Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων & Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Πρόβλεψη Δευτεροταγούς Δομής Πρωτεϊνών με τη Χρήση Νευρωνικών Δικτύων

Διπλωματικη Εργασια

Μπαγάκης Εμμανουήλ

Επιβλέπων: Ανδρέας-Γεώργιος Σταφυλοπάτης, Καθηγητής ΕΜΠ

Αθήνα, Σεπτέμβριος 2020

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων & Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Πρόβλεψη Δευτεροταγούς Δομής Πρωτεϊνών με τη Χρήση Νευρωνικών Δικτύων

Διπλωματικη Εργασια

Εμμανουήλ Μπαγάκης

Επιβλέπων: Ανδρέας-Γεώργιος Σταφυλοπάτης, Καθηγητής ΕΜΠ

Εγκρίθηκε από την κάτωθι τριμελή επιτροπή τη 29η Σεπτεμβρίου 2020.

Ανδρέας-Γεώργιος Σταφυλοπάτης Καθηγητής	Στέφανος Κόλλιας Καθηγητής	Γιώργος Στάμου Αναπληρωτής Καθηγητής

Μπαγάκης Εμμανουήλ
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών ΕΜΠ

Περίληψη

Ένα από τα σημαντικότερα ζητήματα στην επιστήμη της βιολογίας αφορά την μελέτη και κατανόηση σε βάθος των πρωτεϊνών. Σε κάθε λειτουργία του ανθρώπινου οργανισμού, οι πρωτεΐνες έχουν καταλυτικό ρόλο και συμβάλλουν στην ομαλή λειτουργία του. Οι ιδιότητες τους σχετίζονται άμεσα με τη δομή τους και για αυτό, το πρόβλημα της αναδίπλωσης τους, αποτελεί βασικό αντικείμενο ανάλυσης.

Επειδή η μελέτη της δομής των πρωτεϊνών έχει μεγάλο χρονικό και χρηματικό κόστος, η επιστημονική κοινότητα ασχολείται ενεργά με την πρόβλεψη των δομών. Η πρόβλεψη της δευτεροταγούς δομής είναι το πρώτο βήμα για την συσχέτιση των ιδιοτήτων των πρωτεϊνών με τον τρόπο που αναδιπλώνονται και αποκτούν σχήμα. Για το σκοπό αυτό έχουν χρησιμοποιηθεί διάφορα μοντέλα ανά τα έτη, με διαφορετικές φιλοσοφίες που αντικατοπτρίζουν την ανάπτυξη των αλγοριθμικών μεθόδων τις τελευταίες δεκαετίες.

Στην προσπάθειά μας να συνεισφέρουμε σε αυτό το έργο, χρησιμοποιήσαμε εφαρμογές Τεχνητής Νοημοσύνης, για να κατασκευάσουμε ένα Νευρωνικό Μοντέλο Μηχανικής Μάθησης το οποίο θα προβλέπει τη δευτεροταγή δομή των πρωτεϊνών. Πιο συγκεκριμένα, στήθηκαν δύο μοντέλα Συνελικτικών Νευρωνικών Δικτύων τα οποία κατηγοριοποιούν τα αμινοξέα των πρωτεϊνών σε 3 και 8 κλάσσεις αντίστοιχα.

Στα τελευταία δύο κεφάλαια αναφερόμαστε λεπτομερώς στη διαδικασία κατασκευής του μοντέλου, εξηγώντας κάθε αρχιτεκτονική επιλογή μας. Περιγράφουμε το πείραμα που διεξάγαμε και τα προβλήματα που συναντήσαμε. Τέλος παρουσιάζουμε τα αποτελέσματα των μοντέλων, τα συγκρίνουμε με τα υπόλοιπα μοντέλα που επικρατούν στη βιοπληροφορική και σχολιάζουμε τις διαφορές και τις ομοιότητες μεταξύ τους.

Λέξεις-Κλειδιά: Μηχανική Μάθηση, Τεχνητά Νευρωνικά Δίκτυα, Συνελικτικά Δίκτυα, Πρόβλεψη Δευτεροταγούς Δομής Πρωτεϊνών, Πρωτεΐνες

Abstract

Deep understanding of Proteins has been one of the most important issues in modern Biology. Proteins are considered a vital factor in every function of the human body. Their features and roles are deeply dependent on their shape, so the protein folding problem is deemed of high importance.

Since the analysis of a protein structure is a costly procedure, both time-wise and money-wise, scientists are actively involved in the prediction of this analysis. Secondary structure prediction is the first step towards the prediction of a Protein's shape and how it reaches its finite state. For this reason, many prediction models have been developed throughout the last 60 years, reflecting the ever improving algorithms that are used in Protein Prediction.

In our effort to contribute to this field, we developed through Artificial Intelligence applications, an Artificial Neural Network for predicting the secondary structure. We came up with two Convolutional Neural Networks to predict the 3-way and 8-way classification of amino-acids within a protein.

In the last two chapters we explain the procedure of building the model, explaining every decision we made along the way. We then describe the experiment and the problems we faced. Finally, we present the models' results, and we compare them with the state of the art models, commenting on the differences and similarities between them.

Keywords: Machine Learning, Artificial Neural Networks, Convolutional Networks,Protein Secondary Structure Prediction, Proteins

Περιεχόμενα

Πίνακας Σχημάτων

Πίνακας Πινάκων

Κεφάλαιο 1

Εισαγωγή

Το πρόβλημα της αναδίπλωσης των πρωτεϊνών, αποτελεί ένα από τα σπουδαιότερα ζητήματα στο κλάδο της Βιολογίας τα τελευταία 60 χρόνια. Το 1962, το βραβείο Nobel Χημείας απονεμήθηκε στους Max Perutz και John Kendrew, για την πρωτοπόρα δουλειά τους πάνω στην δομή των πρωτεϊνών [1]. Το έργο τους έθεσε τα θεμέλια για τον κλάδο της δομικής βιολογίας [2], η οποία ερμηνεύει τις ιδιότητες που φέρουν οι πρωτεΐνες σε σχέση με τη δομή τους [3]. Στην διπλωματική αυτή, θα γίνει αρχικά μία αναφορά στη σημασία του προβλήματος καθώς και στο απαιτούμενο θεωρητικό υπόβαθρο (Κεφάλαιο 1). Στη συνέχεια θα ακολουθήσει μια σύντομη ιστορική αναδρομή με τις μεθόδους που έχει αντιμετωπιστεί το πρόβλημα έως σήμερα (Κεφάλαιο 2). Στο τρίτο και τέταρτο κεφάλαιο θα δει κανείς τη δικιά μας προσέγγιση, το νευρωνικό μοντέλο που προτείνουμε για την πρόβλεψη της δομής των πρωτεινών, καθώς και όλες τις λεπτομέριες σχετικά με τα δεδομένα, την αρχιτεκτονική και τα αποτελέσματα. Τέλος γίνεται σχολιασμός στα συμπεράσματα που προκύπτουν από τη παρούσα διπλωματική και μια σύντομη αναφορά σε μελλοντικές επεκτάσεις της.

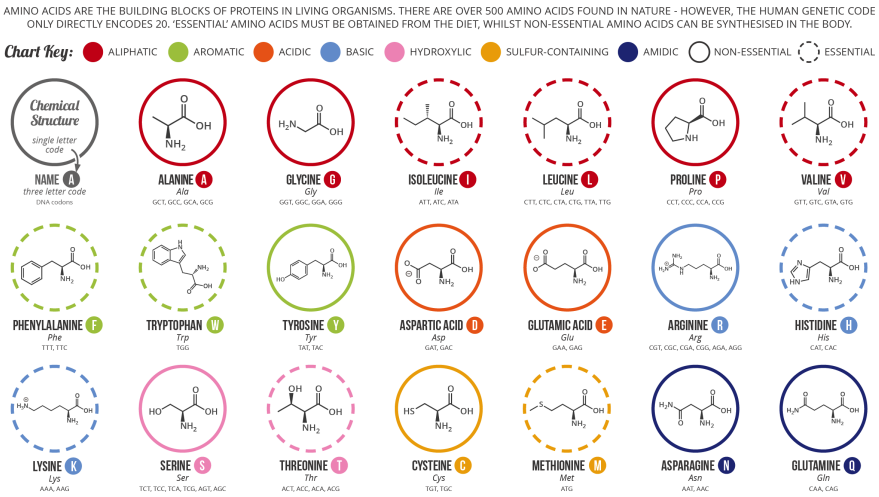
1 Η Πρωτεΐνη και τα Δομικά της Στοιχεία

Η πρωτεΐνη θεωρείται ένα από τα σημαντικότερα μόρια όλων των κυττάρων και αποτελεί το βασικό υλικό συστατικό όλων των ζωντανών οργναισμών. Οι λειτουργίες τους συναντόται σε όλα τα βήματα της ζωής. Από τη είσοδο και έξοδο ουσιών σε κύτταρα, μέχρι την λειτουργία των ζωτικών οργάνων [4], οι πρωτεΐνες συντελούν στην ομαλή λειτουργία του οργανισμνού [5][6]. Αν και ο ακριβής αριθμός των πρωτεϊνών στον άνθρωπο δεν έχει υπολογιστεί, εκτιμάται ότι είναι πάνω από 20.000 έως και εκατομμύρια, δεδομένου ότι

από ένα γονίδιο μπορεί να προκύψουν και 100 διαφορετικές πρωτεΐνες [7]. Μία πρωτεΐνη αποτελείται κατά βάση από μια αλληλουχία αμινοξέων. Αμινοξέα χαρακτηρίζουμε τις χημικές ενώσεις που περιέχουν τουλάχιστον μία καρβονική ομάδα (από τα καρβονικά οξέα RCOOH) και μία τουλάχιστον αμινομάδα NH_2. Δύο αμινοξέα συνδέονται με έναν χημικό δεσμό ο οποίος ονομάζεται πεπτιδικός δεσμός. Πρόκειται για έναν ομοιοπολικό δεσμό, άνθρακα (C)- αζώτου (N) που συνδέει την ομάδα καρβοξυλίου (COOH) ενός αμινοξέος με την αμινική ομάδα (NH_2) ενός διπλανού του, ελκύοντας ένα μόριο ύδατος (H_20). Με πεπτιδικούς δεσμούς μπορούν να ενωθούν πολλά αμινοξέα προκειμένου να δημιουργήσουν τελικά πρωτεΐνες. Στη φύση συνολικά έχουν παρατηρηθεί περίπου 500 διαφορετικά αμινοξέα, ωστόσο 20 είναι αυτά που εμφανίζονται στον γενετικό κώδικα του ανθρώπου και θα μας απασχολήσουν στην εργασία. Τα γράμματα που χρησιμοποιούμε ως κωδική ονομασία είναι τα εξής:

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

Ορισμένες φορές, για λόγους πληρότητας χρησιμοποιείται το γράμμα 'X' για να υποδηλώσει οποιο- δήποτε άλλο ή κάποιο άγνωστο αμινοξύ. Από τα 20, τα 9 θεωρούνται στοιχειώδη, καθώς το ανθρώπινο σώμα δεν μπορεί να τα συνθέσει και λαμβάνονται αποκλειστικά από τις τροφές [8]. Στον άνθρωπο το μέσο μήκος μιας πρωτεΐνης ανέρχεται στα 375 αμινοξέα [9], με το μήκος αυτό να κυμαίνεται από μερικές δεκάδες αμινοξέα έως και δεκάδες χιλιάδες.



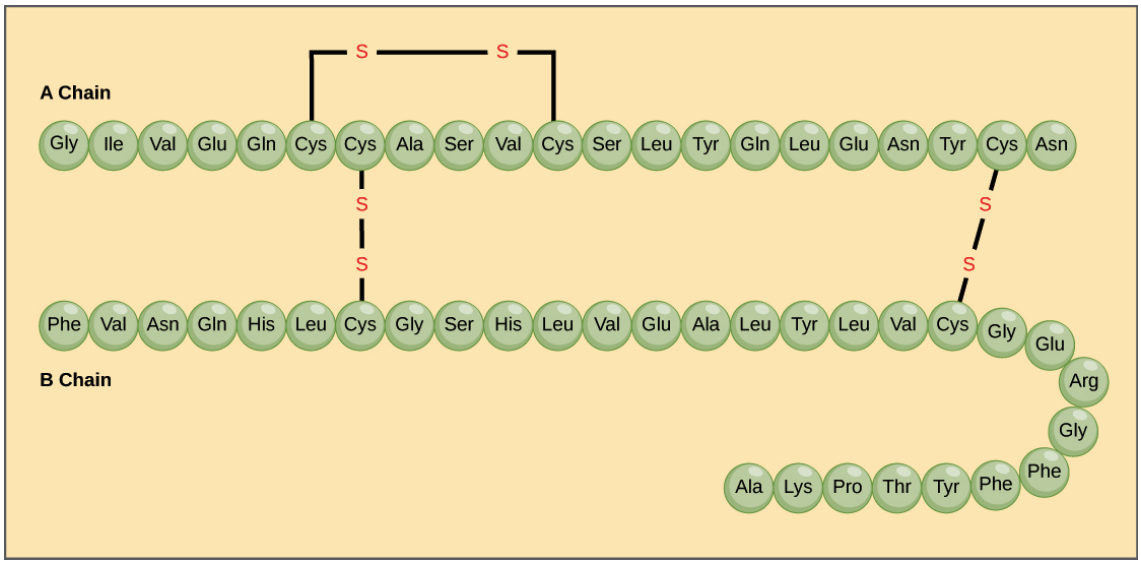
Σχήμα 1.1: Τα Αμινοξέα των που Συνθέτουν τις Πρωτεΐνες.

2 Τα Επίπεδα Δομής της Πρωτεΐνης

Προσπαθώντας να μελετήσουν την πρωτεΐνη, οι επιστήμονες διακρίνουν 4 δομικά στάδια. Τα στάδια αυτά αναφέρονται τόσο στη διαδικασία δημιουργίας μιας πρωτεΐνης, όσο και στη κλίμακα μεγέθους με την οποία τις μελετάμε.

2.1 Η Πρωτοταγής δομή

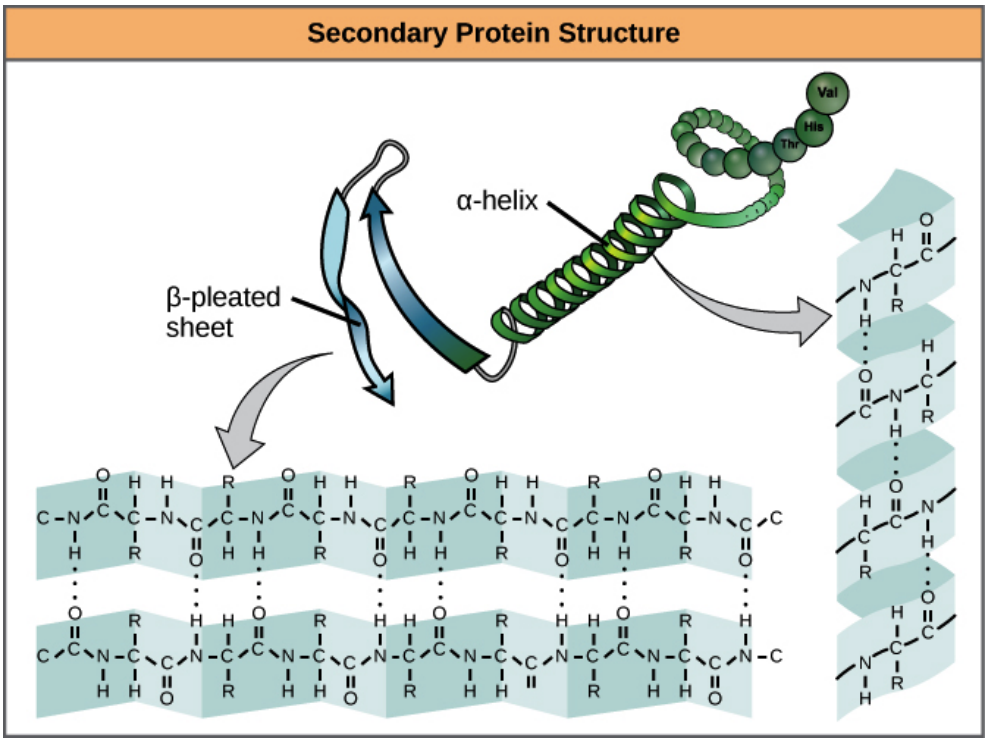
Όπως αναφέραμε και παραπάνω η πρωτοταγής δομής αναφέρεται στην αλληλουχία των αμινο- ξέων που είναι ενωμένα με πεπτιδικούς δεσμούς, δημιουργώντας έτσι μία αλυσίδα πολυπεπτιδίων ώστε να σχηματιστεί η πρωτεΐνη. Μια πρωτεΐνη μπορεί να αποτελείται από μια ή περισσότερες τέτοιες αλυσίδες.



Σχήμα 1.2: Η Πρωτοταγής Δομή 2 Πρωτεΐνης με 2 Αλυσίδες

2.2 Η Δευτεροταγής δομή

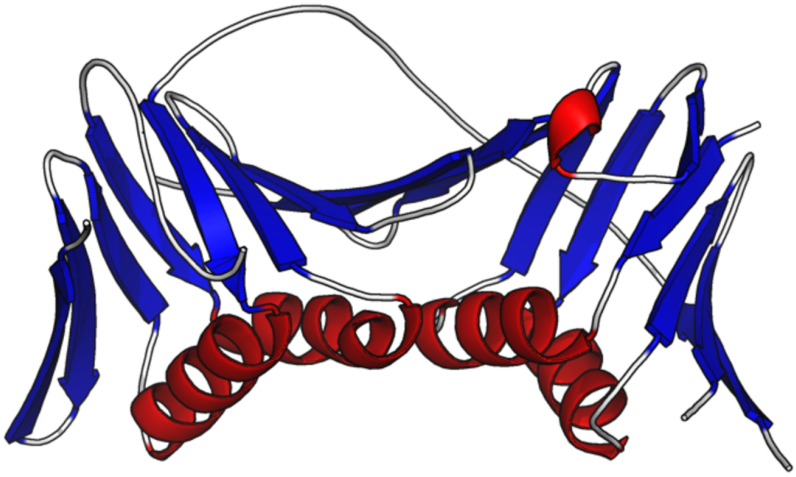
Η δευτεροταγής δομή αποτελεί το δεύτερο στάδιο. Αναφέρεται στις τοπικές αναδιπλώσεις μέσα σε μια αλληλουχία. Υπάρχουν 3 τύποι τοπικών αναδιπλώσεων [10]. Η **α-έλικα**, μία σπειροειδής αναδίπλωση της αλυσίδας με ελικοειδή δομή με 3 έως 6 αμινοξέα ανά στροφή της έλικας. Οι διαδοχικές αυτές στροφές της άλφα έλικας συνδέονται με ασθενείς δεσμούς υδρογόνου με συνέπεια η δομή να είναι περισσότερο σταθερή από μια μη σπειροειδή αλυσίδα πολυπεπτιδίων. Η **β-πτύχωση** είναι το δεύτερο μοτίβο που συναντάται στη δευτεροταγή δομή. Όταν συνδέονται β-πτυχώσεις μεταξύ τους δημιουργούνται β-επιφάνειες. Συνήθως 4-5 στον αριθμό. Συνήθως οι πτυχώσεις εμπεριέχουν 3 έως 10 αμινοξέα. Τέλος αξίζει να αναφερθούμε στα **loops** την τρίτη κατηγορία που μπορούν να ανήκουν τα αμινοξέα, που είναι κατά κύριο λόγο ακανόνιστες δομές που ενώνουν δύο άλλες δευτεροταγείς δομές μεταξύ τους [11].



Σχήμα 1.3: Η Δευτεροταγής Δομή Πρωτεΐνης με τις δύο Χαρακτηριστικές Δομές που περιέχει

2.3 Η Τριτοταγής δομή

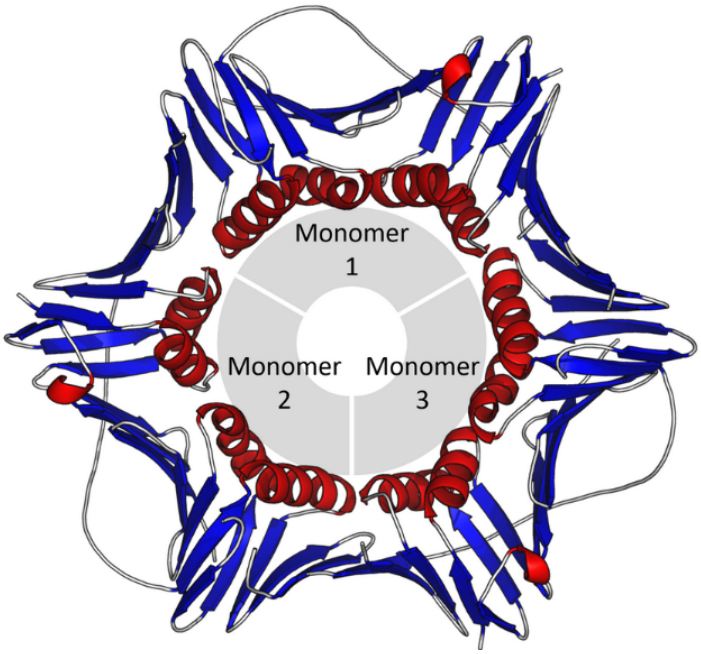
Σε αυτό το σημείο, τα πολυπεπτίδια υφίστανται ακόμα ένα δίπλωμα όπως συνέβη από την πρωτοταγή στη δευτεροταγή, στο αποτέλεσμα του οποίου αναφερόμαστε ως τριτοταγής δομή. Με τον όρο τριτοταγή δομή, εννοούμε το τελικό σχήμα που αποκτά το πολυπεπτίδιο, το οποίο πλέον θεωρείται μία πλήρως λειτουργική πρωτεΐνη. Αυτή η αναδίπλωση πραγματοποιείται από την αλληλεπίδραση των πλευρικών ομάδων των αμινοξέων (π.χ. σχηματισμός δισουλφιδικών δεσμών μεταξύ δύο κυστεϊνικών καταλοίπων). Ο χρόνος που χρειάζεται για το δίπλωμα και τελικά για την "δημιουργία" της πρωτεΐνης εξαρτάται από την πολυπλοκότητα της πρωτεΐνης και τα είδη των δεσμών της. Γενικά οι χρόνοι κυμάινονται σε τάξεις milliseconds, microseconds.



Σχήμα 1.4: Η Τριτοταγής Δομή Πρωτεΐνης

2.4 Η Τεταρτοταγής δομή

Πολλές πρωτεΐνες κατασκευάζονται από μία πολυπεπτιδική αλυσίδα και επομένως τα τρία παραπάνω δομικά επίπεδα. Ορισμένες πρωτεΐνες αποτελούνται από περισσότερες από μία αλυσίδες (υπομονάδες). Όταν αυτές οι υπομονάδες ενώνονται, δίνουν στην πρωτεΐνη την τεταρτοταγή δομή της.



Σχήμα 1.5: Η Τεταρτοταγής Δομή Πρωτεΐνης

3 Τι Προσφέρει η Μελέτη της Δομής των Πρωτεϊνών και η πρόβλεψη τους;

Για να γίνει αντιληπτή η σημασία του ζητουμένου, πρέπει να απαντήσουμε σε δύο σημαντικά ερωτήματα:

- Γιατί είναι σημαντικό να γνωρίζουμε **πως** αναδιπλώνονται οι πρωτεΐνες;
- Κατά **πόσο** είναι χρήσιμη η πρόβλεψη της δομής καθαυτής;

Όσο δύσκολο είναι το πρόβλημα της αναδίπλωσης των πρωτεϊνών, άλλο τόσο εύκολη είναι η απάντηση στο πρώτο ερώτημα. Οι ιδιότητες και τα χαρακτηριστικά μίας πρωτεΐνης καθορίζονται από το σχήμα της. Αυτό συνεπάγεται ότι αν οι πρωτεΐνες είναι πράγματι άξιες μελέτης, τότε ο τρόπος που αποκτούν την δομή τους και συνεπώς τις ιδιότητές τους αποτελούν ένα πολύ σημαντικό κεφάλαιο ανάλυσης. Μένει να διερευνήσουμε κατά πόσο είναι σημαντική η συνεισφορά τους στους οργανισμούς.

Σε ένα τυπικό ανθρώπινο κύτταρο υπάρχουν από 20.000 έως 100.000 μοναδικές πρωτεΐνες. Η κάθε μια εκτελεί ένα συγκεκριμένο καθήκον. Κάποιες είναι δομικές, και δίνουν τα χαρακτηριστικά ακαμψίας και ελαστικότητας σε μύες και νευρώνες. Άλλες προσδένονται σε συγκεκριμένου τύπου μόρια και αποστέλλονται σε νέες τοποθεσίες και άλλες επιταγχύνουν αντιδράσεις που επιτρέπουν στα κύτταρα να διαιρούνται και να αναπτύσσονται. Πίσω από κάθε διεργασία του οργανισμού μας, κρύβεται μία ομάδα πρωτεϊνών που εξασφαλίζει την ομαλή λειτουργία της διεργασίας αυτής. Η τεράστια ποικιλία και σημασία των πρωτεϊνών φαίνεται να πηγάζει από μία απλή ιδιότητά τους: Αναδιπλώνονται [12].

Η δημιουργία των πρωτεϊνών ωστόσο δεν κυλάει πάντα ομαλά. Υπό ορισμένες συνθήκες η αναδίπλωση των πρωτεϊνών αποτυγχάνει. Γονιδιακές μεταλλάξεις, φυσικά λάθη κατά τη δημιουργία αλυσίδων αμινοξέων, και εχθρικές περιβαλλοντικές συνθήκες κατά τη δημιουργία τους, είναι οι σημαντικότερες αιτίες που οι πρωτεΐνες δεν αναδιπλώνονται με τον προσδοκούμενο τρόπο. Με αυτό τον τρόπο προκύπτουν πολλές γνωστές ασθένειες, με τους επιστήμονες να υποστηρίζουν ότι είναι πολύ περισσότερες εκείνους που τελικά σχετίζονται με το πρόβλημα αυτό. Η αποτυχία αυτή δημιουργεί δύο είδους προβλήματα.

Το πρώτο αναφέρεται ως "Απώλεια λειτουργίας", το οποίο συμβαίνει όταν υπάρχει έλλειψη σε πλήθος κάποιων ομάδων πρωτεϊνών που χρειάζονται για ένα συγκεκριμένο καθήκον. Έτσι μπορεί μια λειτουργία να μην εκτελείται κατά το προσδοκόμενο όπως η διάσπαση τοξινών στα κύτταρα, ο μεταβολισμός ζάχαρης, αλκόολης κλπ. Ασθένειες όπως Κυστική Ίνωση, το σύνδρομο Marfan και ορισμένες μορφές καρκίνου, είναι παραδείγματα ασθενειών που προκαλούνται όταν ένας τύπος πρωτεΐνης δεν μπορεί να φέρει εις πέρας την λειτουργία του. Ας αναλογιστούμε ότι αρκεί μία από τις δεκάδες χιλιάδες πρωτεΐνες να μην αναδιπλώνεται σωστά για να συμβεί αυτό.

Το δεύτερο πρόβλημα έχει να κάνει με την επιρροή της υγείας ενός κυττάρου ανεξάρτητα από τη δομή της πρωτεΐνης. Συμβαίνει όταν λόγο της εσφαλμένης αναδίπλωσης, υδρο-

φοβικά σημεία της πρωτεΐνης καταλήγουν στην εξωτερικής της πλευρά και ενώνονται μεταξύ τους δημιουργώντας μία συσσώρευση. Τέτοιες ανωμαλίες συνδέονται με διάφορες νευρολογικές παθήσεις όπως η νόσος Alzheimer's,Parkinson's και Lou Gehrig's (ALS).

Όλα τα παραπάνω, μας δείχνουν ότι η σε βάθος αντίληψη της λειτουργίας των πρωτεϊνών θα μας οδηγήσει στην καλύτερη αντιμετώπιση των σημαντικότερων ασθενειών που αφορούν τον Άνθρωπο. Ακόμα περισσότερο όμως, θα μας βοηθήσουν στη κατανόηση των τρόπων με τους οποίους ο ανθρώπινος οργανισμός λειτουργεί και θα μας οδηγήσει στη δημιουργία τεχνητών πρωτεϊνών που θα καλύπτουν και θα ενισχύουν τις τρέχουσες λειτουργίες των οργανισμών.

Από την πειραματική σκοπιά, η μελέτη της δομής μιας πρωτεΐνης χρειάζεται χρόνο, εξειδικευμένο προσωπικό και κοστίζει ακριβά. Η μέθοδος που χρησιμοποιείται είναι η κρυσταλλογραφία ακτίνων-Χ. Με την εκπομπή ακτίνων-Χ προς διάφορες κατευθύνσεις και διαφορετικές γωνίες, η κρυσταλλογραφία ακτίνων-Χ παράγει μια τρισδιάστατη εικόνα πυκνότητας ηλεκτρονίων. Από την πυκνότητα των ηλεκτρονίων στο χώρο, υπολογίζεται η θέση των ατόμων καθώς και των χημικών δεσμών. Η μέθοδος Η λεπτομέρια που εμπεριέχει η διαδικασία αναδίπλωσης, καθιστά την προσομοίωση του πρόβληματος ένα εξαρετικά σύνθετο πρόβλημα. Τα βήματα και οι αλλαγές που συμβαίνουν μεταξύ των μορίων συμβαίνουν σε χρονικές κλίμακες μικρότερες από nanosecond. Αυτό σημαίνει ότι τα ιδρύματα και οι εταιρείες που ασχολούνται με το πρόβλημα αυτό καλούνται να επενδύουν τεράστια ποσά σε υπολογιστική δύναμη. Ακόμα και έτσι τα αποτελέσματα χρειάζονται σημαντικό χρόνο και δεν είναι πάντα με απόλυτη ακρίβεια. Το πλήθος των πρωτεϊνών που ανακαλύπτονται κάθε χρόνο είναι πολλές φορές περισσότερο από εκείνες που μελετόνται πλήρως κάθε χρόνο [13]. Η διαφορά αυτή όλο και μεγαλώνει κάθε χρόνο καθώς ο ρυθμός αύξησης εύρεσης νέων δομών είναι μεγαλύτερος από το ρυθμό αύξησης των μελετημένων. Για αυτό το λόγο η πρόβλεψη της δομής προσφέρει μια εξαιρετικά γρηγορότερη, φτηνότερη λύση στο πρόβλημα της αναδίπλωσης των πρωτεϊνών.

4 Βιβλιογραφία

[1] J. C. Kendrewet al.,Nature181, 662 (1958)

[2] J. C. Kendrewet al.,Nature185, 422 (1960).

[3] M. F. Perutzet al.,Nature185, 416 (1960).

[4] https://courses.lumenlearning.com/wm-biology1/chapter/reading-function-of-proteins/

[5] Lin Tang Liu, Wen Dong, Shaowen Yao, et al., An overview of topic modeling and its current applications in bioinformatics, Springerplus 5 (1) (2016) 1608.

[6] A. Saini, J. Hou, Progressive clustering based method for protein function

[7] Ponomarenko EA, Poverennaya EV, Ilgisonis EV, et al. The Size of the Human Proteome: The Width and Depth. Int J Anal Chem. 2016;2016:7436849. doi:10.1155/2016/7436849 prediction, Bull. Math. Biol. 75 (2) (2013) 331–350.

[8] Wikipedia: Aminoacid Occurrence and functions in biochemistry

[9] http://book.bionumbers.org/how-big-is-the-average-protein/

[10] Kendrew JC, Dickerson RE, Strandberg BE, Hart RG, Davies DR, Phillips DC, Shore VC (February 1960). "Structure of myoglobin: A three-dimensional Fourier synthesis at 2 Å resolution". Nature. 185 (4711): 422–7. Bibcode:1960Natur.185..422K.

[11] Choi Y, Agarwal S, Deane CM. How long is a piece of loop?. PeerJ. 2013;1:e1. Published 2013 Feb 12. doi:10.7717/peerj.1

[12] http://sitn.hms.harvard.edu/flash/2010/issue65/

[13] https://www.dnastar.com/blog/structural-biology/why-structure-prediction-matters/

Κεφάλαιο 2

Ιστορική Αναδρομή PSSP

1 Οι Τρεις Εποχές Πρόβλεψης των Πρωτεϊνών

Στη παρούσα διπλωματική εργασία θα ασχοληθούμε με το πρόβλημα **Πρόβλεψη Δευτεροταγούς δομής Πρωτεϊνών** ή **Protein Secondary Structure Prediction** ή **PSSP**. Το PSSP ανφέρεται στην πρόβλεψη της δευτεροταγούς δομής από την πρωτοταγή. Με άλλα λόγια δεδομένου της ακολουθίας αμινοξέων (παραδείγματος χάρη NPVVHFF), να προβλεφθεί σε τι κατηγορία θα ανήκει το κάθε αμινοξύ κατα τη δευτεροταγή δομή. Οι επιστημονική κοινότητα χωρίζει της κατηγορίες της δευτεροταγούς δομής σε 3: Helix, Strand, Loop ή αργότερα αναλυτικότερα σε 8: HGI, EB, STC υποκατηγορίες των 3 αντίστοιχα [1]. Στο κεφάλαιο θα δούμε τους τρόπους με τους οποίους έχει μελετηθεί το πρόβλημα έως σήμερα καθώς και τα αποτελέσματα του.

2 Η Πρώτη Γενιά Πρόβλεψης

Η πρώτη εποχή πρόβλεψης της δομής των πρωτείνών, εντοπίζεται από τα τέλη της δεκαετίας του '60 έως τα τέλη της δεκαετίας του '80. Οι πρώτες μεθόδοι ήταν περιορισμένες στην πρόβλεψη των τριών κατηγοριών. Ήταν βασισμένες στην τάση ορισμένων αμινοξέων να δημιουργούν έλικες ή πτυχώσεις, ενώ ορισμένες φορές στην πρόβλεψη εμπεριέχονταν κανόνες σχτεικά με την απαιτούμενη ενέργεια που χρειάζεται για να δημιουργηθούν οι δευτεροταγείς δομές [2]. Ως δεδομένα οι πρώτες μέθοδοι λάμβαναν την ακολουθία των αμινοξέων και οι προβλέψεις βασιζόντουσαν κυρίως σε στατιστική ανάλυση δομών που ήδη υπήρχαν. Αυτό σήμαινε ότι σημαντικό μέρος της επιτυχίας της πρόβλεψης βασιζόταν στο ότι έχουν μελετηθεί ήδη πρωτεΐνες με παρόμοια δομή και λειτουργίες. Σε περίπτωση μίας πρωτεΐνης με πολύ χαμηλό ποσοστό ομολογίας με άλλες, τότε η πρόβλεψη βασιζό-

ταν σε απλά στατιστικά δείγματα των προηγουμένων, κάτι το οποίο δεν είχε αποδοτικά αποτελέσματα. Η πιο αντιπροσωπευτική μέθοδος της πρώτης γενιάς είναι εκείνη των Chou-Fasman [3].

2.1 Το Μοντέλο Chou-Fasman

Η μέθοδος αυτή ανέλυε τις σχετικές συχνότητες κάθε αμινοξέου στις τρεις κατηγορίες της δευτεροταγούς δομής που εμφανίζονται σε μελετημένες πρωτεΐνες από κρυσταλλογραφία ακτίνων Χ. Από τις συχνότητες παράγεται ένα σύνολο παραμέτρων, οι οποίες με τη σειρά τους χρησιμοποιούνται για να υπολογιστεί αν μια υπακολουθία αμινοξέων θα ανήκει σε α-έλικα, β-πτύχωση ή loop [4]. Η μέθοδος Chou-Fasman προβλέπει έλικες και πτυχώσεις με παρόμοιο τρόπο. Πρώτα αναζητά γραμμικά για έναν "πυρήνα" αμινοξέων με υψηλή πιθανότητα για έλικα ή πτύχωση. Στη συνέχεια επεκτείνεται γύρω από τον πυρήνα αυτό μέχρι 4 συνεχόμενα αμινοξέα να μην ξεπερνούν πάνω από το πιθανοτικό όριο της έλικας ή της πτύχωσης. Σχετικά με τα loops, αυτά προβλέπονται όταν σε μια ομάδα 4 ή περισσοτέρων αμινοξέων η πιθανότητα να εμφανιστεί loop είναι μεγαλύτερη από την πιθανότητα να εμφανιστεί έλικα ή πτύχωση ενώ είναι απαραίτητο το γινόμενο των πιθανοτήτων των επιμέρους αμινοξέων να βρίσκονται σε loop να ξεπερνάει ένα πιθανοτικό όριο [5] της τάξεως του 0.0075. Η μέθοδος αυτή υπολογίστηκε ότι έδινε ποσοστά επιτυχούς πρόβλεψης κοντά στο 60%. Ωστόσο στις αρχές της δεκαετίας του '80 το ποσοστό αυτό υπολογίστηκε λίγο πάνω από 50% [6].

3 Η Βελτίωση των Τεχνικών Πρόβλεψης

Κατά την δεκαετία του '80, οι μέθοδοι πρόβλεψης της δομής των πρωτεϊνών σημείωσαν βελτίωση, που ωστόσο ακόμα δεν ήταν ικανοποιητική για να χρησιμοποιείται ως αξιόπιστο εργαλείο. Σε αντίθεση με την πρώτη γενιά, τα μοντέλα πλέον όχι απλά χρησιμοποιούσαν πληροφορίες των γειτονικών αμινοξέων αλλά συμπεριλάμβαναν μαθηματικοποιημένα, φυσικοχημικές πληροφορίες των αμινοξέων [7]. Η χαρακτηριστική μέθοδος που σηματοδότησε την βελτίωση των προβλέψεων, είναι εκείνη των *Garnier-Osguthorpe-Robson* ή GOR [8].

3.1 Το Μοντέλο GOR

Η μέθοδος GOR είναι μία μέθοδος που αναπτύχθηκε στα τέλη της δεκαετίας του '70. Αυτό που κάνει την διαφορά είναι ότι η μέθοδος αυτή περιλαμβάνει και τις υπό συνθήκη πιθανότητες τα γειτονικά αμινοξέα να ανήκουν στον ίδιο τύπο δομής [9]. Πρακτικά εμπεριέχται η ουσία της Μπεϋζιανής φιλοσοφίας στο μοντέλο, κάτι αρκετά πρωτοπόρο αν αναλογιστεί κανείς πως κατά τη δεκαετία του '70, τα πορίσματα του Μπέυζ ήταν ακόμη αμφιλεγόμενα. Στις πρώτες μορφές του, το μοντέλο χρησιμοποιούσε ένα κινούμενο παράθυρο μήκους 17 αμινοξέων. Το μήκος αυτό θα συνεχίσει να χρησιμοποιείται μέχρι και σήμερα καθώς προκύπτει από στατιστικά δεδομένα ότι η κατηγοριοποίηση εξαρτάται σε μεγάλο βαθμό από τα 16 γειτονικά αμινοξέα. Υπήρχαν πίνακες σκορ 17x20 με τις πιθανότητες να βρεθεί το δεδομένο αμινοξέο σε κάθε θέση στην ακολουθία των 17 [10]. Αν και αποτέλεσε βελτίωση της μεθόδου των Chou-Fasman, οι επιτυχία του κυμαινόταν κοντά στο 65%.

4 Η Τρέχουσα Γενιά Μοντέλων

Από τις αρχές της δεκαετίας του '90 και έπειτα, οι περισσότερες τεχνικές άρχισαν να εμπεριέχουν όλο και πιο πολλά χαρακτηριστικά στα δεδομένα εισόδου, κάτι το οποίο έφερε αξιοσημείωτες βελτιώσεις στα ποσοστά πρόβλεψης. Συγκεκριμένα χρησιμοποιούν-

ται εξελικτικές πληροφορίες από πολλαπλές ομόλογες ακολουθίες [11]. Πολλά μοντέλα βασιζόμενα στις προηγούμενες τεχνικές, υλοποιημένα με Μπεϋζιανά ή κρυφά Μαρκοβιανά δίκτυα, έφεραν αύξηση στα ποσοστά πρόβλεψης της δευτεροταγούς δομής. 30 χρόνια αργότερα και έχοντας μια πλήρη εικόνα για τα μοντέλα που αναπτύχθηκαν αυτά τα χρόνια, παρατηρούμε ότι τα μοντέλα της μηχανικής μάθησης ήταν εκείνα που επικράτησαν, έχοντας σαφώς μεγαλύτερα ποσοστά ορθότητας. [12].

4.1 Αντιπροσωπευτικά Μοντέλα

Παρατίθονται παρακάτω ορισμένα μοντέλα τα οποία βασίζονται σε διαφορετικές φιλοσοφίες μηχανικής μάθησης. Στο επόμενο κεφάλαιο θα μιλήσουμε αναλυτικά για τη μηχανική μάθηση και θα εξηγήσουμε τις ιδιαιτερότητες που τα διαφοροποιούν από τα υπόλοιπα σύγχρονα μοντέλα.

Έτος	Μέθοδος	Q3 Ορθότητα	Αναφορά
1994	NN(3 layers)	72%	[13]
2006	BRNN	73.1%	[14]
2005	RBFNN	77.4%	[15]
2015	DBN	80.7%	[16]
2020	DeepLearning*	83%	[17]

Πίνακας 2.1: Τα Ποσοστά Πρόβλεψης Νευρωνικών τα Τελευταία 26 χρόνια.

Όπως βλέπουμε και στον πίνακα, τα πρώτα χρόνια της γενιάς των Νευρωνικών Μοντέλων, οι υλοποιήσεις ήταν απλές σε σχέση με αυτές των τελευταίων ετών. Συγκεκριμένα τα πρώτα μοντέλα αποτελούνταν από πυκνά στρώματα που όπως θα δούμε και στο επόμενο κεφάλαιο, δεν ήταν η βέλτιστη επιλογή.

Στα επόμενα χρόνια παρατηρήθηκε μια μεγάλη αύξηση στην χρήση **Αναδρομικών Νευρωνικών Δικτύων** ή **Recurrent Neural Networks** και συγκεκριμένα των **Αμφίδρομων Αναδρομικών Νευρωνικών Δικτύων** ή **Bidirectional Recurrent Neural Networks**. Η βασική ιδέα των αμφίδρομων αναδρομικών δικτύων είναι να

χρησιμοποιούν τα δεδομένα εισόδου τόσο προς τα εμπρός όσο και προς τα πίσω κάνοντας χρήση δύο χωριστών αναδρομικών κρυφών στρωμάτων, τα οποία αμφότερα είναι συνδεδεμένα με το ίδιο στρώμα εξόδου. Αυτή η δομή με αυτόν τον τρόπο, παρέχει στο επίπεδο εξόδου ένα πλήρες στιγμιότυπο των παρελθοντικών και των μελλοντικών εισόδων για κάθε χρονικό βήμα στην ακολουθία εισόδου. Αυτή η ιδιότητα αποδείχθηκε πολύ χρήσιμη για το πρόβλημα αυτό, όπως και για κάθε πρόβλημα του οποίου η είσοδος εξαρτάται και από προηγούμενα αλλά και από επόμενα δεδομένα. Στην προκειμένη περίπτωση αναφερόμαστε στη σχέση των αμινοξέων μεταξύ τους.

Παράλληλα μια άλλη κατηογρία μοντέλων σημείναν ενθαρρυντικά αποτελέσματα, εκείνη των **Νευρωνικών Δικτύων Ακτινικών Συναρτήσεων Βάσης** ή **Radial Basis Function Neural Network**. Τα δίκτυα αυτά χρησιμοποιούν ειδικές συναρτήσεις σε συνδυασμό με τις υπόλοιπες παραμέτρους του δικτύου προκειμένου να κατηγοριοποιήσουν τα δεδομένα.

Τα τελευταία χρόνια ωστόσο, τα περισσότερα Νευρωνικά Μοντέλα αποτελούν συνδυασμό περισσοτέρων νευρωνικών συνδεδεμένων μεταξύ τους. Χρησιμοποιώντας **Βαθιά Μάθηση** ή **Deep Learning**, μοντέλα όπως το **Δίκτυο Βαθιάς Πεποιθήσεως** ή **Deep Belief Network** στον πίνακα (4), συνδυάζουν πολλαπλά δίκτυα για την επιτυχή κατηγοριοποίηση. Η ιδιαιτερότητα των DBN είναι ότι ενώ τα κρυφά στρώματα συνδέονται μεταξύ τους, οι μονάδες εντός στρώματος δεν ενώνονται μεταξύ τους.

Τέλος υπάρχουν και ορισμένα μοντέλα τα οποία χρησιμοποιούν τεχνικές **Αυτόματης Κωδικοποιήσης** ή **Auto-Encoded Models**, τα οποία στοχεύουν στην εκμάθηση μίας αναπαράστασης για ένα σύνολο δεδομένων με σκοπό τη μείωση της διαστατικότητας των δεδομένων ή την αφαίρεση "θορύβου".

5 Βιβλιογραφία

[1] RostB. Review: protein secondary structure prediction continues to rise. J Struct Biol 2001;134:204–18.

[2] Scheraga HA. Structural studies of ribonuclease.3. A model for the secondary and tertiary structure. J Am Chem Soc 1960;82:3847–52.

[3] Chou PY, Fasman GD (Jan 1974). "Prediction of protein conformation". Biochemistry. 13 (2): 222–45.

[4] Chou PY, Fasman GD (1978). "Empirical predictions of protein conformation". Annual Review of Biochemistry. 47: 251–76.

[5] Chou PY, Fasman GD (1978). "Prediction of the secondary structure of proteins from their amino acid sequence". Advances in Enzymology and Related Areas of Molecular Biology.

[6] Kabsch W, Sander C (1983). "How good are predictions of protein secondary structure?". FEBS Lett. 155 (2): 179–82.

[7] Yuedong Yang, Jianzhao Gao, Jihua Wang, Rhys Heffernan, Jack Hanson, Kuldip Paliwal, Yaoqi Zhou, Sixty-five years of the long march in protein secondary structure prediction: the final stretch?, Briefings in Bioinformatics, Volume 19, Issue 3, May 2018, Pages 482–494

[8] Garnier J, Osguthorpe DJ, Robson B (March 1978). "Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins". Journal of Molecular Biology. 120 (1): 97–120

[9] Garnier J, Osguthorpe DJ, Robson B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J Mol Biol 1978; 120:97–120.

[10] Mitchell, E. M., Artymiuk, P. J., Rice, D. W., Willett, P. (1990). Use of techniques derived from graph theory to compare secondary structure motifs in proteins. Journal of Molecular Biology, 212(1), 151-166.

[11] Rost, B., & Sander, C. (1993). Improved prediction of protein secondary structure by use of sequence profiles and neural networks. Proceedings of the National Academy of Sciences, 90(16), 7558-7562.

[12] Rost, B., & Sander, C. (1993). Improved prediction of protein secondary structure by use of sequence profiles and neural networks. Proceedings of the National Academy of Sciences, 90(16), 7558-7562.

[13] Rost B, Sander C, Schneider R. Redefining the goals of protein secondary structure prediction. J Mol Biol. 1994;235(1):13-26. doi:10.1016/s0022-2836(05)80007-5

[14] Chen, J., Chaudhari, N. Bidirectional segmented-memory recurrent neural network for protein secondary structure prediction. Soft Comput 10, 315–324 (2006). https://doi.org/10.1007/s00500-005-0489-5

[15] G.Z. Zhang, D.S. Huang, Y.P. Zhu, et al., Improving protein secondary structure prediction by using the residue conformational classes, Pattern Recognit. Lett. 26 (15) (2005) 2346–2352

[16] M. Spencer, J. Eickholt, J. Cheng, A deep learning network approach to ab initio protein secondary structure prediction, IEEE/ACM Trans. Comput. Biol. Bioinf. 12 (1) (2014) 103–112.

[17] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, Burkhard Rost ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing.

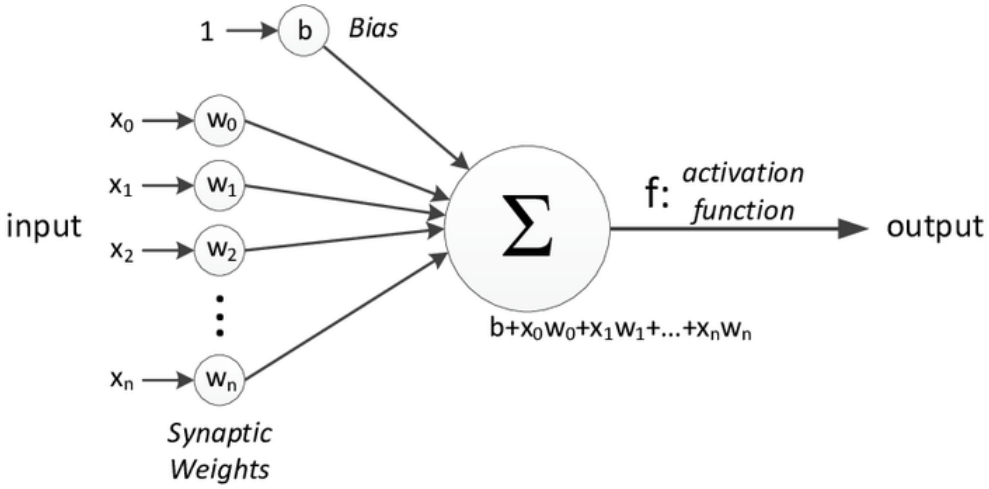
Κεφάλαιο 3

Τα Τεχνητά Νευρωνικά Δίκτυα και το Μοντέλο της Εργασίας

Τα Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks), αποτελούν ένα σύνολο αφηρημένων αλγοριθμικών κατασκευασμάτων το οποίο ανήκει στον ευρύτερο κλάδο της Τεχνητής Νοημοσύνης (Artificial Intelligence). Τα δίκτυα αυτά, εμπνέονται από τη δομή του Κεντρικού Νευρικού Συστήματος του ανθρώπου, δηλαδή τον τρόπο με τον οποίο ο εγκέφαλος επεξεργάζεται τις πληροφορίες, προκειμένου να προσομοιώσουν τη λειτουργία τους.

1 Η λειτουργία των νευρωνικών

Σύμφωνα με το μοντέλο των *McCulloch-Pitts* [1], τα Τεχνητά Νευρωνικά Δίκτυα αποτελούνται από διασυνδεδεμένους νευρώνες, οι οποίοι ανταλλάσουν πληροφορίες μεταξύ τους, τις περισσότερες φορές προωθώντας τιμές. Κάθε νευρώνας δέχεται ένα διάνυσμα $\mathbf{x}$, το οποίο πολλαπλασιάζεται με ένα διάνυσμα βαρών $\mathbf{w}$. Το διανυσμά βαρών δρα ως συντελεστής που δίνει διαφορετική αξία στην κάθε πληροφορία εισόδου. Στο γινόμενο συμπεριλαμβάνεται συνήθως ένα συναπτικό βάρος $\mathbf{b}$, το οποίο λειτουργεί ως την τιμή κατωφλίου που πρέπει να ξεπεράσει το γινόμενο. Το αποτέλεσματα αυτό οδηγείται σε μια Συνάρτηση Ενεργοποίησης, απο την οποία παράγεται τελικά η έξοδος $\mathbf{y}$ του νευρώνα. Η Συνάρτηση Ενεργοποίησης υιοθετεί τον ρόλο του συγκριτή προκειμένου να ενεργοποιηθεί ο νευρώνας και να μεταφέρει την πληροφορία στα επόμενα στρώματα.



Σχήμα 3.1: Η Αρχιτεκτοννική ενός απλού Νευρώνα

1.1 Η Συνάρτηση Ενεργοποίησης

Τα Νευρωνικά Δίκτυα χαρακτηρίζονται από την αρχιτεκτονική τους, τον τρόπο που μεταδίδεται η πληροφορία μέσα στο δίκτυο. Πέρα από αυτή, τα δίκτυα και η λειτουργία τους εξαρτάται από διάφορες παραμέτρους. Μία από αυτές είναι η Συνάρτηση Ενεργοποίησης την οποία αναφέραμε παραπάνω. Ενδεικτικά αναφέρουμε τις τέσσερις που βρίσκουν την ευρύτερη εφαρμογή:

Σιγμοειδής Συνάρτηση

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{3.1}$$

Υπερβολική Εφαπτομένη

$$f(x) = \tanh x = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{3.2}$$

Μονάδα Γραμμικής Ανόρθωσης

$$ReLU(x) = (0, max) \tag{3.3}$$

Υπερβολική Εφαπτομένη

$$Softmax(x_i) = \frac{e^{x_i}}{\sum_{i=0}^{N} e^{x_i}} \tag{3.4}$$

1.2 Είδη Εκμάθησης

Ένα κύριο χαρακτηριστικό των Τεχνητών Νευρωνικών Δικτύων είναι η δυνατότητα που έχουν να βελτιώνουν την ικανότητα τους για την επίλυση ενός προβλήματος. Η εκμάθηση των Δικτύων πραγματοποιείται μέσω της επαναληπτικής εκπαίδευσής τους και τον συνεχή επαναπροσδιορισμό των παραμέτρων τους. Τελικός στόχος είναι η επιτυχής γενίκευση, δηλαδή να μπορούν να κάνουν ορθές προβλέψεις για δεδομένα εισόδου, τα οποία δεν έχουν ξανασυναντήσει. Χωρίζουμε τις μεθόδους με τις οποίες πραγματοποιείται η εκμάθηση στις παρακάτω:

- **Επιβλεπόμενη μάθηση**

  Η Επιβλεπόμενη Μηχανική Μάθηση [2] χρησιμοποιεί αλγορίθμους που βασίζονται στην εκπαίδευση πάνω σε δεδομένα για τα οποία ήδη ξέρουμε την απάντηση του προβλήματος (ετικέτες). Για παράδειγμα στην εργασία αυτή θα χρησιμοποιήσουμε πρωτεΐνες των οποίων ξέρουμε ήδη τη δευτεροταγή δομή, προκειμένου να προβλέψουμε τη δομή πρωτεϊνών που δνε έχουν μελετηθεί ακόμα [3]. Συνήθως αρχικοποιούμε τυχαίες τιμές στα βάρη των νευρώνων και κατά την εκπαίδευση αυτά διορθώνονται, τροποποιούνται με βάση το σφάλμα, το πόσο απέχουμε από τη σωστή,

προσδοκόμενη απάντηση.

- **Μη Επιβλεπόμενη Μάθηση**

  Κατά την Μη Επιβλεπόμενη Μάθηση τα δεδομένα τα οποία δίνουμε στο Δίκτυο δεν έχουν κάποια ετικέτα, απάντηση στο πρόβλημα μας. Οι αλγόριθμοι εκπαιδεύονται με τέτοιο τρόπο ώστε να ανακαλύπτουν τα μοτίβα που υπάρχουν στα δεδομένα εκπαίδευσης και να βρίσκουν συσχετίσεις μεταξύ τους [4]. Δύο χαρακτηριστικές κατηγορίες προβλημάτων είναι το *Clustering*, στο οποίο το Δίκτυο καλείται να ομαδοποιήσει τα δεδομένα με τέτοιο τρόπο ώστε κάθε ομάδα να έχει παρόμοιες ιδιότητες στα μέλη της, αλλά διαφορετικά σε σχέση με της άλλες (π.χ. ομαδοποίηση κειμένων με βάση το περιεχόμενο τους). Η δεύτερη κατηγορία είναι η μείωση διαστάσεων, πρόβλημα σύμφωνα με το οποίο το δίκτυο καλείται να συμπτύξει το πλήθος των χαρακτηριστικών των δεδομένων μέσω της αφαίρεσης-ένωσης χαρακτηριστικών χωρίς να χάνεται η σημασία τους.

- **Ενισχυτική Μάθηση**

  Με την Ενισχυτική [5], το δίκτυο αλληλεπιδρά με ένα δυναμικό πριβάλλον, στο οποίο πράκτορες (software agents) καλούνται να πάρουν αποφάσεις στο περιβάλλον αυτό με σκοπό να μεγιστοποιήσουν τις ανταμοιβές που θέτονται, για παράδειγμα την αναμενόμενη τιμή του σήματος ενίσχυσης στο επόμενο βήμα. Το σύστημα δεν καθοδηγείται από κάποιον εξωτερικό επιβλέποντα για το ποια ενέργεια θα πρέπει να ακολουθήσει αλλά πρέπει να ανακαλύψει μόνο του ποιες ενέργειες είναι αυτές που θα του αποφέρουν το μεγαλύτερο κέρδος. Κλασσικά παραδείγματα Ενισχυτικής μάθησης, είναι η κίνηση των ρομποτ και η εκμάθηση δικτύων σε επιτραπέζια και ηλεκτρονικά παιχνίδια.

2 Τα Συνελικτικά Νευρωνικά Δίκτυα

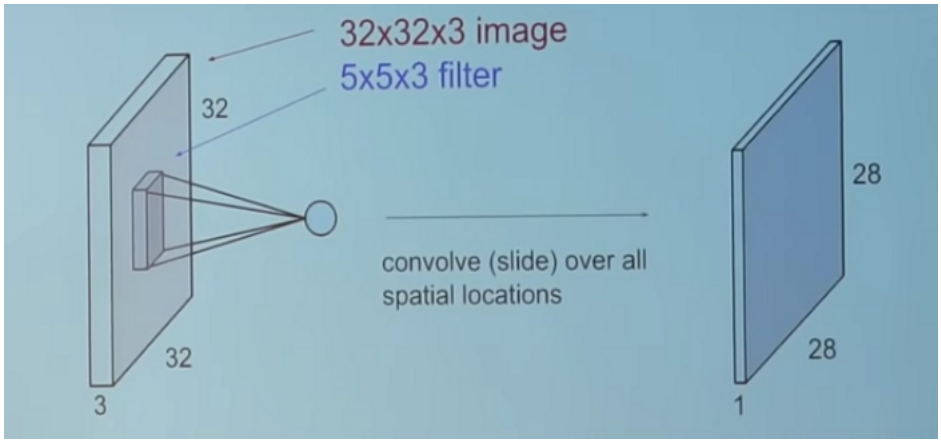
Όπως αναφέραμε παραπάνω, υπάρχουν διάφορα είδη Νευρωνικών Δικτύων τα οποία ενώ έχουν κοινή φιλοσοφία, διαφοροποιούνται ως προς τις δυνατότητές τους και η διαφορά αυτή έγκειται στην αρχιτεκτονική τους. Για τους σκοπούς της εργασίας αυτής αποφασίσαμε να υλοποιήσουμε ένα δίκτυο που ανήκει στην κατηγορία των Συνελικτικών Νευρωνικών Δικτύων (ΣΝΔ), ή Convolutional Neural Networks (CNN). Παρακάτω, θα κάνουμε μια αναφορά στα ΣΝΔ και στα χαρακτηριστικά τους, συγκρίνοντας τα με τα υπόλοιπα είδη Νευρωνικών Δικτύων. Στο επόμενο κομμάτι, θα γίνει εκτενής περιγραφή του συγκεκριμένου μοντέλου της εργασίας αυτής, θα εξηγήσουμε τους λόγους που πήραμε ορισμένες αποφάσεις σχετικά με τις παραμέτρους και την αρχιτεκτονική του. Τέλος θα μιήσουμε για κάποιες ιδέες που απορρίφθηκαν πάνω στο στήσιμο του μοντέλου καθώς και τους λόγους που οδηγηθήκαμε σε αυτή την απόφαση.

2.1 Η Πράξη της Συνέλιξης

Ένα ΣΝΔ αποτελεί ένα αλγόριθμό Βαθιάς Μηχανικής Μάθησης ο οποίος μπορεί να δεχθεί ως είσοδο δεδομένα σε μορφή εικόνας, να εντοπίσει χαρακτηριστικά και διαφορές αντικειμένων στην εικόνα και να είναι σε θέση τα αναγνωρίζει ένα αντικείμενο σε σχέση με ένα άλλο. Η αρχιτεκτονική ενός ΣΝΔ, είναι ανάλογη με την συνδετικότητα που παρουσιάζουν οι νευρώνες στον εγκέφαλο, και συγκεκριμένα έχουν επηρεαστεί σε μεγάλο βαθμό από τον οπτικό φλοιό. Οι νευρώνες δέχονται μονάχα κομμάτι της εικόνας που λαμβάνει το μάτι, και η υπέρθεση τους συνθέτει την εικόνα ολόκληρη [6]. Μια ερώτηση που χρειάζεται να απαντήσουμε είναι γιατί απλά πυκνά στρώματα νευρώνων δεν είναι αποτελεσματικά. Γιατί δηλαδή να μην χρησιμοποιηθεί ένα απλό δίκτυο με πολλά στρώματα **Αντίληπτρων ή Multi Layer Perceptron**; Στην προκειμένη περίπτωση θα χρειαζόταν να σειριοποιήσουμε τα δεδομένα. Αν λοιπόν έχουμε ως είσοδο μια εικόνα 28*28, θα χρειαζόταν να την περάσουμε ως ένα διάνυσμα 784*1. Η μετατροπή αυτή των δεδομένων, διαισθητικά

αφαιρεί την σχέση που έχουν τα γειτονικά πίξελ στην εικόνα μεταξύ τους. Φυσίκα το μεγάλο πλήθος νευρώνων θα είναι σε θέση να εντοπίσει αυτές τις σχέσεις, αλλά δεν είναι εγγυημένη η διατήρηση της συσχέτισης τους. Όταν αντιμετωπίζουμε προβλήματα που η είσοδος έχει τα χαρακτηριστικά και τις συσχετίσεις μιάς εικόνας τα Συνελικτικά Νευρωνικά Δίκτυα αποτελούν μια στοχευμένη λύση. Το ΣΝΔ με τη χρήση της συνέλιξης καταφέρνει με επιτυχία να αιχμαλωτίζει τις τοπικές και χρονικές σχέσεις σε μια εικόνα με τη συνέλιξη της εικόνας με φίλτρα σε σχέση με την σειριοποίηση των δεδομένων. Επίσης αποφεύγουμε όπως θα δούμε, το εξαιρετικό μεγάλο αριθμο παραμέτρων που θα χρειάζονταν σε ένα "πυκνό" δίκτυο. Ας αναλογιστούμε ότι ένα MLP για εικόνες ανάλυσης 4K, θα χρειαστεί περίπου 25 εκατομμύρια παραμέτρους μονάχα για το στρώμα εισόδου! Η επιλογή των ΣΝΔ λοιπόν, μειώνει τη μορφή της εικόνας, διατηρώντας τα κρίσιμα χαρακτηριστικά της, παρέχοντας τη δυνατότητα κλιμάκωσης σε μεγάλες εικόνες, datasets, αρχιτεκτονικές.

Ας υποθέσουμε ότι έχουμε μια **εικόνα-εισόδου** με $32 * 32 * 3$ pixels. Ορίζουμε ως **φίλτρο** ένα πίνακα με μικρότερες διαστάσεις $(5 * 5 * 3)$.



Σχήμα 3.2: Η Εικόνα Εισόδου, το Φίλτρο και ο Πίνακας Χαρακτηριστικών.

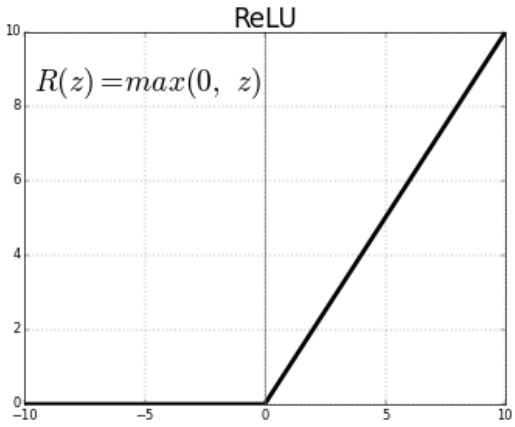
Για την εκτέλεση της συνέλιξης τοποθετούμε το φίλτρο πάνω στον πίνακα και το μετακινούμε σιγά σιγά κατά μήκος όλης της εικόνας-εισόδου. Σε κάθε μετακίνηση υπολογίζουμε το γινόμενο των δύο πινάκων και προκύπτει ο **Πίνακας Χαρακτηρηστικών**

ή **Feature Map**. Τα χαρακτηριστικά του φίλτρου είναι πολύ σημαντικά για το τι θέλουμε να πετύχουμε. Οι τιμές του φίλτρου θα ορίσουν πιθανά χαρακτηριστικά που θέλουμε να εντοπίσουμε στην εικόνα. Το **βάθος** του φίλτρου αντιστοιχεί στο πλήθος των διαφορετικών φίλτρων που θέλουμε να χρησιμοποιήσουμε στο συγκεκριμένο στρώμα. Με αυτό τον τρόπο προκύπτουν πολλαπλοί Πίνακες χαρακτηριστικών. Η **δρασκελιά** του φίλτρου ή **stride**, αναφέρεται στο πλήθος των πίξελ που προχωράει σε κάθε βήμα υπολογισμού της συνέλιξης. Όσο μεγαλύτερη η δρασκελιά τόσο μικρότεροι οι Πίνακες Χαρακτηριστικών που παράγονται. Μία ακόμη επιλογή είναι το **Γέμισμα μηδενικών** ή **zero-padding**, κατά το οποίο η εικόνα είσοδος γεμίζει περιμετρικά με μηδενικά, ώστε να μπορούμε να φιλτράρουμε καλύτερα τα συνοριακά στοιχεία του πίνακα και να έχουμε έλεγχο στις διαστάσεις του Πίνακα (wide/narrow Convolution).

2.2 Οι Συναρτήσεις Ενεργοποίησης

Όπως αναφέραμε στο προηγούμενο κεφάλαιο, οι Συναρτήσεις Ενεργοποίησης, επιτελούν πολύ σημαντικό ρόλο στη διάδοση πληροφοριών στο Νευρωνικό Δίκτυο, δρώντας ως έμμεσος συγκριτής που αποφασίζει ποιές τιμές θα μεταδωθούν στα μετέπειτα στάδια και ποιές όχι. Ανάλογα με τη επιλογή μας μπορούμε να προσδώσουμε σπουδαία χαρακτηρηστικά στο Δίκτυο. Στην περίπτωση των ΣΝΔ, η πιο αποτελεσματική συνάρτηση είναι τα τελευταία χρόνια [7] η **Μονάδα Γραμμικής Ανόρθωσης**:

$$ReLU(x) = (0, max) \tag{3.5}$$

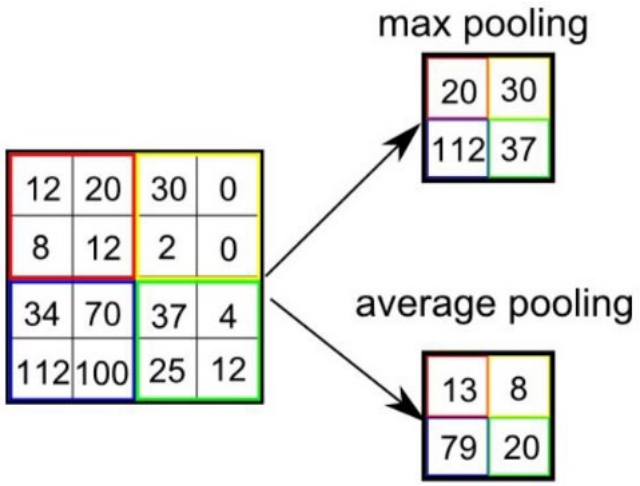


Σχήμα 3.3: Η Συνάρτηση Μονάδας Γραμμικής Ανόρθωσης.

Η συγκεκριμένη συνάρτηση, μας επιτρέπει να προσδώσουμε έναμη γραμμικό χαρακτήρα στον τρόπο λήψης αποφάσεων του Δικτύου [8]. Στην πράξη, αντικαθιστά στον Πίνακα Ενεργοποίησης όλα τα πίξελ αρνητικών τιμών με μηδενικά. Ένα σημαντικό προτέρημα σε σχέση με άλλες μη-γραμμικές όπως η υπερβολική εφαπτομένη και τη σιγμοειδή συνάρτηση, ανταποκρίνεται καλύτερα στο πρόβλημα της εξαφανιζόμενης κλίσης (Vanishing Gradient Problem)[9], ενώ ταυτόχρονα ο υπολογισμός της συνάρτησης είναι πολύ πιο γρήγορος και επιτρέπει πιο γρήγορους χρόνους εκπαίδευσης.

2.3 Συγκέντρωση (Pooling)

Ως τώρα λοιπόν έχουμε την Εικόνα Είσοδο, στην οποία εφαρμόζονται ορισμένα Φίλτρα, με το αποτέλεσμα να περνάει μέσα από μια Συνάρτηση Ενεργοποίησης. Το επόμενο είναι το βήμα της **Συγκέντρωσης** ή **Pooling** [10]. Κατά τη διαδικασία της Συγκέντρωσης, ορίζουμε ένα τοπικό παράθυρο και χωρίζουμε τον Πίνακα Χαρακτηρηστικών σε τοπικά παράθυρα.



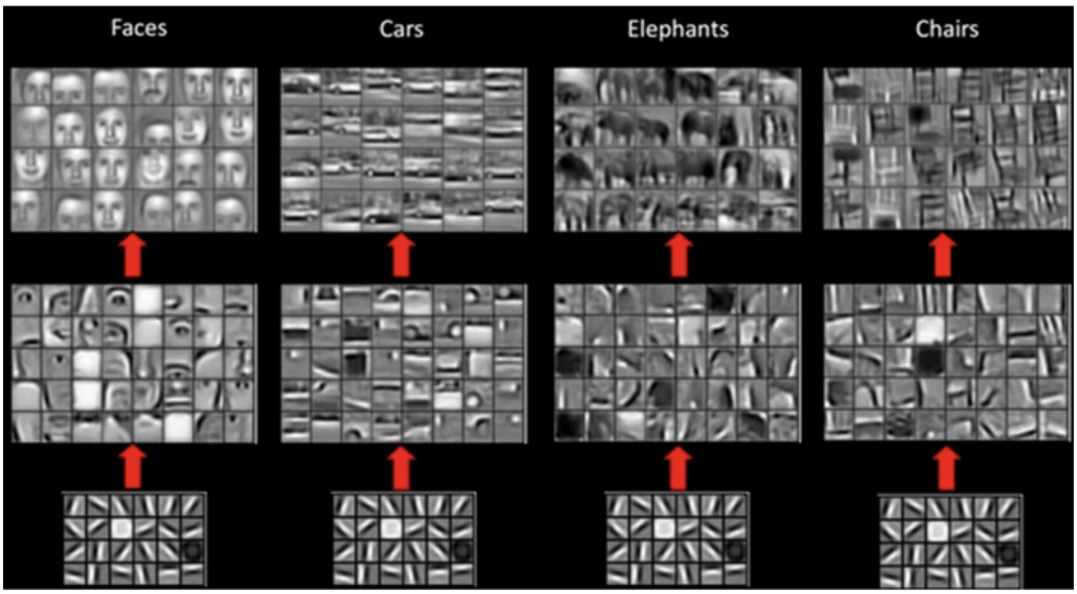
Σχήμα 3.4: Η Συγκέντρωση από τον Πίνακα Χαρακτηριστικών.

Τότε, ανάλογα με το είδος της Συγκέντρωσης που θέλουμε να πετύχουμε (min, max, average κλπ) κρατάμε μόνο αντίστοιχα το μικρότερο/μεγαλύτερο/μέσο όρο από τα στοιχεία του παραθύρου. Αν για παράδειγμα έχουμε ένα 4*4 Πίνακα Χαρακτηριστικών και επιλέγουμε max pooling 2*2 με stride = 2, Τότε χωρίζουμε στην ουσία τον Πίνακα σε 4 2*2 υποπίνακες από τον οποίους κρατάμε μονάχα το μέγιστο στοιχείο καθενός. Η συνεισφορά της Συγκέντρωσης εντοπίζεται στα εξής στοιχεία:

- Μείωση διαστατικότητας και συνεπώς γρηγορότερη διαχείριση των δεδομένων
- Μείωση πλήθους παραμέτρων και υπολογισμών, συνεπώς καλύτερος έλεγχος του φαινομένου overfitting
- Δίνει στο δίκτυο μια "ανοσία" όσον αφορά τους μικρούς μετασχηματισμούς, πιθανό θόρυβο σε μια εικόνα.

2.4 Η έξοδος των Συνελικτικών Στρωμάτων και η Ενημέρωση των Βαρών

Τα παραπάνω στοιχεία αποτελούν την κύρια δομή των ΣΝΔ. Όλη η επεξεργασία των εικόνων, η εξαγωγή χαρακτηριστικών με τη χρήση Φίλτρων, Συναρτήσεων Ενεργοποίησης συμβαίνει σε αυτά τα στρώματα. Η διαδικασία που αναφέρθηκε μπορεί να συμβαίνει παραπάνω από μια φορές, με σκοπό να εξάγονται κάθε φορά μεγαλύτερου επιπέδου ή αφαιρετικότητας αν θέλετε, χαρακτηριστικά από τις εικόνες.



Σχήμα 3.5: Τα Επίπεδα Χαρακτηριστικών που Εξάγονται από Διαδοχικά Στρώματα ΣΝΔ.

Τελικά η έξοδος των Συνελικτικών Στρωμάτων θα εξέλθει σε ένα MLP δίκτυο. Το MLP είναι ένα πλήρως συνδεδεμένο στρώμα, δηλαδή κάθε νευρώνας εξόδου του προηγούμενου στρώματος, συνδέεται με κάθε ένα του επόμενου. Η έξοδος από τα ΣΝΔ αποτελεί ηψυλού επιπέδου χαρακτηριστικά της Εικόνας Εισόδου. Ο σκοπός πλέον είναι να χρησιμοποιήσουμε αυτά τα χαρακτηριστικά για να κατηγοριοποιήσουμετην είσοδο στις κλάσσεις που επιθυμούμε. Φυσικά στο τελικό στάδιο το πλήθος των νευρώνων είναι όσες και οι κατηγορίες που θέλουμε να χωρίσουμε τα δεοδομένα μας. Πέρα από το σκοπό της

κατηγοριοποίησης η πρόσθεση ενός ή περισσοτέρων πλήρως συνδεδεμένων στρωμάτων στην έξοδο είναι ένας εύκολος τρόπος να εκπαιδευτεί σε μη γραμμικούς συνδυαμσούς των χαρακτηριστικών που προκύπτουν από τα ΣΝΔ. Μπορεί από μόνα τους να παρέχουν χρήσιμες πληροφορίες για την κατηγοριοποίηση, αλλά έχουν συνήθως περισσότερα να πουν, όταν συνδέονται μη-γραμμικά! Τα πλήρη συνδεδεμένα δίκτυα του τελικού στρώματος λειτουργούν με την Συνάρτηση Ενεργοποίησης **softmax**:

$$Softmax(x_i) = \frac{e^{x_i}}{\sum_{i=0}^{N} e^{x_i}} \tag{3.6}$$

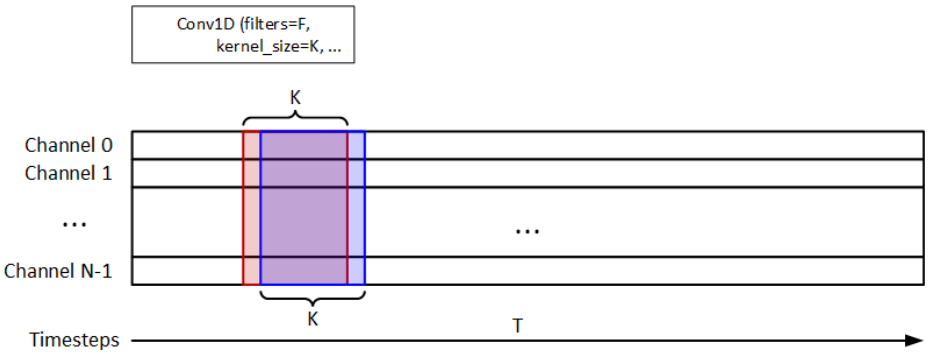
Η Συνάρτηση αυτή επιτρέπει στο δίκτυο να "κανονικοποιήσει" τις τιμές σε μορφή πιθανοτήτων (πραγματικές τιμές μεταξύ 0 και 1 οι οποίες αθροίζουν στο 1). Με αυτό τον τρόπο επιλέγεται από το δίκτυο η κατηγορία με την μεγαλύτερη πιθανότητα. Όταν η κατηγοριοποίηση κατά την εκπαίδευση κάνει λάθος τότε με την διαδικασία του Back-propagation, υπολογίζεται η κλίση του σφάλματος και ενημερώνονται οι τιμές των βαρών του δικτύου. Το πόσο αλλάζουν οι τιμές των βαρών είναι αντίστοιχο με την συνεισφορά τους στο συνολικό σφάλμα. Επειδή η διαδικασία διόρθωσης των βαρών είναι μια υπολογιστικά δύσκολη και χρονοβόρα διαδικασία, ενημερώνουμε τα βάρη αφού έχει εξεταστεί ένα πακέτο δεδομένων (και όχι μετά από κάθε δεδομένο εισόδου) το οποίο ονομάζουμε **batch**.

3 Το Μοντέλο μας

3.1 Η αρχιτεκτονική του μοντέλου

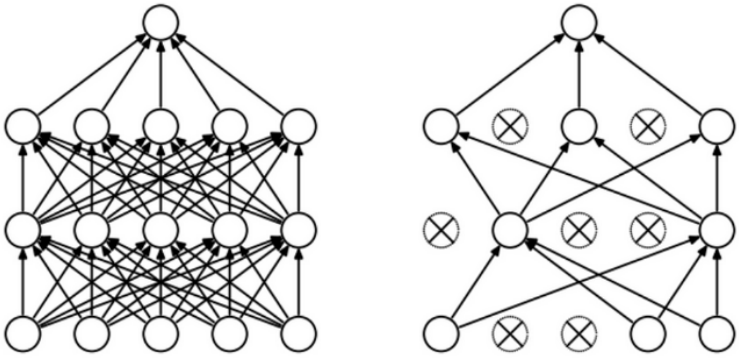
Η αρχιτεκτονική του μοντέλου αποτελείται από **τρία Συνελικτικά στρώματα**. Αρχικά να αναφέρουμε ότι η είσοδος είναι σε μορφή πίνακα 17*21 στοιχείων. Το πρώτο στρώμα αποτελείται από ένα ΣΝΔ (Conv1D) με 128 φίλτρα και διαστάσεις φίλτρου 5*21. Χρησιμοποιούμε γέμισμα με μηδενικά πριν και μετά την είσοδο ώστε το φίλτρο να κρατή-

σει τις ίδες διαστάσεις με την είσοδο. Αυτό σημαίνει ότι το πρώτο στρώμα θα έχει ως παραμέτρους τα βάρη των φίλτρων καθώς και το διάνυσμα bias του δικτύου, δηλαδή $128 * (5 * 21 + 1) = 13568$ παραμέτρους. Όπως αναφέραμε και παραπάνω ως Συνάρτηση Ενεργοποίησης θα χρησιμοποιήσουμε σε κάθε ΣΝΔ στρώμα την **ReLU**.



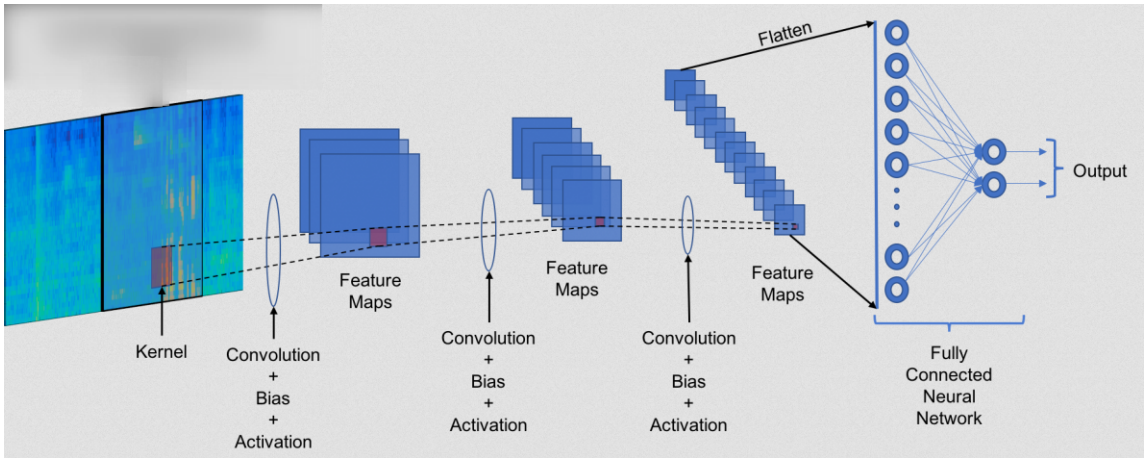
Σχήμα 3.6: Εφαρμογή Συνέλιξης στην Είσοδο.

Μετά τη χρήση του πρώτου φίλτρου, χρησιμοποιούμε δύο σημαντικές διαδικασίες. Η πρώτη είναι η **Κανονικοποίηση των Δεδομένων** ή αλλιώς **Batch Normalization**. Με αυτή το μέθοδο τα δεδομένα που παράγει το στρώμα μας κανονικοποιούνται με βάση το μέσο όρο και την τυπική απόκλιση που έχουν ανά σύνολα (Batches). Αυτό μας προσφέρει ανοχή στις τυχαίες αρχικοποιήσεις των βαρών κατά την εκκίνηση του δικτύου, κάνει την εκμάθηση γρηγορότερη και πιο σταθερή, αφού επιτρέπει να χρησιμοποιούνται μεγαλύτεροι ρυθμοί εκμάθησης με καλύτερα αποτελέσματα [11]. Η δευτερη σημαντική διαδικασία, είναι εκείνη της **Εγκατάλειψης Δεδομένων** ή **Dropout**. Κατά την Εγκατάλειψη Δεδομένων, κάθε κρυφός νευρώνας, σε κάθε δείγμα εισόδου έχει μια πιθανότητα p να αγνοηθεί και να μην μεταφέρει πρόσθια πληροφορία.



Σχήμα 3.7: Η Εγκατάλειψη Δεδομένων σε ένα Νευρωνικό Δίκτυο.

Με την εγκατάλειψη δεδομένων μειώνεται το φαινόμενο της **Υπερ-Εκπαίδευσης** σε ορισμένα δεδομένα ή **Overfitting** και αναγκάζουμε το δίκτυο να μάθει πιο σημαντικά χαρακτηριστικά που είναι χρήσιμα σε συνδυασμό με διάφορα τυχαία δείγματα ενεργών νευρώνων. Είναι σημαντικό να επισημάνουμε ότι με την Εγκατάλειψη, η ταχύτητα σύγκλισης μειώνεται αφού χρειάζονται περισσότερα περάσματα δεδομένων, αλλά αυτό αντισταθμίζεται σε σημαντικό βαθμό από την ταχύτερη εκπαίδευση ανά πέρασμα.

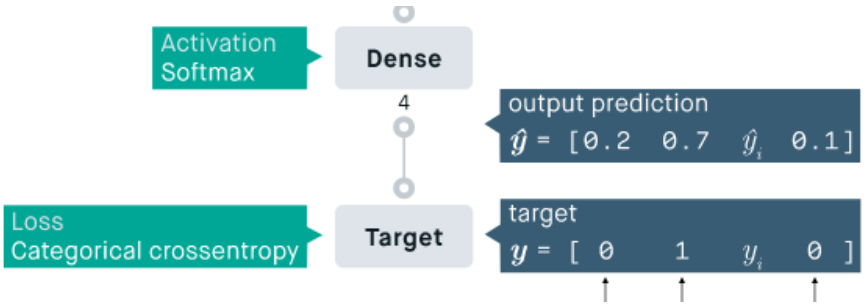


Σχήμα 3.8: Σχέδιο Αρχιτεκτονικής του Νευρωνικού Μοντέλου

Τα επόμενα δύο στρώματα είναι και αυτά ΣΝΔ, με 128 και 63 φίλτρα αντίστοιχα με δι-

αστάσεις 3*21 το καθένα. Όπως και στο πρώτο υπάρχει Κανονικοποίηση των Δεδομένων και Εγκατάλειψη Δεδομένων. Το ποσοστό Εγκατάλειψης είναι για τα τρία στρώματα $0.40, 0.35, 0.30$ αντίστοιχα. Το ποσοστό μειώνεται με το σκεπτικό ότι τα υψηλού επιπέδου χαρακτηριστικά είναι σημαντικά και δυσκολότερο να βρεθούν. Στην προηγούμενη υποενότητα αναφερθήκαμε στη χρησιμότητα διασύνδεσης των ΣΝΔ με κάποιο πλήρως συνδεδεμένο στρώμα στο τέλος του Δικτύου. Προκειμένου να ταιριάζουν οι τύποι δεδομένων, στην έξοδο του τελευταίου ΣΝΔ, εφαρμόζουμε την μέθοδο της **Εξομάλυνσης** ή **Flatten** ώστε να φέρουμε τα δεδομένα σε μονοδιάστατη μορφή. Στη συνέχεια ακολουθούν δύο **Πυκνά Στρώματα** ή **Dense Layers** με 128 και 32 νευρώνες αντίστοιχα και Συνάρτηση Ενεργοποίησης την **ReLU**. Τα δύο αυτά στρώματα θα μας βοηθήσουν να κάνουμε την κατηγοριοποίηση των χαρακτηριστικών που εξάγαμε με τα 3 στρώματα των ΣΝΔ. Τέλος χρησιμοποιούμε ένα τελικό Πυκνό Στρώμα με πλήθος όσες και οι κατηγορίες μας. Κατά την μετρική Q8 έχουμε 8 νευρώνες, ενώ στο Q3 έχουμε 3. Φυσικά η Συνάρτηση Ενεργοποίησης αυτή τη φορά δεν είναι άλλη από την **Softmax**, έτσι ώστε τα βάρη να αντιστοιχούν σε πιθανότητες να ανήκει το κάθε δείγμα στην αντίστοιχη κλάση.

Όσον αφορά την συνάρτηση κόστους, χρησιμοποιείται η **Categorical Cross-Entropy Loss**. Η συγκεκριμένη χρησιμοποιείται σε προβλήματα κατηγοριοποίησης και δίνεται από τον εξής τύπο:



Σχήμα 3.9: Τελευταία Έξοδος του Δικτύου, οι Πιθανότητες που Δίνει το Μοντέλο

$$Loss = \sum_{i=1}^{No.Classes} y_i * \log \hat{y}_i \tag{3.7}$$

Συνολικά το Δίκτυο της εργασίας αυτής έχει 232,552 παραμέτρους. Στο τέταρτο κεφάλιο θα μιλήσουμε πιο αναλυτικά για κάποιες από τις υπερπαραμέτρους του Δικτύου και θα εξηγήσουμε τη διαδικασία από τη λήψη των δεδομένων, την επεξεργασία τους και την εκπαίδευση τους, μέχρι τη συγκέντρωση αποτελεσμάτων και τη σύγκρισή τους με άλλά state of the art μοντέλα.

3.2 Εναλλακτικές ιδέες που απορρίφθηκαν

Κατά το στήσιμο του παραπάνω δικτύου λάβαμε υπόψιν δύο ιδέες οι οποίες εν τέλει απορρίφθηκαν.

Η πρώτη ιδέα ήταν η χρήση ενός **Αναδρομικού Νευρωνικού Δικτύου** ή **Recurrent Neural Network** ως κύριο μοντέλο της εργασίας προκειμένου να γίνει η ταξινόμηση. Κατά την προσπάθεια στησίματος συμπεράναμε ότι τα RNN δεν έχουν την πολυπλοκότητα που απαιτείται για να εξάγουν υψηλού επιπέδου χαρακτηριστικά, σε σχέση με τα CNN. Τα πρώτα αποτελέσματα δεν ήταν ενθαρρυντικά και για αυτό χρειαστηκέ να μελετηθεί το CNN ως αρχιτεκτονικό μοντέλο. Όπως θα δούμε και στο επόμενο κεφάλαιο, τα μοντέλα των τελευταίων χρόνων που περιέχουν RNN αρχιτεκτονικές είναι η μειοψηφία και όλα αυτά είναι πάντα ένας σύνθετος συνδυασμών πολλαπλών αρχιτεκτονικών. Διαισθητικά, κρύβονται οι αδυναμίες των RNN στο συγκεκριμένο πρόβλημα από το μεγάλο μέγεθος (και αντοίστιχα υπολογιστικής δύναμης) που έχουν αυτά τα δίκτυα. Μία άλλη ιδέα που απορρίφθηκε ήταν να θεωρούμε ως είσοδο ολόκληρη την πρωτεινή και να γίνεται μαζικά πρόβλεψη για το κάθε αμινοζύ που την περιέχει. Τα αποτελέσματα ήταν συγκριτικά χαμηλότερα και για αυτό πάρθηκε η απόφαση να δίνεται ως είσοδο μονάχα μια κινούμενη-στην-πρωτεΐνη ακολουθία 17 αμινοξέων μελετώντας κάθε φορά το κεντρικό αμινοξύ.

4 Βιβλιογραφία

[1] Ling Zhang and Bo Zhang. “A geometrical representation of McCulloch-Pitts neuralmodel and its applications”. In:IEEE Transactions on Neural Networks10.4 (July1999), pp. 925–929.

[2] Rich Caruana and Alexandru Niculescu-Mizil. “An Empirical Comparison of Supervised Learning Algorithms”. In:Proceedings of the 23rd International Conference on Machine Learning. ICML ’06. Pittsburgh, Pennsylvania, USA: ACM, 2006,pp. 161–168

[3] Martin Riedmiller. “Advanced supervised learning in multi-layer perceptrons-Frombackpropagation to adaptive learning algorithms”. In:Computer Standards Inter-faces16.3 (1994)

[4] .B. Barlow. “Unsupervised Learning”. In:Neural Computation1.3 (1989), pp. 295–311.

[5] Richard S. Sutton and Andrew G. Barto.Introduction to Reinforcement Learning.1st. Cambridge, MA, USA: MIT Press, 1998

[6] "Convolutional Neural Networks (LeNet) – DeepLearning 0.1 documentation". DeepLearning 0.1. LISA Lab. Retrieved 31 August 2013.

[7] Xavier Glorot, Antoine Bordes and Yoshua Bengio (2011). Deep sparse rectifier neural networks (PDF). AISTATS. "Rectifier and softplus activation functions. The second one is a smooth version of the first."

[8] Ramachandran, Prajit; Barret, Zoph; Quoc, V. Le (October 16, 2017). "Searching for Activation Functions". arXiv:1710.05941

[9] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In S. C. Kremer and J. F. Kolen, editors, A Field Guide to Dynamical Recurrent Neural Networks. IEEE Press, 2001.

[10] Géron, Aurélien (2019). Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow. Sebastopol, CA: O'Reilly Media. ISBN 9780226484648., pp. 448

[11] Ioffe, S., Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.

Κεφάλαιο 4

Τα Δεδομένα, το Πείραμα και τα Αποτελέσματα

1 Τα Δεδομένα Εισόδου του Προβλήματος

Όπως είχαμε αναφέρει και στο εισαγωγικό κεφάλαιο, τα δεδομένα του Protein Folding Problem και συγκεκριμένα του PSSP (Protein Secondary Structure Prediciton) δεν είναι τίποτα άλλο παρά μελετημένες πρωτεΐνες. Για την εκπαίδευση του Δικτύου χρησιμοποιήσαμε τη συλλογή πρωτεϊνών **CullPDB** [1]. Το CullPDB αποτελεί ένα σετ από 6133 πρωτεΐνες οι οποίες μεταξύ τους δεν παρουσιάζουν πολλές ομοιότητες μεταξύ τους (δείκτης ομολογίας $<25\%$). Ο δείκτης ομολογίας είναι ένα πολύ σημαντικό κριτήριο για την αξιοπιστία ενός dataset [2]. Ναι φυσικά, σε ένα σετ δεδομένων με παρόμοιες πρωτεΐνες, ένα μοντέλο θα πετυχαίνει εξαιρετικά αποτελέσματα, μονάχα όμως σε συγγενείς πρωτεΐνες. Στοχεύοντας όμως σε ένα ολοκληρωτικό μοντέλο πρόβλεψης, θα ασχοληθούμε με πιο "ποικιλόμορφα" datasets όπως το CullPDB. Σχετικά με τη μορφή δεδομένων, κάθε μια από τις 6133 πρωτεΐνες, αποτελείται από 28 έως 700 αμινοξέα. Για κάθε αμινοξύ έχουμε τα εξής 57 χαρακτηριστικά:

- **[0,21]** One-Hot encoded αμινοξέα 'A', 'C', 'E', 'D', 'G', 'F', 'I', 'H', 'K', 'M', 'L', 'N', 'Q', 'P', 'S', 'R', 'T', 'W', 'V', 'Y', 'X','NoSeq'
- **[22,30]** Οι 8 ετικέτες δευτεροταγούς δομής 'L', 'B', 'E', 'G', 'I', 'H', 'S', 'T','NoSeq'
- **[31,33]** N-, C- τερματικά
- **[34-35]** Σχετική και απόλυτη διαλυτική προσβασιμότητα
- **[36,56]** Sequence Profiles για κάθε αμινοξύ.

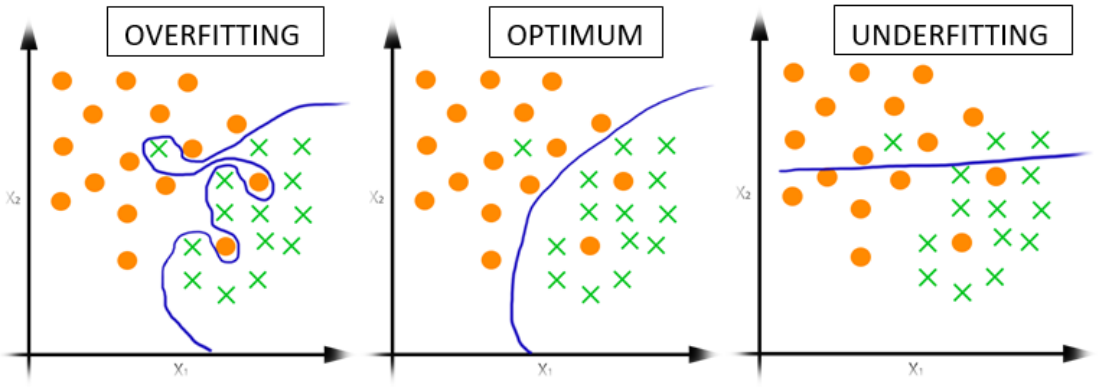
Το τελευταίο από τα features είναι και το σημαντικότερο. Τα **Προφίλ Ακολουθιών** ή **Sequence Profiles** παράγονται από τους **Πίνακες Σκορ Συγκεκριμένης Θέσης** ή **Position Specfic Scoring Matrices** και αναφέρονται στη συχνότητα που έχουν τα αμινοξέα να εμφανίζονται σε αντίστοιχες γειτονικές αλληλουχίες [3]. Σε κάθε θέση ενός αμινοξέος μας δείχνει το σκορ που έχουν όλα τα αμινοξέα με βάση το πόσο συχνά θα μπορούσαν να βρεθούν στη θέση του. Τα προφίλ αυτά θα αποτελέσουν το βασικό χαρακτηριστικό πάνω στο οποίο θα εκπαιδευτεί το δίκτυο.

Ίσως η σημαντικότερη βάση δεδομένων για benchmarks, είναι η **CB513** των Cuff και Barton [4]. Περιέχει 513 πρωτείνες με 84,107 αμινοξέα συνολικά, ενώ κάθε πρωτείνη έχει ποσοστό ομοιότητας μικρότερο από 25%. Περιέχει τις 396 πρωτείνες που εμπεριέχονται στο dataset **CB396** [5] και 117 από τις 126 πρωτείνες που εμπεριέχονται στο **RS126** [6]. Το μέση μήκος πρωτείνης είναι 163 για την CB513, 157 για την CB396 και 185 για την RS126.

2 Η Διαδικασία του Πειράματος

Προκειμένου να εκπαιδεύσουμε το μοντέλο, το πρώτο βήμα ήταν να επεξεργαστούμε τα δεδομένα με ώστε να έρθουν στη σωστή μορφή για εκπαίδευση. Αρχικά μετατρέψαμε τα δεδομένα σε ένα 3-διάστατο πίνακα 6133*700*57 με τα νούμερα αυτά να υποδηλώνουν το πλήθος των πρωτεϊνών, το μέγιστο πλήθος αμινοξέων σε πρωτείνη, και το πλήθος των χαρακτηριστικών. Στη συνέχεια χωρίσαμε τις ετικέτες και αφαιρέσαμε τα χαρακτηριστικά που δεν θα χρησιμοποιήσουμε στην εκπαίδευση. Το dataset χωρίστηκε σε τρία τυχαία μέρη. Το πρώτο μέρος αφορά το **training set** που αποτελεί το 80% των δεδομένων. Τα άλλα δύο μέρη, από 10%, είναι το **validation test** και το **test set**. Σε κάθε εποχή που εκπαιδεύεται το μοντέλο, γίνεται επαλήθευση της εκπαίδευσης στο validation set, ώστε να έχουμε έλεγχο φαινομένων όπως το **overfitting** [7]. Όταν η επιτυχία του training set αυξάνεται και ξεπερνάει την επιτυχία του validation set το οποίο αρχίζει να μένει

στάσιμο, τότε έχουμε overfitting. Αυτό σημαίνει ότι το δίκτυο ερμηνεύει τα μοτίβα με πολύ ειδικό τρόπο μονάχα σε συγκεκριμένα (μελετημένα) δεδομένα. Χάνεται η δυνατότητα γενίκευσης της πρόβλεψης και έτσι το δίκτυο αδυνατεί να προβλέψει επιτυχώς δεδομένα στα οποία δεν έχει εκπαιδευτεί. Μάρτυρας του φαινομένου αυτού είναι η παρόμοια επιτυχία validation/test set που είναι σαφώς χαμηλότερη από την επιτυχία του training set.



Σχήμα 4.1: Underfitting and Overfitting Effect

Κάθε φορά, τα δεδομένα που απαρτίζουν τα τρία set γίνονται shuffle για να εξαλείψουμε την πιθανότητα συσχέτισης των αποτελεσμάτων μας στην επιλογή των set. Όσες φορες έτρεξε το μοντέλο δεν υπήρξε σημαντική απόκλιση στο αποτέλεσμα. Αποφασίσαμε τα δεδομένα να έρχονται σε $Batchsize = 64$ δηλαδή σε πακέτα των 64. Μετά από κάθε πακέτο γίνεται εκπαίδευση του Δικτύου έως ότου δωθούν όλα τα δεδομένα και το δίκτυο να προχωρήσει στην επόμενη εποχή όπου θα επαναληφθεί η διαδικασία. Το πλεονέκτημα της χρήσης πακέτων είναι ότι το σύστημα χρειάζεται λιγότερη μνήμη κατά την εκπαίδευση και η εκπαίδευση γίνεται γρηγορότερα.

Πέρα από τα δεδομένα χρειάζεται να αναφέρουμε και κάποιες σημαντικές παραμέτρους που επηρεάζουν τα αποτελέσματα του πειράματος. Η πιο σημαντική απόφαση είχε να κάνει με το μέγεθος των δεδομένων που θα εισέρχονται στο Δίκτυο σε κάθε βήμα. Επειδή το μέσο μήκος α-έλικας είναι 11 και β-πτύχωσης 6 αποφασίσαμε να δοκιμάσουμε διψήφιες

τιμές κινούμενου παραθύρου. Μετά από κάποιες προσομοιώσεις, επιλέχθηκε ένα κινούμενο παράθυρο μήκους 17 αμινοξέων και κάθε φορά να αποφασίζεται σε ποιά δευτεροταγή δομή ανήκει το μεσαίο από τα 17. Επόμενως **cnn window = 17**.

Σχετικά με το πλήθος εποχών που χρειάζονται για την εκπαίδευση, δεν υπάρχει κάποια πανάκεια λύση. Το μοντέλο ξεκίνησε να εκπαιδεύει για ένα αόριστο αριθμό εποχών και χρησιμοποιήσαμε δύο κριτήρια για το τέρμα της εκπαίδευσης. Το πρώτο έχει να κάνει με τον κορεσμό της μετρικής **accuracy** που μετράει το ποσοστό επιτυχίας στο test set. Το δεύτερο έχει να κάνει με την αύξηση ανά εποχή, της μετρικής **validation loss** που σηματοδοτεί το φαινόμενο overfitting. Και στις δύο περιπτώσεις η εκπαίδευση σταματάει. Στο πείραμα μας για το Q3 χρειάστηκαν **number of epochs = 45** εποχές ενώ για το Q8 **number of epochs = 50**. Το **learning rate = 0.0008** δηλώνει το μέγεθος της αλλαγής των βαρών κατά την ενημέρωση των παραμέτρων του δικτύου.

3 Πειραματικά Αποτελέσματα και Συγκρίσεις

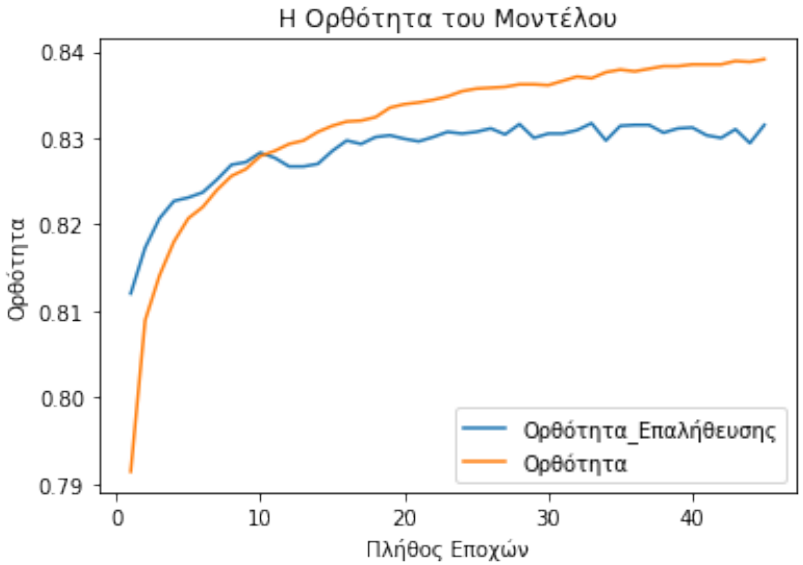
3.1 Τα Αποτελέσματα του Πειράματος

Παρακάτω παρατίθονται τα αποτελέσματα των δύο μοντέλων στο dataset που χρησιμοποιήθηκε για την εκπαίδευση, και στο benchmark dataset.

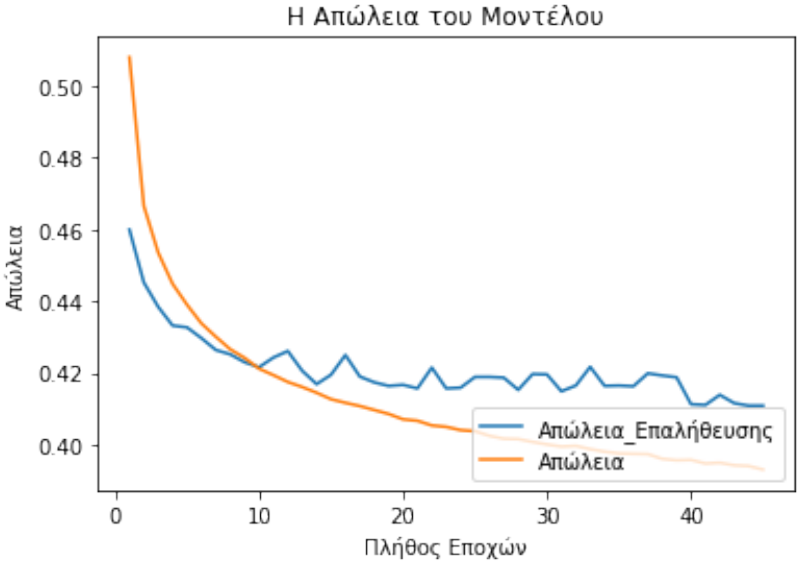
Μοντέλο	Σετ Δεδομένων	Ορθότητα	Απώλεια	MAE
CNN Q3	CullPDB	83.58%	0.4069	0.147
CNN Q3	CB513	82.61%	0.4504	0.157
CNN Q8	CullPDB	72.13%	0.7732	0.091
CNN Q8	CB513	70.09%	0.868	0.097

Πίνακας 4.1: Αποτελέσματα των Μοντέλων Q3, Q8.

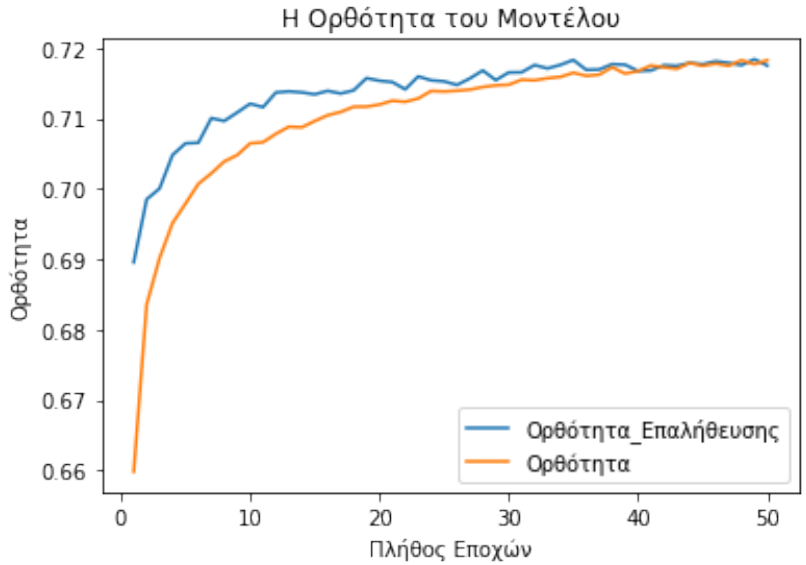
Σχετικά με το Q3 έχουμε για τις μετρικές accuracy, loss:



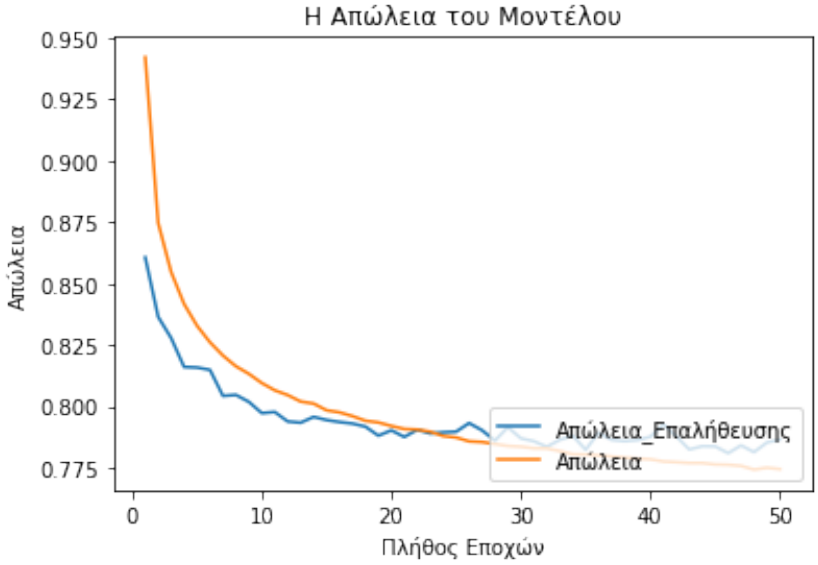
Σχήμα 4.2: Ποσοστό Ορθότητας του Q3 στο CullPDB



Σχήμα 4.3: Ποσοστό Απώλειας του Q3 στο CullPDB



Σχήμα 4.4: Ποσοστό Ορθότητας του Q8 στο CullPDB



Σχήμα 4.5: Ποσοστό Απώλειας του Q8 στο CullPDB

Μοντέλο	Παράμετροι	Χρόνος Εκπαίδευσης	Εποχές	Χρόνος ανα Εποχή
CNN Q3	231,747	5hr 46min 28sec	45	7min 41sec
CNN Q8	231,912	6hr 40min 35sec	50	8min 5sec

Πίνακας 4.2: Χρόνος εκπαίδευσης των Μοντέλων.

3.2 Συγκρίσεις

Πριν σχολιάσουμε τα παραπάνω δεδομένα και για να έχουμε μια καλύτερη εικόνα του τι αντιπροσωπεύουν τα δεδομένα μας, θα τα συγκρίνουμε με άλλα μοντέλα των τελευταίων χρόνων.

Όνομα	CB513	CullPDB	Αναφορά
GSN	66.4%	72.1%	[8]
SSpro	63.5%	66.6%	[9]
RaptorX-SS8	64.9%	69.7%	[10]
SSREDN	68.2%	73.1%	[13]
CNN3	**70.09%**	**72.13%**	-
Deep3l*	71.0%	-	[11]
DeepCNF	72.2%	73.8%	[12]

Πίνακας 4.3: Τα Ποσοστά Πρόβλεψης **Q8** Νευρωνικών.

Όνομα	CB513	CullPDB	Αναφορά
SSpro	78.5%	79.5%	[9]
RaptorX-SS8	78.3%	81.2%	[10]
JPRED	81.7%	82.9%	[14]
CNN3	**82.61%**	**83.58%**	-
Deep3l*	82.8%	-	[11]
DeepCNF	82.3%	85.4%	[12]

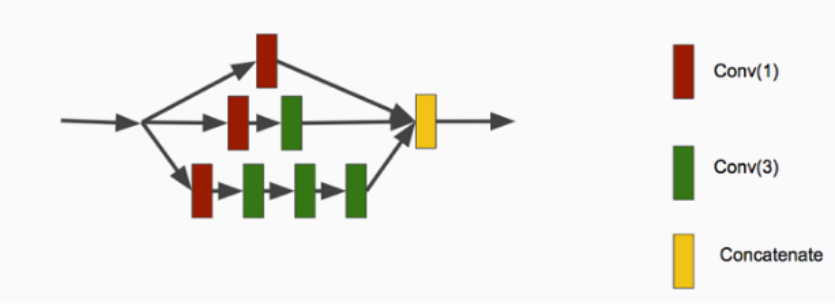
Πίνακας 4.4: Τα Ποσοστά Πρόβλεψης **Q3** Νευρωνικών.

Ας αναφερθούμε αρχικά στα προβαλλόμενα μοντέλα των πινάκων και τις αρχιτεκτονικές επιλογές που έχουν γίνει κατά την υλοποίησή τους.

- **SSpro**: Το SSpro ξεκίνησε την υλοποίησή του στα τέλη της δεκαετίας του 90. Τα συγκεκριμένα νούμερα αφορούν την έκδοση του 2002. Αποτελείται από ένα σύνολο Αμφίδρομων Αναδρομικών Νευρωνικών Δικτύων τα οποία έχουν τοποθετηθεί στη σειρά. Σε αντίθεση με τις πρώτες εκδόσεις του SSpro, αυτό χρησιμοποιεί πλέον δεδομένα από το PSI-BLAST, το οποίο δημιουργεί τα Position Specific Scoring Matrices. Τα Sequence Profiles παράγονται από τα PSSM και αναφέρονται στη συχνότητα που έχουν τα αμινοξέα να εμφανίζονται σε αντίστοιχες γειτονικές αλληλουχίες [3]. Σε κάθε θέση ενός αμινοξέος μας δείχνει το σκορ που έχουν όλα τα αμινοξέα με βάση το πόσο συχνά θα μπορούσαν να βρεθούν στη θέση του. Σε σχέση με το πρώτο σημείωσε αύξηση στα ποσοστά ορθότητας 1-5% ανάλογα το dataset.

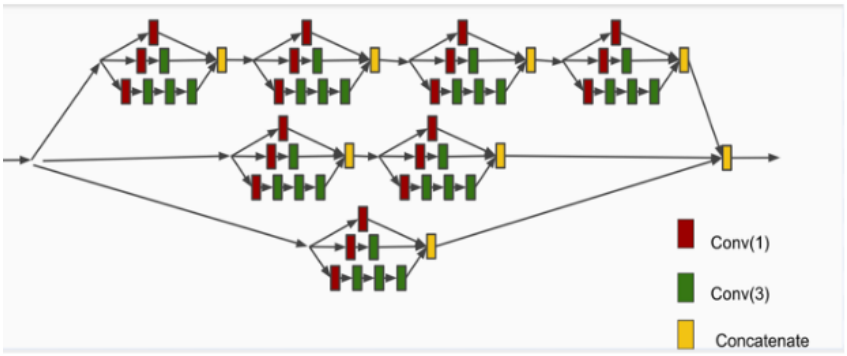
- **RaptorX-SS8**: Το RaptorC-SS8 υλοποιήθηκε στις αρχές του 2010-2011. Κύριο χαρακτηριστικό του είναι η χρήση **Υπό Συνθήκη Τυχαίων Πεδίων** ή **Conditional Random Fields** σε συνδυασμό με Νευρωνικά (Conditional Neural Fields, CNF). Τα CNF, αποτελούν ένα αποτελεσματικό συνδυασμό των Υπο Συνθήκη Τυχαίων Πεδίων και των Νευρωνικών, καθώς τα τελευταία μπορούν να μοντελοποιήσουν τα μη-γραμμικά χαρακτηριστικά μεταξύ των αμινοξέων και τα πρώτα μπορούν εντοπίσουν τη συσχέτιση μεταξύ παρόμοιων ακολουθιών αμινοξέων.

- **Deep3l**: Το συγκεκριμένο δίκτυο βασίζεται σε ένα αρχικό μοντέλο που αποτελείται από διάφορα παράλληλα και σειριακά Συνελικτικά Στρώματα:



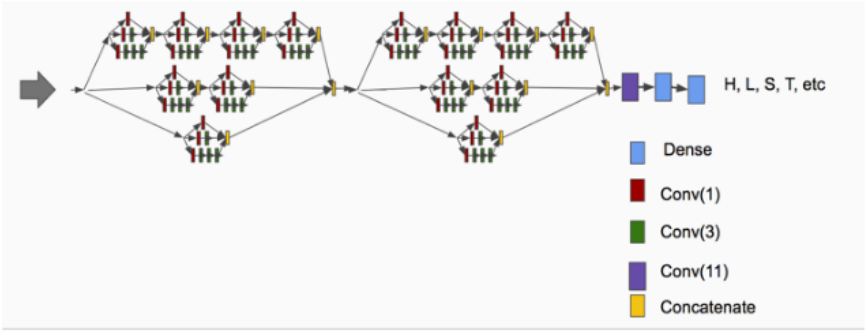
Σχήμα 4.6: Το Βασικό Μοντέλο, αποτελούμενο από Συνελικτικά Στρώματα.

Στη συνέχεια, κάθε στρώμα αντικαθιστάται από ολόκληρο το αρχικό δίκτυο που ονομάζεται Deep3I module:



Σχήμα 4.7: Το Deep3I module.

Τέλος εφαρμόζονται δύο Deep3I modules στη σειρά και στην έξοδο έχουμε ως γνωστόν ένα Πυκνό στρώμα:



Σχήμα 4.8: Το τελικό μοντέλο Deep3I.

- **DeepCNF**: Το DeepCNF, αποτελείται από ένα συνδυασμό CNF και **Βαθιά Συνελικτικά Νευρωνικά Δίκτυα** ή **Deep Convolutional Neural Networks**. Όπως και το RaportX-SS8 παραπάνω, το DeepCNF αξιοποιεί και συνδυάζει τα προτερήματα των CNF και των DCNN.

Από τους παραπάνω Πίνακες βλέπουμε ότι μοντέλο της εργασίας μας έχει αποκτήσει

ορθότητα κοντά στα σύγχρονα μοντέλα. Δίκτυα πρόβλεψης που κυριαρχούσαν τα πρώτα 15 χρόνια του 21ου αιώνα όπως το RaptorX-SS8, το JPRED και το GSN βλέπουμε ότι δίνουν τη θέση τους σε εξαιρετικά πολύπλοκα δίκτυα τα οποία και παρατηρούμε ότι σημειώνουν τα καλύτερα αποτελέσματα. Μελετώντας τα state of the art δίκτυα από τα οποία υστερεί το μοντέλο μας, μπορούμε να εντοπίσουμε ορισμένους λόγους για τους οποίους συμβαίνει αυτό. Αρχικά τα μοντέλα αυτά χρησιμοποιούν πολύπλοκες αρχιτεκτονικές με διάφορα ήδη Νευρωνικών Δικτύων σειριακά. Οι παράμετροι τους είναι πολλαπλάσιες από τις δικές μας, και οι χρόνοι εκπαίδευσης σαφώς μεγαλύτεροι. Φυσικά αυτό δεν σημαίνει ότι η διαφορά έγκειται μονάχα στον υπολογιστικό εξοπλισμό, αλλά και στην πιο σύνθετη φιλοσοφία που ακολουθούν τα μοντέλα αυτά. Επίσης σε ορισμένα από αυτά χρησιμοποιούνται και χαρακτηριστικά των πρωτεϊνών που εμείς δεν μπορούσαμε να χρησιμοποιήσουμε είτε γιατί ήταν ελλειπή είτε γιατί δεν το επέτρεπαν οι χρονικοί περιορισμοί μας.

Το δίκτυο μας όντας σχετικά απλό στην δομή του, καταφέρνει να έχει αξιοσημείωτα αποτελέσματα στη δευτεροταγή πρόβλεψη των πρωτεϊνών. Συγκρίνοντας τη σχέση μεταξύ των ακριβειών των δύο datasets και σε σχέση, με τα υπόλοιπα μοντέλα, παρατηρούμε ότι το μοντέλο μας έχει καλύτερη από την προσδοκόμενη ορθότητα στο CB513. Αυτό είναι ένα ενθαρρυντικό δείγμα για την ικανότητα του μοντέλου να προβλέπει πρωτεΐνες πάνω στις οποίες δεν έχει εκπαιδευτεί.

Εν έτει 2020, υπάρχουν αρκετά μοντέλα που βασίζονται σε όλο και περισσότερες παραμέτρους για να αυξήσουν τα ποσοστά ορθότητας, κάνοντας πολλές φορές το κυνήγι αυτό, ένα διαγωνισμό ισχυρών υπολογιστικών συστημάτων. Είναι πολύ λεπτή η γραμμή μεταξύ αρχιτεκτονικών ιδεών που ταιριάζουν στη φύση του προβλήματος και στην εναπόθεση των ελπιδών σε στρώματα επί στρωμάτων νευρώνων.

4 Βιβλιογραφία

[1] https://www.princeton.edu/ jzthree/datasets/ICML2014/

[2] Jian Z., Olga G. T. Deep Supervised and Convolutional Generative Stochastic Network forProtein Secondary Structure Prediction

[3] Cuff, J. A., Barton, G. J. (2000). Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins: Structure, Function, and Bioinformatics, 40(3), 502-511.

[4] Cuff, J. A., Barton, G. J. (1999). Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. Proteins: Structure, Function, and Bioinformatics, 34(4), 508-519.

[5] J.A. Cuff, G.J. Barton, Evaluation and improvement of multiple sequence methods for protein secondary structure prediction, Proteins Struc. Func. Bioinf. 34 (4) (1999) 508.

[6] Burkhard Rost, Chris Sander, Reinhard Schneider, PHD-an automatic mail server for protein secondary structure prediction, Bioinformatics, Volume 10, Issue 1, February 1994, Pages 53–60, https://doi.org/10.1093/bioinformatics/10.1.53

[7] Burnham, K. P.; Anderson, D. R. (2002), Model Selection and Multimodel Inference (2nd ed.), Springer-Verlag.

[8] J. Zhou, O. G. Troyanskaya, Deep supervised and convolutional generativestochastic network for protein secondary structure prediction, in: Proceed-ings of International Conference on Machine Learning (ICML), 2014

[9] C. N. Magnan, P. Baldi, Sspro/accpro 5: almost perfect prediction of pro-tein secondary structure and relative solvent accessibility using profiles,machine learning and structural similarity, Bioinformatics, 30 (18) (2014)

[10] Z. Wang, F. Zhao, J. Peng, J. Xu, Protein 8-class secondary structureprediction using conditional neural fields, Proteomics, 11 (19) (2011)

[11] Fang, C., Shang, Y., & Xu, D. (2017). MUFold-SS: Protein secondary structure prediction using deep inception-inside-inception networks. arXiv preprint arXiv:1709.06165.

[12] Wang, S., Peng, J., Ma, J. et al. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. Sci Rep 6, 18962 (2016)

[13] Protein secondary structure prediction by using deep learning method YangxuWang, HuaMao, ZhangYi

[14] Cole, C., Barber, J. D., & Barton, G. J. (2008). The Jpred 3 secondary structure prediction server. Nucleic acids research, 36(suppl_2), W197-W201.

Κεφάλαιο 5

Επίλογος

Σε αυτό το κεφάλαιο θα συνοψίσουμε τα αποτελέσματα της μελέτης μας και θα προτείνουμε πιθανούς τρόπους με τους οποίους μπορεί να επεκταθεί μελλοντικά η εργασία αυτή.

1 Συμπεράσματα

Αναπτύξαμε λοιπόν δύο μοντέλα Νευρωνικών Δικτύων που βασίστηκαν στην φιλοσοφία των Συνελικτικών Δικτύων. Στόχος των μοντέλων αυτών ήταν να μπορούν να προβλέψουν με επιτυχία την δευτεροταγή δομή πρωτεϊνών. Για να γίνει αυτό εκπαιδεύσαμε το δίκτυο σε ήδη μελετημένα δεδομένα, που γνωρίζαμε δηλαδή σε ποιά κατηγορία δευτεροταγούς δομής ανήκουν τα αμινοξέα που απαρτίζουν τις πρωτεΐνες. Τα δύο αυτά μοντέλα αναπτύχθηκαν με σκοπό να κατηγοριοποιούν σε 3 και 8 κλάσσεις αντίστοιχα. Οι διαφορές τους δεν ήταν μεγάλες στην αρχιτεκτονική, ωστόσο οι παράμετροι των μοντέλων είχαν αρκετές διαφορές μεταξύ τους. Στο τέλος, αφού λάβαμε τα αποτελέσματα, τα συγκρίναμε με τα υπόλοιπα state of the art μοντέλα. Το αποτέλεσμα ήταν εξαιρετικό, με την έννοια ότι το μοντέλο μας υστερούσε ελάχιστα σε σχέση με τα "μεγαθήρια" του PSSP (Protein Secondary Structure Prediction), ενώ παράλληλα ήταν σαφώς καλύτερο από όλα τα μοντέλα που δημιουργήθηκαν μέχρι το 2016.

2 Μελλοντικές Επεκτάσεις

Σε αυτό το σημείο θα αναφέρουμε ορισμένες μελλοντικές επεκτάσεις της εργασίας αυτής που θα βοηθήσουν τον ενδιαφερόμενο να αποκτήσει μια καλύτερη εικόνα του προβλήματος.

- Το μοντέλο θα μπορούσε να εκπαιδευτεί σε περισσότερα δεδομένα, σε βάσεις δεδομένων που περιέχουν δεκάδες χιλιάδες πρωτεΐνες προκειμένου έχουμε περισσότερα μέτρα σύγκρισης και να διακρίνουμε τις δυνατότητες και αδυναμίες του.

- Ενθαρρύνεται η εξέταση απόδοσης του μοντέλου σε datasets τα οποία είναι γνωστά για τους δύσκολους στόχους. Ενδεικτικά αναφέρουμε το CASP (Critical Assesment of Protein Structure Prediction), ένα παγκοσμίας εμβέλειας διαγωνισμό που κάθε δύο χρόνια οργανώνει διαγωνισμούς πρόβλεψης με πρωτεΐνες στόχους κρυφούς!

- Σε σχέση με την αρχιτεκτονική του δικτύου, θα μπορούσαμε μελλοντικά με αφορμή τα πιο σύνθετα μοντέλα να αναλογιστούμε τι θα μπορούσαμε να προσθέσουμε στο μοντέλο μας ώστε να αποκησει περισσότερα επίπεδα και ενδεχομένως αποτελεσματικότητα.