και CHMM. Σε αυτές συγκαταλέγονται η εφαρμογή segmental k–means για την εκπαίδευση με ετικέτες, κρυφά νευρωνικά δίκτυα με είσοδο προηγούμενων παρατηρήσεων και ημι-επιβλεπόμενες μεθόδους εκπαίδευσης όχι μόνο για δεδομένα που φέρουν ετικέτες εξολοκλήρου, αλλά και για αυτά που φέρουν μερικώς, ή και καθόλου.

Το JUCHHME είναι ένα εκτελέσιμο αρχείο Java που καλείται μέσω της γραμμής εντολών, απαιτεί 32-bit ή 64-bit περιβάλλον Java, έκδοσης 7 ή μεταγενέστερης.

2.3.1. DECODING

Πρόκειται για τη διαδικασία κατά την οποία προσδιορίζεται, με δεδομένη αλληλουχία κρυφών καταστάσεων, από ποια αλληλουχία μεταβλητών είναι πιο πιθανό να απορρέει η συγκεκριμένη ακολουθία παρατηρήσεων. Η διαδικασία της αποκωδικοποίησης παρατηρείται σε κάθε μοντέλο που περιέχει κρυφές μεταβλητές όπως και στα HMM.

Ο πιο απλός τρόπος θα ήταν με εφαρμογή του αλγορίθμου Forward, κατά τον οποίο, για κάθε πιθανή κρυμμένη αλληλουχία καταστάσεων, υπολογίζεται η πιθανότητα της παρατηρήσιμης ακολουθίας, για την εκάστοτε κρυφή. Τελικά, επιλέγεται η κρυφή αλληλουχία καταστάσεων που εμφάνισε τη μέγιστη πιθανότητα απ' όσες υπολογίστηκαν. Όμως κάτι τέτοιο δεν μπορεί να εφαρμοστεί σε προβλήματα του πραγματικού κόσμου, εφόσον ο αριθμός καταστάσεων των ακολουθιών είναι μεγάλος.

Εναλλακτικά, ένας από τους πιο κοινούς αλγορίθμους αποκωδικοποίησης είναι ο Viterbi, ενώ κάποιες πιο αναβαθμισμένες τεχνικές περιλαμβάνουν τον N-Best, Posterior-Viterbi (POSVIT) και Optimal Accuracy Posterior Decoder (PLP).

2.3.1.1. ΑΛΓΟΡΙΘΜΟΣ VITERBI

Όμοια με τον forward αλγόριθμο, αποτελεί ένα είδος δυναμικού προγραμματισμού. Αυτό σημαίνει ότι δεν διέρχεται από όλες τις πιθανές καταστάσεις και κατόπιν υπολογίζει την πιο πιθανή διαδρομή, αλλά σε κάθε κόμβο επιλέγει να μεταβεί στην πιο πιθανή ακολουθία καταστάσεων. Η τιμή κάθε κελιού, υπολογίζεται παίρνοντας αναδρομικά το πιο πιθανό μονοπάτι που μπορεί να οδηγήσει σε αυτό το κελί. Πρακτικά, δεν διαφέρει δραστικά από τον Forward, παρά μόνο στο ότι τα διαδοχικά αθροίσματα αντικαθίστανται από μεγιστοποιήσεις. Δηλαδή, το τελικό μονοπάτι υπολογίζεται από τον τύπο

$$P(x, \pi_{max}/\theta) = max\, u_k(L)a_{kE}\}$$

, όπου u_k οι πιθανότητες μετάβασης από την προηγούμενη κατάσταση και ισχύει ότι για κάθε

$1 \leq i \leq L$:

$$u_l(i) = e_l(x_i)max\, u_k(i\text{-}1)a_{kl}\}.$$

2.3.1.2. ΑΛΓΟΡΙΘΜΟΣ 1-BEST

Ο 1-best αλγόριθμος είναι ευριστικός, και προσπαθεί να εντοπίσει το πιο πιθανό μονοπάτι ετικετών μιας ακολουθίας και όχι το πιο πιθανό μονοπάτι καταστάσεων.

Όμως, συγκριτικά με τον Viterbi, ο χρόνος εκτέλεσης εμφανίζεται σημαντικά αυξημένος και για αυτό δεν επιλέχθηκε να αποκωδικοποιήσει τα δεδομένα.

2.3.1.3. POSTERIOR VITERBI DECODING (POSVIT)

Ο αλγόριθμος αυτός, είναι μια παραλλαγή του Viterbi, στον οποίο οι πιθανότητες γεννήσεως αντικαθίστανται από τις εκ των υστέρων πιθανότητες και οι πιθανότητες μετάβασης από την συνάρτηση δέλτα, σύμφωνα με την οποία ισχύει:

$$\delta(k,l) = \begin{cases} 1, & \alpha\nu\ \alpha(k,l) > 0 \\ 0, & \text{αλλιώς.} \end{cases}$$

Και το μονοπάτι υπολογίζεται από τον τύπο:

$$\pi^{PV} = \operatorname{argmax} \prod_{i=1}^{L} \delta(\pi_i, \pi_{i+1})\, P(\pi_i/x)$$

2.3.1.4. OPTIMAL ACCURACY POSTERIOR DECODER ALGORITHM (PLP)

Ο αλγόριθμος αυτός μοιάζει ιδιαίτερα με τον Posterior-Viterbi και διαφοροποιείται κυρίως στην περίπτωση του CHMM. Οι διαφορές αυτές είναι πως:

1. Χρησιμοποιούνται εκ των υστέρων πιθανότητες των σημάνσεων σε αντίθεση με τη χρήση εκ των υστέρων πιθανότητες των καταστάσεων.
2. Δεν υπολογίζεται το γινόμενο αυτών, αντιθέτως, μεγιστοποιείται το άθροισμά τους. Αυτό οφείλεται στο γεγονός ότι ο PLP υπολογίζει μια αλληλουχία από σημάνσεις που είναι συμβατές με το μοντέλο, αλλά ενδέχεται να περιέχουν πολλά εναλλακτικά μονοπάτια.

Συνεπώς, η τελική πιθανότητα που αποδίδει ο αλγόριθμος αυτός, δεν είναι συγκρίσιμη σε απόλυτες τιμές με τις πιθανότητες των άλλων αλγορίθμων.

2.3.2. ΜΕΘΟΔΟΙ ΕΚΤΙΜΗΣΗΣ

2.3.2.1. MAXIMUM LIKELIHOOD (ML)

Η μέθοδος μέγιστης πιθανοφάνειας βρίσκει ένα σετ τιμών θ^{ML} του μοντέλου, που μεγιστοποιούν την συνάρτηση πιθανοφάνειας, για τη δεδομένη είσοδο. Ουσιαστικά, εκφράζει την από κοινού συνάρτηση κατανομής όλων των παρατηρήσεων, δεδομένων των παραμέτρων του μοντέλου, αν οι παράμετροι θεωρηθούν ως τυχαίες μεταβλητές. Δηλαδή ισχύει ότι:

$$\theta^{ML} = arg_{\theta} max\ P(x|\theta)$$

Έχει καθιερωθεί, να χρησιμοποιείται ο υπολογισμός του λογαρίθμου της πιθανοφάνειας, καθώς μεγιστοποιείται στα ίδια σημεία με την πιθανοφάνεια, άρα έχουμε:

$$l(\mathbf{x}/\theta)=\log P(\mathbf{x}/\theta)$$

Κάθε αλληλουχία θεωρείται ανεξάρτητη από τις υπόλοιπες, και για τον λόγο αυτό η συνολική πιθανοφάνεια εκφράζεται ως το γινόμενο των πιθανοφανειών των επιμέρους αλληλουχιών.

2.3.2.2. CONDITIONAL MAXIMUM LIKELIHOOD (CML)

Ο CML επιλέγεται μόνο όταν οι ακολουθίες διαθέτουν ετικέτες. Συγκεκριμένα, πρόκειται για την αναζήτηση των παραμέτρων θ^{CML} που προκύπτουν από την αφαίρεση της πιθανοφάνειας στην οποία οι ετικέτες λαμβάνονται υπόψιν με την πιθανοφάνεια στην οποία αυτές δεν υπολογίζονται.

Έτσι έχουμε:

$$\theta^{CML}= \arg_{\theta}\max \mathrm{P}(\mathrm{y}|\mathbf{x},\theta)=l_c-l_f,$$

όπου: $l_c = -\log P(\mathbf{x},\mathbf{y}|\theta)$ και
$l_f = -\log P(\mathbf{x}|\theta)$.

Συγκριτικά με τον ML, χρειάζεται τα δεδομένα να έχουν εκπαιδευτεί με τη μέθοδο Gradient-Descent, εμφανίζει μεγαλύτερη ευαισθησία και αποδίδει καλύτερα όταν οι ετικέτες είναι αξιόπιστες. Όμως, εμφανίζει ένα σημαντικό μειονέκτημα, είναι χρονικά ασύμφορος καθώς απαιτεί τον διπλάσιο χρόνο.

2.3.3 ΑΛΓΟΡΙΘΜΟΙ ΕΚΠΑΙΔΕΥΣΗΣ

2.3.3.1. ΑΛΓΟΡΙΘΜΟΣ BAUM-WELCH

Εκτός από την εύρεση της πιθανοφάνειας, πρόκληση αποτελεί η αναζήτηση της αλληλουχίας των καταστάσεων. Όπως γίνεται αντιληπτό, το πρόβλημα γίνεται πιο σύνθετο διότι πρέπει να εκτιμηθούν και οι παράμετροι ταυτόχρονα με τα μονοπάτια. Ένας από τους αλγορίθμους που προτείνεται για τον σκοπό αυτό, είναι ο Baum-Welch. Επιλέγεται για εκτίμηση μέγιστης πιθανοφάνειας όταν υπάρχουν ελλείπουσες τιμές, ενώ αποτελεί μία ειδικότερη προσέγγιση του αλγορίθμου Expectation-Maximisation (EM).

Οι εκτιμητές των πιθανοτήτων μετάβασης, εφόσον υπάρχει πρότερη γνώση του μονοπατιού είναι:

$$\hat{a}_{kl} = \frac{A_{kl}}{\sum_{l'} A_{kl'}}$$

και αντίστοιχα οι εκτιμητές για τις πιθανότητες εκπομπής είναι:

$$\hat{e}_k(b) = \frac{E_k(b)}{\sum_{b'} E_k(b)}$$

Τα αθροίσματα στους παρονομαστές εκτείνονται σε όλο το εύρος των παραμέτρων και για να αντιμετωπιστεί το πρόβλημα μη εμφάνισης παραμέτρου ούτε μια φορά στο σύνολο εκπαίδευσης μπορεί να προστεθούν ψευδοτιμές. Για τις μεταβάσεις ισχυεί ότι:

$$A_{kl} = \sum_{\pi} P(\pi|\mathbf{x},\theta^t) A_{kl}(\pi) = \frac{1}{P(\mathbf{x})} \sum_i f_k(i) a_{kl} e_l(x_{i+1}) b_l(i+1)$$

και για τις εκπομπές ισχύει ότι:

$$E_k(b) = \sum_{\pi} P(\pi|\mathbf{x},\theta^t) E_k(b,\pi) = \frac{1}{P(\mathbf{x})} \sum_{\{i|x_i=b\}} f_k^j(i) b_k^j(i)$$

Ύστερα από τις ανάλογες πράξεις, η τελική συνάρτηση που προκύπτει είναι:

$$Q(\theta \mid \theta^t) = \sum_{k=1} \sum_b E_k(b) \log e_k(b) + \sum_{k=0} \sum_{l=1} A_{kl} \log \alpha_{kl}$$

Η τελευταία συνάρτηση, μεγιστοποιείται από τους εκτιμητές των πιθανοτήτων μετάβασης και εκπομπών. Συνοπτικά, ο αλγόριθμος Baum-Welch απαρτίζεται από το E-βήμα(expectation), που υπολογίζονται οι ποσότητες $f_k(i)$, $b_k(i)$, από τους αλγορίθμους Forward και Backward αντίστοιχα και στη συνέχεια οι αναμενόμενες τιμές για τις πιθανότητες A_{kl}, $E_k(b)$. Το βήμα της βελτιστοποίησης, αφορά την επανα-εισαγωγή των A_{kl}, $E_k(b)$ και στον επανα-υπολογισμό της πιθανοφάνειας του μοντέλου. Αν οι μεταβολές σε αυτήν και στην Q, μετά από κάποια βήματα είναι μικρότερες από ένα κατώφλι, τότε ο αλγόριθμος τερματίζεται (Bagos P. G., 2015).

2.3.3.2. ΑΛΓΟΡΙΘΜΟΣ GRADIENT – DESCENT

Ο αλγόριθμος Baum-Welch, παρά τα προτερήματα που προσφέρει, όπως την σίγουρη σύγκλιση και την ταχύτητα εκτέλεσης, διαθέτει κάποιους σημαντικούς περιορισμούς. Αφενός, αν μια παράμετρος μηδενιστεί, δεν μπορεί να αλλάξει τιμή έπειτα, και αφετέρου απαιτείται ανανέωση των παραμέτρων, την ώρα που όλο το σύνολο εκπαίδευσης έχει ήδη παρουσιαστεί.

Αντίθετα, η μέθοδος Gradient-Descent αποτελεί μια ευριστική μέθοδο ελαχιστοποίησης ενέργειας που προσφέρει «ομαλή» εκπαίδευση. Αυτό σημαίνει ότι δεν απαιτείται ο υπολογισμός βοηθητικών μεταβλητών σε κάποιο ενδιάμεσο βήμα. Εξαλείφει επίσης τον κίνδυνο μηδενισμού κάποιας παραμέτρου και υποστηρίζει την μέθοδο «online training» με ανανέωση των παραμέτρων ανά κατάλοιπο σε αντίθεση με την παρουσίαση ολόκληρου του συνόλου εκπαίδευσης. Μέσω αυτής πραγματοποιείται και η εκπαίδευση Δεσμευμένης Μέγιστης Πιθανοφάνειας (Conditional Maximum Likelihood, CML) (Tamposis, Tsirigos, Theodoropoulou, Kontou, & Bagos, 2019). Η βασική της αδυναμία, είναι η αυθαίρετη επιλογή του ρυθμού μάθησης και κατά επέκταση, αστάθεια στην εκπαίδευση και μικρή ταχύτητα σύγκλισης στον πολυδιάστατο χώρο παραμέτρων.

2.3.4 ΕΠΕΚΤΑΣΕΙΣ

2.3.4.1. ΗΜΙ-ΕΠΙΒΛΕΠΟΜΕΝΗ ΜΑΘΗΣΗ

Το JUCHMME προσφέρει τη δυνατότητα στον χρήστη να χρησιμοποιήσει σημασμένα, μη σημασμένα ή μερικώς σημασμένα δεδομένα. Βασιζόμενο σε μια τροποίηση του Expectation-Maximization (EM) αλγορίθμου, όπου τα δεδομένα που λείπουν θεωρούνται ως χαμένες ετικέτες των μη σημασμένων ή των μερικά σημασμένων δεδομένων. Αυτό συμβαίνει μέσω μιας προσέγγισης self-training όπου το μοντέλο εκπαιδεύεται αρχικά στις πλήρως σημασμένες ακολουθίες, προβλέπονται οι ετικέτες που λείπουν και τα δεδομένα χρησιμοποιούνται για την επανεκπαίδευση του μοντέλου με τις ετικέτες που έχουν προβλεφθεί. Κατόπιν, οι προβλεπόμενες ετικέτες αφαιρούνται και η διαδικασία επαναλαμβάνεται από την πρόβλεψη των ετικετών που λείπουν, μέχρι σύγκλισης των αποτελεσμάτων.

2.3.4.2. ΗΜΜ ΜΕ ΣΥΝΕΙΣΦΟΡΑ ΠΡΟΗΓΟΥΜΕΝΩΝ ΠΑΡΑΤΗΡΗΣΕΩΝ

Πέραν της χρήσης της συνηθισμένης πιθανότητας διανομής εκπομπών, παρέχεται η δυνατότητα λήψης πληροφορίας για n προηγούμενους χαρακτήρες της ακολουθίας, σχηματίζοντας μια ν-οστής τάξης Markov αλυσίδα. Το μείζον πλεονέκτημα της μεθόδου αυτής είναι η ευκολία εφαρμογής, που επιτυγχάνεται μέσω της μεταμόρφωσης της ακολουθίας παρατηρήσεων σε ένα διευρυμένο αλφάβητο. Οι πιθανότητες μετάβασης παραμένουν αυτούσιες, ενώ οι πιθανότητες εκπομπών μεταβάλλονται ανάλογα με την επιλογή encode. [69]

Οι τιμές στην παράμετρο encode που υποστηρίζονται από το JUCHMME για τις πρωτεϊνικές ακολουθίες είναι οι εξής:

1. Κωδικοποίηση με 40 σύμβολα (20x2) που λαμβάνει υπόψιν αν είναι υδρόφοβο (A, F, H, I, L, M, V, W, Y) ή όχι (C, D, E, G, K, N, P, Q, R, S, T) το προηγούμενο κατάλοιπο.
2. Κωδικοποίηση με 80 σύμβολα (20x4) που εκτός από την υδροφοβικότητα ελέγχει αν είναι αρωματικό ή όχι, αλλά και αν είναι πολικό ή φορτισμένο, σε ποια από τις παρακάτω κατηγορίες δηλαδή ανήκει: Hydrophobic–Aromatic (F, H, Y, W), Hydrophobic–non-Aromatic (A, I, L, M, V, G), nonHydrophobic–Charged (D, E, K, R), non-Hydrophobic–Polar (C, N, P, Q, S, T).
3. Κωδικοποίηση με 160 σύμβολα (20x8) που ελέγχει σε ποια από τις παρακάτω 8 κατηγορίες εντάσσεται το αμινοξύ: Υδροφοβικό-Μικρό (A, G), Polar–Special (P, C), Polar–OH (S, T), Polar–NH (N, Q), Φορτισμένο αρνητικά (D, E), Φορτισμένο θετικά (K, R), Υδρόφοβο-Μεγάλο(I, L, M, V) and Υδρόφοβο-Αρωματικό (F, H, Y, W).
4. Κωδικοποίηση με 400 σύμβολα (20x20) που λαμβάνει υπόψη όλους τους πιθανούς συνδυασμούς διπεπτιδίων.

5. Η μεταβλητή ENCODE_TYPE μπορεί να λάβει και την τιμή 0 κατά την οποία ο χρήστης ορίζει την δική του κωδικοποιήση μέσω των παραμέτρων GROUP_SYMBOLS και GROUPING.

2.3.4.3. ΚΡΥΦΑ ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ (HNN)

Πρόκειται για μια μέθοδο της τεχνητής νοημοσύνης που αποτελεί συνδυασμό των μεθόδων Hidden Markov Models και Neural Networks. Το υβριδικό αυτό σύστημα ακολουθεί το πρότυπο των Krogh και Riis. Εν συντομία, οι παράμετροι πιθανοτήτων (emissions) του conditional Hidden Markov Model αντικαθίστανται από την έξοδο των νευρωνικών δικτύων. Τα νευρωνικά δίκτυα, λαμβάνουν ως είσοδο ένα παράθυρο w_i που αντιστοιχεί στην θέση xi και περιλαμβάνει θέσεις γύρω από αυτή. Τα παράθυρα, μπορούν να είναι :

- Συμμετρικά, δηλαδή λαμβάνονται ισόποσες παρατηρήσεις δεξιά(R) και αριστερά(L) του x_i, ενώ το μέγεθος του παραθύρου είναι $R+L+1$, όπου $L>0$, $R>0$.
- Ασύμμετρα, δηλαδή το πλήθος των παρατηρήσεων εκ δεξιών του x_i διαφέρουν από το πλήθος εξ αριστερών αυτού. Το μέγεθος του παραθύρου είναι $R+L+1$, όμως $R \neq L$ και $L>0, R>0$.
- Αριστερό παράθυρο, δηλαδή δεν λαμβάνεται καμία παρατήρηση από τα δεξιά της x_i με μέγεθος $(L+1)$ παρατηρήσεων, όπου $L>0$, $R=0$.
- Δεξί παράθυρο, όπου και λαμβάνονται παρατηρήσεις μόνο από τα δεξιά του x_i, με μέγεθος $(R+1)$ παρατηρήσεων, με $R>0, L=0$.

Η αρχικοποίηση των βαρών γίνεται με τη χρήση της βιβλιοθήκης JOONE (Java Object Oriented Neural Network) (http://www.joone.org/) που χρησιμοποιείται για τη δημιουργία και την εκτέλεση εφαρμογών τεχνητής νοημοσύνης βασιζόμενων σε νευρωνικά δίκτυα.

Το Hidden Neural Network χρησιμοποιεί πρόσθιας τροφοδότησης πολυεπίπεδα νευρωνικά δίκτυα που εκπαιδεύονται με back-propagation αλγόριθμο. Ως δίκτυα πρόσθιας τροφοδότησης (feedforward) ορίζουμε εκείνα στα οποία δεν εμφανίζεται ανατροφοδότηση της εξόδου ενός νευρώνα προς τους νευρώνες από τους οποίους λαμβάνει είσοδο. Δηλαδή η πληροφορία μετακινείται προς μόνο μια κατεύθυνση, από τους νευρώνες εισόδου, σε αυτούς της εξόδου. Ο όρος πολυεπίπεδα, υποδηλώνει την ύπαρξη ενός, τουλάχιστον, κρυφού στρώματος νευρώνων. Εφαρμόζεται πλήρης διασύνδεση μεταξύ των νευρώνων δύο διαδοχικών επιπέδων, χωρίς όμως να συναντώνται διασυνδέσεις που να ανήκουν σε μη διαδοχικά επίπεδα.

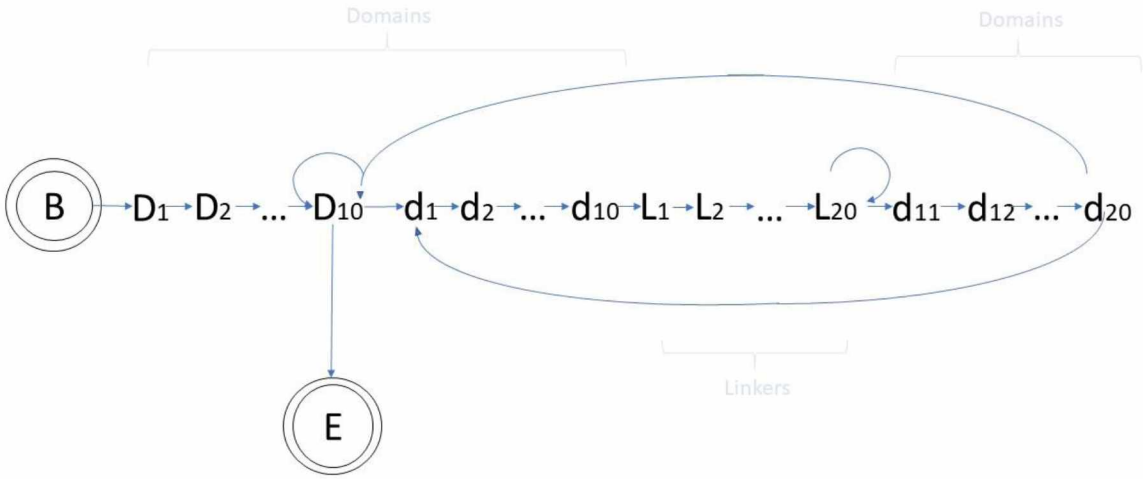
Η αρχιτεκτονική του δικτύου, περιλαμβάνει ένα κρυφό επίπεδο νευρώνων, ένα επίπεδο εξόδου και ένα εισόδου που έχει τόσους νευρώνες όσοι και το μέγεθος του επιλεγόμενου παραθύρου, σύμφωνα με την κωδικοποίηση SPARCE ή άλλη μορφή της κωδικοποίησης BLOSUM.

Ως συνάρτηση ενεργοποίησης του κρυφού επιπέδου νευρώνων, δίνονται οι επιλογές των Sigmoid, modified sigmoid και hyperbolic tangent function, που περιορίζουν το αποτέλεσμα ανάμεσα στις τιμές -1,1. Στο εξωτερικό επίπεδο, η συνάρτηση ενεργοποίησης είναι η σιγμοειδής, που παράγει αποτελέσματα της εμβέλειας (0,1), και συνεπώς μπορεί να απεικονίσει την πιθανότητα εμφάνισης αμινοξέος.

Τέλος, για error functions παρέχονται οι root-mean-square error (RMSE) και Cross Entropy (CE). Το νευρωνικό δίκτυο σταματά τις εποχές εκπαίδευσης όταν ολοκληρώσει ένα ορισμένο πλήθος κύκλων, ή όταν η τιμή του υπολογιζόμενου λάθους είναι μικρότερο από προκαθορισμένο κατώφλι.

2.4. ΜΟΝΤΕΛΟ

Το μοντέλο που χρησιμοποιήθηκε για την πρόβλεψη απεικονίζεται στο παρακάτω σχεδιάγραμμα.



Εικόνα 14. JUCHMME Model for protein linker prediction.

Συγκεκριμένα, έχει 5 ξεχωριστές καταστάσεις B, E, D, d και L. Οι καταστάσεις B και E αναφέρονται σε καταστάσεις έναρξης και λήξης αντίστοιχα, αντικατοπτρίζουν δηλαδή την αρχή και το τέλος της πρωτεϊνικής ακολουθίας.

Στο μοντέλο απεικονίζονται και οι καταστάσεις D,d και L. Τα D,d αντιστοιχούν σε αμινοξέα που ταξινομούνται ως domain ενώ τα L αντιπροσωπεύουν τα αμινοξέα των linker. Τα d, αν και αμινοξέα που κατηγοριοποιούνται ως domain κατάλοιπα, παρατηρήθηκε ότι (καθώς βρίσκονται κοντά στις περιοχές linker) οι συχνότητες αμινοξέων τους διαφέρουν από αυτές των D καταλοίπων.

Έτσι, έγινε εισαγωγή των χαρακτήρων αυτών, 10 αμινοξέα πριν και μετά τους linker, ώστε το μοντέλο εκμεταλλευόμενο αυτή την μεταβολή (κυρίως όταν επρόκειτο για πολλαπλά linker στην ίδια ακολουθία), να δώσει πιο ακριβή αποτελέσματα και καλύτερες προβλέψεις.

Αναλυτικά, ξεκινώντας τα πρώτα 10 αμινοξέα κατηγοριοποιούνται ως domain κατάλοιπα. Η πρακτική αυτή (τα 20 [10D + 10d] πρώτα και τελευταία αμινοξέα να ανήκουν σε domain) υιοθετήθηκε από την βιβλιογραφία, καθώς οι περιοχές linker δεν συναντώνται τόσο κοντά στα άκρα της πολυπεπτιδικής ακολουθίας. Στη συνέχεια, το μοντέλο μπορεί να παραμείνει στην domain κατάσταση μέχρι και το τέλος της ακολουθίας (singledomain proteins). Εάν οι πρωτεΐνες διαθέτουν μία ή περισσότερες περιοχές linker, τότε το μοντέλο εισέρχεται διαδοχικά στις καταστάσεις d1-d10, L1-L20, και εάν το μήκος του linker ξεπερνάει τα 20 αμινοξέα παραμένει στην κατάσταση L-20 μέχρι και το τελευταίο κατάλοιπο του linker. Κατόπιν, διέρχεται από τις καταστάσεις d11-d20 και ακολουθεί η επιστροφή στην κατάσταση D10. Εάν παρατηρείται μόνο μία περιοχή linker στην πολυπεπτιδική ακολουθία, μέχρι το πέρας της ακολουθίας, παραμένει σε αυτήν την κατάσταση. Σε κάθε άλλη περίπτωση, επαναλαμβάνεται η διαδικασία που προαναφέρθηκε, τόσες φορές όσος και ο αριθμός των συνδετικών περιοχών.

Με αυτό τον τρόπο, έγινε εφικτή η δημιουργία ενός μη αυστηρού μοντέλου, που μπορεί να μεταβαίνει από τις καταστάσεις domain σε καταστάσεις linker και το αντίστροφο, όσες φορές και εάν αυτό είναι απαραίτητο.

2.5. ΑΡΧΕΙΑ ΕΙΣΟΔΟΥ

Το JUCHMME για την ορθή λειτουργία του χρειάζεται 5 αρχεία τα οποία περιγράφονται στη συνέχεια.

2.5.1. ΑΡΧΕΙΟ ΑΚΟΛΟΥΘΙΩΝ

Το αρχείο ακολουθιών παίρνει τον τίτλο του από τη μορφή του. Πρόκειται για ένα αρχείο που διαθέτει 3 γραμμές για κάθε υπό εξέταση πρωτεΐνη. Στην πρώτη, αναγράφεται το όνομα της πρωτεΐνης, στη δεύτερη η αμινοξική της ακολουθία, ενώ στην τρίτη, καταγράφεται η κατηγοριοποίηση των αμινοξέων της ακολουθίας σε D ή L ανάλογα με το αν το αμινοξύ βρίσκεται ή όχι, ανάμεσα στα όρια domain. Πρόκειται για μέθοδο επιβλεπόμενης μάθησης, και γι' αυτό χρίζεται αναγκαίο να υπάρχουν και οι ετικέτες των αμινοξέων στο αρχείο εισόδου.

2.5.2. MODEL OPTIONS

Στο πρώτο σκέλος παρατηρούμε ότι ορίστηκε το αλφάβητο των παρατηρήσεων (ESYM) ως ARNDCQEGHILKMFPSTWYV, που συμβολίζει τα 20 αμινοξέα. Το OSYM ορίστηκε ως DdLBE όσες δηλαδή είναι και οι καταστάσεις που μπορεί να υπάρξουν στο μοντέλο (δεσμευμένες καταστάσεις). Το PSYM αφορά το αλφάβητο των ετικετών. Τόσο η κατάσταση d όσο και η D αναφέρονται στην κατηγοριοποίηση σε domain κατάλοιπα, και για το λόγο αυτό το PSYM ορίζεται ως DLBE.

```
# MODEL OPTIONS
MODEL=PREDLINKER
ESYM=ARNDCQEGHILKMFPSTWYV
OSYM=DdLBE
PSYM=DLBE
```

Ακολούθως, ορίζονται οι ετικέτες του μοντέλου (όπου το L αντιστοιχεί σε linker, και το D σε domain).

```
#Model Unique Labels
transmLabels=L
inLabels=D
outLabels=D
```

Συνεχίζοντας παρατηρούνται οι εξής μεταβλητές:

- STATE όπου παρουσιάζεται η διαδοχή των καταστάσεων του μοντέλου γραμμικά
- OSTATE όπου αναπαρίστανται οι «δεσμευμένες» καταστάσεις, αλλά και
- PSTATE όπου παρατηρούνται οι ετικέτες των στοιχείων της OSTATE.

```
#Model states and labels
STATE=B00 D01 D02 D03  D04 D05 D06  D07 D08 D09 D10 d01 d02 d03 d04 d05 d06
d07 d08 d09 d10 L01 L02 L03 L04 L05 L06 L07 L08 L09 L10 L11 L12 L13 L14 L15
L16 L17 L18 L19 L20 d11 d12 d13 d14 d15 d16 d17 d18 d19 d20 E00
OSTATE=B D D D D D D D D D D d d d d d d d d d d L L L L L L L L L L L L L
L L L L L d d d d d d d d d d E
PSTATE=B D D D D D D D D D D D D D D D D D D D D L L L L L L L L L L L L L
L L L L L D D D D D D D D D D E

#MODEL PRIOR for every esym
PRIOR = 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01
0.01 0.01 0.01 0.01 0.01 0.01
# Distribution for each osym
# Each column must have a sum equal to 1
# osym D d L B E
PRIOR1 = 0.99 0.98 0.99 0.0 0.0
PRIOR2 = 0.0 0.0 0.0 0.0 0.0
PRIOR3 = 0.01 0.02 0.01 0.0 0.0
```

2.5.3. ΠΙΝΑΚΑΣ ΜΕΤΑΒΑΣΕΩΝ (TRANSITION TABLE)

Ο πίνακας μεταβάσεων αντιπροσωπεύει την πιθανότητα μετάβασης από μια κατάσταση σε κάποια επόμενη κατάσταση. Συγκεκριμένα, όταν από έναν κόμβο(κατάσταση) εξέρχεται μόνο μία ακμή, η πιθανότητα μετάβασης από την προηγούμενη στην τωρινή είναι 100%. Όταν από έναν κόμβο προκύπτουν δύο ακμές, η πιθανότητα μετάβασης σε αυτές θα είναι 50% για την κάθε μία, όταν προκύπτουν 3 ακμές, η πιθανότητα θα είναι 33.3% κ.ο.κ.

Στο συγκεκριμένο μοντέλο, ο μέγιστος αριθμός εξερχόμενων ακμών είναι 3 (κατάσταση D10) οπότε και οι πιθανότητες μετάβασης κυμαίνονται από 33%-100%.

2.5.4. ΠΙΝΑΚΑΣ ΕΚΠΟΜΠΩΝ (EMISSION TABLE)

Ο πίνακας εκπομπών ορίζει τις emission πιθανότητες για τις καταστάσεις στο HMM. Ενδελεχώς, ο πίνακας εκπομπών έχει πλήθος στηλών ίσο με το μέγεθος του αλφάβητου των παρατηρήσεων και πλήθος γραμμών όσες και οι καταστάσεις του μοντέλου επί το μέγεθος του αλφαβήτου. Συνεπώς, για το συγκεκριμένο μοντέλο, ο πίνακας εκπομπών διαθέτει 20 στήλες και 62 γραμμές(20 Domain + 20 domain + 20 Linker + B + E).

Ως στοιχεία του πίνακα, έχουν τοποθετηθεί τα εκάστοτε ποσοστά εμφάνισης αμινοξέων στην αντίστοιχη κατηγορία. Καθώς η πρώτη και η τελευταία γραμμή αντιστοιχούν στις καταστάσεις Begin και End αντίστοιχα, έχουν όλα τα στοιχεία τους ίσα με το 0. Οι πρώτες 10 γραμμές, που έχουν μη μηδενικές τιμές, θα περιλαμβάνουν τα ποσοστά που αφορούν τις Domain καταστάσεις, οι 10 επόμενες τα ποσοστά για τις ενδιάμεσες domain καταστάσεις, κ.ο.κ.

total Frequencies	Domain	10 before/after linker	Linker	Amino acid
8.16	8.18	8.04	7.98	A
4.95	4.93	5.2	5.22	R
4.47	4.5	4.41	4.18	N
5.73	5.73	5.41	5.75	D
1.41	1.4	1.71	1.58	C
3.88	3.87	4.11	4.05	Q
2.35	2.36	2.07	2.15	H
5.48	5.52	5.54	5.16	I
8.56	8.5	8.88	9.11	L
6.08	6.04	6.49	6.58	K
2.28	2.32	2.12	1.95	M
3.97	3.98	3.71	3.89	F
4.86	4.76	4.42	5.93	P
6.05	6.04	5.89	6.08	S
5.66	5.67	5.65	5.56	T
1.4	1.43	1.43	1.11	W
3.5	3.56	3.13	2.95	Y
6.96	6.99	7	6.66	V
6.61	6.56	6.86	7.06	E
7.62	7.68	7.93	7.05	G

Εικόνα 15. Συχνότητες εμφάνισης αμινοξέων σε όλο το μήκος της πρωτεϊνικής ακολουθίας, στις περιοχές domain, στις ενδιάμεσες περιοχές ±10 καταλοίπων από τα όρια των linker και στις περιοχές των linker αντίστοιχα.

2.5.5. CONFIGURATION FILE

Το configuration file καθορίζει τον τρόπο με τον οποίο θα επεξεργαστούν τα δεδομένα. Πρόκειται για μια σειρά παραμέτρων οι οποίες λαμβάνουν δυαδική τιμή (true/false) ή μία εκ των δεδομένων τιμών που αναγράφονται σε σχόλια πάνω από αυτές. Το αρχείο αυτό, παρέχει τη δυνατότητα στον χρήστη, να προσαρμόσει την εκπαίδευση, την κωδικοποίηση και την μέθοδο αξιολόγησης των αποτελεσμάτων σύμφωνα με τις ανάγκες του βιολογικού προβλήματος που επιθυμεί να διερευνήσει, χωρίς την απαίτηση εξοικείωσής του με κώδικα.

2.6. ΒΑΣΙΚΟΙ ΠΑΡΑΜΕΤΡΟΙ CONFIGURATION FILE

Αρχικά, στην πρώτη ομάδα (TRAINING OPTIONS) επιλέγεται ο αλγόριθμος της εκπαίδευσης ανάμεσα από τους Maximal Likelihood, Conditional Maximum Likelihood, αλλά και των κρυφών νευρωνικών δικτύων.

```
# TRAINING OPTIONS
RUN_CML=false
RUN_GRADIENT=false
HNN=false
ALLOW_BEGIN=true
ALLOW_END=true
RUN_ViterbiTraining=false
threshold=0.02
maxIter=200
```

```
# TRAINING OPTIONS
RUN_CML=true
RUN_GRADIENT=true
HNN=false
ALLOW_BEGIN=true
ALLOW_END=true
RUN_ViterbiTraining=false
threshold=0.02
maxIter=200
```

Εικόνα 16. Οι παράμετροι που αφορούν τα κριτήρια εκπαίδευσης. Όταν στις μεταβλητές RUN_CML,RUN_GRADIENT ανατίθεται false, γίνεται χρήση του κριτηρίου Maximal Likelihood(εικόνα στα αριστερά). Στη δεξιά εικόνα εμφανίζονται οι τιμές παραμέτρων για εκπαίδευση με CML.

Η παράμετρος RUN_ViterbiTraining αναφέρεται στην χρήση της μεθόδου Viterbi, ενώ οι μεταβλητές threshold και maxIter, αφορούν τον τερματισμό του αλγορίθμου εκπαίδευσης, όταν αυτός επιτύχει καλύτερο σκόρ από το ορισμένο κατώφλι ή όταν ολοκληρώσει τον μέγιστο αριθμό επαναλήψεων, αντίστοιχα. Η μέγιστη τιμή που μπορεί να λάβει η μεταβλητή maxIter είναι 200 επαναλήψεις, ενώ το προτεινόμενο όριο επαναλήψεων για την εκπαίδευση με κρυφά νευρωνικά δίκτυα είναι 50, καθώς όντας μέθοδος με μεγαλύτερη ευαισθησία, συγκλίνει συντομότερα η log likelihood.

Συνεχίζοντας, στο τμήμα (PROBABILITIES) ορίζεται ο τρόπος που θα αντληθούν οι πιθανότητες, οι μεταβάσεις και οι εκπομπές του μοντέλου, με επιλογές να είναι είτε μέσω αρχείων, είτε να γίνει υπολογισμός τους σύμφωνα με έναν από τους ήδη διαθέσιμους αλγορίθμους (RANDOM, UNIFORM, VITERBI).

Υπάρχει η δυνατότητα ρύθμισης της παραλληλοποίησης καθώς και πόσοι πυρήνες θα αξιοποιηθούν για τους υπολογισμούς.

Στο επόμενο τμήμα που μας ενδιαφέρει (EXTENDED PAST OBSERVATIONS), ορίζεται το εάν θα χρησιμοποιηθεί η πληροφορία από το προηγούμενο κατάλοιπο (PAST_OBS_EXTENSION=true) καθώς και τον τρόπο με τον οποίο αυτό θα επηρεάσει την πρόβλεψη (ENCODE_TYPE). Οι επιλογές της μεταβλητής encode επεξηγούνται στην αντίστοιχη ενότητα (βλ. κεφ. 2.3.4).

Σημαντικά επίσης είναι τα Refine OPTIONS που καθορίζουν την ύπαρξη καθώς και το μέγεθος της flanking region σε κατάλοιπα προς κάθε κατεύθυνση. Στην ενότητα DECODING OPTIONS μπορούν να επιλεγούν κάποιοι ή και όλοι οι αλγόριθμοι αποκωδικοποίησης σύμφωνα με τους οποίους θα προκύψουν οι μετρικές αξιοπιστίας.

```
# DECODING OPTIONS
VITERBI=true
NBEST=false
DYNAMIC=false
POSVIT=true
PLP=true
CONSTRAINT=false
```

Εικόνα 17. Επιλογές αποκωδικοποίησης. Χρήση όσων ανατίθενται ως true, απενεργοποίηση των υπολοίπων. Στη δεδομένη περίπτωση γίνεται χρήση των αλγορίθμων Viterbi,POSVIT και PLP.

Στην ενότητα PRIOR OPTIONS καθορίζεται η εκ των προτέρων πιθανότητα που θα προστεθεί στις μεταβάσεις ή στις εκπομπές. Η μεταβλητή στην PRIOR_TRANS λαμβάνει τιμές εύρους [0:1]. Παρά το γεγονός ότι η προεπιλεγμένη πιθανότητα, είναι 0.001, στην παρούσα εργασία αποσκοπούμε να βρεθούν οι καλύτερες παράμετροι που βελτιστοποιούν την απόδοση του μοντέλου και για το λόγο αυτό δοκιμάστηκε ένα εύρος prior πιθανοτήτων.

2.6.2. HIDDEN NEURAl NETWORKS & ΠΑΡΑΜΕΤΡΟΙ CONFIGURATION FILE

Στο τμήμα GRADIENT DESCENT OPTIONS περιλαμβάνεται η επιλογή αλγορίθμου για προσαρμογή του ρυθμού μάθησης, ανάμεσα από τους SILVA και RPROP. Γενικά, προτείνεται ο δεύτερος καθώς δείχνει να αποδίδει καλύτερα. Δίνεται επίσης η επιλογή χρήσης του απλού gradient descent αλγορίθμου. Οι υπόλοιπες μεταβλητές ορίζουν τον ρυθμό μάθησης για τις πιθανότητες εκπομπής και μετάβασης, προσθήκη momentum για την αποφυγή τοπικών ελαχίστων, αλλά και τις μέγιστες και ελάχιστες τιμές που μπορούν οι ρυθμοί να αποκτήσουν.

Η ενότητα HNN OPTIONS περιλαμβάνει ρυθμίσεις βασικών παραμέτρων μια από τις οποίες είναι το μέγεθος του παραθύρου. Συγκεκριμένα, οι μεταβλητές 'windowLeft' και 'windowRight', ορίζουν αντίστοιχα, το πλήθος καταλοίπων δεξιά και αριστερά του καταλοίπου ενδιαφέροντος προσδιορισμού. Καθορίζεται επίσης το πλήθος των κρυφών νευρώνων, το πόσο θα φθίνουν τα βάρη των νευρώνων προς το 0 αλλά και η συνάρτηση του κρυφού στρώματος που θα χρησιμοποιηθεί μέσω της τιμής που αποδίδεται στις μεταβλητές 'nhidden', 'DECAY' και 'hiddenLayerFunction'.

Μέσω της BOOTSTRAP OPTIONS ρυθμίζονται μεταβλητές που σχετίζονται με την λειτουργία του αλγορίθμου, σύμφωνα με τον οποίο με τυχαία δειγματοληψία των δεδομένων, δημιουργεί καινούργια υποσύνολα, τα οποία σταθμίζονται για να δώσουν τις τελικές εκτιμήσεις.

Τέλος η παράγραφος RPROPNN OPTIONS αναφέρεται στον ομώνυμο αλγόριθμο βελτιστοποίησης και ρυθμίζει παραμέτρους όπως τον καθορισμό του ρυθμού εκπαίδευσης, τον ρυθμό αύξησης ή μείωσης των βαρών εκπαίδευσης, το πλήθος των επαναλήψεων που θα πραγματοποιηθεί μέχρι τον τερματισμό του αλγόριθμου εκπαίδευσης αλλά και το αν θα χρησιμοποιηθεί η μέθοδος cross-validation ή όχι.

Αξίζει να αναφερθεί πως παρέχονται μεταβλητές και για τη ρύθμιση των ακόλουθων:

- EARLY STOPPING OPTIONS: Η μέθοδος πρόωρου τερματισμού εφαρμόζεται για να εμποδίσει το μοντέλο να υπερπροσαρμοστεί στα δεδομένα εκπαίδευσης, διακόπτοντας την εκπαίδευση, όταν αυτή υπερβαίνει κάποιο από τα όρια που έχουν τεθεί.
- SEMI-SUPERVISED LEARNING OPTIONS: Πρόκειται για μεταβλητές που αφορούν την μερικώς επιβλεπόμενη μάθηση.
- DYNAMIC OPTIONS: Οι παράμετροι αυτές ρυθμίζουν τον αλγόριθμο δυναμικού προγραμματισμού που δημιουργεί τους περιορισμούς για την πρόβλεψη της κωδικοποίησης Posterior. Όμως θεωρείται απαρχαιωμένος και δεν χρησιμοποιείται.

2.7. ΔΙΑΔΙΚΑΣΙΑ ΑΠΟΔΟΣΗΣ ΑΞΙΟΠΙΣΤΙΑΣ ΜΟΝΤΕΛΟΥ

2.7.1. SELF-CONSISTENCY

Για το self-consistency test το ίδιο dataset χρησιμοποιείται τόσο για την εκπαίδευση, όσο και για την επικύρωση των αποτελεσμάτων. Ωστόσο, είναι πιθανό να δώσει μεροληπτικά αποτελέσματα γιατί ελέγχεται στα δεδομένα που έχει εκπαιδευτεί.

2.7.2. K-FOLD CROSS VALIDATION

Η μέθοδος Cross-validation είναι μία από τις ευρέως γνωστές για την αποτίμηση της ικανότητας γενίκευσης του μοντέλου πρόβλεψης και για να αποφευχθεί το φαινόμενο της υπερπροσαρμογής (overfitting). Συνεπώς, χρησιμοποιείται για την εκτίμηση της απόδοσης του τελικού μοντέλου στα νέα δεδομένα.

Σύμφωνα με αυτό, το σύνολο δεδομένων χωρίζεται σε υποκατηγορίες k δειγμάτων. Το μοντέλο εκπαιδεύεται στα v-k δείγματα και ελέγχει τα αποτελέσματα της πρόβλεψης στην ομάδα δειγμάτων που αποκλείστηκε από την εκπαίδευση. Η διαδικασία επαναλαμβάνεται για όλες τις ομάδες k δειγμάτων που προκύπτουν ενώ οι μετρικές του μοντέλου υπολογίζονται ως ο μέσος όρος των μετρήσεων του κάθε ελέγχου. Στην συγκεκριμένη περίπτωση, εφαρμόζεται cross-validation με k=10.

2.7.3. JACKKNIFE

Πρόκειται για μία μέθοδο δειγματοληψίας που χρησιμοποιείται για τον υπολογισμό της διακύμανσης και της μεροληψίας σε έναν μεγάλο πληθυσμό. Αποτελεί παραλλαγή του cross-validation, δίνει καλύτερα αποτελέσματα όταν δεν υπάρχουν μεγάλες μεταβολές μεταξύ των παρατηρήσεων αλλά είναι υπολογιστικά και χρονικά πιο απαιτητική.

Η μέθοδος δειγματοληψίας απαιτεί τον αποκλεισμό από το κάθε υποσύνολο δεδομένων μιας παρατήρησης. Κατ' αυτόν τον τρόπο, για ένα σύνολο δεδομένων n δειγμάτων, δημιουργούνται n υποσύνολα, καθένα από τα οποία περιλαμβάνει n-1 δείγματα. Έτσι, ο αναλυτής θα λύσει το μοντέλο n φορές επιτρέποντάς του να εκτιμήσει τόσο τις τιμές παραμέτρων, όσο και το αντίστοιχο standard error, ενώ οι τελικές μετρικές προκύπτουν με τον μέσο όρο των αποτελεσμάτων.

2.8. ΜΕΤΡΑ ΑΠΟΤΙΜΗΣΗΣ ΑΞΙΟΠΙΣΤΙΑΣ ΜΟΝΤΕΛΟΥ

Προκειμένου να διαπιστώσουμε την αξιοπιστία και την εγκυρότητα των αποτελεσμάτων πρόβλεψης υπολογίζονται ορισμένες μετρικές, που χρησιμοποιούνται κατά κόρον, σε τέτοιου είδους προβλήματα. Οι μετρικές αυτές υπολογίζονται με το τρέξιμο του JUCHMME, και καταγράφονται στις τελευταίες σειρές των αποτελεσμάτων, ενώ υπολογίζονται για κάθε αλγόριθμο κωδικοποίησης που έχει επιλεχθεί (Viterbi, PLP, POSVIT, κ.α.).

Έτσι, στα αρχικά TP αντιστοιχούν τα True Positive (Αληθώς θετικά) αμινοξέα, δηλαδή τα κατάλοιπα που έχουν προβλεφθεί ως θετικά, και η πρόβλεψη είναι σωστή. Με τα αρχικά TN, αναφερόμαστε στα True Negative (Αληθώς αρνητικά) κατάλοιπα, δηλαδή τα κατάλοιπα που έχουν προβλεφθεί ορθώς ως αρνητικά. Ακόμα, εμφανίζονται τα αρχικά FP (False Positives) που αναφέρονται στο πλήθος καταλοίπων που εσφαλμένα έχουν κατηγοριοποιηθεί ως θετικά. Στη δεδομένη εφαρμογή, οι παρατηρήσεις δεν αντιστοιχούν σε μεμονωμένα κατάλοιπα αλλά σε ολόκληρες πρωτεΐνες, και στο εάν αυτές έχουν προβλεφθεί σωστά ή όχι. Η μετρική που υποδεικνύει το ποσοστό των καταλοίπων που έχουν προβλεφθεί σωστά είναι το Q, όπου:

$$Q = \frac{TP+TN}{TP+TN+FP+FN} * 100\%$$

Επίσης, υπολογίζεται ο συντελεστής Matthews (MCC), που λαμβάνει τιμές μεταξύ των (-1, 1) όπου το 0 υποδεικνύει τυχαία πρόγνωση. Για προβλήματα πολλών κλάσεων (k), εφαρμόζουμε τον δείκτη Q, αλλά ο C, πρέπει να εφαρμοστεί για κάθε ομάδα, αγνοώντας τις υπόλοιπες. Σαν πιο αξιόπιστη λύση, χρησιμοποιείται ο δείκτης SOV (segment's overlap measure), για αλγορίθμους πρόγνωσης δευτεροταγούς δομής, με τιμές συνεχείς στο διάστημα 0-1.

Κάποια μέτρα που χρησιμοποιούνται επίσης, τα οποία δεν προκύπτουν άμεσα με το τρέξιμο του μοντέλου, αλλά έμμεσα, από τις μετρικές που καταγράφονται είναι η ειδικότητα (specificity) και η ευαισθησία (sensitivity). Οι μετρικές αυτές ορίζονται ως εξής:

$$\text{Specificity} = \frac{TP}{TP+FP},$$

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

ΚΕΦΑΛΑΙΟ 3 ^{Ο}: ΑΠΟΤΕΛΕΣΜΑΤΑ

3.1. ΑΠΟΤΕΛΕΣΜΑΤΑ

Τα αποτελέσματα προέκυψαν μέσω του JUCHMME από την εφαρμογή τριών διαφορετικών test:

(1) Cross – validation test

(2) Self - consistency test

(3) Jacknife Test

Τα κριτήρια εκπαίδευσης που χρησιμοποιήθηκαν είναι τα Maximal Likelihood (ML), Conditional Maximum Likelihood (CML) και Hidden Neural Networks. Το μοντέλο που χρησιμοποιήθηκε για την εκπαίδευση σύμφωνα με τις παραμέτρους που αναγράφονται στο configuration file αποκωδικοποιήθηκε με τους αλγορίθμους αποκωδικοποίησης Viterbi, Posterior-Viterbi (POSVIT) και Optimal accuracy posterior decoder (βλ. 2.3.1) (PLP). Οι εντολές οι οποίες χρησιμοποιήθηκαν, για την κλήση του προγράμματος από τη γραμμή εντολών αναγράφονται ακολούθως:

Cross – validation test

- `java hmm/Juchmme -a ../tables/A_LINKER.txt -e ../tables/E_LINKER.txt -c ../conf/conf.tmbb -m ../models/linker.mdel -t ../input/3linefile.txt -v 10 > linkersCross.txt`

Self - consistency test

- `java hmm/Juchmme -a ../tables/A_LINKER.txt -e ../tables/E_LINKER.txt -c ../conf/conf.tmbbCML -m ../models/linker.mdel -t ../input/3linefile.txt -s > linkersSelf.txt`

Jacknife Test

- `java hmm/Juchmme -a ../tables/A_LINKER.txt -e ../tables/E_LINKER.txt -c ../conf/conf.tmbb -m ../models/linker.MDEL -t ../input/3linefile.txt -j > linkersJack.txt`

Επεξήγηση παραμέτρων

- Αρχικά ο χαρακτήρας «>» υποδηλώνει ανακατεύθυνση και υλοποιεί την εκτύπωση των αποτελεσμάτων σε αρχείο, που έχει τίτλο ό,τι βρίσκεται δεξιότερα αυτού. Το αρχείο εάν δεν υπάρχει ήδη, δημιουργείται, ενώ αν έχει προηγούμενα περιεχόμενα, αυτά χάνονται. Εάν παραληφθεί ο συγκεκριμένος χαρακτήρας και ό,τι έπεται, το πρόγραμμα θα εκτελεστεί κανονικά και τα αποτελέσματα θα εμφανιστούν στην γραμμή εντολών.
- Έλλειψη σε οποιοδήποτε άλλο στοιχείο της εντολής, θα οδηγήσει σε αδυναμία τρεξίματος και την εμφάνιση μηνύματος λάθους στη γραμμή εντολών. Παρατηρούμε ότι και οι 3 εντολές έχουν παρόμοια δομή με μόνη βασική διαφορά το τελευταίο όρισμα (`v, s, j`) που καθορίζει και το τεστ που θα εφαρμοστεί κάθε φορά. Στο cross validation test αναγράφεται επίσης και το μέγεθος των ομάδων που θα χωριστεί το συνολικό dataset. Δηλαδή για n=10 δημιουργούνται ομάδες των 10 ακολουθιών η κάθεμια, εκτός της τελευταίας που θα περιέχει τόσες όσο το πηλίκο της διαίρεσης του πλήθους των ακολουθιών με το n (4).
- Τα πρώτα δύο ορίσματα και `../tables/E_LINKER.txt`) παρέχουν το σχετικό μονοπάτι για τα αρχεία του πίνακα μεταβάσεων και εκπομπών αντίστοιχα. Η τρίτη παράμετρος `../conf/conf.tmbbCML`) περιέχει το σχετικό μονοπάτι για τα configuration files. Ο μόνος λόγος το cml test έχει διαφορετικό conf αρχείο, είναι γιατί αλλάζουν οι παράμετροι που καθορίζουν τα κριτήρια εκπαίδευσης, ωστόσο, με τις κατάλληλες μετατροπές, μπορεί να χρησιμοποιηθεί και το αρχείο conf.tmbb.
- Τα επόμενα ορίσματα, (`-m ../models/linker.MDEL, -t ../input/3linefile.txt`) αναφέρονται στα μονοπάτια για το μοντέλο της πρόβλεψης και το αρχείο εισόδου, που περιέχει τις πρωτεϊνικές ακολουθίες.

Οι εντολές για τα νευρωνικά δίκτυα είναι οι εξής:

Cross – validation test

- `java hmm/Juchmme -a ../tables/A_LINKER.txt -x ../tables/SPARCE -c ../conf/conf.tmbbHNN -m ../models/linker.mdel -t ../input/3linefile.txt -v 10 > linkersCrossNN.txt`

Self - consistency test

- `java hmm/Juchmme -a ../tables/A_LINKER.txt -x ../tables/SPARCE -c ../conf/conf.tmbbHNN -m ../models/linker.mdel -t ../input/3linefile.txt -s > linkersSelfNN.txt`

Jacknife Test

- `java hmm/Juchmme -a ../tables/A_LINKER.txt -x ../tables/SPARCE -c ../conf/conf.tmbbHNN -m ../models/linker.mdel -t ../input/3linefile.txt -j > linkersJackNN.txt`

  - Παρατηρούμε πως εκ πρώτης όψεως, οι εντολές δεν διαφέρουν σημαντικά με αυτές που παρατέθηκαν παραπάνω. Η μόνη διαφορά που διαφαίνεται είναι το δεύτερο όρισμα όπου πλέον δεν εισάγεται το μονοπάτι του πίνακα εκπομπών, όπως παραπάνω αλλά αυτό του πίνακα SPARCE. Κάτι τέτοιο είναι λογικό και αναμενόμενο, εφόσον όπως αναφέρθηκε και προηγουμένως στην σχετική ενότητα (βλ.), οι πιθανότητες εκπομπών υπολογίζονται από το νευρωνικό δίκτυο.

Αντιθέτως, αυτό που καθορίζει την λειτουργία του προγράμματος είναι οι μεταβολές στο conf αρχείο που χρησιμοποιείται. Για την αποσαφήνιση των αλλαγών, θα παρατεθεί το περιεχόμενο του αρχείου `conf.tmbbHNN` Σε αυτό το σημείο κρίνεται σκόπιμο να αναφερθεί, πως το ίδιο αποτέλεσμα θα προκύψει απ' όλα τα conf αρχεία εάν στις αντίστοιχες θέσεις, τοποθετηθούν οι τιμές που εμφανίζονται σε αυτό. Το αρχείο `conf.tmbbHNN` συνεπώς, φτιάχτηκε χάριν ευκολίας και η δομή του δεν διαφέρει από των υπολοίπων αρχείων διαμόρφωσης.

```
# TRAINING OPTIONS
RUN_CML=true
RUN_GRADIENT=true
HNN=true
ALLOW_BEGIN=true
ALLOW_END=true
```

```
RUN_ViterbiTraining=false
threshold=0.02
maxIter=200
#PROBABILITIES
#FILE, RANDOM, UNIFORM, VITERBI
TRANSITIONS=FILE
#FILE, RANDOM, UNIFORM, VITERBI
EMISSIONS=FILE
#FILE, RANDOM_NORMAL, RANDOM_UNIFORM, RPROP, BOOT
WEIGHTS=RPROP

# Multithreaded parallelization for multicores
PARALLEL=false
defCPU=true
nCPU=10

#SEMI-SUPERVISED LEARNING OPTIONS
SSL_ENABLED=false
# SSL (standard Semi-supervised Method) or GEM (Generalized EM)
SSL_METHOD=SSL
#1: Use all, 2: Use weight (Constant) for each sequence, 3: Use weight (Reliability) for each sequence, 4: Use a few most confident
SSL_ADD_METHOD=4
#1:VITERBI, 2:NBEST, 3:POSVIT, 4:PLP
SSL_USING_METHOD=4
SSL_THRESHOLD=0.000002
SSL_maxIter=200
SSL_relscore= 0.95
SSL_WEIGHT=0.2

#EXTENDED PAST OBSERVATIONS
PAST_OBS_EXTENSION=false
#1=40, 2=80, 3=160, 4=400, 0=Your Encoding
ENCODE_TYPE=1
GROUP_SYMBOLS = 10
GROUPING=10001011011000000111
PAST_OBS_NO = 1

#DYNAMIC OPTIONS
MINHLEN=7
MINLLEN=1
```

MAXHLEN=17

MAXNSTRAND=32

MINSSC=3

STRDIV=9

#Refine OPTIONS

FLANK=3

REFINE=true

ML_INIT=false

DECODING OPTIONS

VITERBI=true

NBEST=false

DYNAMIC=false

POSVIT=false

PLP=true

CONSTRAINT=false

#EARLY STOPPING OPTIONS

EARLY=false

CUSTOM_STOP=0.0

NTRAIN=15

NROUND=5

ITER=2

#GRADIENT DESCENT OPTIONS

JACOBI=false

RPROP=true

SILVA=true

momentum=0.0

kappaA=0.01

kappaAmin=1E-20

kappaAmax=1.0

kappaE=0.01

kappaEmin=1E-20

kappaEmax=1.0

NPLUS=1.2

NMINUS=0.5

ni=50.0

#PRIOR OPTIONS

NOISE_TR=true

NOISE_EM=true

PRIOR_TRANS=0.001

#HNN OPTIONS

windowLeft=5

windowRight=5

nhidden=9

ADD_GRAD=0.0

DECAY=0.001

#1: Sigmoid, 2: Sigmoid Modified, 3: Tanh

hiddenLayerFunction=2

#BOOTSTRAP OPTIONS (HNN ONLY)

BOOT=0

STDEV=1.5

RANGE=5.0

SEED=568381

WEIGHT_RAND=0.0

WEIGHT_TIME=false

#RPROPNN OPTIONS (HNN ONLY)

numberOfCycles=50

doCrossVal=true

crossVallter=5

minGEdiff=0

#RMSE, CE: Cross Entropy

globalError=CE

initialDelta=0.1

maxDelta=50.0

minDelta=1e-6

etaInc=1.2

etaDec=0.5

#RANDOM SEQUENCE UTILITY

- Στο σημείο αυτό, πρέπει να επισημανθεί πως για την σωστή ρύθμιση των παραμέτρων για HNN πρέπει να οριστούν ως true όλες οι παράμετροι RUN_GRADIENT, RUN_CML και HNN αλλά και το μέγιστο πλήθος επαναλήψεων (maxIter) να οριστεί ως 50.

Το μοντέλο δοκιμάστηκε με διαφορετικούς συνδυασμούς παραμέτρων. Παρατίθεται παρακάτω ποιες παράμετροι μεταβλήθηκαν, αλλά και το πως επηρεάζει η εκάστοτε παράμετρος τον υπολογισμό των αποτελεσμάτων.

1. PAST_OBS_EXTENSION=false

   Αυτή η μεταβλητή για τιμή ίση με false ορίζει πως το τρέχον παρατηρούμενο σύμβολο (αμινοξικό κατάλοιπο) θα επηρεάζεται μόνο από την τωρινή κατάσταση (ανεξάρτητο από τις προηγούμενες παρατηρήσεις). Στην περίπτωση αυτή, εφαρμόζεται ο κλασσικός αλγόριθμος Class Hidden Markov Model.

2. PAST_OBS_EXTENSION=true

   Όταν η τιμή ισούται με true, εφαρμόζεται μια παραλλαγή του αλγορίθμου Class Hidden Markov Model. Σύμφωνα με αυτή, το τρέχον παρατηρούμενο αμινοξικό κατάλοιπο, εξαρτάται τόσο από την τρέχουσα κατάσταση, όσο και από μια σειρά προηγούμενων παρατηρούμενων συμβόλων, μεταμορφώνοντας την ακολουθία των παρατηρήσεων.

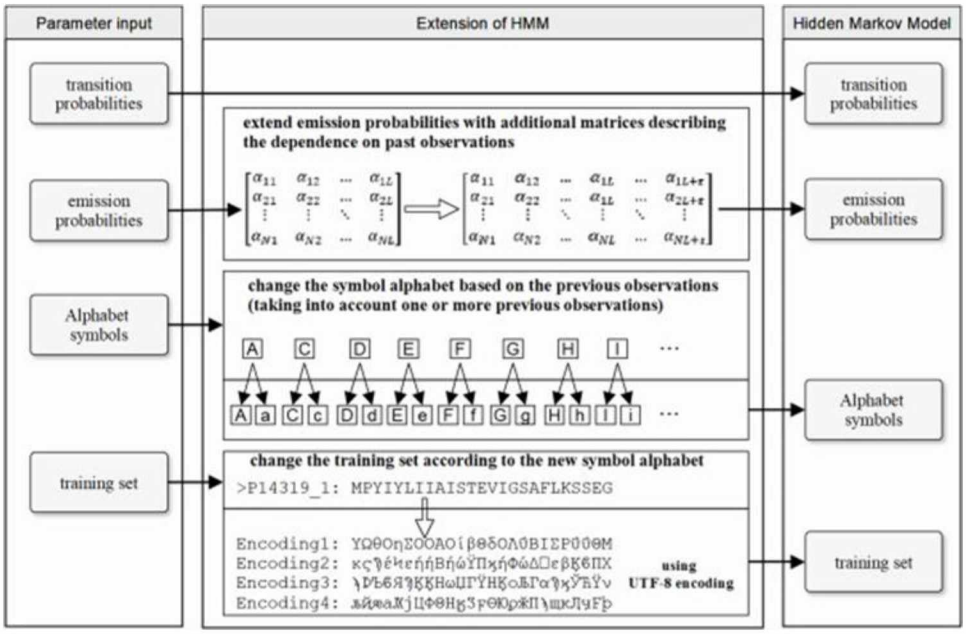
   Η καινούργια πληροφορία αυτή, ενσωματώνεται στους υπολογισμούς, μέσω μεταβολών του αλφαβήτου και των πιθανοτήτων emissions. Συγκεκριμένα, στο αρχικό αλφάβητο της εργασίας αυτής, κάθε σύμβολο εκ των 20 που αναγράφονται, κωδικοποιεί ένα αμινοξύ.

   A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y.

   Το καινούργιο αλφάβητο θα αποτελείται από τόσα σύμβολα όσα είναι το πλήθος του αρχικού αλφαβήτου επί του πλήθους των φυσικοχημικών ιδιοτήτων των αμινοξέων που εξετάζονται (βλ. 2.3.4.2) ενώ οι διαστάσεις του πίνακα εκπομπών μεταβάλλονται σε N * s_n, (όπου το N εκφράζει το πλήθος των καταστάσεων του μοντέλου και s_n εκφράζει το μέγεθος του εμπλουτισμένου αλφαβήτου). Αντιθέτως, ο πίνακας μεταβάσεων παραμένει αμετάβλητος.

   Για την υλοποίηση της νέας μεθόδου κωδικοποίησης, γίνεται αυτόματη μετατροπή των αρχικών συμβόλων, σύμφωνα με το «standard character encoding Unicode», μέσω του UTF-8. Το UTF-8 ευνοεί την χρήση λιγότερης μνήμης και μπορεί να χρησιμοποιηθεί από λειτουργικά συστήματα και από γλώσσες προγραμματισμού. Ξεκινώντας από το λατινικό κεφαλαίο γράμμα A, η μέθοδος που προτάθηκε χρησιμοποιεί σειριακά συνεχείς Unicode χαρακτήρες για να δημιουργήσει ένα νέο αλφάβητο.

Με την καινούργια κωδικοποίηση, το τρέχον παρατηρούμενο σύμβολο μιας ακολουθίας συσχετίζεται με ένα πλήθος πρότερα παρατηρούμενων συμβόλων, αυξάνοντας έτσι την αξιοπιστία της πρόβλεψης.

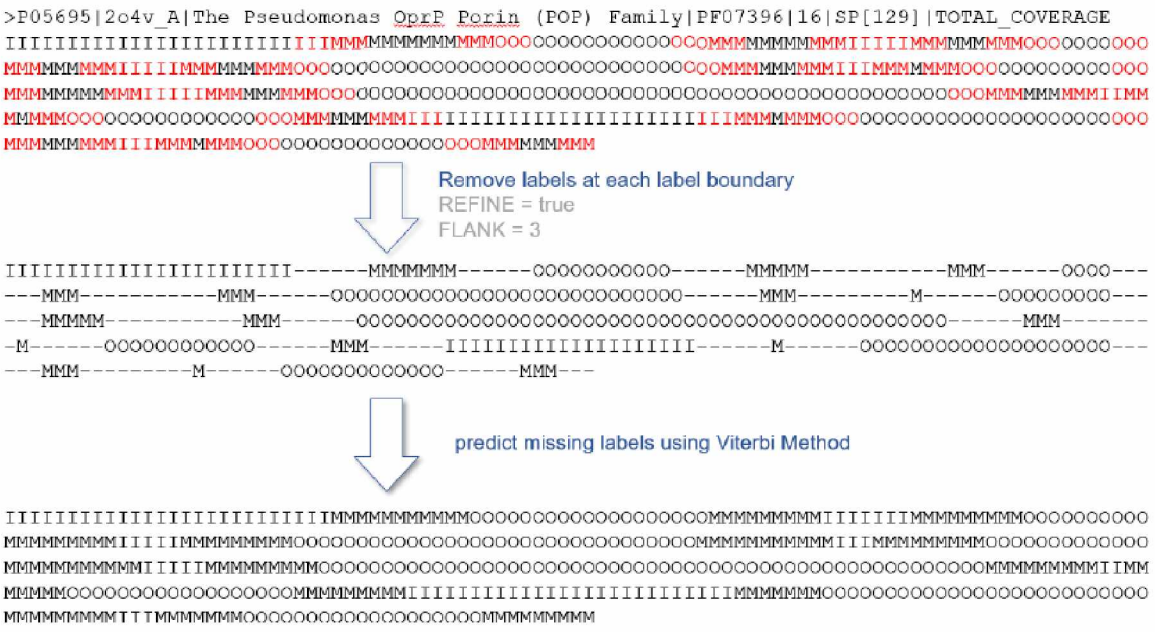


Εικόνα 18. Σχεδιάγραμμα κωδικοποίησης. Η encode κωδικοποίηση θα χρησιμοποιηθεί τόσο με το CML κριτήριο, όσο και με το ML κριτήριο, ενσωματώνοντας πληροφορία και από τις προηγούμενες πιθανότητες. (Tamposis et al., 2018)

3. FLANK=3

Πρόκειται για μια μεταβλητή που λαμβάνεται υπόψιν μόνο όταν ισχύει REFINE=true. Αφορά μια τεχνική για την αντιμετώπιση των λανθασμένα σηματοδοτημένων ορίων στα δεδομένα εκπαίδευσης. Σύμφωνα με την μέθοδο αυτή, αφαιρούνται οι ετικέτες μερικών παρατηρήσεων στα άκρα κάθε ετικέτας (πχ στα άκρα linker ακολουθιών) με στόχο τον επαναυπολογισμό και τη διόρθωση των λανθασμένα σηματοδοτημένων περιοχών. Μετά από την αρχική εκτίμηση, οι ετικέτες των ακολουθιών διαγράφονται σε μια περιοχή ορισμένων καταλοίπων (πχ προς κάθε κατεύθυνση του πρώτου και τελευταίου αμινοξέως που ανήκει σε linker) και γίνεται εκ νέου πρόβλεψη των θέσεων αυτών. Με τον τρόπο αυτό, επιτυγχάνεται μια απόδοση ετικετών πιο συνεπής σε σχέση με την συνολική δομή της πρωτεΐνης, με την μετακίνηση των ορίων κατά τέτοιον τρόπο, που αυτά να ταιριάζουν καλύτερα στο συγκεκριμένο μοντέλο.

Οι τιμές που χρησιμοποιούνται για την μεταβλητή flank είναι 3,4,5 και δείχνουν το πλήθος των αμινοξέων στα οποία θα αφαιρεθούν οι ετικέτες, προς κάθε κατεύθυνση από το σημείο ενδιαφέροντος (σημείο διαχωρισμού linker- domain περιοχών) με σκοπό τον επαναυπολογισμό τους.



Εικόνα 19. Παράδειγμα εφαρμογής της μεθόδου refine. Η πρώτη ακολουθία δείχνει την αρχική πρόβλεψη, η δεύτερη ακολουθία δείχνει τις αφαιρέσεις των ετικετών στις flanking regions ενώ στην τελευταία ακολουθία φαίνεται η τελική πρόβλεψη μετά τον επαναυπολογισμό των τμημάτων που έχουν αφαιρεθεί [68]..

4. PRIOR_TRANS=0. 1

Το JUCHMME παρέχει έναν ευέλικτο τρόπο ώστε να ενσωματωθούν οι εκ των προτέρων πιθανότητες από τις ακολουθίες εισόδου στην διαδικασία της πρόβλεψης. Σύμφωνα με αυτόν, μία χαρτογράφηση ανάμεσα στις ετικέτες των καταστάσεων σε ένα Hidden Markov Model και των διαφορετικών τύπων πρότερης γνώσης πρέπει να οριστεί. Αυτό εφαρμόζεται προσθέτοντας θόρυβο στις παραμέτρους του μοντέλου, και κατόπιν μειώνοντας τα επίπεδα του θορύβου αυτού κατά μια prior τιμή ανά εκτέλεση. Για τις μεταβάσεις, η μεταβλητή PRIOR_TRANS ορίζει αν το σετ των ετικετών αποδέχεται μία εκ των προτέρων πιθανότητα στο [0:1].

3.2. ΓΡΑΦΗΜΑΤΑ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

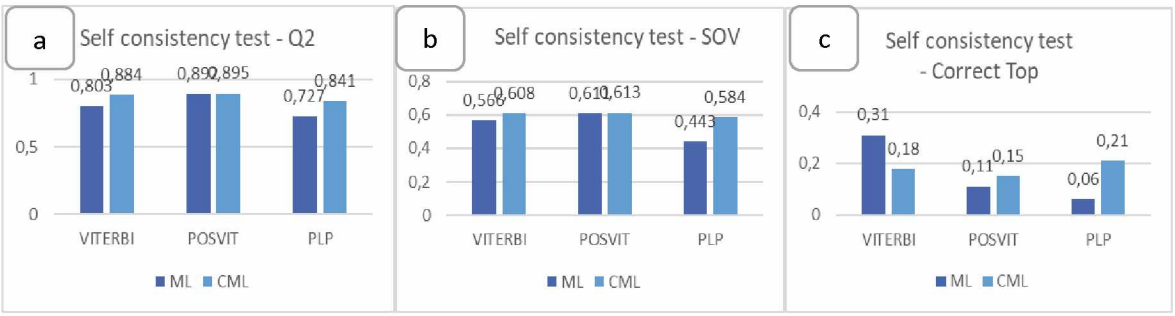
Από τα αποτελέσματα που προέκυψαν, εφαρμόζοντας το μοντέλο με διαφορετικές παραμέτρους έγινε η διαλογή εκείνων, που αποδίδουν καλύτερα σε μία ή περισσότερες μετρικές.

Q2	SOV	True Positive	False Positive	Correct Top	False Negative	SM	tp/tp+fn	tp/tp+fp
0.558	0.278	449	2090	1	2	0.412	0.995	0.18
0.810	0.594	335	735	45	116	0.479	0.74	0.31
0.744	0.484	299	152	16	152	0.384	0.66	0.66
0.905	0.687	116	103	54	335	0.369	0.257	0.53
0.876	0.646	111	248	97	340	0.273	0.24	0.31

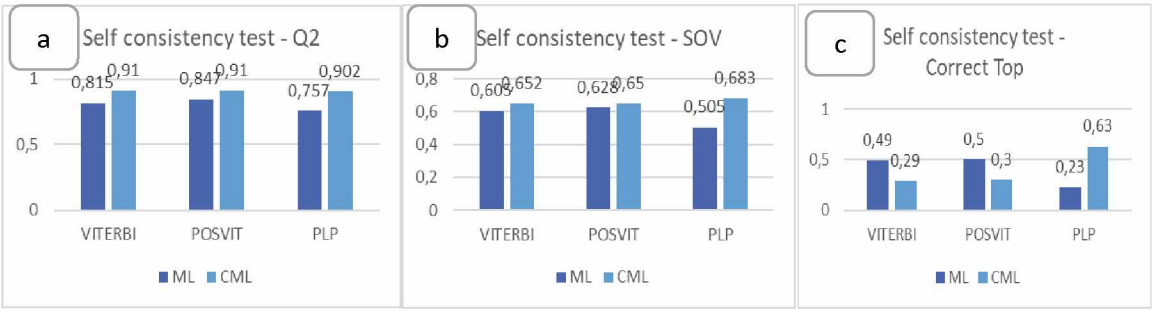
Εικόνα 20. Μερικά από τα κορυφαία αποτελέσματα του μοντέλου. Με μωβ εμφανίζονται οι μεγαλύτερες τιμές ανά μετρική, ενώ με μπλε εμφανίζεται το καλύτερο συνολικά αποτέλεσμα.

Στη συνέχεια, θέλοντας να διαπιστώσουμε αν οι αντίστοιχες παράμετροι δίνουν εξίσου καλά αποτελέσματα εάν αλλάξει η μέθοδος εκτίμησης της πιθανοφάνειας (βλ. 2.3.2), συγκρίναμε τα αποτελέσματά τους. Μερικά από αυτά, παρατίθενται στα παρακάτω διαγράμματα. Οι μετρικές που απεικονίζονται είναι οι Q2 (διαγράμματα a), SOV (διαγράμματα b) και Correct Top (διαγράμματα c) (βλ. 2.8) για τους αλγορίθμους αποκωδικοποίησης Viterbi, Posterior-Viterbi (POSVIT) και Optimal accuracy posterior decoder (βλ. 2.3.1) (PLP).

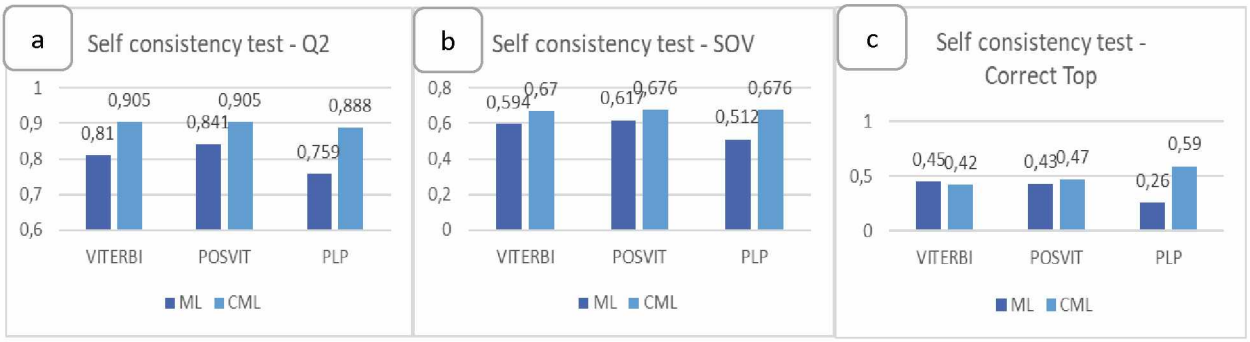
Αρχικά εικονίζονται οι συγκρίσεις με τα καλύτερα αποτελέσματα που δίνονται από την Maximum Likelihood, με τα αντίστοιχα αποτελέσματα που προκύπτουν με τις ίδιες παραμέτρους για τη μέθοδο Conditional Maximum Likelihood, έχοντας ως μέθοδο απόδοσης αξιοπιστίας μοντέλου το Self consistency test.



Η αποκωδικοποίηση έγινε μέσω των αλγορίθμων Viterbi, Posterior Viterbi (POSVIT) και Optimal Accuracy Posterior Probabilities (PLP), με μεθόδους εύρεσης πιθανοφάνειας Maximum Likelihood (μπλε σκούρο) και Conditional Maximum Likelihood (γαλάζιο). Οι παράμετροι με τις οποίες προέκυψαν τα αποτελέσματα αυτά είναι flank 4, prior 0.1 και encode 1. Δηλαδή, μετά την πρόβλεψη της ακολουθίας αφαιρέθηκαν 4 αμινοξεά προς κάθε κατεύθυνση από τα όρια των linker και επαναλήφθηκε η πρόβλεψη για τα αμινοξέα αυτά, ενώ το encode = 1 υποδηλώνει ότι λήφθηκε υπόψιν και η υδροφοβικότητα των αμινοξέων για τον υπολογισμό των ετικετών.

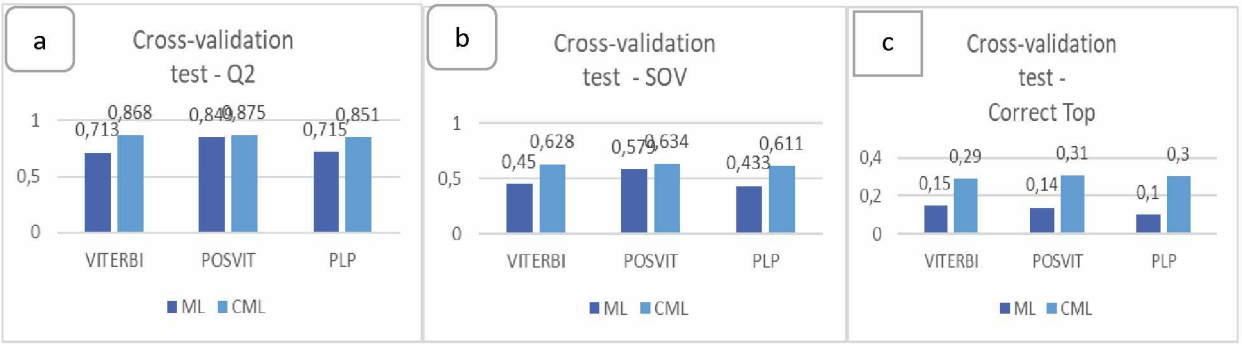


Όμοια με προηγουμένως, στα παραπάνω γραφήματα, παρατίθενται τα αποτελέσματα Self-consistency Test των μετρικών Q2, SOV και Correct Top (a, b και c αντίστοιχα), με αποκωδικοποίηση μέσω των αλγορίθμων Viterbi, Posterior Viterbi και Optimal Accuracy Posterior Probabilities, με τις προσεγγίσεις Maximum Likelihood και Conditional Maximum Likelihood. Οι παράμετροι με τις οποίες προέκυψαν τα αποτελέσματα αυτά είναι flank 3, prior 0.1 και encode 4. Δηλαδή, τα αμινοξέα που αφαιρέθηκαν προς κάθε κατεύθυνση στα όρια αλλαγής ετικετών ήταν 3, ενώ για την πρόβλεψη χρησιμοποιήθηκαν όλοι οι πιθανοί συνδυασμοί αμινοξέων (encode = 4).

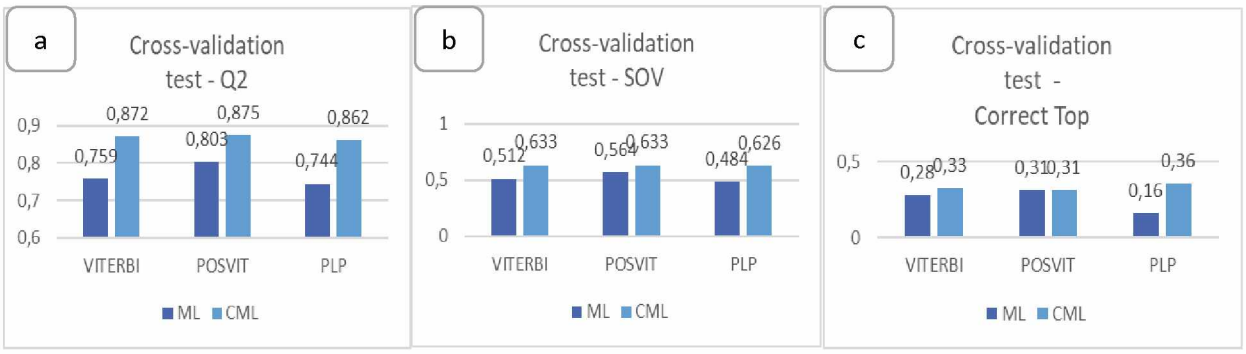


Στα παραπάνω γραφήματα, παρατίθενται τα αποτελέσματα Self-consistency Test των μετρικών Q2, SOV και Correct Top (a, b και c αντίστοιχα), με αποκωδικοποίηση μέσω των αλγορίθμων Viterbi, Posterior Viterbi και Optimal Accuracy Posterior Probabilities, με τις προσεγγίσεις ML και CML. Οι παράμετροι με τις οποίες προέκυψαν τα αποτελέσματα αυτά είναι flank 4, prior 0.1 και encode 4.

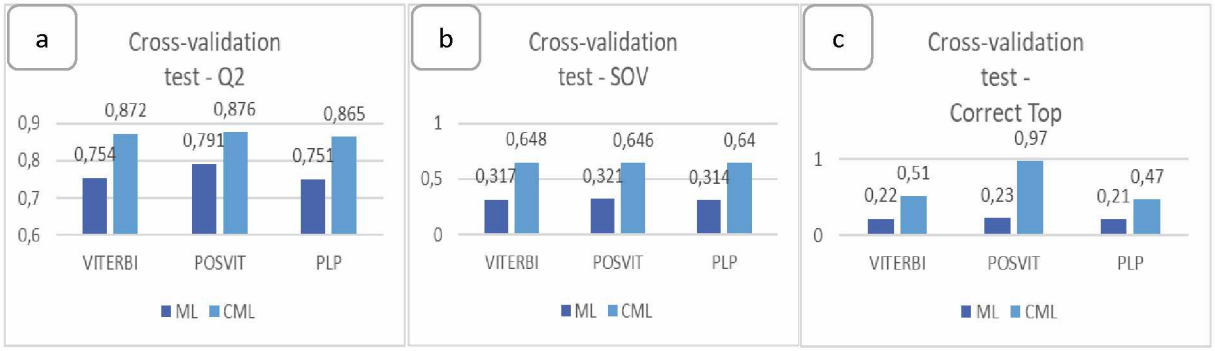
Έπειτα διαφαίνονται οι συγκρίσεις με τα καλύτερα αποτελέσματα που δίνονται από την Maximum Likelihood μέθοδο με τα αντίστοιχα αποτελέσματα που προκύπτουν με τις ίδιες παραμέτρους για τη μέθοδο Conditional Maximum Likelihood έχοντας ως μέθοδο απόδοσης αξιοπιστίας μοντέλου το Cross-Validation test.



Στα παραπάνω γραφήματα, παρατίθενται τα αποτελέσματα Cross-validation Test των μετρικών Q2, SOV και Correct Top (a, b και c αντίστοιχα), με αποκωδικοποίηση μέσω των αλγορίθμων Viterbi, Posterior Viterbi και Optimal Accuracy Posterior Probabilities, με υπολογισμούς πιθανοφάνειας με Maximum Likelihood (μπλε) και Conditional Maximum Likelihood (γαλάζιο). Οι παράμετροι με τις οποίες προέκυψαν τα αποτελέσματα αυτά είναι flank 3, prior 0.1 και encode -. Δηλαδή, το εύρος της flanking region που εκτελείται το refine είναι 3 αμινοξέα προς κάθε κατεύθυνση ενώ δεν χρησιμοποιείται πρότερη γνώση για την πρόβλεψη.

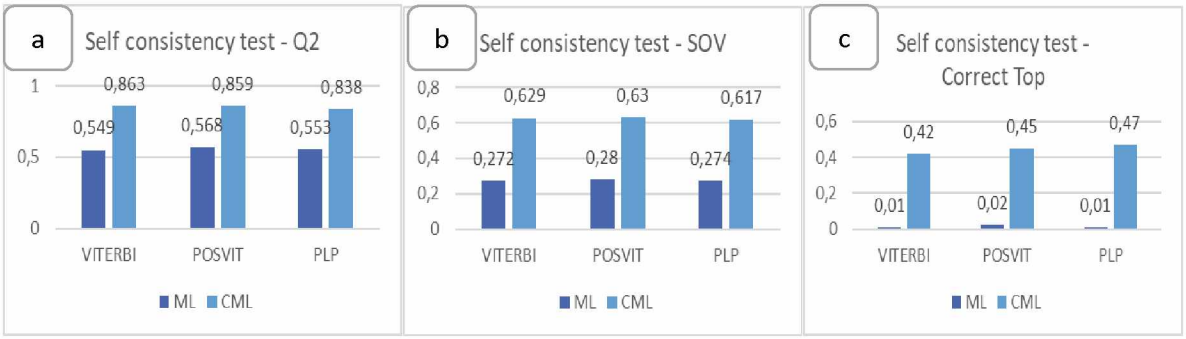


Στα παραπάνω γραφήματα, παρατίθενται τα αποτελέσματα Cross-validation Test των μετρικών Q2, SOV και Correct Top (a, b και c αντίστοιχα), με αποκωδικοποίηση μέσω των αλγορίθμων Viterbi, Posterior Viterbi και Optimal Accuracy Posterior Probabilities, με υπολογισμούς πιθανοφάνειας με Maximum Likelihood (μπλε) και Conditional Maximum Likelihood (γαλάζιο). Οι παράμετροι με τις οποίες προέκυψαν τα αποτελέσματα αυτά είναι flank 3, prior 0.1 και encode 4. Συνεπώς το flanking region ισούται με 3 κατάλοιπα προς κάθε κατεύθυνση του σημείου εναλλαγής μεταξύ των περιοχών, και λαμβάνονται υπόψιν όλοι οι δυνατοί συνδυασμοί διπεπτιδίων.

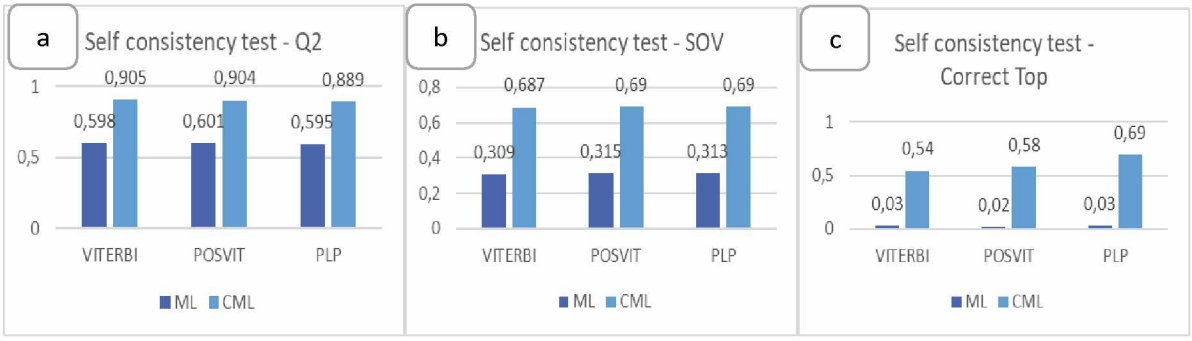


Στα παραπάνω γραφήματα, παρατίθενται τα αποτελέσματα Cross-validation Test των μετρικών Q2, SOV και Correct Top (a, b και c αντίστοιχα), με αποκωδικοποίηση μέσω των αλγορίθμων Viterbi, Posterior Viterbi και Optimal Accuracy Posterior Probabilities, με υπολογισμούς πιθανοφάνειας με Maximum Likelihood (μπλε) και Conditional Maximum Likelihood (γαλάζιο). Οι παράμετροι με τις οποίες προέκυψαν τα αποτελέσματα αυτά είναι flank 5, prior 0.1 και encode 4. Συνεπώς, το μέγεθος της flanking περιοχής είναι 10 αμινοξέα (5 προς κάθε κατεύθυνση) και λαμβάνονται υπόψιν όλοι οι συνδυασμοί των διπεπτιδίων.

Συνεχίζοντας εμφανίζονται οι συγκρίσεις με τα καλύτερα αποτελέσματα που δίνονται από την Conditional Maximum Likelihood μέθοδο με τα αντίστοιχα αποτελέσματα που προκύπτουν με τις ίδιες παραμέτρους για τη μέθοδο Maximum Likelihood. Η απόδοση αξιοπιστίας του μοντέλου γίνεται με Self consistency test.



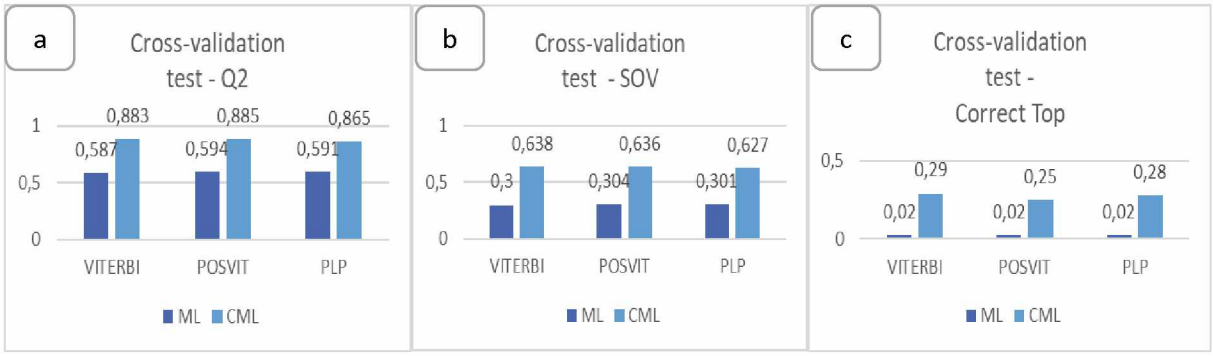
Στα παραπάνω γραφήματα, παρατίθενται τα αποτελέσματα Cross-validation Test των μετρικών Q2, SOV και Correct Top (a, b και c αντίστοιχα), με αποκωδικοποίηση μέσω των αλγορίθμων Viterbi, Posterior Viterbi και Optimal Accuracy Posterior Probabilities, με υπολογισμούς πιθανοφάνειας με Maximum Likelihood (μπλε) και Conditional Maximum Likelihood (γαλάζιο). Οι παράμετροι με τις οποίες προέκυψαν τα αποτελέσματα αυτά είναι flank 3, prior 0.5 και encode -. Δηλαδή το εύρος της flanking region είναι 6 αμινοξέα, ενώ δεν λαμβάνεται πρότερη γνώση για τον υπολογισμό της πρόβλεψης.



Στα παραπάνω γραφήματα, παρατίθενται τα αποτελέσματα Cross-validation Test των μετρικών Q2, SOV και Correct Top (a, b και c αντίστοιχα), με αποκωδικοποίηση μέσω των αλγορίθμων Viterbi, Posterior Viterbi και Optimal Accuracy Posterior Probabilities, με τις προσεγγίσεις Maximum Likelihood και Conditional Maximum Likelihood. Οι παράμετροι με τις οποίες προέκυψαν τα αποτελέσματα αυτά είναι

flank 5, prior 0.5 και encode 4. Συνεπώς, το flanking region αποτελείται από 10 αμινοξέα, και λαμβάνονται υπόψιν όλοι οι συνδυασμοί των διπεπτιδίων.

Τέλος, παρατίθεται η σύγκριση με τα καλύτερα αποτελέσματα που δίνονται από την την Conditional Maximum Likelihood μέθοδο με τα αντίστοιχα αποτελέσματα που προκύπτουν με τις ίδιες παραμέτρους για τη μέθοδο Maximum Likelihood και απόδοση αξιοπιστίας μοντέλου με Cross-Validation test.



Στα παραπάνω γραφήματα, παρατίθενται τα αποτελέσματα Cross-validation Test των μετρικών Q2, SOV και Correct Top (a, b και c αντίστοιχα), με αποκωδικοποίηση μέσω των αλγορίθμων Viterbi, Posterior Viterbi και Optimal Accuracy Posterior Probabilities, με τις προσεγγίσεις Maximum Likelihood και Conditional Maximum Likelihood. Οι παράμετροι με τις οποίες προέκυψαν τα αποτελέσματα αυτά είναι flank 4, prior 0.5 και encode 3. Δηλαδή το μήκος της flanking region θα είναι 8 αμινοξέα (4 προς κάθε κατεύθυνση) και οι φυσικοχημικές ιδιότητες που λαμβάνονται υπόψιν είναι η υδροφοβικότητα, η πολικότητα, το μέγεθος και η αρωματικότητα των αμινοξέων.

ΚΕΦΑΛΑΙΟ 4^{Ο}: ΣΥΖΗΤΗΣΗ

4.1. ΒΕΛΤΙΣΤΑ ΑΠΟΤΕΛΕΣΜΑΤΑ ΔΟΚΙΜΩΝ

Το μοντέλο δοκιμάστηκε σε διαφορετικούς συνδυασμούς παραμέτρων ούτως ώστε να εξακριβωθεί πρωτίστως ποιες από αυτές βελτιστοποιούν την απόδοσή του, μεγιστοποιώντας τις παραγόμενες μετρικές, αλλά και για να διαπιστωθεί το πώς επιδρούν αυτές στο τελικό αποτέλεσμα.

Αρχικά, το μοντέλο δοκιμάστηκε για τον αλγόριθμο προσδιορισμού πιθανοφάνειας Maximum Likelihood. Συγκεκριμένα, εξετάσθηκε η συμπεριφορά του μοντέλου όταν:

- δεν γινόταν χρήση encode ή όταν encode $\in [1,4]$
    - Δηλαδή εξετάζεται η συμπεριφορά του ως έχει (χωρίς πληροφορία προηγούμενων παρατηρήσεων) αλλά και το πως επιδρά με διευρυμένο αλφάβητο, που ανταποκρίνεται στις εκάστοτε ιδιότητες των προηγούμενων παρατηρήσεων που μελετώνται.
- οι τιμές prior probability ανήκουν στο εύρος [0.1 ,0.5] και
- οι τιμές της μεταβλητής flank region $\in [3,5]$. (βλ. 3.1)

Ο αλγόριθμος Maximum Likelihood είναι σημαντικά ταχύτερος από τον Conditional Maximum Likelihood και για τον λόγο αυτό επιλέχθηκε στο συγκεκριμένο εγχείρημα εύρεσης των βέλτιστων συνόλων. Μετά τις δοκιμές, προέκυψε πως τα βέλτιστα αποτελέσματα δίνονται για prior probability ίσο με 0.1 ή 0.5. Συνεπώς, ο αλγόριθμος Conditional Maximum Likelihood δοκιμάστηκε μόνο για αυτές τις τιμές της παραμέτρου. Επίσης, δίνεται η ένδειξη πως όσο μεγαλώνει η τιμή του encode (δηλαδή όσες περισσότερες ιδιότητες προηγούμενων παρατηρήσεων υπολογίζονται στην πρόβλεψη του μοντέλου), τόσο καλύτερα είναι τα αποτελέσματα. Όμως, η βελτίωση αυτή δεν ήταν πάγια, δηλαδή κάθε αποτέλεσμα για encode_1 να δίνει χειρότερα αποτελέσματα από ότι για encode_2, όπου $\text{encode}_1 < \text{encode}_2$. Έτσι, ο αλγόριθμος Conditional Maximum Likelihood δοκιμάστηκε για όλες τις τιμές της παραμέτρου encode.

Από τον πίνακα των παραγόμενων αποτελεσμάτων, έγινε επιλογή των καλύτερων από αυτών, που χρησιμοποιήθηκαν για τα παραπάνω διαγράμματα (βλ. Κεφάλαιο 3.2.). Από τους τρεις αλγορίθμους αποκωδικοποίησης (βλ. 2.3.1) που εφαρμόστηκαν, (Viterbi, PLP, POSVIT) δεν ξεχώρισε κανείς συντριπτικά σε σχέση με τους υπόλοιπους, ούτε βρέθηκε να υστερεί σημαντικά. Στις περισσότερες περιπτώσεις τα αποτελέσματα ήταν παραπλήσια, ωστόσο οι καλύτερες τιμές αυτών προέκυψαν από διαφορετικές παραμέτρους. Άρα, δεν κρίνεται ουσιώδης η απομάκρυνση κάποιου από αυτούς ως περιττός. Όμως, λόγω του ότι το καλύτερο συνολικά αποτέλεσμα προκύπτει από την εφαρμογή του αλγορίθμου Optimal Accuracy Posterior Probabilities, θεωρείται ως αυτός που αποδίδει καλύτερα για το μοντέλο πρόβλεψης που αναπτύχθηκε.

Γενικά, τα αποτελέσματα των δύο αλγορίθμων εκτίμησης πιθανοφάνειας (Maximum Likelihood, Conditional Maximum Likelihood) παρουσίαζαν διαφορές. Ο πρώτος, βρίσκει περισσότερα True Positive και σχετικά λίγα False Negative, όμως τα False Positive είναι ιδιαίτερα υψηλά. Συνεπώς, ο Maximum Likelihood έχει μεγαλύτερη ευαισθησία, που ανέρχεται μέχρι και 99,5%, ενώ ο Conditional Maximum Likelihood εμφανίζει μεγαλύτερη ειδικότητα με μέγιστο ποσοστό τα 68%. Επίσης, στον τελευταίο, υψηλότερες τιμές παρατηρούνται και στις μετρικές Q2 , SOV και Correct Top, που ανέρχονται στα ποσοστά 91,1%, 69% και 97% αντίστοιχα. Κάτι τέτοιο φάνηκε άλλωστε και από τα διαγράμματα, καθώς στην σύγκριση των βέλτιστων αποτελεσμάτων για τον αλγόριθμο Maximum Likelihood με τα αντίστοιχα του Conditional Maximum Likelihood, αυτά του δεύτερου παρουσιάζονται υψηλότερα. Τέλος, ενώ η μέγιστη τιμή της μετρικής SM παρατηρείται στον Conditional Maximum Likelihood αλγόριθμο και ισούται με 53%, συνολικά, τα αποτελέσματα του Maximum Likelihood εμφανίζουν υψηλότερα ποσοστά.

Παρά το γεγονός ότι έχουν προκύψει αρκετά υψηλά ποσοστά, σπάνια συναντώνται παράμετροι που να δίνουν καλά αποτελέσματα για όλες τις μετρικές που μας ενδιαφέρουν. Έτσι, το βέλτιστο αποτέλεσμα, ως συνολικά πιο ολοκληρωμένο, δίνεται από τις παραμέτρους flank=3, prior=0.1, encode=4 με εκτίμηση πιθανοφάνειας από τον αλγόριθμο Maximum Likelihood, κωδικοποίηση με εφαρμογή του Optimal Accuracy Posterior Probabilities (PLP) και έλεγχο αξιοπιστίας αποτελεσμάτων μέσω του k-Cross-Validation (όπου k=10). Ειδικότερα, οι τιμές που δίνει για τις μετρικές Q2, SOV, SM, sensitivity και specificity είναι 74,4%, 48,4%, 38.4%, 66% και 66% αντιστοίχως (βλ. 2.8).

Όμως, θα παρατεθούν και οι παράμετροι που μεγιστοποιούν τις εκάστοτε μετρικές.

- ✓ Η μεγαλύτερη ευαισθησία 99.5% δίνεται για flank=3, prior=0.5, encode=2 με τον αλγόριθμο εκτίμησης πιθανοφάνειας Maximum Likelihood, αποκωδικοποίηση με εφαρμογή Viterbi και έλεγχο αξιοπιστίας με self-consistency test.
- ✓ Η μεγαλύτερη ακρίβεια 68% δίνεται για flank=4, prior=0.1, encode=4 με τον αλγόριθμο εκτίμησης πιθανοφάνειας Conditional Maximum Likelihood με αποκωδικοποίηση μέσω Viterbi και έλεγχο αξιοπιστίας με Self-consistency test.
- ✓ Το μέγιστο Correct Top 97% δίνεται από flank=5, prior=0.1, encode=4 με τον αλγόριθμο εκτίμησης πιθανοφάνειας Conditional Maximum Likelihood με αλγόριθμο αποκωδικοποίησης Posterior Viterbi και μέθοδο αξιολόγησης Cross Validation test.
- ✓ Το βέλτιστο Q2 91,1% προκύπτει για flank=3, prior=0.1, encode=2 με τον αλγόριθμο εκτίμησης πιθανοφάνειας Conditional Maximum Likelihood με αλγόριθμο αποκωδικοποίησης Optimal Accuracy Posterior Probabilities και μέθοδο αξιολόγησης το Self-consistency test.
- ✓ Επίσης, το βέλτιστο SM, 53%, και το μέγιστο SOV 69%, προκύπτει από flank=5, prior=0.5, encode=4 με τον αλγόριθμο Conditional Maximum Likelihood με αποκωδικοποίηση μέσω Viterbi και έλεγχο αξιοπιστίας με Self-consistency test.

	Flank	Prior probability	Encode	Likelihood Estimation	Validation	Decoding
best overal *sensitivity* 66%, *specificity* 66%	3	0.1	4	ML	CROSS VALIDATION	PLP
sensitivity 99.5%	3	0.5	2	ML	SELF-CONSISTENCY	Viterbi
specificity 68%	4	0.1	4	CML	CROSS VALIDATION	POSVIT
Correct Top 97%	5	0.1	4	CML	CROSS VALIDATION	POSVIT
Q2 91,1%	3	0.1	2	CML	SELF-CONSISTENCY	POSVIT/PLP
SOV & SM 69% 35%	5	0.5	4	CML	SELF-CONSISTENCY	Viterbi

Εικόνα 21. Συγκεντρωτικά οι παράμετροι των καλύτερων αποτελεσμάτων

Όσον αφορά τα κρυφά νευρωνικά δίκτυα, τα αποτελέσματα δεν βελτιώθηκαν όσο ήταν αναμενόμενο. Οι δοκιμές έγιναν για flank = 3 και prior probability ίσο με 0,1 και 0,5. Κατά τις δοκιμές, τα κρυφά νευρωνικά δίκτυα είχαν ως είσοδο κυλιόμενα συμμετρικά παράθυρα διαφορετικών μεγεθών (5-9 αμινοξέα δεξιά και αριστερά του κατάλοιπου ενδιαφέροντος) αλλά διερευνήθηκαν και τα ασύμμετρα παράθυρα (15 αμινοξέα αριστερά, 5 αμινοξέα δεξιά). Τα Hidden Neural Networks που χρησιμοποιήθηκαν ήταν πολυεπίπεδα με 7 έως 11 κρυφούς νευρώνες.

Από τα αποτελέσματα που προέκυψαν, έχουν ξεχωρίσει δύο σύνολα παραμέτρων που δίνουν Q2 90,9%, 91,4% και SOV 64,4%, 62% με αρκετά χαμηλότερη ευαισθησία και ειδικότητα συγκριτικά με την εφαρμογή των απλών Hidden Markov Models. Το πρώτο σύνολο παραμέτρων χαρακτηρίζεται από flank=3, prior=0.1, window=5,5 , nhidden = 8, ενώ το δεύτερο σύνολο παραμέτρων χαρακτηρίζεται από flank=3, prior=0.5, window=15,5, nhidden=11.

Η εξήγηση πίσω από τα χαμηλά αποτελέσματα που δίνουν τα κρυφά νευρωνικά δίκτυα δεν είναι προφανής και ενδεχομένως να οφείλεται στην απουσία της πληροφορίας που φέρει ο πίνακας εκπομπών. Λόγω της μικρής διαφοράς μεταξύ των στοιχείων που χαρακτηρίζουν τις περιοχές linker από αυτές των domain, ίσως δεν είναι ικανό να τις διακρίνει, οριοθετώντας λανθασμένα τις περιοχές αυτές. Μία επιπλέον δυσκολία στην πρόβλεψη αυτή, αποτελεί το μικρό μέγεθος των linker συγκριτικά με αυτό των domain.

4.1. ΣΥΓΚΡΙΣΗ ΑΠΟΔΟΣΗΣ ΜΕ ΑΝΤΙΣΤΟΙΧΑ ΠΡΟΓΡΑΜΜΑΤΑ

Η απευθείας σύγκριση του μοντέλου με άλλα προγράμματα που επιχειρούν να λύσουν το ίδιο βιολογικό πρόβλημα δεν είναι δυνατή, καθώς διαφέρουν τα dataset μεταξύ τους, τόσο ποιοτικά όσο και ποσοτικά σε ακολουθίες. Συγκεκριμένα, κάποια από τα προγράμματα της βιβλιογραφίας είναι τόσο παλιά, που ορισμένες από τις πρωτεΐνες που έχουν χρησιμοποιηθεί για την εκπαίδευσή τους, θεωρούνται απαρχαιωμένες και δεν εμφανίζονται στην PDB (ή έχουν αλλάξει).

Ένα άλλο πρόβλημα που παρατηρείται με κάποιες από τις εφαρμογές που έχουν δημοσιευτεί, είναι η χρήση διαφορετικών μετρικών για την αξιολόγηση των αποτελεσμάτων και συνεπώς η αδυναμία σύγκρισης των αποτελεσμάτων.

Ωστόσο, συγκρίνοντας τα δημοσιευμένα αποτελέσματά των συγγραφέων με τις μετρικές sensitivity και specificity, με αυτά που έχουν προκύψει από την εφαρμογή του μοντέλου προκύπτουν τα εξής:

Article	Sensitivity	Specificity
best overall approach -JUCHMME linker	66%	66%
DomNet: Protein Domain Boundary Prediction Using Enhanced General Regression Network and New Profiles	72%	70%
Prediction of protein interdomain linker regions by a hidden Markov model	63.3%	92.9%
Domain boundary prediction based on profile domain linker propensity index	72%	35%
Identification of putative domain linkers by a neural network - application to a large sequence database	10,30%	81,80%

Εικόνα 222. Πίνακας σύγκρισης των αποτελεσμάτων, με αντίστοιχες εφαρμογές που έχουν δημοσιευτεί με αποτελέσματα υψηλότερα του παρόντος μοντέλου σε μία ή και στις δύο μετρικές αξιολόγησης..

Article	Sensitivity	Specificity
Characteristics and prediction of domain linker sequences in multi-domain proteins	37,00%	42,00%
Loop-Length-Dependent SVM Prediction of Domain Linkers for High-Throughput Structural Proteomics	59,70%	43,60%
Armadillo: Domain Boundary Prediction by Amino Acid Composition	37,00%	36,00%
DOMpro: Protein Domain Prediction Using Profiles, Secondary Structure, Relative Solvent Accessibility, and Recursive Neural Networks	59,00%	38,00%
Characterization and prediction of linker sequences of multi-domain proteins by a neural network	36%	58%
Improvement of domain linker prediction by incorporating loop-length-dependent characteristics	36,10%	40,60%
best overall approach -JUCHMME linker	66%	66%
DomCut: prediction of inter-domain linker regions in amino acid sequences	53.5%	50.1%

Εικόνα 23. Πίνακας σύγκρισης των αποτελεσμάτων, με αντίστοιχες εφαρμογές που έχουν δημοσιευτεί με αποτελέσματα χαμηλότερα του παρόντος μοντέλου .

Αξίζει να αναφέρουμε πως από τις εφαρμογές της βιβλιογραφίας, μόνο το DomCut διαθέτει online server για πρόβλεψη περιοχών linker τη δεδομένη στιγμή, καθώς τα υπόλοιπα links είναι ανενεργά. Παρατηρούμε πως το μοντέλο δίνει καλά αποτελέσματα σε σχέση με αυτά της βιβλιογραφίας. Ακόμα και με την παράμετρο best overall approach (που δεν αντιπροσωπεύει το μέγιστο sensitivity που έχει παρατηρηθεί) δίνει ικανοποιητικά ποσοστά σε sensitivity και specificity δεδομένης της δυσκολίας του προβλήματος. Ενώ, εάν δημιουργηθεί online server θα είναι η online μέθοδος με την καλύτερη απόδοση.

4.2. ΜΕΛΛΟΝΤΙΚΕΣ ΒΕΛΤΙΩΣΕΙΣ ΚΑΙ ΕΠΕΚΤΑΣΕΙΣ

Το μοντέλο παρατηρούμε ότι παράγει ικανοποιητικά αποτελέσματα. Η καλύτερη παράμετρος δίνει sensitivity και specificity ίσο με 66%, ενώ σε διαφορετικούς συνδυασμούς παραμέτρων επιτυγχάνεται sensitivity που φτάνει το ποσοστό των 99.9%. Όμως, υπάρχουν αναντίρρητα περιθώρια βελτίωσης ώστε να μειωθούν τα σχετικά υψηλά FP που προκύπτουν και κατ' επέκταση να αυξηθεί το specificity, που μέχρι στιγμής εμφανίζει μέγιστη τιμή μόλις 68%.

Πρωτίστως, συμπληρωματικά με το μοντέλο, θα μπορούσε να γίνει μια διερεύνηση για ακολουθίες consensus. Πρόκειται για ακολουθίες συντηρημένες, δηλαδή ακολουθίες μοτίβα που σχετίζονται με κάποια βιολογική διαδικασία (πχ. miRNA binding site). Παρά το γεγονός ότι δεν υπάρχουν ισχυρά συντηρημένες περιοχές που να διαχωρίζουν τα domains, μελλοντικά ίσως θα μπορούσε αυτή η εφαρμογή να λειτουργήσει συμπληρωματικά με το μοντέλο, βελτιώνοντας την ποιότητα της πρόβλεψης.

Τέλος, θα μπορούσε να διερευνηθεί και ο λόγος αδυναμίας ισχυρής πρόβλεψης (στο συγκεκριμένο μοντέλο) με την χρήση Hidden Neural Networks. Ενδεχομένως με διαφορετικά αρχεία εισόδου για το Hidden Neural Network ή με εκτενέστερη μελέτη και δοκιμή περισσότερων παραμέτρων που έχει αποδειχτεί ότι δίνουν καλά αποτελέσματα (όπως encode ίσο με 4) να υπάρξει βελτίωση των παραγόμενων ποσοστών.

ΚΕΦΑΛΑΙΟ 5^{Ο}: ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Britannica, T.E.o.E. "Learn about the structure and function of proteins".
2. Kessel, A.a.N.B.-T., *Introduction to proteins: structure, function, and motion.* 2018: Chapman and Hall/CRC.
3. Whitford, D., *Proteins: structure and function*. 2013: John Wiley & Sons.
4. CRICK, F., *Central Dogma of Molecular Biology* NATURE, 1970. **227**.
5. *Central Dogma: Dna to Rna to protein*. Available from: https://byjus.com/neet-questions/what-are-the-three-post-transcriptional-modifications/.
6. Watson, J.D. and F.H.C. Crick, *Molecular Structure of Nucleic Acids A Structure for Deoxyribose Nucleic Acid* Nature, 1953. **171**: p. 737-738.
7. Christopher P. Austin, M.D. *Deoxyribonucleic Acid (DNA)*. Available from: https://www.genome.gov/genetics-glossary/Deoxyribonucleic-Acid.
8. *Transcription / RNA Synthesis*. Available from: https://socratic.org/biology/dna-structure-and-function/transcription--rna-synthesis.
9. *Protein Synthesis : Translation*. Available from: https://sites.google.com/site/proteinsynthesistranslation/home/terms.
10. *Turns and Loops - Lecture Notes - Biochemistry I | Chem 471, Study notes for Biochemistry*. Available from: https://www.docsity.com/en/turns-and-loops-lecture-notes-biochemistry-i-chem-471/6778442/.
11. *Proteins*. Available from: https://bio.libretexts.org/Courses/Portland_Community_College/Cascade_Microbiology/21%3A_Appendix_A_-_Biochemistry_Review/21.4%3A_Proteins.
12. Σπηλιόπουλος, Ι. and Μ. Ξαπλαντέρη, *Πρωτεΐνες*, Ι. Σαρηγιάννης, Editor. 2015.
13. *Πρωτεΐνες*. Available from: https://blogs.sch.gr/geortsolbio/tag/%CE%B4%CE%B5%CF%85%CF%84%CE%B5%CF%81%CE%BF%CF%84%CE%B1%CE%B3%CE%AE%CF%82-%CE%B4%CE%BF%CE%BC%CE%AE/.
14. *What is X-ray Crystallography?* 2018; Available from: https://macromoltek.medium.com/what-is-x-ray-crystallography-1e186bc3d180.
15. Γλυκός, Ν.Μ., *Μία μη μαθηματική εισαγωγή στην κρυσταλλογραφία πρωτεϊνών*. 2015.
16. *X-ray Crystallography*. Available from: https://www.creativebiomart.net/resource/principle-protocol-x-ray-crystallography-393.htm.
17. Dessau, M.A. and Y. Modis, *Protein crystallization for X-ray crystallography.* J Vis Exp, 2011(47).
18. Zhou, R.-B., et al., *A review on recent advances for nucleants and nucleation in protein crystallization.* CrystEngComm, 2017. **19**(8): p. 1143-1155.
19. Keeler, J., *Understanding NMR spectroscopy*. 2011: John Wiley & Sons.
20. Wüthrich, K., *"NMR with proteins and nucleic acids."*. Europhysics News, 1986: p. 11-13.
21. Au - Loquet, A., et al., *Atomic Scale Structural Studies of Macromolecular Assemblies by Solid-state Nuclear Magnetic Resonance Spectroscopy.* JoVE, 2017(127): p. e55779.
22. Σπυρούλιας, Γ. *«Φασματοσκοπία Πυρηνικού Μαγνητικού Συντονισμού »*. 2015; 1.0.:[Available from: https://eclass.upatras.gr/courses/PHA1614/
23. Suyama, M. and O. Ohara, *DomCut: prediction of inter-domain linker regions in amino acid sequences.* Bioinformatics, 2003. **19**(5): p. 673-4.
24. 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.
25. Udwary, D.W., M. Merski, and C.A. Townsend, *A method for prediction of the locations of linker regions within large multifunctional proteins, and application to a type I polyketide synthase.* J Mol Biol, 2002. **323**(3): p. 585-98.
26. Miyazaki, S., Y. Kuroda, and S. Yokoyama, *Identification of putative domain linkers by a neural network - application to a large sequence database.* BMC Bioinformatics, 2006. **7**: p. 323.

27. Bae, K., B.K. Mallick, and C.G. Elsik, *Prediction of protein interdomain linker regions by a hidden Markov model.* Bioinformatics, 2005. **21**(10): p. 2264-70.
28. Liu, J. and B. Rost, *Sequence-based prediction of protein domains.* Nucleic Acids Res, 2004. **32**(12): p. 3522-30.
29. Miyazaki, S., Y. Kuroda, and S. Yokoyama, *Characterization and prediction of linker sequences of multi-domain proteins by a neural network.* J Struct Funct Genomics, 2002. **2**(1): p. 37-51.
30. George, R.A. and J. Heringa, *Protein domain identification and improved sequence similarity searching using PSI-BLAST.* Proteins, 2002. **48**(4): p. 672-81.
31. Nagarajan, N. and G. Yona, *Automatic prediction of protein domains from sequence information using a hybrid learning system.* Bioinformatics, 2004. **20**(9): p. 1335-60.
32. Basak, A., et al.,, *High-resolution X-ray crystal structures of human gammaD crystallin (1.25 A) and the R58H mutant (1.15 A) associated with aculeiform cataract*. 2003.
33. Basak, A., et al., *High-resolution X-ray crystal structures of human gammaD crystallin (1.25 A) and the R58H mutant (1.15 A) associated with aculeiform cataract.* J Mol Biol, 2003. **328**(5): p. 1137-47.
34. Chatterjee, P., et al., *PDP-CON: prediction of domain/linker residues in protein sequences using a consensus approach.* J Mol Model, 2016. **22**(4): p. 72.
35. Andreeva, A., et al., *SCOP database in 2004: refinements integrate structure and sequence family data.* Nucleic Acids Res, 2004. **32**(Database issue): p. D226-9.
36. Swindells, M.B., *A procedure for detecting structural domains in proteins.* Protein Sci, 1995. **4**(1): p. 103-12.
37. Swindells, M.B., *A procedure for the automatic determination of hydrophobic cores in protein structures.* Protein Sci, 1995. **4**(1): p. 93-102.
38. Sowdhamini, R. and T.L. Blundell, *An automatic method involving cluster analysis of secondary structures for the identification of domains in proteins.* Protein Sci, 1995. **4**(3): p. 506-20.
39. Holm, L.S., C., *Parser for protein folding units.* Proteins: Struct. Funct. Genet. , 1994(19): p. 256–268.
40. Xu, Y., D. Xu, and H.N. Gabow, *Protein domain decomposition using a graph-theoretic approach.* Bioinformatics, 2000. **16**(12): p. 1091-104.
41. Jones, S., et al., *Domain assignment for protein structures using a consensus approach: characterization and analysis.* Protein Sci, 1998. **7**(2): p. 233-42.
42. Orengo, C.A., et al., *CATH--a hierarchic classification of protein domain structures.* Structure, 1997. **5**(8): p. 1093-108.
43. Holm, L. and C. Sander, *Touring protein fold space with Dali/FSSP.* Nucleic Acids Res, 1998. **26**(1): p. 316-9.
44. Marchler-Bauer, A., et al., *MMDB: Entrez's 3D structure database.* Nucleic Acids Res, 1999. **27**(1): p. 240-3.
45. Yang, A.S. and B. Honig, *An integrated approach to the analysis and modeling of protein sequences and structures. III. A comparative study of sequence conservation in protein structural families using multiple structural alignments.* J Mol Biol, 2000. **301**(3): p. 691-711.
46. Dumontier, M., et al., *Armadillo: domain boundary prediction by amino acid composition.* J Mol Biol, 2005. **350**(5): p. 1061-73.
47. Tanaka, T., Y. Kuroda, and S. Yokoyama, *Characteristics and prediction of domain linker sequences in multi-domain proteins.* J Struct Funct Genomics, 2003. **4**(2-3): p. 79-85.
48. Vlahovicek, K., et al., *The SBASE domain sequence library, release 10: domain architecture prediction.* Nucleic Acids Res, 2003. **31**(1): p. 403-5.
49. Gough, J., et al., *Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure.* J Mol Biol, 2001. **313**(4): p. 903-19.

50. Contreras-Moreira, B. and P.A. Bates, *Domain fishing: a first step in protein comparative modelling.* Bioinformatics, 2002. **18**(8): p. 1141-2.
51. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
52. Wheelan, S.J., A. Marchler-Bauer, and S.H. Bryant, *Domain size distributions can predict domain boundaries.* Bioinformatics, 2000. **16**(7): p. 613-8.
53. Sim, J., S.Y. Kim, and J. Lee, *PPRODO: prediction of protein domain boundaries using neural networks.* Proteins, 2005. **59**(3): p. 627-32.
54. Cheng, J., M.l.J. Sweredoski, and P. Baldi, *DOMpro: Protein Domain Prediction Using Profiles, Secondary Structure, Relative Solvent Accessibility and Recursive Neural Networks.* Data Mining and Knowledge Discovery, 2006 **13** (1): p. 1–10.
55. Yoo, P.D., et al., *DomNet: protein domain boundary prediction using enhanced general regression network and new profiles.* IEEE Trans Nanobioscience, 2008. **7**(2): p. 172-81.
56. 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.
57. Eickholt, J., X. Deng, and J. & Cheng, *DoBo: Protein domain boundary prediction by integrating evolutionary signals and machine learning.* BMC Bioinformatics, 2011. **12**(43).
58. Servant, F., et al., *ProDom: automated clustering of homologous domains.* Brief Bioinform, 2002. **3**(3): p. 246-51.
59. Heger, A., et al., *ADDA: a domain database with global coverage of the protein universe.* Nucleic Acids Res, 2005. **33**(Database issue): p. D188-91.
60. Portugaly, E., et al., *EVEREST: automatic identification and classification of protein domains in all protein sequences.* BMC Bioinformatics, 2006. **7**: p. 277.
61. George, R.A. and J. Heringa, *SnapDRAGON: a method to delineate protein structural domains from sequence data.* J Mol Biol, 2002. **316**(3): p. 839-51.
62. Kim, D.E., et al., *Automated prediction of domain boundaries in CASP6 targets using Ginzu and RosettaDOM.* Proteins, 2005. **61 Suppl 7**: p. 193-200.
63. Wu, Y., et al., *OPUS-Dom: applying the folding-based method VECFOLD to determine protein domain boundaries.* J Mol Biol, 2009. **385**(4): p. 1314-29.
64. Binaykiya, T. *Hidden Markov Models*. 2018; Available from: https://tanmaybinaykiya.github.io/hmm-applications.
65. Yoon, B.J., *Hidden Markov Models and their Applications in Biological Sequence Analysis.* Curr Genomics, 2009. **10**(6): p. 402-15.
66. Bekesiene, S., R. Smaliukiene, and R. Vaicaitiene, *Using Artificial Neural Networks in Predicting the Level of Stress among Military Conscripts.* Mathematics, 2021. **9**(6): p. 626.
67. Hapudeniya, M., *Artificial neural networks in bioinformatics.* Sri Lanka Journal of Bio-Medical Informatics 2010.
68. Tamposis, I.A., Tsirigos, K. D., Theodoropoulou, M. C., Kontou, P. I., Tsaousis, G. N., Sarantopoulou, D., . . . Bagos, P. G. , *JUCHMME: a Java Utility for Class Hidden Markov Models and Extensions for biological sequence analysis. Bioinformatics, 35(24), 5309-5312.* 2019.
69. Tamposis, I.A., et al., *Extending hidden Markov models to allow conditioning on previous observations.* J Bioinform Comput Biol, 2018. **16**(5): p. 1850019.