ΠΑΝΕΠΙΣΤΗΜΙΟ
ΔΥΤΙΚΗΣ ΜΑΚΕΔΟΝΙΑΣ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ
ΠΛΗΡΟΦΟΡΙΚΗΣ & ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΧΡΗΣΗ ΑΚΟΛΟΥΘΙΑΚΩΝ ΠΡΟΤΥΠΩΝ ΓΙΑ ΤΑΞΙΝΟΜΗΣΗ ΔΟΜΗΣ ΠΡΩΤΕΪΝΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ
ΤΗΣ

ΝΤΑΓΙΟΥ ΑΝΝΑΣ

ΕΠΙΒΛΕΠΟΝΤΕΣ ΚΑΘΗΓΗΤΕΣ

Αγγελίδης Παντελης
Τσιπουρας Μαρκος

Κοζανη, Ιουλιος 2017

ΠΑΝΕΠΙΣΤΗΜΙΟ
ΔΥΤΙΚΗΣ ΜΑΚΕΔΟΝΙΑΣ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ
ΠΛΗΡΟΦΟΡΙΚΗΣ & ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΧΡΗΣΗ ΑΚΟΛΟΥΘΙΑΚΩΝ ΠΡΟΤΥΠΩΝ ΓΙΑ ΤΑΞΙΝΟΜΗΣΗ ΔΟΜΗΣ ΠΡΩΤΕΪΝΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ
ΤΗΣ

ΝΤΑΓΙΟΥ ΑΝΝΑΣ

ΕΠΙΒΛΕΠΟΝΤΕΣ ΚΑΘΗΓΗΤΕΣ

Αγγελίδης Παντελης
Τσιπουρας Μαρκος

Κοζανη, Ιουλιος 2017

Περίληψη

Η ταξινόμηση δεδομένων είναι ένα σημαντικό θέμα στον τομέα της εξόρυξης δεδομένων εξαιτίας της ευρείας κλίμακας εφαρμογών που μπορεί να χρησιμοποιηθεί. Υπάρχει ένας μεγάλος αριθμός μεθόδων που έχουν προταθεί για την επίλυση τέτοιου είδους προβλημάτων, οι οποίοι βασίζονται σε γνωστές τεχνικές εξόρυξης δεδομένων όπως τα δέντρα απόφασης ή τα νευρωνικά δίκτυα. Ωστόσο, αυτοί οι τύποι αλγόριθμων ταξινόμησης δεν μπορούν να εφαρμοστούν και να αποδώσουν σε ακολουθιακά δεδομένα όπως πρωτεϊνικές βάσεις δεδομένων που αποτελούνται από ακολουθίες αμινοξέων.

Στη Διπλωματική Εργασία μελετήθηκαν καινοτόμες αλγοριθμικές τεχνικές ταξινόμησης μέσω ακολουθιακών προτύπων για την πρόβλεψη της δευτεροταγούς δομής πρωτεϊνών. Συνοπτικά, επιλέχτηκε ένας βασικός αλγόριθμος στο κομματι της εξαγωγής ακολουθιακών προτύπων και αναπτύχθηκε κώδικας για την χρήση αυτών των προτύπων στην πρόβλεψη πρωτεϊνικών δομών σε ένα σύνολο από αλληλουχίες αμινοξέων. Σε ότι αφορά την πρόβλεψη πρωτεϊνικών δομών και την βαθμολόγηση ακολουθιακών προυτύπων υλοποιήθηκε μία πληθώρα καινοτόμων μεθοδολογιών, οι οποίες θεωρητικά και πειραματικά ξεπερνούν τα μειονεκτήματα των υπαρχόντων αλγορίθμων.

Λέξεις - Κλειδιά

Ακολουθίες αμινοξέων, Ακολουθιακά πρότυπα, Δομή πρωτεϊνών, Ταξινόμηση, Κατηγοριοποίηση, Πρόβλεψη, Εξόρυξη δεδομένων

Abstract

Data classification is an important issue in the field of Data Mining due to the broad applications that can be used. There is a number of methods that have been proposed to solve such problems, which are based on well-known data mining techniques such as decision trees or neural networks. However, these types of classification algorithms can not be applied and attribute to sequential data such as protein databases consisting of amino acid sequences.

In this Diploma Thesis, innovative classification algorithmic techniques through sequential patterns were researched to predict the secondary structure of protein. In summary, a basic algorithm was selected for the extraction of the sequential patterns and another algorithm was developed which employs these patterns and a set of amino acid sequences for protein structure prediction. In the matter of predicting protein structures and scoring sequential patterns, a multitude of innovative methodologies has been implemented that theoretically and experimentally overcome the disadvantages of existing algorithms.

Δήλωση Πνευματικών Δικαιωμάτων

Δηλώνω ρητά ότι, σύμφωνα με το άρθρο 8 του Ν. 1599/1986 και τα άρθρα 2,4,6 παρ. 3 του Ν. 1256/1982, η παρούσα εργασία αποτελεί αποκλειστικά προϊόν προσωπικής εργασίας και δεν προσβάλλει κάθε μορφής πνευματικά δικαιώματα τρίτων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο ή/και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα. Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και μόνο.

Ευχαριστίες

Αρχικά θα ήθελα να ευχαριστήσω θερμά τους καθηγητές του Τμήματος Μηχανικών Πληροφορικής και Τηλεπικοινωνιών του Πανεπιστημίου Δυτικής Μακεδονίας και συγκεκριμένα τον επιβλέποντα καθηγητή μου, κ.Μάρκο Τσίπουρα, για την καθοδήγησή του καθ' όλη τη διάρκεια της εκπόνησης της διπλωματικής εργασίας και για το χρόνο που αφιέρωσε για την επίβλεψή της. Υπήρξε σύμβουλος και βοηθός σε όλα τα στάδια της εργασίας και ήταν από τους πιο αφοσιωμένους συνεργάτες στην εργασία αυτή.

Ιδιαιτέρως, θα ήθελα να ευχαριστήσω όλα τα μέλη της οικογένειας μου για την αμέριστη συμπαράσταση, την βοήθεια και την υποστήριξη τους.

Τέλος θερμές ευχαριστίες οφείλω στους φίλους, τους κοντινούς μου ανθρώπους και όλους όσους ήταν δίπλα μου όλα αυτά τα χρόνια για τη στήριξη που μου παρείχαν, υλική και συναισθηματική, και με βοήθησαν να φέρω εις πέρας τους στόχους μου και να πραγματοποιήσω αυτή μου την επιθυμία.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Πρωτεΐνες και η τρισδιάστατη δομή τους

1.1 Εισαγωγή στις πρωτεΐνες

Οι πρωτεΐνες αποτελούν τα πιο άφθονα και πολυδιάστατα, τόσο στη μορφή όσο και στη λειτουργία τους, βιολογικά μακρομόρια γιατί υπάρχουν σε όλα τα κύτταρα και σε όλα τα μέρη των κυττάρων. Χαρακτηρίζονται από μεγάλη ετερογένεια διότι στο ίδιο κύτταρο μπορούν να υπάρχουν χιλιάδες διαφορετικά είδη, με το καθένα εξ αυτών να έχει ιδιαίτερο ρόλο. Λειτουργούν ως καταλύτες, μεταφορείς και αποθηκευτές άλλων μορίων, όπως το οξυγόνο, παρέχουν μηχανική στήριξη και ανοσοπροστασία, δημιουργούν κίνηση, διαβιβάζουν νευρικές ώσεις και ρυθμίζουν την ανάπτυξη και τη διαφοροποίηση.

Όλες οι πρωτεΐνες, είτε προέρχονται από τα πιο αρχέγονα βακτήρια είτε προέρχονται από τις πιο περίπλοκες μορφές ζωής, κατασκευάζονται από το ίδιο οικουμενικό σύνολο 20 αμινοξέων, τα οποία ενώνονται μεταξύ τους με πεπτιδικούς δεσμούς σχηματίζοντας χαρακτηριστικές γραμμικές αλληλουχίες. Αυτή η ομάδα των 20 μορίων μπορεί να θεωρηθεί ως το αλφάβητο στο οποίο γράφεται η γλώσσα της πρωτεϊνικής δομής [1].

Αμινοξύ	Σύντμηση	Σύμβολο
Αλανίνη	A	Ala
Αργινίνη	R	Arg
Ασπαραγίνη	N	Asn
Ασπαρτικό	D	Asp
Βαλίνη	V	Val
Γλουταμικό	E	Glu
Γλουταμίνη	Q	Gln
Γλυκίνη	G	Gly
Θρεονίνη	T	Thr
Ισολευκίνη	I	Ile
Ιστιδίνη	H	His
Κυστεΐνη	C	Cys
Λευκίνη	L	Leu
Λυσίνη	K	Lys
Μεθειονίνη	M	Met
Προλίνη	P	Pro
Σερίνη	S	Ser
Τρυπτοφάνη	W	Trp
Τυροσίνη	Y	Tyr
Φαινυλαλανίνη	F	Phe

Πίνακας 1.1: Συμβολισμοί και σύντμηση των 20 αμινοξέων

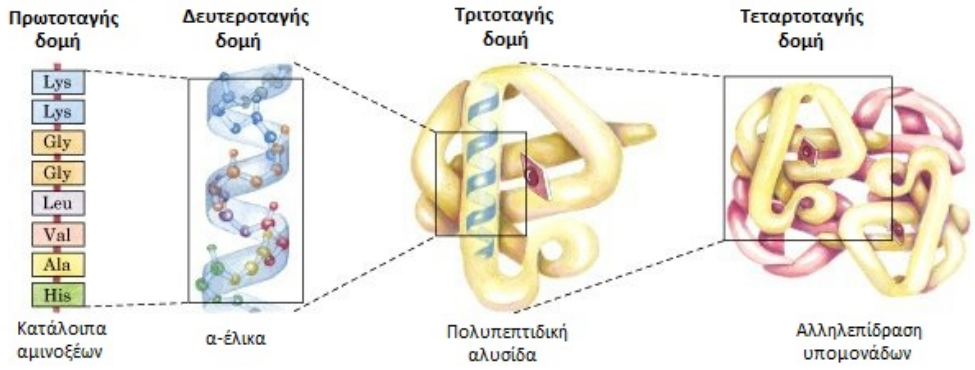
Το πιο αξιοσημείωτο είναι ότι τα κύτταρα παράγουν πρωτεΐνες μ΄ εντυπωσιακά διαφορετικές ιδιότητες συνδέοντας τα ίδια 20 αμινοξέα σε πολλούς διαφορετικούς συνδιασμούς και αλληλουχίες. Η καταπληκτική ποικιλία πρωτεϊνικών λειτουργιών είναι αποτέλεσμα της ποικιλότητας και ποικιλομορφίας αυτών των 20 δομικών στοιχείων. Μια τυπική πρωτεΐνη μπορεί να περιέχει 300 ή περισσότερα αμινοξέα συνδεδεμένα μεταξύ τους με πεπτιδικούς δεσμούς. Κάθε πρωτεΐνη έχει τον δικό της αριθμό και τη δική της αλληλουχία αμινοξέων.

Στον πίνακα 1.1 δίνεται μία αναλυτική περιγραφή των κοινών αμινοξέων των πρωτεϊνών καθώς και ο συμβολισμός τους [1].

1.2 Επίπεδα οργάνωσης πρωτεϊνών

Οι πρωτεΐνες είναι μία σημαντική κατηγορία μεγαλομορίων γι΄ αυτό και η ταξινόμηση τους γίνεται σε διαφορετικά επίπεδα πολυπλοκότητας με ιεραρχική σειρά. Έτσι λοιπόν τις διακρίνουμε σε τέσσερα επίπεδα πρωτεϊνικής δομής και συγκεκριμένα την πρωτοταγή, την δευτεροταγή, την τριτοταγή και την τεταρτοταγή δομή όπως φαίνεται και στο σχήμα 1.1.

Η πρωτοταγής δομή δείχνει την αλληλουχία των αμινοξέων, δηλαδή περιγράφει τους δεσμούς που συνδέουν τα κατάλοιπα των αμινοξέων σε μία πολυπεπτιδική αλυσίδα. Η δευτεροταγής δείχνει την θέση που λαμβάνει η πρωτεϊνική αλυσίδα στο χώρο και πιο συγκεκριμένα αναφέρεται σε ιδιαίτερες, σταθερές διατάξεις αμινοξέων που αποδίδουν ευδιάκριτα δομικά πρότυπα. Η τριτοταγής δομή περιγράφει συνολικά, όλες τις παραμέτρους που αφορούν την αναδίπλωση του πρωτεϊνικού μορίου στο χώρο. Η τεταρτοταγής δομή περιγράφει την αμοιβαία θέση των επιμέρους πρωτεϊνικών αλυσίδων από τις οποίες μπορεί να αποτελείται μία πρωτεΐνη.

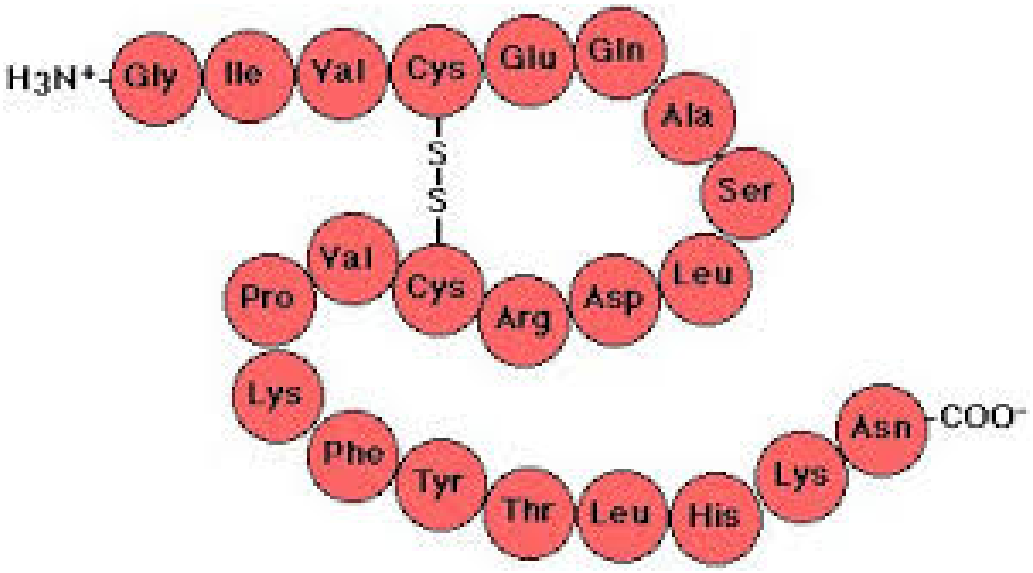


Σχήμα 1.1: Επίπεδα δομικής οργάνωσης των πρωτεϊνών [1]

1.2.1 Πρωτοταγής δομή

Η πρωτοταγής δομή αναφέρεται στην αλληλουχία των αμινοξέων, όπως φαίνεται και στην εικόνα 1.2, που συνδέονται με πεπτιδικούς δεσμούς για να δημιουργήσουν την πρωτεϊνική αλυσίδα. Επίσης η πρωτοταγής ακολουθία μίας πρωτεΐνης καθορίζει πως θα διπλωθεί σε μία μοναδική τρισδιάστατη δομή, δηλαδή καθορίζει την δομή της στον χώρο καθώς και τον λειτουργικό της ρόλο.

Η ορθή διάταξη των αμινοξέων της πρωτεΐνης συνδέονται με την ορθή δομή και σωστή λειτουργία της και αντιστοίχως, μια εσφαλμένη διάταξη συνδέεται με την ακατάλληλη δομή και κακή λειτουργία της. Έτσι εάν αλλάξει έστω ένα αμινοξύ, η πρωτεΐνη μπορεί να χάσει την λειτουργία της, όπως συμβαίνει στην περίπτωση της αναιμίας.



Σχήμα 1.2: Αναπαράσταση μέρους της αλληλουχίας αμινοξέων μίας πρωτεΐνης

1.2.2 Δευτεροταγής δομή

Το επόμενο επίπεδο οργάνωσης των πρωτεϊνών, μετά την πρωτοταγή δομή , είναι η δευτεροταγής διαμόρφωση του μορίου της. Αυτή η διαμόρφωση ανα-

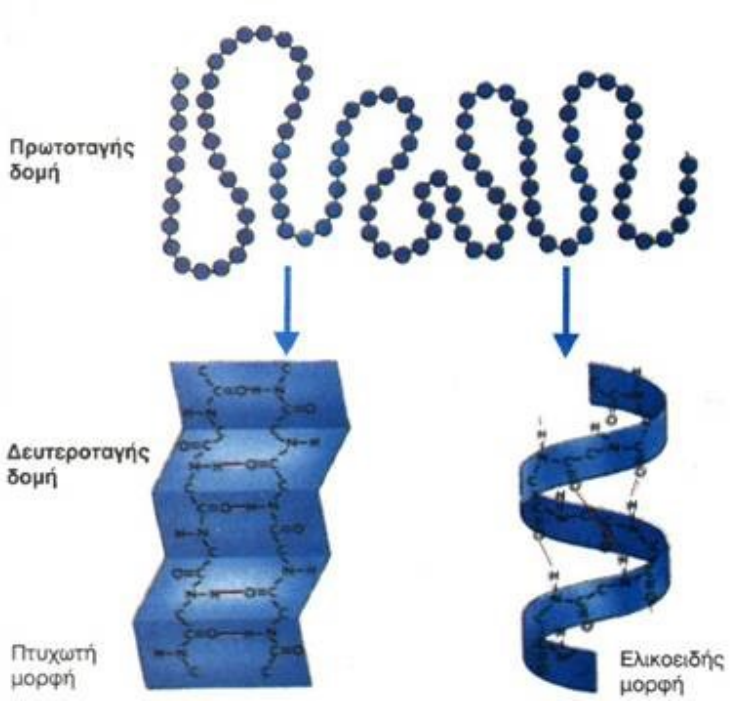
φέρεται στον τρισδιάστατο τρόπο διάταξης των ατόμων που απαρτίζουν την πρωτεϊνική αλυσίδα. Είναι η διάταξη που λαμβάνει στο χώρο η πολυπεπτιδική αλυσίδα γύρω από τον άξονα που σχηματίζεται από τον πεπτιδικό δεσμό.

Είναι αξιοσημείωτο να αναφερθεί, ότι οι Pauling και Corey είχαν προβλέψει την ύπαρξη δευτεροταγών δομών το 1951, αρκετά χρόνια προτού περιγραφεί η πρώτη ολοκληρωμένη πρωτεϊνική δομή, χρησιμοποιώντας θεμελιώδης χημικές αρχές[1]. Έτσι φτάνουμε σήμερα να παρουσιάσουμε ότι ανάλογα με τον τρόπο που αναπτύσσονται οι δεσμοί συναντούμε τις δύο εξής βασικές δευτεροταγείς δομές στις πρωτεΐνες[2].

- Δομή α-έλικας (α-helix)
- Δομή β-πτυχωτό φύλλο ή β-φύλλο (β-plated sheet)

Οι παραπάνω δομές α-έλικα και β-φύλλο αποτελούν τις δύο κοινές δευτεροταγής δομές που συναντώνται στις πρωτεΐνες. Αλλά υπάρχουν και διατάξεις στις οποίες απαντούνται οι δομές α-έλικα και β-φύλλο και συνδέονται μεταξύ τους με άλλες δομές όπως αυτήν των στροφών. Οι στροφές αποτελούνται από ένα σχετικά μικρό αριθμό αμινοξέων και έχουν σημαντικό δομικό και πρωτεϊνικό ρόλο.

Παρόλα αυτά στην εργασία αυτήν χρησιμοποιήθηκαν δύο κλάσεις πρωτεΐνών, αυτές που έχουν μόνο δομή α-έλικας και αυτές που έχουν μόνο δομή β-φύλλου.



Σχήμα 1.3: Δύο κοινά πρότυπα δευτεροταγούς δομής, β-πτυχωτό φύλλο και α-έλικα

1.3 Τεχνικές προσδιορισμού της πρωτεϊνικής δομής

Είναι πλέον γνωστό ότι η σημασία που έχουν οι πρωτεΐνες για τους οργανισμούς είναι σπουδαία καθώς και ότι έχουν τεράστια ποικιλία βιολογικών δράσεων. Για να γίνει γνωστή η λειτουργικότητα, δηλαδή η βιολογική δράση μίας πρωτεΐνης, είναι αναγκαία η γνώση της στερεοδιάστασης της, ειδικότερα πρέπει να παρέχεται η γνώση του τρόπου με τον οποίο η γραμμική αμινοξική της ακολουθία αναδιπλώνεται στον χώρο. Τις τελευταίες δεκαετίες, το ερευνητικό κοινό έχει στραφεί στον καθορισμό της δομής των πρωτεϊνών κυρίως για τον σχεδιασμό φαρμάκων. Παρακάτω αναφέρονται οι βασικότεροι λόγοι για τους οποίους οι ερευνητές κάνουν προσπάθειες για την ανάπτυξη τεχνικών προσδιορισμού

αλλά και κατανόησης της δόμησης των πρωτεϊνών.

- Λειτουργία
- Ορθολογικός σχεδιασμός φαρμάκων
- Πρόβλεψη πρωτεϊνικών αλληλεπιδράσεων
- Κατανόηση μηχανισμών λειτουργίας
- Επίδραση μεταλλάξεων
- Κατανόηση γενετικών ασθενειών
- Κατανόηση των αρχών της πρωτεϊνικής αρχιτεκτονικής
- Σχεδιασμό πρωτεϊνών με προκαθορισμένες ιδιότητες

Γι΄ αυτό το λόγο έχουν αναπτυχτεί πειραματικοί μεθόδοι με τις οποίες μπορεί να γίνει ο εργαστηριακός προσδιορισμός της δομής των πρωτεϊνών. Από τις πιο διαδεδομένες είναι η κρυσταλλογραφική ανάλυση με περίθλαση ακτίνων - Χ (X-ray crystallography), η φασματοσκοπία πυρηνικού, μαγνητικού συντονισμού (NMR spectroscopy) και η κρυοηλεκτρονιακή μικροσκοπία (cryoelectron microscopy)[3]. Αυτές οι μέθοδοι παράγουν σημαντικές πληροφορίες για τον καθορισμό της δομής και με την μεγαλύτερη ακρίβεια από άλλες τεχνικές. Παρόλα αυτά χαρακτηρίζονται ως δαπανηρές, επίπονες και χρονοβόρες. Επομένως, προέκυψε η ανάγκη να ληφθούν πληροφορίες για την φυσική στερεοδιάταξη με άλλες μεθόδους.

Τα τελευταία χρόνια γίνονται γνωστές ολοένα και πιο πολλές αμινοξικές αλληλουχίες πρωτεϊνών. Επιπρόσθετα τα δεδομένα δείχνουν ότι όλη η αναγκαία πληροφορία για το δίπλωμα της πρωτεΐνης στην απαραίτητη δομή βρίσκεται στην γραμμική αμινοξική της αλυσίδα, έχουν γίνει πολλές προσπάθειες ανάπτυξης μεθόδων πρόβλεψης της στερεοδομής μιας πρωτεΐνης από την γραμμική αμινοξική της ακολουθία. Αυτές οι τεχνικές βιοπληροφορικής βασίζονται στην επεξεργασία διάφορων βιολογικών δεδομένων, που βρίσκονται αποθηκευμένα σε βάσεις δεδομένων, τα οποία έχουν εξαχθεί, είτε από εργαστηριακές, είτε από υπολογιστικές μεθόδους. Αυτός ο κλάδος της βιοπληροφορικής στοχεύει στην

αντικατάσταση των υπάρχοντων, δαπανηρών και χρονοβόρων εργαστηριακών τεχνικών αλλά δεν είναι αρκετά αξιόπιστος, διότι η ακρίβεια των υπολογιστικών τεχνικών δεν έχει ξεπεράσει αυτή των εργαστηριακών, εάν και είναι πολύ πιο φθήνοι, τόσο σε χρόνο όσο και σε χρήμα.

1.4 Προγνωστικές μέθοδοι της πρωτεϊνικής δομής από την αμινοξική ακολουθία

Στόχος της σύγχρονης βιοπληροφορικής είναι να αναπτύξει υπολογιστικές τεχνικές με τις οποίες μπορεί να γίνει η πρόβλεψη της αρχικής στερεοδιάστασης των πρωτεϊνών με τέτοια ακρίβεια έτσι ώστε να προχωρήσει η αντικατάσταση των εργαστηριακών μεθόδων. Τα πειραματικά δεδομένα οδηγούν με όλο και μεγαλύτερη βεβαιότητα στη διαπίστωση ότι η απαραίτητη πληροφορία για το δίπλωμα μιας πρωτεΐνης υπάρχει στην πρωτοταγή της δομή, πολλές προσπάθειες έχουν γίνει μέχρι σήμερα για την πρόγνωση της 3D-δομής των πρωτεϊνών από την ακολουθία και μόνο. Η βιολογική έρευνα απαιτεί λοιπόν περισσότερες μεθοδολογίες με της οποίες το θέμα της εύρεσης της πρωτεϊνικής δομής στο χώρο να επιλυθεί με την χρήση της αμινοξικής ακολουθίας της πρωτεΐνης.

Έχουν επινοηθεί διάφοροι εμπειρικοί αλγόριθμοι πρόβλεψης δευτεροταγούς δομής μιας πρωτεΐνης και με την βελτιστοποίηση αυτών παρέχεται μία αρχική εικόνα της πρωτεϊνικής δομής η οποία κατέπεκταση οδηγεί στην τελική φυσική στερεοδομή της. Οι αλγόριθμοι αυτοί βασίζονται σε μεθόδους που ταξινομούνται σε τρεις μεγάλες κατηγορίες και αυτές είναι οι homology modeling μέθοδοι, οι folding recognition μέθοδοι και οι ab initio μέθοδοι [3].

1.4.1 Μοντελοποίηση με βάση την ομολογία

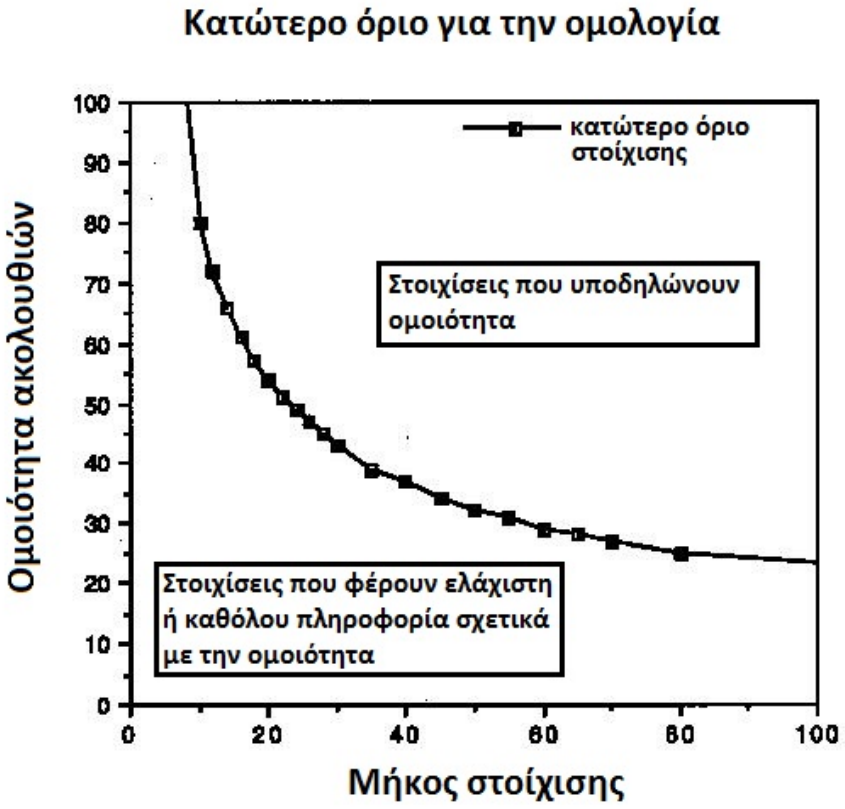
Η μοντελοποίηση με βάση την ομολογία (homology modeling) μέθοδοι εφαρμόζονται σε πρωτεΐνες άγνωστης δομής, αν βρεθεί ομόλογη της με λυμένη, γνωστή δομή[3]. Τα τελευταία χρόνια, ο αριθμός των λυμένων κρυσταλλογραφικά πρωτεϊνικών δομών που βρίσκονται κατατεθειμένες σε ελεύθερα προσβάσιμες τράπεζες δεδομένων (κατατεθειμένων στην Protein Data Bank (PDB)[4])

διαρκώς αυξάνεται.

Η ανάλυση των στοιχίσεων μεταξύ ακολουθιών με γνωστές δομές αναδεικνύει ότι ζευγάρια πρωτεϊνών με μεγαλύτερη ομοιότητα από 40% (για μήκη στοίχισης μεγαλύτερα των 100 καταλοίπων) έχουν ίδιες δομές, δηλαδή το βασικό δίπλωμα τους είναι παρόμοιο. Αν, λοιπόν, διαπιστωθεί ομολογία ανάμεσα σε μια πρωτεΐνη Α άγνωστης δομής με μία γνωστής Β τότε μπορεί να μοντελοποιηθεί μια δομή για την Α με την χρήση της Β. Το σχήμα 1.4 που απεικονίζει την ασφαλή περιοχή, στην οποία δύο πρωτεΐνες θεωρούνται ομόλογες, συναρτήση του μήκους στοίχησης των καταλοίπων τους.

Συνοπτικά τα βήματα τις μεθόδου είναι τα εξής:

- Εύρεση του πρότυπου και πραγματοποίηση της στοίχισης
- Διόρθωση της στοίχισης
- Κατασκευή του σκελετού της κύριας αλυσίδας
- Μοντελοποίηση των βρόχων και των πλευρικών αλυσίδων
- Βελτιστοποίηση του μοντέλου
- Έλεγχος ποιότητας του μοντέλου



Σχήμα 1.4: Η σχέση της ομοιότητας στοίχισης, με το μήκος και την ποιότητα της στοίχισης

Όπως είπαμε, η μοντελοποίηση με βάση την ομολογία, είναι η ενδεικνυόμενη μέθοδος για τις περιπτώσεις στις οποίες μια ομόλογη πρωτεΐνη με γνωστή δομή μπορεί να αναγνωριστεί εύκολα με μεθόδους στοίχισης αλληλουχιών. Με την χρήση του homology modeling έγινε δυνατόν να μεγαλώσει το πλήθος των γνωστών δομών, να εισάγονται όλο και περισσότερες πρωτεϊνικές δομές στην Protein Data Bank και αυτό να βοηθήσει στην μετέπειτα αναγνώριση πρωτεϊνών με άγνωστη δομή.

1.4.2 Αναγνώριση διπλώματος και ύφανση

Η ύφανση (threading) ή αλλιώς αναγνώριση διπλώματος (fold recognition) είναι μια τεχνική που χρησιμοποιείται σε περιπτώσεις κατά τις οποίες η πρωτεΐνη-στόχος δεν έχει ανιχνεύσιμη ομοιότητα σε επίπεδο αλληλουχίας με κάποια πρωτεΐνη γνωστής δομής, αλλά μοιράζεται το ίδιο δίπλωμα με αυτές.

Με τη διαδικασία αυτή, γίνεται έλεγχος αν η αμινοξική ακολουθία είναι συμβατή με κάποια από τις γνωστές δομές και μετά κατασκευάζεται η στοίχιση με την δομή αυτή. Η βασική διαφορά από την μοντελοποίηση με βάση την ομολογία, στην οποία το πρότυπο το αντιμετωπίζεται ως αλληλουχία, είναι ότι στην ύφανση το πρότυπο χρησιμοποιείται σαν δομή. Οι μέθοδοι αναγνώρισης διπλώματος έχουν αποκτήσει μεγάλη δημοφιλία, λόγω της γνωστής αρχής ότι η δομή συντηρείται περισσότερο από την αλληλουχία και, κατά συνέπεια, από την παρατήρηση ότι ακόμα και διαφορετικές πρωτεΐνες μπορεί να έχουν παρόμοια δομή. Επιπλέον πιστεύεται γενικά ότι ο αριθμός των δομών είναι πεπερασμένος και το threading γίνεται σε ένα σύνολο βασικών δομών που αντιπροσωπεύουν τις κυριότερες πρωτεϊνικές δομές. Έτσι, μια πρωτεΐνη με μη ανιχνεύσιμη ομοιότητα σε επίπεδο αλληλουχίας αμινοξέων, είναι παρ' όλα αυτά πολύ πιθανό να μπορεί να ταυτιστεί με κάποια γνωστή στερεοδιάταξη.

1.4.3 Ab initio

Η ab initio μέθοδος χρησιμοποιείται όταν η πρωτεΐνη στόχος δεν μπορεί να ταυτοποιηθεί ούτε με βάση την ομολογία αλλά ούτε και με βάση το δίπλωμα, έχει σαν είσοδο την αμινοξική ακολουθία και κάνει χρήση βασικών αρχών της φυσικής (αλληλεπιδράσεις ατόμων και υπολογισμοί ενέργειας).

Ειδικότερα για την πρωτεΐνη-στόχο, δηλαδή την πρωτεΐνη με άγνωστη δομή αλλά γνωστή αλληλουχία αμινοξεών, γίνεται αναζήτηση όλων των δυνατών διαμορφώσεων στον χώρο της αλυσίδας της, έτσι ώστε να βρεθεί η πιο σταθερή δομή που μπορεί να έχει η πρωτεΐνη στον χώρο. Για την τεχνική αυτή είναι απαραίτητοι οι φυσικοχημικοί κανόνες που ορίζουν την ενέργεια της εκάστοτε πρωτεΐνης.

1.5 Ανάγκη ανάπτυξης προγνωστικών μεθόδων για τη δευτεροταγή πρωτεϊνική δομή

Είναι πλέον γνωστό ότι η βιολογική δράση των πρωτεϊνών εξαρτάται κυρίως από την δομή τους στο χώρο, από τον τρόπο δηλαδή με τον οποίο η γραμμική αμινοξική τους ακολουθία αναδιπλώνεται στο χώρο. Επίσης συνδυασμοί στοιχείων δευτεροταγούς δομής εμφανίζονται συχνά σε δομές πρωτεϊνών. Έτσι, επιτυχής πρόγνωση των στοιχείων δευτεροταγούς δομής με την χρήση της πρωτεϊνικής αμινοξικής αλληλουχίας (πρωτοταγής δομή) δύναται να βοηθήσει στον καθορισμό της φυσικής δομής της πρωτεΐνης (τριτοταγής δομή).

Έχουν αναπτυχθεί πολλές τεχνικές καθορισμού της πρωτεϊνικής δομής αλλά ακόμη κι αν θεωρήσουμε πως υπάρχει μεγάλο ποσοστό επιτυχίας για αυτές τις τεχνικές, ένα σφάλμα κατά την εκτέλεση τους μπορεί να αποφέρει αδιόρθωτες επιπτώσεις στην μοντελοποίηση των πρωτεϊνών. Το πρόβλημα του πρωτεϊνικού διπλώματος, της πρόγνωσης δηλαδή της δομής απευθείας από την αμινοξική αλληλουχία είναι στην ουσία ένα από τα μεγαλύτερα προβλήματα της σύγχρονης βιολογίας, με μεγάλες προσδοκίες και δεκάδες ερευνητές έχουν ασχοληθεί.

Η κεντρική ιδέα, γύρω από την οποία περιστρέφονται οι περισσότερες μέθοδοι πρόγνωσης δευτεροταγούς δομής, είναι το γεγονός ότι τμήματα διαδοχικών καταλοίπων έχουν προτιμήσεις να βρίσκονται σε συγκεκριμένες καταστάσεις δευτεροταγούς δομής. Επομένως, το πρόβλημα στην ουσία ανάγεται σε πρόβλημα αναγνώρισης προτύπων (pattern recognition) και αντιμετωπίζεται με τη χρήση αντίστοιχων αλγόριθμων. Στο επόμενο κεφάλαιο θα αναλύσουμε τους κύριους, γνωστούς pattern recognition αλγόριθμους που χρησιμοποιούνται γενικότερα στην εύρεση επαναλαμβανόμενων προτύπων και θα μπορούσαν να χρησιμοποιηθούν για την πρόγνωση δευτεροταγούς δομής στην βιοπληροφορική.

Κεφάλαιο 2

Αλγόριθμοι προσδιορισμού της δευτεροταγούς δομής των πρωτεϊνών

2.1 Εμπειρικοί αλγόριθμοι πρόγνωσης δευτεροταγούς δομής

Η κεντρική ιδέα, γύρω απο την οποία περιστρέφονται οι περισσότερες μέθοδοι πρόγνωσης δευτεροταγούς δομής, είναι να μπορεί, μέσω μίας αρχικής στερεοδιάταξης να αναπτυχθεί η φυσική στερεοδομή της πρωτεΐνης. Έτσι η πρόβλεψη της τρισδιάστατης δομής της πρωτεΐνης γίνεται αρχικά σε ένα απλοποιημένο επίπεδο. Η ανάγκη ανάπτυξης τέτοιων μεθόδων έγκειται στο γεγονός ότι οι τέχνικες που αναφέρθηκαν στο προηγούμενο κεφάλαιο (homology,threading,ab initio) δεν εξάγουν σίγουρα συμπεράσματα από μια αναζήτηση ομοιότητας και μόνο και έτσι συνεχίζουν να υπάρχουν πρωτεΐνες αγνώστης δομής και λειτουργίας.

Το πρόβλημα αυτό λοιπόν έρχονται να λύσουν μέθοδοι πρόγνωσης, οι οποίοι χρησιμοποιώντας μόνο την ακολουθία της πρωτεΐνης, μπορούν να προβλέψουν τα δομικά ή λειτουργικά χαρακτηριστικά της. Μονοδιάστατες προγνώσεις όπως

αυτήν της δευτεροταγούς δομής είναι δυνατές, με αξιόλογα πολλές φορές αποτελέσματα. Κάποιες από τις μεθόδους αναγράφονται στον παρακάτω πίνακα 2.1 [5].

Αλγόριθμος	Γενιά
GORI Chou & Fasman Lim	1
ALB GORIII COMBINE	2
PHD JPRED PSIPRED	3

Πίνακας 2.1: Βασικοί εμπειρικοί αλγόριθμοι πρόβλεψης της δευτεροταγούς δομής της πρωτεΐνης

Στην συνέχεια παρουσιάζονται αναλυτικότερα κάποιοι από τους αλγόριθμους του πίνακα 2.1 και συγκεκριμένα οι αλγόριθμοι τρίτης γενιάς εφόσον έχουν μεγαλύτερο ποσοστό ακρίβειας από τους υπόλοιπους.

2.1.1 PHD

Η μέθοδος PHD[6] προβλέπει την πρωτεϊνική δευτεροταγή δομή με ακρίβεια μεγαλύτερη του 70%, αναπτύχθηκε από τους Rost & Sander, και χρησιμοποιεί πολλαπλή στοίχιση ακολουθιών. Οι προβλέψεις του PHD ήταν πρωτοποριακές γιατί η μέθοδος χρησιμοποιεί ένα σύστημα νευρωνικών δικτύων (jury of networks), μεγάλο σύνολο δεδομένων, ένα δεύτερο δίκτυο για φιλτράρισμα των αποτελεσμάτων (structure-to-structure network) αλλά και το πιο σημαντικό από όλα έγινε χρήση της εξελικτικής πληροφορίας από πολλαπλή στοίχιση ακολουθιών.

Τα κυριότερα στοιχεία που οδήγησαν στην επιτυχία αυτής της μεθόδου και στην μεγάλη ακρίβεια της είναι το πολυ-επίπεδο σύστημα που χρησιμοποιείται,

η πρόβλεψη β-φύλλων μέσω μίας ισορροπημένης διαδικασίας μάθησης καθώς όπως έχει αναφερθεί και η εκμετάλλευση της εξελικτικής πληροφορίας. Από την ανάπτυξη του PHD είναι ευρέως γνωστό στην κοινότητα της βιοπληροφορικής οτι τέτοιο ποσοστό ακρίβειας μπορεί να επιτευχθεί μόνο με την χρήση πολλαπλών στοιχίσεων.

2.1.2 JRRED

Το Jpred[7] που αναπτύχθηκε το 1998 ήταν μία από τις πρώτες μεθόδους που δεν είναι ένα μεμονωμένο πρόγραμμα, αλλά ένας δικτυακός συνδυασμός μεθόδων πρόβλεψης δευτεροταγούς δομής. Μαζί με την σύγκριση των αποτελεσμάτων όλων των μεθόδων, χρησιμοποιεί και την εξελικτική πληροφορία γι΄ αυτό το λόγο παράγει και ένα υψηλής ποιότητας, συνδυαστικό αποτέλεσμα πρόβλεψης.

Αρχικά χρησιμοποίησε μία πληθώρα αλγορίθμων της εποχής όπως οι NNSSP, DSC, PREDATOR, MULPRED, PHD, ZPRED και αυτό επέφερε μία αυξημένη απόδοση. Το 2015, η μέθοδος έφτασε την τέταρτη έκδοση η οποία παρέχεται μέσω server στους χρήστες και μπορεί να παρουσιάσει τα αποτελέσματα πρόβλεψης της δευτεροταγούς δομής της πρωτεΐνης και με γραφικές απεικονήσεις.

2.1.3 PSIRRED

Το PSIPRED[8] αναπτύχθηκε το 1999 και θεωρείται μία απλή μέθοδος πρόγνωσης δευτεροταγούς δομής που χρησιμοποίησε τα αποτελέσματα του PSIBLAST (Position Specific Iterated - BLAST)[9]. Ενσωματώνει δύο διαδοχικά νευρωνικά δίκτυα με προς τα εμπρός τροφοδότηση τα οποία προσφέρουν καλύτερη απόδοση και γι΄ αυτό η μέθοδος PSIPRED φτάνει το ποσοστό του 77% ακρίβειας.

Θεωρείται μια αξιόπιστη μέθοδος εφόσον χρησιμοποιεί και μεγάλο σύνολο δεδομένων εκπαίδευσης. Τον Απρίλιο του 2010 είχε αναπτυχθεί η τρίτη έκδοση του PSIPRED που προσφέρει ακόμη μεγαλύτερες τιμές ακρίβειας.

2.2 Αλγόριθμοι αναγνώρισης προτύπων

Οι εμπειρικοί αλγόριθμοι πρόγνωσης δευτεροταγούς δομής μπορούν να δώσουν μια αρχική στερεοδιάταξη, περιορίζοντας έτσι την έρευνα για τη φυσική δομή αλλά οι σύγχρονοι μέθοδοι βιοπληροφορικής για την πρόγνωση της δευτεροταγούς δομής περιστρέφονται στο γεγονός ότι τμήματα διαδοχικών αμινοξέων έχουν προτιμήσεις να βρίσκονται σε συγκεκριμένες καταστάσεις δευτεροταγούς δομής. Ως εκ τούτου, μπορούν να χρησιμοποιηθούν αλγόριθμοι αναγνώρισης προτύπων, οι οποίοι αρχικά δημιουργήθηκαν για εύρεση προτύπων σε διαφορετικά σύνολα δεδομένων. Έτσι το πρόβλημα πρόγνωσης μπορεί να επιλυθεί μέσω προτύπων, εντοπίζοντας αλληλουχίες καταλοίπων κατάλληλου μήκους για την επίτευξη όσο το δυνατόν μεγαλύτερης τιμής ακρίβειας.

Υπάρχουν αρκετοί αλγόριθμοι που εφαρμόζονται στο πεδίο αναγνώρισης - εξόρυξης προτύπων όπως οι PrefixSpan [10], GSP[11], SPADE[12], SPAM [13], LAPIN [14], ClaSP [15], BIDE+ [16], MaxSP[17] όπου ο καθένας από αυτούς έχει πλεονεκτήματα και μειονεκτήματα. Επίσης έχουν ένα εύρος εφαρμογών, χρησιμοπιούνται σε πολλά επιστημονικά πεδία και όχι μόνο στις βιολογικές ακολουθίες. Κάποιες από τις εφαρμογές τους εμφανίζονται παρακάτω.

- Αναγνώριση της καταναλωτικής συμπεριφοράς
  - Κατανόηση του τι, πότε και πως αγοράζουν οι πελάτες
  - Εφαρμογή στα στοιχεία πωλήσεων σούπερ μάρκετ
- Αναγνώριση της ηλεκτρονικής συμπεριφοράς
  - Κατανόηση του τι και πότε κάνουν ΄κλικ΄ οι ηλεκτρονικοί χρηστες
  - Εφαρμογή στις αλληλουχίες ΄κλικ΄ χρηστών όταν πλοηγούνται στο ίντερνετ
- Ιατρικοφαρμακευτικές θεραπευτικές αγωγές
- Φυσικές καταστροφές
- Χρηματιστήριο
- Μηχανικές διαδικασίες

- Μηχανική λογισμικού σε αλληλουχίες εκτελέσεων προγραμμάτων λογισμικού
- Βιολογικές αλληλουχίες όπως πρωτεϊνικές και DNA

Στην συνέχεια θα γίνει μία περιγραφή των βασικότερων από τους βασικούς αλγόριθμους που αναφέρθηκαν παραπάνω και θα αναλυθούν τα χαρακτηριστικά λειτουργίας, εκτέλεσης αλλά και απόδοσης τους.

2.2.1 GSP

Ο Generalized Sequential Patterns (GSP), είναι ένας αλγόριθμος στην λογική του Apriori[11] που χρησιμοποιεί χρονικούς περιορισμούς με τεχνικές κινούμενου χρονικά παραθύρου. Η αξιολόγιση δείχνει ότι ο GSP κλιμακώνεται γραμμικά με τον αριθμό του συνόλου δεδομένων και έχει πολύ καλές ιδιότητες κλιμάκωσης σε σχέση με τον αριθμό των στοιχείων της κάθε αλληλουχίας. Παρακάτω δίνονται τα βασικά βήματα[18] καθώς και ψευδοκώδικας του αλγόριθμου στο σχήμα 2.1.

- Αρχικά όλα τα στοιχεία της βάσης δεδομένων είναι υποψήφια πρότυπα μήκους-1
- Για κάθε επίπεδο (π.χ. αλληλουχίες μήκους-ν)
  - Σκανάρισμα της βάσης δεδομένων για τον υπολογισμό της υποστήριξης κάθε υποψήφιου προτύπου
  - Δημιουργία υποψήφιων προτύπων μήκους-(ν+1) από της συχνά πρότυπα μήκους-ν χρησιμοποιώντας τον αλγόριθμο Apriori
- Επανάληψη μέχρι να μην υπάρχουν συχνά πρότυπα ή υποψήφια πρότυπα

```
GSP(In: Database 𝒟,In: Integer min_supp, In/Out: Set F )
 1: ℱ1 ← {frequent 1-sequences}
 2: for k ← 2; ℱk−1 ≠ 0; k ← k + 1 do
 3:    ℱk ← ∅
 4:    Ck ← candidates created from ℱk−1
 5:    for all β ∈ Ck do
 6:       β.support ← support of β in 𝒟
 7:       if β.support ≥ min_supp then
 8:          ℱk ← ℱk ⋃ β
 9:       end if
10:    end for
11:    F ← F ⋃ ℱk
12: end for
```

Σχήμα 2.1: Ψευδοκώδικας του αλγόριθμου GSP

Ο αλγόριθμος GSP έχει οφέλη από το κλάδεμα που πραγματοποιείται από τον αλγόριθμο Apriori διότι μειώνει το πεδίο αναζήτησης παρόλα αυτά σαρώνει την βάση δεδομένων πολλές φορές και εξάγει ένα τεράστιο σύνολο υποψήφιων προτύπων, στοιχεία τα οποία δημιουργούν την ανάγκη ανάπτυξης μεθόδων αποτελεσματικότερης εξόρυξης προτύπων.

2.2.2 SPADE

Ο SPADE (Sequential PAttern Discovery using Equivalent Class) είναι ένας αλγόριθμος, ο οποίος αναπτύχθηκε από τον Zaki το 2001[12]. Είναι μία προσέγγιση με κάθετη μορφή των αλληλουχιών της βάσης δεδομένων διότι μία ακολουθία έχει αντιστοιχιστεί με την μορφή ʿSID, EIDʾ όπου SID είναι το sequence_id και EID το event_id. Η εξόρυξη των προτύπων γίνεται αυξάνοντας το κατά ένα στοιχείο την φορά τα πρότυπα χρησιμοποιώντας τον αλγόριθμο Apriori[11], δηλαδή τα πρότυπα μήκους-ν γίνεται από τα πρότυπα μήκους-(ν-1). Παρακάτω δίνεται ο ψευδοκώδικας του αλγόριθμου στο σχήμα 2.2.

SPADE(**In**: AtomSet ϵ,**In**: Integer min_supp, **In/Out**: Set $\mathcal{F}$)

for all atoms $A_i \in \epsilon$ **do**
  $T_i \leftarrow \{\}$
  for all atoms $A_j \in \epsilon, j \geq i$ and all combinations α of A_i, A_j **do**
    $\mathcal{L}(\alpha) =$ temporal TID list join of $\mathcal{L}(A_i)$ with $\mathcal{L}(A_j)$
    if $Supp(\alpha) \geq min\_supp$ **then**
      $T_i \leftarrow T_i \bigcup \{\alpha\}$
      $F = F \bigcup \alpha$
    end if
  end for
  Spade(T_i, *min_supp*, $\mathcal{F}$)
end for

Σχήμα 2.2: Ψευδοκώδικας του αλγόριθμου SPADE

Υπάρχουν αρκετοί λόγοι για του οποίους ο αλγόριθμος SPADE υπερτερεί του GSP. Ένας από αυτούς είναι ότι η απόδοση του SPADE είναι σχεδόν διπλάσια του GSP σε χαμηλές τιμές υποστήριξης. Επίσης ο SPADE χρησιμοποιεί μόνο μία προσωρινή ένωση μεταξύ των λιστών με τα id. Όσο το μήκος των συχνών προτύπων αυξάνεται τόσο το μέγεθος των λιστών με τα id μειώνεται έχοντας ως αποτέλεσμα γρήγορες ενώσεις. Τέλος όσο η ελάχιστη υποστήριξη μειώνεται, περισσότερα και μεγαλύτερα σε μήκος, συχνά πρότυπα εξάγονται. Ο GSP κάνει μία σάρωση ολόκληρης της βάσης δεδομένων σε κάθε επανάληψη. Ο SPADE από την άλλη πλευρά περιορίζεται σε μόνο τρείς σαρώσεις και έτσι είναι πιο αποδοτικός σε ότι αφορά την είσοδο/έξοδο δεδομένων.

Παρόλα αυτά υπάρχουν κάποια σημεία τα οποία πρέπει να προσεχτούν έτσι ώστε ο αλγόριθμος να χρησιμοποιηθεί στο κατάλληλο πρόβλημα. Ο SPADE εξάγει πολλά πρότυπα και ιδιάιτερα πρότυπα μήκους-2. Έπειτα το μήκος των προτύπων αυξάνεται σε κάθε σάρωση της βάσης δεδομένων και όπως αναφέρθηκε παραπάνω πραγματοποιούνται 3 συνολικές σαρώσεις. Γι΄ αυτό το λόγο, ο SPADE είναι ανεπαρκείς στην εξόρυξη προτύπων μεγάλου μήκους[19].

2.2.3 PrefixSpan

Ο αλγόριθμος PrefixSpan (Prefix-Projected Sequential Pattern Growth) αναπτύχθηκε το 2001[10] και ακολουθεί την μεθοδολογία διαίρει και βασίλευε. Δηλαδή παράγει ακολουθιακά πρότυπα με προοδευτική κατάτμηση των βάσεων δεδομένων σε μικρότερες υπο-βάσεις δεδομένων χρησιμοποιώντας προθεματικές (Prefix) και μεταθεματικές (Suffix) προβολές.

Παρακάτω δίνονται τα βασικά βήματα[20] καθώς και ψευδοκώδικας του αλγόριθμου στο σχήμα 2.3.

- Αρχικά όλες οι ακολουθίες σε κάθε ομάδα έχουν το ίδιο πρόθεμα(prefix)
- Σκανάρισμα της κάθε ομάδας για την εύρεση των συχνές ακολουθίες μήκους-1
- Διαχωρισμός της βάσης δεδομένων σε ομάδες σύμφωνα με τις εν λόγω συχνές ακολουθίες
  - Κάθε ομάδα είναι η προβολή της βάσης δεδομένων σε σχέση με την αντίστοιχη μήκους-1 ακολουθία
  - Εάν έχουμε χ προθέματα τότε θα υπάρχουν και χ ομάδες με στοιχεία της βάσης
- Εύρεση των υποσύνολων των ακολουθιακών προτύπων
- Κατασκευή των αντίστοιχων προβολών της βάσης και αναδρομική εξόρυξη κάθεμιας

PREFIXSPAN-RECURSIVE(**In**: Database $\mathcal{D}_\alpha$, **In**: Sequence α, **In**: Integer min_supp, **In/Out**: Set $\mathcal{F}$)

1: $\mathcal{F}_1 \leftarrow$ {frequent items in $\mathcal{D}_\alpha$}
2: **for all** items $b_i \in \mathcal{F}_1$ **do**
3: $\beta = (\alpha_1 \rightarrow \cdots \rightarrow (\alpha_n \bigcup \{b_i\}))$
4: $\gamma = (\alpha_1 \rightarrow \cdots \rightarrow \alpha_n \rightarrow (b_i))$
5: **if** *Supp*$(\beta, \mathcal{D}_\alpha) \geq$min_supp **then**
6: $\mathcal{F} \leftarrow \mathcal{F} \bigcup \{\beta\}$
7: $\mathcal{D}' \leftarrow (\mathcal{D}_\alpha)|_\beta$
8: Prefixspan-Recursive($\mathcal{D}'$, β, min_supp, $\mathcal{F}$)
9: **end if**
10: **if** *Supp*$(\gamma, \mathcal{D}_\alpha) \geq$min_supp **then**
11: $\mathcal{F} \leftarrow \mathcal{F} \bigcup \{\gamma\}$
12: $\mathcal{D}' \leftarrow (\mathcal{D}_\alpha)|_\gamma$
13: Prefixspan-Recursive($\mathcal{D}'$, γ, min_supp, $\mathcal{F}$)
14: **end if**
15: **end for**

Σχήμα 2.3: Ψευδοκώδικας του αλγόριθμου PrefixSpan

Ο PrefixSpan χρειάζεται αρκέτο χρόνο διότι πρέπει να κατασκευάσει και να σκανάρει όλες τις ομάδες που δημιουργούνται με τις προβολές της βάσης δεδομένων και αυτό είναι το μεγαλύτερο μειονέκτημα του αλγορίθμου. Τα πλεονετήματα είναι ότι δεν χρειάζεται να δημιουργηθούν υποψήφια πρότυπα και ότι οι ομάδες που δημιουργούνται με τις προβολές συρρικνώνονται.

Κεφάλαιο 3

Βιβλιογραφική ανασκόπηση

3.1 Παλαιότερες έρευνες

Στην βιβλιογραφία έχουν παρουσιαστεί αρκετές έρευνες που παρουσιάζουν αλγοριθμικές τεχνικές για την πρόβλεψη της δευτεροταγούς δομής πρωτεϊνών. Κάποιες από αυτές χρησιμοποιούν μεθόδους εξόρυξης προτύπων που αναφέρθηκαν στο Κεφάλαιο 2. Κάποιες άλλες διαφέρουν στον τρόπο ταξινόμησης μέσω ακολουθιακών προτύπων. Παρακάτω γίνεται αναφορά σε μερικές από αυτές.

3.1.1 A two-stage methodology for sequence classification based on sequential pattern mining and optimization

Αυτή η έρευνα παρουσιάζει μία μεθοδολογία για ταξινόμηση ακολουθιών, η οποία εφαρμόζει SPM (Sequential Pattern Mining) μεθόδους και βελτιστοποίηση, σε μία διαδικασία δύο σταδίων [21]. Στο πρώτο στάδιο ορίζεται ένα μοντέλο ταξινόμησης ακολουθιών, βασισμένο σε ένα σύνολο απο ακολυθιακά πρότυπα, και εισάγονται δύο σύνολα από βάρη, ένα για τα πρότυπα και ένα για τις κλάσεις.

Στο δεύτερο στάδιο χρησιμοποιείται μία τεχνική βελτιστοποίησης για να εκτιμήσει τις τιμές των βαρών και να επιτευχθεί η βέλτιστη ακρίβεια ταξινόμησης. Πραγματοποιείται εκτεταμένη αξιολόγηση της μεθοδολογίας, μεταβάλλοντας τον αριθμό των ακολουθιών, τον αριθμό των προτύπων και τον αριθμό των κλάσεων.

3.1.2 An optimized sequential pattern matching methodology for sequence classification

Σε αυτήν την μελέτη παρουσιάζεται μια μεθοδολογία ταξινόμησης ακολουθιών βασισμένη στην εξόρυξη ακολουθιακών προτύπων και σε αλγόριθμους βελτιστοποίησης[22]. Η προτεινόμενη μεθοδολογία δημιουργεί αυτόματα ένα μοντέλο ταξινόμησης ακολουθιών, που βασίζεται σε δύο στάδια. Στο πρώτο στάδιο ένας αλγόριθμος εξόρυξης ακολουθιακών προτύπων εφαρμόζεται σε ένα σύνολο ακολουθιών και με αυτόν τον τρόπο εξάγονται ακολουθιακά πρότυπα. Έπειτα, η βαθμολογία του κάθε προτύπου υπολογίζεται σε σχέση με κάθε ακολουθία χρησιμοποιώντας μία συνάρτηση βαθμολόγησης και η βαθμολογία κάθε κλάσης εκτιμάται από το άθροισμα των βαθμολογιών συγκεκριμένων προτύπων. Κάθε βαθμολογία ενημερώνεται πολλαπλασιάζοντας ένα βάρος και έτσι το αποτέλεσμα του πρώτου στάδιου είναι η μήτρα σύγχυσης (Confusion Matrix).

Στο δεύτερο στάδιο μία τεχνική βελτιστοποιήσης, αποσκοπεί στην εύρεση ενός συνόλου με βάρη που ελαχιστοποιούν την αντικειμενική συνάρτηση, χρησιμοποιώντας τον confusion matrix. Το σύνολο των προτύπων που έχουν εξαχθεί καθώς και τα βέλτιστα βάρη των κλάσεων συνθέτουν το μοντέλο ταξινόμησης ακολουθιών. Επίσης διεξήχθη εκτενής αξιολόγηση της μεθοδολογίας στην ταξινόμηση δομής πρωτεϊνών, μεταβάλλοντας τον αριθμό των δεδομένων εκπαίδευσης (training sequences), εξέτασης (test sequences), προτύπων και κλάσεων. Η μεθοδολογία συγκρίνεται με άλλες όμοιες προσεγγίσεις ταξινόμησης ακολουθιών. Η προτεινόμενη μεθοδολογία επιδεικνύει πλεονεκτήματα, όπως η αυτόματη ανάθεση βάρους σε κλάσεις χρησιμοποιώντας τεχνικές βελτιστοποίησης.

3.1.3 Protein structure prediction by means of sequential pattern mining

Σε αυτήν την έρευνα γίνεται μελέτη για την ταξινόμηση ακολουθιών η οποία συνδυάζει δύο τεχνικές εξόρυξης δεδομένων [23]. Αυτές είναι η SPM (Sequential Pattern Mining) και οι αλγόριθμοι ταξινόμησης, οι οποίες παρέχουν την κατάλληλη δυνατότητα επιλογής για τα ακολουθιακά δεδομένα. Σε αυτήν την προσέγγιση για την αναγνώριση προτύπων χρησιμοποιήθηκε ο αλγόριθμος PrefixSpan. Επίσης χρησιμοποιήθηκε ένα σύνολο δεδομένων SCOPE, το οποίο είναι ένα ταξινομημένο σύνολο δεδομένων, αλλά και ένα σύνολο δεδομένων ASTRAL, το οποίο είναι ένα σύνολο ακολουθιακών δεδομένων του SCOPE.

Η προτεινόμενη μεθοδολογία έχει κάποια στάδια. Αρχικά επιλέγονται οι παράμετροι για να ενσωματωθούν στον αλγόριθμο PrefixSpan και στην συνέχεια εξάγονται πρότυπα από τα ακολουθιακά πρωτεϊνικά δεδομένα. Υπάρχει μία συνάρτηση βαθμολογίας, η οποία υπολογίζει την βαθμολογία μίας πρωτεϊνικής ακολουθίας με άγνωστη κλάση, για κάθε πρότυπο. Έπειτα η τελική βαθμολογία την άγνωστης αυτής πρωτεΐνης υπολογίζεται σε σχέση με μία κλάση, οδηγώντας στην ταξινόμηση της πρωτεΐνης.

3.1.4 A PSO-AB classifier for solving sequence classification problems

Αυτή η έρευνα προτείνει μία αποτελεσματική μεθοδο ταξινόμησης, δύο βημάτων, βασισμένη σε SPM μεθόδους για να επιλύσει σύνθετα προβλήματα ταξινόμησης ακολουθιών [24]. Στο πρώτο βήμα, κατά την διάρκεια εξόρυξης διαδοχικών προτύπων, τα ανώφελα επαναλαμβανόμενα πρότυπα αναγνωρίζονται όταν το πρότυπο είναι μία υπό-ακολουθία των άλλων ακολουθιών. Έτσι δημιουργείται μία λίστα από μικρά ακολουθιακά πρότυπα, εξαιρώντας περιττά πρότυπα, η οποία χρησιμοποιείται ως είσοδος για το δεύτερο βήμα.

Στο δεύτερο βήμα γίνεται χρήση μίας μετρικής ομοιότητας ακολουθιών για να εκτιμήσει την μερική ομοιότητα μεταξύ των προτύπων και των ακολουθιών. Τέλος αναπτύχθηκε ο ακολουθιακός ταξινομητής Particle Swarm Optimization-AdaBoost (PSO-AB) για να βελτιώσει την ακρίβεια της ταξινόμησης των ακο-

λουθιών. Στον PSO-AB ταξινομητή, χρησιμοποιείται ο αλγόριθμος PSO για να αλλάξει προσαρμοστικά την κατανομή των προτύπων που είναι δύσκολο να ταξινομηθούν.

3.1.5 A New Classification Approach using Gapped Subsequences

Σε αυτήν την έρευνα παρουσιάζεται το πρόβλημα της ανάλυσης ακολουθιών με υπο-ακολουθίες αυτών, οι οποίες έχουν κενά [25]. Σκοπός αυτής της μελέτης είναι να βρει αποδοτικά πρότυπα και να παρέχει μία ταξινόμηση για αυτά. Η ακολουθία αντιμετωπίζεται ως μία διατεταγμένη λίστα από στοιχεία και τα ιδανικά πρότυπα ονομάζονται ως επαναλαμβανόμενες υπο-ακολουθίες με κενά.

Η τεχνική έυρεσης των προτύπων χρησιμοποιεί την έννοια της συχνής υποστήριξης για να υπολογίστει πόσο συχνά επαναλαμβάνεται ένα πρότυπο σε μία ακολουθία. Επίσης δεν γίνεται μόνο η εκτίμηση της επανάληψης των προτύπων σε άλλες ακολουθίες του συνόλου αλλά μετράται και η επανάληψη του προτύπου μέσα στην ίδια την ακολουθία.

Κεφάλαιο 4

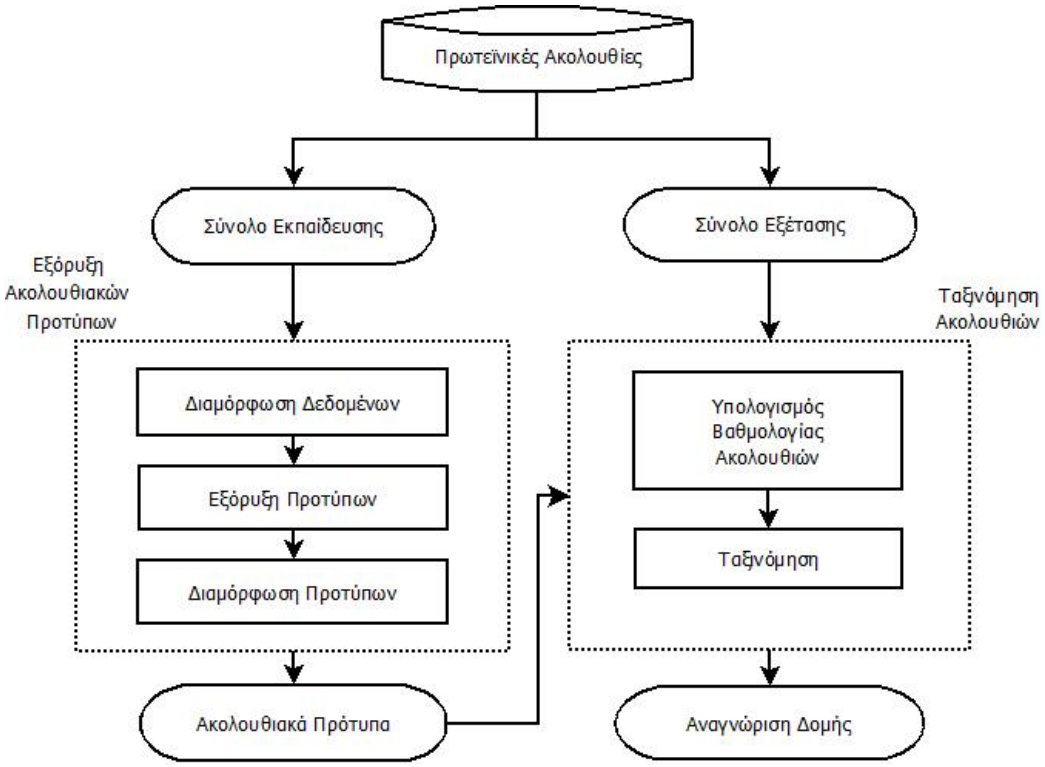
Προτεινόμενη μεθοδολογία

4.1 Περιγραφή μεθοδολογίας

Η προτεινόμενη μεθοδολογία αποτελείται από δύο στάδια όπως φαίνεται και στην εικόνα 4.1. Κατά την διάρκεια του πρώτου σταδίου γίνεται η εξόρυξη ακολουθιακών πρωτύπων και στο δεύτερο εκτελείται ο αλγόριθμος ταξινόμησης. Ειδικότερα στο πρώτο στάδιο απαιτείται η διαμόρφωση του συνόλου δεδομένων, έτσι ώστε να είναι συμβατό με την είσοδο που εξυπηρετεί τον αλγόριθμο τύπου SPM (Sequential Pattern Mining). Εφόσον τα κατάλληλα διαμορφωμένα δεδομένα εισαχθούν στον αλγόριθμο εξόρυξης προτύπων, πρότυπα διαφορετικού μήκους θα παραχθούν από τη τεχνική αυτή. Έπειτα γίνεται μία επιπλέον διαμόρφωση, αυτή την φορά των προτύπων, έτσι ώστε να διαγραφούν διπλοεγγραφές και να πάρουν την κατάλληλη μορφή για να ενσωματωθούν στο δεύτερο στάδιο.

Στο στάδιο ταξινόμησης ακολουθιών εκτελείται μία τεχνική βαθμονόμησης των ακολουθιών, οι οποία βασίζεται στον εντοπισμό των ακολουθιακών προτύπων, που έχουν εξαχθεί από το πρώτο στάδιο, μέσα στο σύνολο ακολουθιακών δεδομένων. Έπειτα εφαρμόζεται ο αλγόριθμος ταξινόμησης για την αναγνώριση της δομής των πρωτεϊνών και υπολογίζονται πλήθος μέτρων αξιολόγησης την μεθοδολογίας. Ωστόσο, επειδή για την κατηγοριοποίηση των ακολουθιών έχει επιλεγεί ένας αλγόριθμος επιβλεπόμενης μάθησης, είναι απαραίτητο να χωρι-

στούν τα δεδομένα σε δύο σύνολα, εκπαίδευσης και εξέτασης. Γι΄ αυτό το λόγο πριν την εκκίνηση του πρώτου σταδίου, επιλέγονται ποια δεδομένα από την βάση δεδομένων πρωτεϊνικών ακολουθιακών θα προχωρήσουν στην εκπαίδευση (1 στάδιο) και ποια στη εξέταση (2 στάδιο).



Σχήμα 4.1: Γενικό διάγραμμα της προτεινόμενης μεθοδολογίας

Στις επόμενες ενότητες θα γίνει μια εκτενέστερη περιγραφή του συνόλου δεδομένων, της μεθοδολογίας καθώς και των μέτρων αξιολόγησής της.

4.2 Σύνολο δεδομένων

Η προτεινόμενη μεθοδολογία αξιολογήθηκε χρησιμοποιώντας ένα σύνολο δεδομένων 200 πρωτεϊνικών ακολουθιών από την βάση δεδομένων SCOPe [26] και πιο συγκεκριμένα την Astral SCOPe 2.06. Ο βασικότερος στόχος της βάσης SCOPe είναι η ανάλυση των δομικών και εξελικτικών σχέσεων που παρατηρούνται μεταξύ όλων των πρωτεϊνών γνωστής δομής καταχωρημένων στην PDB [4]. Η ταξινόμηση των πρωτεϊνών πραγματοποιείται με βάση αυτών των εξελικτικών και δομικών σχέσεων. Τα βασικά επίπεδα ταξινόμησης είναι τέσσερα: η οικογένεια (Family), η υπερ-οικογένεια (Superfamily), το δίπλωμα (Fold) και η τάξη (Class). Στην παρούσα εργασία το ενδιαφέρον επικεντρώθηκε στην τάξη, όπου κάθε τάξη ορίζεται ως μία κλάση της μεθοδολογίας.

Στην τάξη, η ταξινόμηση γίνεται βάσει του διπλώματος των στοιχείων δευτεροταγούς δομής των πρωτεϊνών σε τέσσερις κύριες δομικές κατηγορίες:

- all-α, όπου η δομή σχηματίζεται από α-έλικες
- all-β, όπου η δομή αποτελείται από β-πτυχωτές επιφάνειες
- α/β, όπου στην δομή της πρωτεΐνης εναλλάσσονται α-έλικες και β-πτυχωτές επιφάνειες
- α+β, όπου σε διακριτές περιοχές της δομής βρίσκονται α-έλικες και β-πτυχωτές επιφάνειες

Στην προτεινόμενη μεθοδολογία έγινε πρόβλεψη δύο κλάσεων, γι΄ αύτο το λόγο χρησιμοποιήθηκαν πρωτεϊνικές ακολουθίες που ανήκουν στις δυο πρώτες τάξεις, την all-α και την all-β. Επίσης οι πρωτεΐνες έχουν μικρότερη από 40% ομοιότητα μεταξύ τους γεγονός που ελαχιστοποιεί την εξόρυξη ίδιων προτύπων και την τυχαία αναγνώριση τους στις πρωτεϊνικές ακολουθίες. Στην παρακάτω εικόνα 4.3 δίνεται η μορφή των ακολουθιών των πρωτεϊνών όπως παρέχεται από την βάση δεδομένων Astral SCOPe.

```
>d1dlwa_ a.1.1.1 (A:) Protozoan/bacterial hemoglobin {Ciliate (Paramecium caudatum) [TaxId: 5885]}
slfeqlggqaavqavtaqfyaniqadatvatffngidmpnqtnktaaflcaalggpnawt
grnlkevhanmgvsnaqfttvighlrsaltgagvaaalveqtvavaetvrgdvvtv
>d2gkma_ a.1.1.1 (A:) Protozoan/bacterial hemoglobin {Mycobacterium tuberculosis, HbN [TaxId: 1773]}
gllsrlrkrepisiydkiggheaievvvedffvrvladdqlsaffsgtnmsrlkgkqvef
faaalggpepytgapmkqvhqgrgitmhhfslvaghladaltaagvpsetiteilgviap
lavdvts
>d2qrwa_ a.1.1.1 (A:) Protozoan/bacterial hemoglobin {Mycobacterium tuberculosis, HbO [TaxId: 1773]}
ksfydavggaktfdaivsrfyaqvaedevlrrvypeddlagaeerlrmfleqywggprty
seqrghprlrmrhapfrislierdaflrcmhtavasidsetlddehrrelldylemaahs
lvnspf
>d2bkma_ a.1.1.1 (A:) automated matches {Geobacillus stearothermophilus [TaxId: 1422]}
eqwqtlyeaiggeetvaklveafyrrvaahpdlrpifpddltetahkqkqfltqylggpp
lytaehghpmlrarhlrfeitpkraeawlacmraamdeiglsgpareqfyhrlvltahhm
vntpdhld
>d4i0va_ a.1.1.1 (A:) automated matches {Synechococcus sp. [TaxId: 32049]}
aslyeklggaaavdlavekfygkvladervnrffvntdmakqkqhqkdfmtyafggtdrf
pgrsmraahqdlvenagltdvhfdaiaenlvltlqelnvsqdlidevvtivgsvqhrndv
lnr
```

Σχήμα 4.2: Η μορφή των πρωτεϊνικών ακολουθιών της βάσης δεδομένων Astral SCOPe

Αξιοσημείωτο είναι να αναφερθεί ότι η ταξινόμηση βάσει των σχέσεων μεταξύ των πρωτεϊνών πραγματοποιείται αποκλειστικά από ειδικούς επιστήμονες μετά από λεπτομερή μελέτη και σύγκριση των πρωτεϊνικών δομών. Αυτοματοποιημένες μέθοδοι χρησιμοποιούνται μόνο για την ομοιογένεια των δεδομένων που περιέχονται στη βάση.

4.3 Στάδιο 1: Εξόρυξη ακολουθιακών προτύπων

Η κύρια διαδικασία του πρώτου σταδίου είναι ο αλγόριθμος εξόρυξης προτύπων και θα μπορούσαν να έχουν χρησιμοποιηθεί αρκέτοι αλγόριθμοι που έχουν αναφερθεί στο Κεφάλαιο 2. Ωστόσο, θα ήταν πιο αποτελεσματικό εάν υπήρχε η δυνατότητα δήλωσης κάποιων περιορισμών έτσι ώστε να γίνονται αλλαγές στα κένα που επιτρέπονται ή στην υποστήριξη που θα έχουν τα εξαγόμενα πρότυπα. Ένας αποδοτικός αλγόριθμος που χρησιμοποιεί περιορισμούς στην εξόρυξη ακολουθιακών προτύπων είναι ο αλγόριθμος cSPADE. Ο cSPADE βασίζεται στον αλγόριθμο SPADE, ο οποίος χρησιμοποιεί αποτελεσματικές τεχνικές αναζήτησης πλέγματος και απλές λειτουργίες ένωσης σε λίστες με ID. Όλα τα ακολουθιακά πρότυπα εντοπίζονται μόνο με τρεις σαρώσεις του συνόλου δεδομένων, μειώνοντας τον χρόνο εκτέλεσης του αλγορίθμου. Η πρώτη εξυπη-

ρετεί στην εξόρυξη συχνών προτύπων μήκους-1, η δεύτερη συχνών προτύπων μήκους-2 και η τελευταία παράγει όλες τις συχνές ακολουθίες μήκους ν.

Όπως αναφέρθηκε και στην ενότητα 2.2.2 ο αλγόριθμος SPADE όπως και ο cSPADE είναι μία μέθοδος εξόρυξης ακολουθιακών προτύπων που απαιτεί την κάθετη μορφή των αλληλουχιών του συνόλου δεδομένων τύπου <SID (Sequence ID), EID (Event ID)>. Στον πίνακα 4.1 δίνεται ένα παράδειγμα μορφής μίας βάσης δεδομένων με ακολουθίες πρωτεϊνών και στον πίνακα 4.2 η αντίστοιχη μορφή των ίδιων δεδομένων διαμορφωμένα κατάλληλα για τον αλγόριθμο cSPADE.

Sequence ID	Sequence
1	YEFPW
2	IVG
3	KSSD

Πίνακας 4.1: Η μορφή βάσης δεδομένων με ακολουθίες πρωτεϊνών

Sequence ID	Event ID	Item
1	1	Y
1	2	E
1	3	F
1	4	P
1	5	W
2	1	I
2	2	V
2	3	G
3	1	K
3	2	S
3	3	S
3	4	D

Πίνακας 4.2: Η κάθετη μορφή των πρωτεϊνικών ακολουθίων του πίνακα 4.1

Εφόσον η διαμόρφωση των δεδομένων έχει ολοκληρωθεί οι ακολουθίες και των δύο κλάσεων (all-α και all-β) με την κάθετη πλέον μορφή τους εισάγονται στον

αλγόριθμο εξόρυξης προτύπων cSPADE. Επίσης έχουν επιλεγεί οι περιορισμοί που θα βελτιστοποιήσουν την ποιότητα των προτύπων, και ειδικότερα αυτοί είναι η υποστήριξη (support) και ο μέγιστος αριθμός κενών (maxgap). Η υποστήριξη ένος προτύπου X εκφράζει το ποσοστό των ακολουθιών στην εξεταζόμενη βάση δεδομένων στις οποίες περιέχεται το πρότυπο X και μπορεί να υπολογιστεί σύμφωνα με την ακόλουθη σχέση:

$$supp(X) = \frac{N_x}{N}$$

όπου:
N_x, είναι το πλήθος των ακολουθιών που περιέχεται το σύνολο X
N, είναι ο συνολικός αριθμός των ακολουθιών του συνόλου δεδομένων

Έπειτα με τον όρο μέγιστο αριθμό κενών ορίζεται ο μεγαλύτερος αριθμός επιτρεπόμενων κενών σε διαδοχικά ακολουθιακά στοιχεία τα οποία στην προκειμένη περίπτωση είναι τα αμινοξέα. Στον παρακάτω πίνακα 4.3 μπορούμε να δούμε τις δυνατές ακολουθίες στις οποίες ένα πρότυπο μπορεί να περιέχεται.

Ακολουθία	Κενά
LTE	LTE
LATE	L_TE
LTYE	LT_E
LATYE	L_T_E
LATYFE	L_T__E
LAKTYE	L__T_E
LAKTE	L__TE
LTYFE	LT__E
LAKTYFE	L__T__E

Πίνακας 4.3: Εντοπισμοί του προτύπου LTE σε ακολουθίες με περιορισμό μέγιστου αριθμού κενών ίσο με 2

Εφόσον οι περιορισμοί έχουν τεθεί, ο αλγόριθμος cSPADE εκτελείται και παράγει ακολουθιακά πρότυπα για κάθε κλάση. Όπως αναφέρθηκε και στην ενότητα 2.2.2 στον cSPADE το πρότυπα μεγαλώνουν σε μήκος με κάθε σάρωση, και σε αυτήν την μεθοδολογία πραγματοποιήθηκαν τρεις σαρώσεις, επομένως και τα μέγιστα προτύπα έχουν τρία αμινοξέα. Το επόμενο βήμα περιλαμβάνει

την διαμόρφωση των προτύπων, δηλαδή την διαγραφή προτύπων που έχουν επαναληφθεί αλλά και την διαγραφή χαρακτήρων που δεν είναι απαραίτητοι και εξάγονται από τον αλγόριθμο. Στους πίνακες 4.4 και 4.5 φαίνονται τα αρχικά πρότυπα που παράγονται από τον cSPADE και η τελική μορφή τους.

Αριθμός	Ακολουθία	Υποστήριξη
1	$'<\{h\}>'$	0,9
2	$'<\{i\}>'$	1
3	$'<\{g\},\{v\}>'$	0,9
4	$'<\{a\},\{q\}>'$	0,8
5	$'<\{l\},\{s\},\{e\}>'$	0,8
6	$'<\{e\},\{l\},\{r\}>'$	0,9

Πίνακας 4.4: Ακολουθιακά πρότυπα που εξορύχθηκαν από τον αλγόριθμο cSPADE

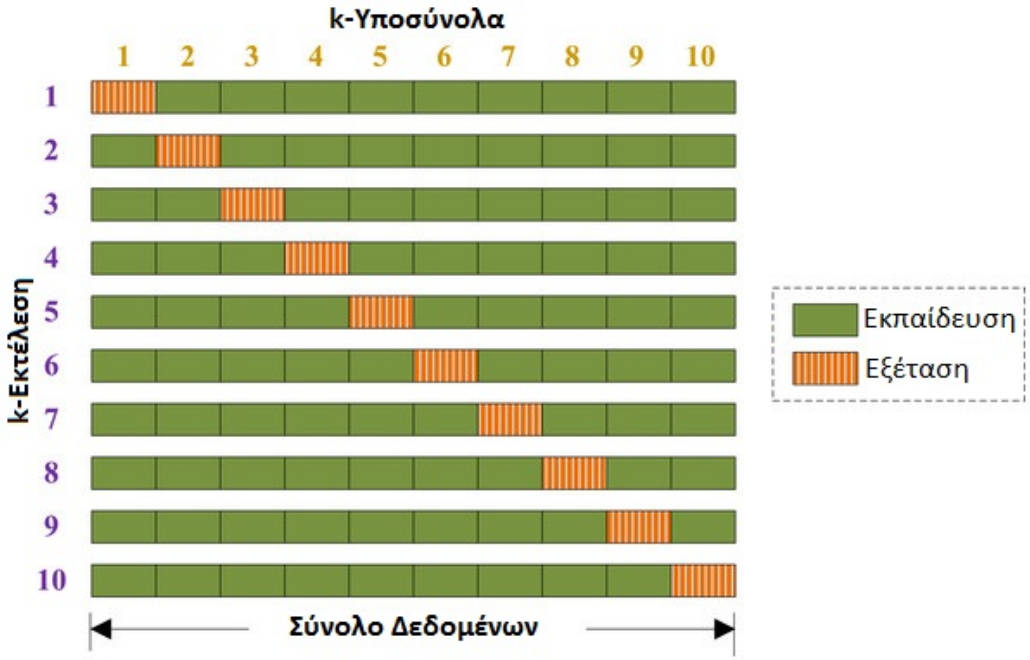
Ακολουθία	Υποστήριξη
h	0,9
i	1
gv	0,9
aq	0,8
lse	0,8
elr	0,9

Πίνακας 4.5: Τελική διαμόρφωση των προτύπων του πίνακα 4.4

Έτσι το πρώτο στάδιο της προτεινόμενης μεθοδολογίας αλοκληρώνεται και στο πέρας αυτού έχουν παραχθεί τα ακολουθιακά πρότυπα από τις δύο κλάσεις του συνόλου εκπαίδευσης των πρωτεϊνικών ακολουθιών. Τα ίδια αυτά πρότυπα θα χρησιμοποιηθούν ως είσοδος μαζί με το σύνολο εξέτασης, στο δεύτερο και τελευταίο στάδιο, όπου θα πραγματοποιηθεί η ταξινόμηση ή αλλιώς κατηγοριοποίηση των πρωτεϊνικών δεδομένων.

4.4 Στάδιο 2: Ταξινόμηση ακολουθιών

Στο δεύτερο στάδιο γίνεται η ταξινόμηση των πρωτεϊνικών ακολουθιών και η βασική διαδικασία είναι ο αλγόριθμος κατηγοριοποίησης που χρησιμοποιείται. Θα μπορούσαν να έχουν επιλεγεί αρκετοί αλγόριθμοι αλλά αυτός που έχει προτιμηθεί είναι η διασταυρωτική αξιολόγηση (Cross Validation) [27] και ειδικότερα η 10-πλη διασταυρωτική αξιολόγηση (10-fold cross validation). Στην περίπτωση του αλγορίθμου αυτού, το σύνολο δεδομένων διαχωρίζονται με τυχαίο τρόπο σε 10 υποσύνολα $D_1, D_2, ..., D_10$ καθένα από αυτά περιέχει ισάριθμο πλήθος ακολουθιών. Στην συνέχεια η διαδικασία της εκπαίδευσης του μοντέλου κατηγοριοποίησης και της εξέτασης λαμβάνει χώρα 10 φορές, όπου σε κάθε επανάληψη i, το υποσύνολο D_i χρησιμοποιείται για αξιολόγηση ενώ τα υπόλοιπα 9 υποσύνολα αποτελούν τα δεδομένα εκπαίδευσης, όπως φαίνεται και στο σχήμα 4.3 [28].



Σχήμα 4.3: Μέθοδος 10-πλης διασταυρωτικής αξιολόγησης (10-fold cross validation)

Η συνολική αξιολόγηση του ταξινομητή προκύπτει αθροίζοντας τις επιμέρους αποδόσεις του μοντέλου για κάθε μία από τις 10 επαναλήψεις. Σε αντίθεση με άλλες τεχνικές ταξινόμησης, με την χρήση της 10-πλής διασταυρωτικής αξιολόγησης εξασφαλίζεται ότι κάθε στιγμιότυπο θα χρησιμοποιηθεί 9 φορές για την εκπαίδευση του μοντέλου και μία φορά για την εξέταση της απόδοσης του [28]. Επίσης χρησιμοποιήθηκαν 10 επαναλήψεις διότι για επαναλήψεις Κ λιγότερες των 10, ένα μεγάλο ποσοστό του συνόλου δεδομένων δεν θα χρησιμοποιούνταν για εκπαίδευση σε καθεμία από τις Κ επαναλήψεις, ενώ για Κ μεγαλύτερο του 10 υπέρχει το ενδεχόμενο οι ακολουθίες λιγότερο εκπροσωπούμενων κλάσεων να συμπεριληφθούν σε ορισμένες από τις Κ επαναλήψεις μόνο για εκπαίδευση του μοντέλου ή μόνο για τον έλεγχο της απόδοσης του.

Για την εξέταση της απόδοσης του αλγορίθμου χρησιμοποιείται ένας πίνακας ο οποίος συμπεριλαμβάνει τα αποτελέσματα του αλγορίθμου ταξινόμησης. Προφανώς για καθεμία από τις ακολουθίες του συνόλου εξέτασης, ο αλγόριθμος ταξινόμησης θα τις έχει κατηγοριοποιήσει είτε σωστά, δηλαδή θα το έχει αναθέσει στην προβλεπόμενη κλάση, είτε λάθος. Έτσι για δύο κλάσεις υπάρχουν οι εξής περιπτώσεις κατηγοριοποίησης των ακολουθιών:

- Αληθώς θετικά True Positives - TP: Αριθμός των ακολουθιών της κλάσης + που κατηγοριοποιήθηκαν σωστά ως + απο τον ταξινομητή
- Αληθώς αρνητικά True Negatives - TN: Αριθμός των ακολουθιών που ανήκουν στην κλάση - και ο ταξινομητής κατηγοριοποίησε ως -
- Ψευδώς θετικά False Positives - FP: Αριθμός των ακολουθιών της κλάσης - που εσφαλμένα κατηγοριοποιήθηκαν ως +
- Ψευδώς αρνητικά False Negatives - FN: Αριθμός των ακολουθιών της κλάσης + που κατηγοριοποιήθηκαν λάθος ως - απο τον ταξινομητή

Έτσι με βάση τις παραπάνω περιπτώσεις ταξινόμησης των ακολουθιών δημιουργείται ο πίνακας (μήτρα) σύγχυχης (Confusion Matrix), ο οποίος φαίνεται και στην εικόνα 4.4.

Σχήμα 4.4: Πίνακας σύγχυχης (Confusion Matrix)

Επομένως εάν ως δεδομένα χρησιμοποιηθούν 100 πρωτεϊνικές ακολουθίες και έτσι σύμφωνα με τον αλγόριθμο 10-πλης διασταυρωτικής αξιολόγησης τα συνολικά δεδομένα χωρίζονται σε 10 ισομερή, υποσύνολα των 10 ακολουθιών. Τα εννέα υποσύνολα συνθέτουν το σύνολο εκπαίδευσης, το οποίο στην προκειμένη περίπτωση περιέχει 90 ακολουθίες πρωτεϊνών, και το ένα υποσύνολο των 10 ακολουθιών λειτουργεί ως το σύνολο εξέτασης της απόδοσης του αλγορίθμου ταξινόμησης. Αυτή η διαδικασία εκπαίδευσης και εξέτασης των δεδομένων επαναλαμβάνεται 10 φόρες με αποτέλεσμα να δημιουργούνται 10 πίνακες σύγχυσης, ένας για κάθε σύνολο εξέτασης. Στην ολοκλήρωση του αλγόριθμου ταξινόμησης όλοι οι πίνακες σύγχυσης συγκεντρώνονται αθροιστικά σε έναν για την εξέταση τών συνολικών αποτελεσμάτων της μεθοδολογίας.

Παρόλα αυτά ο αλγόριθμος πρέπει να εφαρμόζει ένα μοντέλο έτσι ώστε να κατηγοριοποιήσει την κάθε ακολουθία εξέτασης στην προβλεπόμενη κλάση της. Γι΄ αυτόν τον λόγο εφαρμόστηκαν τρία μοντέλα βαθμολόγησης των ακολουθιών έτσι ώστε να επιτευχθεί η μέγιστη ακρίβεια και η καλυτερη απόδοση του αλγορίθμου ταξινόμησης. Το πρώτο βασίζεται στον αριθμό εμφανίσεων των προτύπων στις ακολουθίες, το δεύτερο βασίζεται στην εμφάνιση ή μη εμφάνιση των προτύπων στις ακολουθίες και τέλος το τρίτο αντιστοιχίζει βελτιστοποιημένο βάρος στα προτύπα.

Και τα τρία μοντέλα εφαρμόζουν τεχνικές για την βαθμολογία της κάθε ακολουθίας του συνόλου εξέτασης και εξάγουν τα αποτελέσματα σε πίνακες που περιέχουν τις ακολουθίες, τις προβλεπόμενες κλάσεις και άλλα στοιχεία όπως φαίνεται σε ένα παράδειγμα εκτέλεσης στον πίνακα 4.6. Στον πίνακα 4.7 αναφέρονται τα περιεχόμενα της κάθε στήλης του πίνακα 4.6.

Στήλη 1	Στήλη 2	Στήλη 3	Στήλη 4	Στήλη 5	Στήλη 6	Στήλη 7
mktiviedkqries...	B	338	2,2092	351	2,25	B
deaaelmqqvnv...	A	151	0,9869	145	0,9295	A
mlilstekepnfey...	A	335	2,1895	315	2,0192	A
vvmdniidvsipv...	A	212	1,3856	229	1,4679	B

Πίνακας 4.6: Πίνακας αποτελεσμάτων του συνόλου εξέτασης με πλήθος προτύπων A=153 και πλήθος προτύπων B=156

Στήλη	Επεξήγηση
1	Ακολουθία
2	Πραγματική κλάση
3	Μετρική μοντέλου της κλάσης A
4	Κανονικοποίηση στήλης 3
5	Μετρική μοντέλου της κλάσης B
6	Κανονικοποίηση στήλης 5
7	Προβλεπόμενη κλάση

Πίνακας 4.7: Περιεχόμενα των στηλών του πίνακα 4.6

4.4.1 Μοντέλο 1: Πλήθος εμφανίσεων των προτύπων στις ακολουθίες

Με το μοντέλο αυτό γίνεται χρήση των προτύπων που έχουν εξορυχθεί στο πρώτο στάδιο της μεθοδολογίας. Τα πρότυπα της κλάσης A (all-α) εντοπίζονται στις ακολουθίες του συνόλου εξέτασης και αντίστοιχα αυτό γίνεται και για τα πρότυπα της κλάσης B (all-β). Έπειτα δημιουργείται ο πίνακας 4.6 όπου στην στήλη 3 και 5 αναφέρει τις συνολικές εμφανίσεις των προτύπων των κλάσεων

Α και Β αντίστοιχα καθώς και τις κανονικοποιήσεις αυτών των αθροισμάτων στις στήλες 4 και 6. Για την εύρεση της κλάσης που προβλέπεται από τον αλγόριθμο συγκρίνονται οι στήλες 4 και 6, και η μεγαλύτερη ορίζει και την κλάση της ακολουθίας από το σύνολο εξέτασης. Η πρόβλεψη του αλγόριθμου αποθηκεύεται στην 7 στήλη του πίνακα 4.6.

4.4.2 Μοντέλο 2: Εμφανίση ή μη, των προτύπων στις ακολουθίες

Αυτό το μοντέλο θυμίζει το μοντέλο 1 μόνο που υπάρχει μια βασική διαφορά στην μετρική των προτύπων. Εδώ για την βαθμονόμηση των ακολουθιών του συνόλου εξέτασης δεν γίνεται η καταμέτρηση των εμφανίσεων των προτύπων στις ακολουθίες του συνόλου εξέτασης αλλά το μοντέλο λαμβάνει υπ΄ όψιν το γέγονος εάν βρέθηκε το πρότυπο στην ακολουθία ή όχι. Με αυτόν τον τρόπο οι στήλες 3 και 5 περιέχουν τον αριθμό των προτύπων των κλάσεων Α και Β αντίστοιχα που βρέθηκαν σε κάθε ακολουθία του συνόλου εξέτασης. Έπειτα οι στήλες 4 και 6 περιέχουν τις κανονικοποιήσεις των στηλών 3 και 5, και η στήλη 7 την προβλεπόμενη κλάση της ακολουθίας.

4.4.3 Μοντέλο 3: Βελτιστοποίηση με βαθμονόμηση προτύπων

Για την μέθοδο αυτήν εφαρμόστηκε αρχικά ένα τυχαίο βάρος στα εξαγόμενα πρότυπα που προκύπτουν από το δεύτερο στάδιο και έπειτα χρησιμοποιήθηκε η μέθοδος βελτιστοποίησης του προγράμματος matlab, fminsearch, η οποία παράγει τα καλύτερα βάρη για τα πρότυπα έτσι ώστε να ειτευχθεί η μεγαλυτερη απόδοση. Η μέθοδος αυτή βρίσκει το ελάχιστο μίας συνάρτησης με πολλές παραμέτρους και χωρίς περιορισμούς χρησιμοποιώντας μία τεχνική χωρίς παράγωγους, δηλαδή ψάχνει το ελάχιστο μίας συνάρτησης το οποίο καθορίζεται ως εξής:

$$min_x f(x)$$

όπου:
$f(x)$, είναι η συνάρτηση για βελτιστοποίηση

Η συνάρτηση σε αυτήν την μέθοδο ορίζεται ως τον ίδιο αλγόριθμο ταξινόμησης του δεύτερου σταδίου μόνο που ως σύνολο εξέτασης τώρα χρησιμοποιείται ένα υποσύνολο του συνόλου εκπαίδευσης. Στην ουσία η συνάρτηση εκπαιδεύει τα πρότυπά της βαθμονομώντας τα και χρησιμοποιώντας το σύνολο εκπαίδευσης. Ως έξοδος της συνάρτησης ορίζεται το σφάλμα (1-απόδοση) , το οποίο πρέπει να ελαχιστοποιηθεί.

Έπειτα το βάρος αυτό συμμετέχει στην βαθμολογία του συνόλου εξέτασης. Συγκεκριμένα έαν ένα πρότυπο M έχει μεγαλύτερο βάρος απο ένα άλλο N, τότε οι ακολουθίες του συνόλου εξέτασης που περιέχουν το πρότυπο M θα έχουν μεγαλύτερη βαθμολογία από τις ακολουθίες που περιέχουν το πρότυπο N.

Κεφάλαιο 5

Μετρικές απόδοσης και αποτελέσματα

5.1 Μετρικές απόδοσης

Γι΄ αυτήν την μεθοδολογία χρησιμοποιήθηκε ένα υποσύνολο 200 ακολουθιών της βάσης δεδομένων Astral SCOPe 2.06,η οποία περιγράφτηκε στην ενότητα 4.2. Έτσι το σύνολο εκπαίδευσης αποτελείται από 180 ακολουθίες και το σύνολο εξέτασης από 20 ακολουθίες σε κάθεμία από τις 10 επαναληψεις του αλγορίθμου 10-πλης διασταυρωτικής αξιολόγησης. Για την εξαγωγή των προτύπων ο περιορισμος που τέθηκε είναι 80% υποστήριξη. Στο τρίτο μοντέλο για την βελτιστοποίηση των βαρών των προτύπων χρησιμοποιούνται οι 40 απο τις 180 ακολουθιες του συνόλου εκπαίδευσης ως σύνολο εξέτασης της συνάρτησης βελτιστοποίησης. Και για τις τρεις τεχνικές έγινε μια παραμετροποίηση για το μεγιστο αριθμό κενών ετσι ώστε να εξεταστεί ο βέλτιστος αριθμός κενών.

Όσον αφορά για την μέτρηση της απόδοσης της προτεινόμενης μεθοδολογίας χρησιμοποιήθηκαν κάποιες μετρικές απόδοσης οι οποίες βασίζονται στον πίνακα σύγχυσης της ενότητας 4.4. Αυτός ο πίνακας λοιπόν περιέχει τα αληθώς θετικά (TP), αληθώς αρνητικά (TN), ψευδώς θετικά (FP), ψευδώς αρνητικά (FN) αποτελέσματα. Με βάση αυτους τους αριθμούς μπορουν να υπολογιστούν οι

μετρικες της ακρίβειας, της ευαισθησίας και της εξειδικευτικότητας.

Η ακρίβεια είναι το συνηθέστερο και πλέον χρησιμοποιούμενο μέτρο αξιολόγησης, εκφράζει τον βαθμό των σωστών κατηγοριοποιήσεων του ταξινομητή και μπορεί να υπολογιστεί από την ακόλουθη σχέση:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Η ευαισθησία αξιολογεί την ικανότητα του ταξινομητή να κατηγοριοποιήσει σωστά όλα τα θετικά στοιχεία του συνόλου εξέτασης και ορίζεται απο την παρακάτω σχέση:

$$sensitivity = \frac{TP}{TP + FP}$$

Η εξειδικευτικότητα εκτιμά την απόδοση του ταξινομητή για την κατηγοριοποίηση των αρνητικών παραδειγμάτων και υπολογίζεται απο την ακόλουθη σχέση:

$$specificity = \frac{TN}{TN + FN}$$

Οι παρακάτω πίνακες παρουσιάζουν τα αποτελέσματα, χρησιμοποιώντας τα παραπάνω μέτρα αξιολόγησης, του καθε μοντέλου της μεθοδολογίας παραμετροποιημένο με βάση τον μέγιστο αριθμό κενών.

5.2 Πίνακες αποτελεσμάτων

Μοντέλο 1, Μέγιστος Αριθμός Κενών 0							
Αριθμός		Εκπαίδευση			Εξέταση		
Υποσύνολο	Πρότυπα	Ευαισθ.	Εξειδικ.	Ακρίβεια	Ευαισθ.	Εξειδικ.	Ακρίβεια
1	39	1.1	100.0	50.1	0.0	100.0	50.0
2	39	1.1	100.0	50.1	0.0	100.0	50.0
3	39	1.1	100.0	50.1	0.0	100.0	50.0
4	39	1.1	100.0	50.1	0.0	100.0	50.0
5	39	1.1	100.0	50.1	0.0	100.0	50.0
6	39	0.0	100.0	50.0	10.0	100.0	55.0
7	39	1.1	100.0	50.1	0.0	100.0	50.0
8	39	1.1	100.0	50.1	0.0	100.0	50.0
9	39	1.1	100.0	50.1	0.0	100.0	50.0
10	39	1.1	100.0	50.1	0.0	100.0	50.0
Σύνολο	390	1.0	100.0	50.5	1.0	100.0	50.5

Πίνακας 5.1: Αριθμός των προτύπων που έχουν εξαχθεί και μέτρα αξιολόγησης για το σύνολο εκπάιδευσης και εξέτασης, με βάση το μοντέλο 1 και μέγιστο αριθμό κενων 0

Μοντέλο 1, Μέγιστος Αριθμός Κενών 1							
Αριθμός		Εκπαίδευση			Εξέταση		
Υποσύνολο	Πρότυπα	Ευαισθ.	Εξειδικ.	Ακρίβεια	Ευαισθ.	Εξειδικ.	Ακρίβεια
1	120	93.3	68.2	80.9	100.0	70.0	85.0
2	122	70.0	94.4	82.2	80.0	90.0	85.0
3	123	98.9	68.9	83.9	90.0	40.0	65.0
4	122	63.3	96.7	80.0	70.0	90.0	80.0
5	121	75.6	85.6	80.6	90.0	70.0	80.0
6	122	92.0	71.3	81.6	88.9	70.0	79.0
7	123	90.0	76.7	83.3	80.0	50.0	65.0
8	119	87.8	74.4	81.1	70.0	90.0	80.0
9	116	90.0	73.3	81.7	80.0	60.0	70.0
10	123	98.9	55.6	77.2	80.0	70.0	75.0
Σύνολο	1211	86.0	76.6	81.3	82.9	70.0	76.4

Πίνακας 5.2: Αριθμός των προτύπων που έχουν εξαχθεί και μέτρα αξιολόγησης για το σύνολο εκπάιδευσης και εξέτασης, με βάση το μοντέλο 1 και μέγιστο αριθμό κενων 1

Μοντέλο 1, Μέγιστος Αριθμός Κενών 2							
Αριθμός		Εκπαίδευση			Εξέταση		
Υποσύνολο	Πρότυπα	Ευαισθ.	Εξειδικ.	Ακρίβεια	Ευαισθ.	Εξειδικ.	Ακρίβεια
1	294	94.4	80.0	87.2	60.0	70.0	65.0
2	303	90.0	86.7	88.3	90.0	70.0	80.0
3	303	88.9	85.6	87.2	80.0	70.0	75.0
4	307	91.1	83.3	87.2	100.0	90.0	95.0
5	305	92.2	70.0	81.1	100.0	80.0	90.0
6	302	91.1	80.0	85.6	90.0	90.0	90.0
7	301	86.7	87.8	87.2	90.0	90.0	90.0
8	289	78.9	85.6	82.2	70.0	90.0	80.0
9	303	78.9	88.9	83.9	80.0	100.0	90.0
10	309	77.8	91.1	84.4	90.0	100.0	95.0
Σύνολο	3016	87.0	83.9	85.4	85.0	85.0	85.0

Πίνακας 5.3: Αριθμός των προτύπων που έχουν εξαχθεί και μέτρα αξιολόγησης για το σύνολο εκπάιδευσης και εξέτασης, με βάση το μοντέλο 1 και μέγιστο αριθμό κενων 2

Μοντέλο 1, Μέγιστος Αριθμός Κενών 3							
Αριθμός		Εκπαίδευση			Εξέταση		
Υποσύνολο	Πρότυπα	Ευαισθ.	Εξειδικ.	Ακρίβεια	Ευαισθ.	Εξειδικ.	Ακρίβεια
1	537	71.1	97.8	84.4	90.0	90.0	90.0
2	526	68.9	97.8	83.3	50.0	100.0	75.0
3	544	57.8	97.8	77.8	50.0	100.0	75.0
4	528	67.8	97.8	82.8	60.0	100.0	80.0
5	533	86.7	90.0	88.1	70.0	90.0	80.0
6	539	82.2	95.6	88.9	80.0	100.0	90.0
7	549	60.0	97.8	78.9	70.0	100.0	85.0
8	526	75.6	95.6	85.6	70.0	100.0	85.0
9	598	73.3	97.8	85.6	60.0	90.0	75.0
10	591	83.3	95.6	89.4	60.0	90.0	75.0
Σύνολο	5471	72.7	96.3	84.5	66.0	96.0	81.0

Πίνακας 5.4: Αριθμός των προτύπων που έχουν εξαχθεί και μέτρα αξιολόγησης για το σύνολο εκπάιδευσης και εξέτασης, με βάση το μοντέλο 1 και μέγιστο αριθμό κενων 3

Μοντέλο 2, Μέγιστος Αριθμός Κενών 0							
Αριθμός		Εκπαίδευση			Εξέταση		
Υποσύνολο	Πρότυπα	Ευαισθ.	Εξειδικ.	Ακρίβεια	Ευαισθ.	Εξειδικ.	Ακρίβεια
1	39	46.2	60.3	53.9	50.0	100.0	70.0
2	39	49.1	63.9	57.0	20.0	50.0	36.4
3	39	48.0	64.4	56.9	37.5	50.0	43.8
4	39	48.1	64.4	56.8	33.3	50.0	42.9
5	39	48.1	61.0	55.0	33.3	75.0	57.1
6	39	45.5	63.3	54.8	66.7	57.1	60.0
7	39	44.4	66.1	55.8	75.0	37.5	50.0
8	39	42.3	60.7	52.2	83.3	83.3	83.3
9	39	47.1	62.3	55.4	42.9	66.7	53.9
10	39	47.6	60.7	54.5	42.9	83.3	61.6
Σύνολο	390	46.6	62.7	55.2	46.6	62.7	55.2

Πίνακας 5.5: Αριθμός των προτύπων που έχουν εξαχθεί και μέτρα αξιολόγησης για το σύνολο εκπάιδευσης και εξέτασης, με βάση το μοντέλο 2 και μέγιστο αριθμό κενων 0

Μοντέλο 2, Μέγιστος Αριθμός Κενών 1							
Αριθμός		Εκπαίδευση			Εξέταση		
Υποσύνολο	Πρότυπα	Ευαισθ.	Εξειδικ.	Ακρίβεια	Ευαισθ.	Εξειδικ.	Ακρίβεια
1	125	76.7	76.4	76.6	90.0	80.0	85.0
2	133	81.1	75.6	78.3	70.0	60.0	65.0
3	127	73.3	77.8	75.6	70.0	70.0	70.0
4	126	66.7	88.9	77.8	50.0	80.0	65.0
5	126	82.2	77.8	80.0	80.0	60.0	70.0
6	121	81.1	78.9	80.0	70.0	80.0	75.0
7	131	81.1	83.3	82.2	30.0	60.0	45.0
8	133	76.7	80.0	78.3	80.0	90.0	85.0
9	126	82.0	74.4	78.2	70.0	70.0	70.0
10	129	81.1	73.3	77.2	90.0	80.0	85.0
Σύνολο	390	78.2	78.7	78.5	70.0	73.0	71.5

Πίνακας 5.6: Αριθμός των προτύπων που έχουν εξαχθεί και μέτρα αξιολόγησης για το σύνολο εκπάιδευσης και εξέτασης, με βάση το μοντέλο 2 και μέγιστο αριθμό κενων 1

Μοντέλο 2, Μέγιστος Αριθμός Κενών 2							
Αριθμός		Εκπαίδευση			Εξέταση		
Υποσύνολο	Πρότυπα	Ευαισθ.	Εξειδικ.	Ακρίβεια	Ευαισθ.	Εξειδικ.	Ακρίβεια
1	294	81.1	88.9	85.0	60.0	90.0	75.0
2	303	74.4	93.3	83.9	80.0	70.0	75.0
3	303	78.7	95.6	87.2	70.0	90.0	80.0
4	307	74.4	88.9	81.7	100.0	100.0	100.0
5	305	84.4	90.0	87.2	40.0	100.0	70.0
6	302	77.8	92.2	85.0	80.0	80.0	80.0
7	301	75.6	90.0	82.8	90.0	80.0	85.0
8	299	75.6	93.3	84.4	40.0	90.0	65.0
9	303	76.4	94.4	85.4	60.0	100.0	80.0
10	309	75.6	93.3	84.4	80.0	100.0	90.0
Σύνολο	3026	77.4	92.0	84.7	70.0	90.0	80.0

Πίνακας 5.7: Αριθμός των προτύπων που έχουν εξαχθεί και μέτρα αξιολόγησης για το σύνολο εκπάιδευσης και εξέτασης, με βάση το μοντέλο 2 και μέγιστο αριθμό κενων 2

Μοντέλο 2, Μέγιστος Αριθμός Κενών 3							
Αριθμός		Εκπαίδευση			Εξέταση		
Υποσύνολο	Πρότυπα	Ευαισθ.	Εξειδικ.	Ακρίβεια	Ευαισθ.	Εξειδικ.	Ακρίβεια
1	514	90.0	87.8	88.9	100.0	80.0	90.0
2	530	83.3	94.4	88.9	60.0	90.0	75.0
3	536	83.3	95.6	89.4	50.0	90.0	70.0
4	533	80.0	94.4	87.2	90.0	80.0	85.0
5	530	76.7	96.7	86.7	70.0	100.0	85.0
6	537	77.8	92.2	85.0	80.0	100.0	90.0
7	537	80.0	94.4	87.2	60.0	80.0	70.0
8	528	77.8	94.4	86.1	90.0	100.0	95.0
9	520	72.2	95.6	83.9	100.0	90.0	95.0
10	533	80.0	94.4	87.2	60.0	100.0	80.0
Σύνολο	5298	80.1	94.0	87.1	76.0	91.0	83.5

Πίνακας 5.8: Αριθμός των προτύπων που έχουν εξαχθεί και μέτρα αξιολόγησης για το σύνολο εκπάιδευσης και εξέτασης, με βάση το μοντέλο 2 και μέγιστο αριθμό κενων 3

Μοντέλο 3, Μέγιστος Αριθμός Κενών 0							
Αριθμός		Εκπαίδευση			Εξέταση		
Υποσύνολο	Πρότυπα	Ευαισθ.	Εξειδικ.	Ακρίβεια	Ευαισθ.	Εξειδικ.	Ακρίβεια
1	39	3.3	100.0	51.7	0.0	100.0	50.0
2	39	0.0	100.0	50.0	0.0	100.0	50.0
3	39	0.0	100.0	50.0	0.0	100.0	50.0
4	39	0.0	100.0	50.0	0.0	100.0	50.0
5	39	0.0	100.0	50.0	0.0	100.0	50.0
6	39	0.0	100.0	50.0	0.0	100.0	50.0
7	39	0.0	100.0	50.0	0.0	100.0	50.0
8	39	0.0	100.0	50.0	0.0	100.0	50.0
9	39	75.6	27.8	51.7	80.0	30.0	55.0
10	39	0.0	100.0	50.0	0.0	100.0	50.0
Σύνολο	390	17.9	82.8	50.3	18.0	83.0	50.5

Πίνακας 5.9: Αριθμός των προτύπων που έχουν εξαχθεί και μέτρα αξιολόγησης για το σύνολο εκπάιδευσης και εξέτασης, με βάση το μοντέλο 3 και μέγιστο αριθμό κενων 0

Μοντέλο 3, Μέγιστος Αριθμός Κενών 1							
Αριθμός		Εκπαίδευση			Εξέταση		
Υποσύνολο	Πρότυπα	Ευαισθ.	Εξειδικ.	Ακρίβεια	Ευαισθ.	Εξειδικ.	Ακρίβεια
1	120	90.0	55.5	72.8	90.0	80.0	85.0
2	122	60.0	81.1	70.6	50.0	50.0	50.0
3	126	72.2	73.3	72.8	70.0	60.0	65.0
4	124	94.4	26.7	60.6	80.0	50.0	65.0
5	121	73.3	58.9	66.1	80.0	20.0	50.0
6	123	100.0	6.7	53.3	100.0	0.0	50.0
7	122	97.8	8.9	53.3	100.0	0.0	50.0
8	118	7.8	98.9	53.3	0.0	100.0	50.0
9	121	60.0	74.4	67.2	40.0	80.0	60.0
10	117	98.9	22.2	60.6	100.0	40.0	70.0
Σύνολο	1214	75.4	50.7	63.1	71.0	48.0	59.5

Πίνακας 5.10: Αριθμός των προτύπων που έχουν εξαχθεί και μέτρα αξιολόγησης για το σύνολο εκπάιδευσης και εξέτασης, με βάση το μοντέλο 3 και μέγιστο αριθμό κενων 1

Μοντέλο 3, Μέγιστος Αριθμός Κενών 2							
Αριθμός		Εκπαίδευση			Εξέταση		
Υποσύνολο	Πρότυπα	Ευαισθ.	Εξειδικ.	Ακρίβεια	Ευαισθ.	Εξειδικ.	Ακρίβεια
1	294	66.7	88.9	77.8	40.0	100.0	70.0
2	303	47.8	94.4	71.1	40.0	80.0	60.0
3	303	47.8	98.9	73.3	30.0	100.0	65.0
4	307	54.4	90.0	72.2	70.0	80.0	75.0
5	305	55.6	93.3	74.4	10.0	100.0	55.0
6	302	82.2	78.9	80.6	80.0	60.0	70.0
7	301	43.3	97.8	70.6	70.0	80.0	75.0
8	299	92.2	58.9	75.6	90.0	60.0	75.0
9	303	62.2	95.6	78.9	70.0	100.0	85.0
10	309	8.9	100.0	54.4	10.0	100.0	55.0
Σύνολο	3026	56.1	89.7	72.9	51.0	86.0	68.5

Πίνακας 5.11: Αριθμός των προτύπων που έχουν εξαχθεί και μέτρα αξιολόγησης για το σύνολο εκπάιδευσης και εξέτασης, με βάση το μοντέλο 3 και μέγιστο αριθμό κενων 2

Μοντέλο 3, Μέγιστος Αριθμός Κενών 3							
Αριθμός		Εκπαίδευση			Εξέταση		
Υποσύνολο	Πρότυπα	Ευαισθ.	Εξειδικ.	Ακρίβεια	Ευαισθ.	Εξειδικ.	Ακρίβεια
1	574	21.1	100.0	60.6	10.0	90.0	50.0
2	552	12.2	100.0	56.1	0.0	100.0	50.0
3	565	55.6	95.6	75.6	60.0	100.0	80.0
4	554	65.6	92.2	78.9	60.0	70.0	65.0
5	572	86.7	83.3	85.0	80.0	60.0	70.0
6	557	94.4	65.6	80.0	100.0	90.0	95.0
7	563	30.0	100.0	65.0	0.0	100.0	50.0
8	568	80.0	83.3	81.7	80.0	80.0	80.0
9	560	68.9	92.2	80.6	60.0	90.0	75.0
10	540	77.8	84.4	81.1	70.0	80.0	75.0
Σύνολο	5605	59.2	89.7	74.4	52.0	86.0	69.0

Πίνακας 5.12: Αριθμός των προτύπων που έχουν εξαχθεί και μέτρα αξιολόγησης για το σύνολο εκπάιδευσης και εξέτασης, με βάση το μοντέλο 3 και μέγιστο αριθμό κενων 3

Κεφάλαιο 6

Συγκρίσεις μεταξύ μοντέλων και παλαιότερων ερευνών

6.1 Συγκρίσεις μεταξύ των προτεινόμενων μοντέλων

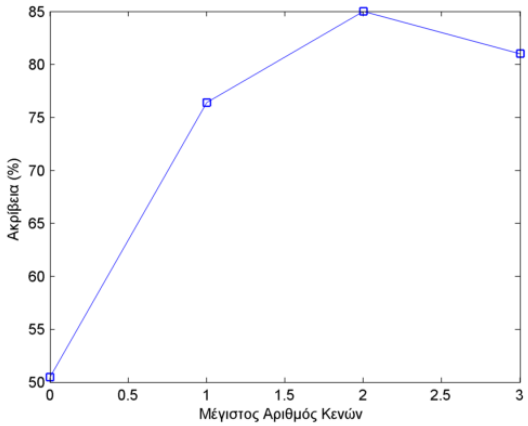
Στην παρούσα εργασία αναπτύχθηκε μία μεθοδολογία πρόβλεψης της δευτεροταγούς δομής πρωτεϊνών χρησιμοποιώντας αλγοριθμικές τεχνικές βασιζόμενες στην εξόρυξη προτύπων. Ο αλγόριθμος χωρίζεται σε δύο στάδια, αρχικά γίνεται η εξόρυξη των ακολουθιακών προτύπων μέσα απο προεπεξεργασμένες πρωτεϊνικές ακολουθίες και στο δεύτερο στάδιο εκτελείται ο αλγόριθμος ταξινόμησης και εξάγονται τα αποτελέσματα και οι μετρικές απόδοσης της πρόβλεψης της δομής των πρωτεϊνων.

Όπως παρουσιάστηκε και στο προηγούμενο κεφάλαιο κάθε μοντέλο της προτεινόμενης μεθοδολογίας έχει εξαιρετικά υψηλή απόδοση ακρίβειας. Στον παρακάτω πίνακα παρουσιάζονται αναλυτικά τα ποσοστά ακρίβειας για το συνολο εκπαίδευσης και εξέτασης της προτεινόμενης μεθοδολογίας όλων των μοντέλων με τα τέσσερα πειράματα για κάθε μέγιστο αριθμό κενών.

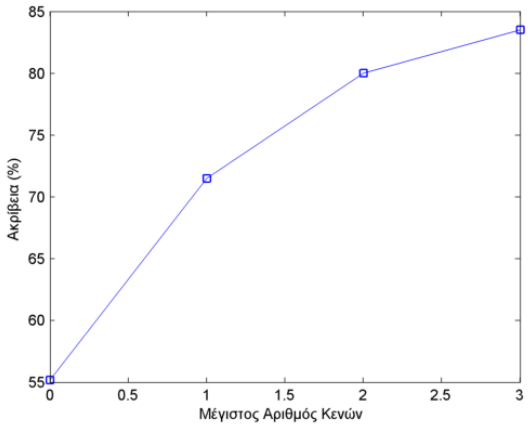
M.A.K.	Σύνολο Εκπαίδευσης			Σύνολο Εξέτασης		
-	Μοντέλο 1	Μοντέλο 2	Μοντέλο 3	Μοντέλο 1	Μοντέλο 2	Μοντέλο 3
0	50.5	55.2	50.3	50.5	55.2	50.5
1	81.3	78.5	63.1	76.4	71.5	59.5
2	**85.4**	84.7	72.9	**85.0**	80.0	68.5
3	84.5	**87.1**	74.4	81.0	**83.5**	69.0

Πίνακας 6.1: Ποσοστά ακρίβειας για το συνολο εκπαίδευσης και εξέτασης της προτεινόμενης μεθοδολογίας όλων των μοντέλων με τα τέσσερα πειράματα για κάθε μέγιστο αριθμό κενών (M.A.K.)

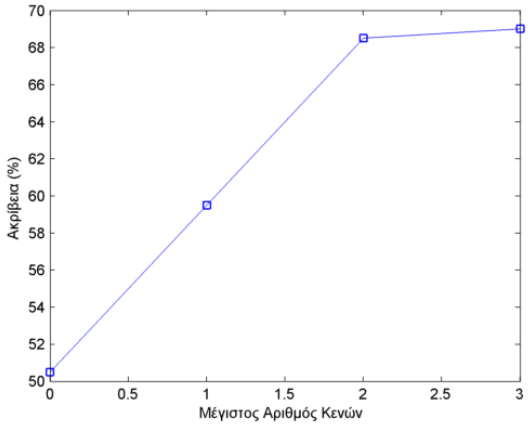
Οι παρακάτω εικόνες παρουσιάζουν γραφικά τα ποσοστά ακρίβειας του κάθε μοντέλου συγκριτικά με τον μέγιστο αριθμό κενών καθώς υπάρχει και ένα συγκεντρωτικό γράφημά που συγκρίνει όλα τα αποτελέσματα ακρίβειας όλων των μοντέλων και για κάθε μέγιστο αριθμό κενών. Όπως διακρίνεται το ποσοστό ακρίβειας είναι αρκετά υψηλό και βασίζεται στον αλγόριθμο 10-πλής διασταυρωτικής αξιολόγησης ο οποίος εξασφαλίζει ότι τα δεδομένα θα χρησιμοποιηθούν 10 φορές για την εξαγωγή των αποτελεσμάτων. Έτσι εξαλείφονται οι πιθανότητες των τυχαίων υψηλών ποσοστών ακρίβειας.



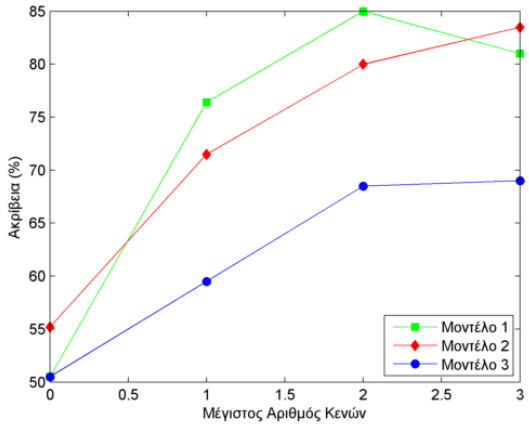
Σχήμα 6.1: Γραφική αναπαράσταση της ακρίβειας της προτεινόμενης μεθοδολογίας του μοντέλου 1 με τα τέσσερα πειράματα για κάθε μέγιστο αριθμό κενών



Σχήμα 6.2: Γραφική αναπαράσταση της ακρίβειας της προτεινόμενης μεθοδολογίας του μοντέλου 2 με τα τέσσερα πειράματα για κάθε μέγιστο αριθμό κενών



Σχήμα 6.3: Γραφική αναπαράσταση της ακρίβειας της προτεινόμενης μεθοδολογίας του μοντέλου 3 με τα τέσσερα πειράματα για κάθε μέγιστο αριθμό κενών



Σχήμα 6.4: Γραφική αναπαράσταση της ακρίβειας της προτεινόμενης μεθοδολογίας όλων των μοντέλων με τα τέσσερα πειράματα για κάθε μέγιστο αριθμό κενών

6.2 Συγκριτική μελέτη

Ενα από τα βασικότερα συμπεράσματα της προτεινόμενης μεθοδολογίας είναι ότι τα αποτελέσματα της ακρίβειας της πρόβλεψης της δευτεροταγούς δομής της πρωτεΐνης είναι εξαιρετικά υψηλά φτάνοντας το ποσοστό του 85.0% στο μοντέλο 1 και 83.5% στο μοντέλο 2. Τέτοιας κλίμακας ποσοστά δεν έχουν εξαχθεί με καμία απο τις προηγούμενες αλγοριθμικές τεχνικές που εξετάστηκαν στο κεφάλαιο 3. Στον παρακάτω πίνακα 6.2 παρουσιάζεται μια ποιοτική σύγκριση όλων των μεθοδολογιών πρόβλεψης της δομής της πρωτεΐνης.

Στην προτεινόμενη μεθοδολογία χρησιμοποιήθηκαν δύο κλάσεις, η πρώτη ήταν οι πρωτεΐνες με δομή α-έλικας και η δεύτερη οι πρωτεΐνες με δομή β-φύλλο όπως το ίδιο γίνεται και στις μεθοδολογίες 1, 2, 3 και 4. Τα δεδομένα που χρησιμοποιήθηκαν είναι 200 πρωτεΐνες ενώ στις παραπάνω μεθοδολογίες χρησιμοποιουνται 1000 πρωτεΐνες. Έπειτα οι τεχνικές εξόρυξης προτύπων διαφέρουν εκτός από την μεθοδολογία 1 και 2 όπου χρησιμοποιείται ο cSpade αλλά και πάλι ο περιορισμός για την υποστήριξη στην προτεινόμενη μεθοδολογία είναι 80% σε αντίθεση με τις προαναφερθείσες που είναι 50%.

Συγγραφέας	Πρόβλημα	Δεδομένα	Μέθοδος	Αποτέλεσμα
Anna N. Ntagiou, Markos G. Tsipouras, Pantelis Angelidis	2 categories (class A and class B folds)	200 proteins, ASTRAL SCOP (version 2.06)	cSpade, 10-fold cross validation, min 80% support, Optimization weights for the paterns	85.0
Themis P. Exarchos, Markos G. Tsipouras, Costas Papaloukas, Dimitrios I. Fotiadis	Experiment 4 : 2 categories (class A and class B folds)	1000 proteins, SCOP	cSpade, min 50% support, Optimization weights for the patterns and for the classes	77.8
Themis P. Exarchos, Markos G. Tsipouras, Costas Papaloukas, Dimitrios I. Fotiadis	Experiment 4 : 2 categories (class A and class B folds)	1000 proteins, SCOP	cSpade, min 50% support, Optimization weights for the classes	79.0
Maral Azizi, Mohammad Saniee Abade	Experiment 4 : 2 categories (class A and class B folds)	1000 proteins, ASTRAL SCOP (version 2.05)	PrefixSpan, min 50% support, Genetic algorithm	70.7
Tsai Chieh-Yuan, Chen Chih-Jung	3-4 classes	7 synthetic datasets (200 - 12.000 proteins) and 1 dataset consists of 17 categories (class A and class B folds, 1000 proteins, SCOP)	ClosedMining, PSO-AB	78.44
Kusum Sharma, Asha Ambhaikar	Experiment 4 : 2 categories (class A and class B folds)	1002 proteins, SCOP	MinSupComp, PATgrow, GSgrow, ClosePATgrow, min support is defined by user, Optimization weights for the classes	68.1

Πίνακας 6.2: Σύγκριση μεθοδολογιών για την πρόβλεψη της δευτεροταγούς δομής της πρωτεΐνης

Το πιο αξιοσημείωτο είναι η τεχνική ταξινόμησης που εφαρμόζεται. Σε κάθε μεθοδολογία τα δεδομένα χωρίζονται ως εξής: 66% των πρωτεϊνών για εκπαίδευση και 33% για εξέταση. Έτσι κάποιες ακολουθίες δεν χρησιμοποιούνται καθόλου για την εκπαίδευση με αποτέλεσμα το ποσοστό ακρίβειας να μην είναι αξιόπιστο. Στην προτεινόμενη μεθοδολογία με την χρήση της 10-πλής διασταυρωτικής αξιολόγησης εξασφαλίζεται ότι κάθε στιγμιότυπο θα χρησιμοποιηθεί 9 φορές για την εκπαίδευση του μοντέλου και μία φορά για την εξέταση της απόδοσης του. Έτσι τα αποτελέσματα είναι προιόντα ενός μέσου όρου και έτσι εκμηδενίζονται οι περιπτώσεις της τυχαίας κατηγοριοποίησης. Αυτό είναι ένα από τα μεγαλύτερα πλεονεκτήματα αυτής της μεθοδολογίας, έτσι τα υψηλά ποσοστά ακρίβειας που αγγίζουν το ποσοστό του 85% είναι έγκυρα.

6.3 Μελλοντικές επεκτάσεις

Σαν μελλοντική επέκταση αυτής της διπλωματικής εργασίας είναι η βελτίωση της μεθοδολογίας ώστε να υπάρχουν όσο το δυνατό λιγότερα λανθασμένα αποτελέσματα και μεγαλύτερα ποσοστά ακρίβειας. Επίσης μια επέκταση που θα μπορούσε να υλοποιηθεί είναι η χρήση δεδομένων πολλαπλών μεταβλητών και δεδομένων χρονοσειρών αλλά και η χρήση μεγαλύτερης βάσης δεδομένων (άνω των 1000 πρωτεϊνών). Έπειτα θα μπορούσε να δωθεί προσοχή στον αλγόριθμο εξόρυξης προτύπων και να χρησιμοποιηθούν αλγόριθμοι που εξάγουν μεγαλύτερου μήκους πρότυπα σε μικρότερο χρόνο.

Βιβλιογραφία

[1] David L. Nelson and Michael M. Cox. *Principles of Biochemistry*. English. W.H. Freeman and Company, 2005. ISBN: 0-7167-4339-6.

[2] Ι. Γ. Γεωργάτσου. *Εισαγωγή στην Βιοχημεία*. Εκδόσεις Γιαχούδη, 2013. ΙΣΒΝ: 960-7425-02-2.

[3] Philip E. Bourne and Helge Weissig. *Structural bioinformatics*. English. Wiley- Liss, 2003. ISBN: 0-471-20200-2.

[4] Helen M. Berman et al. "The Protein Data Bank". English. In: *Nucleic Acids Research* 28.1 (2000), p. 235. DOI: 10.1093/nar/28.1.235. eprint: /oup/backfile/Content_public/Journal/nar/28/1/10.1093_nar_28.1.235/1/280235.pdf. URL: http://dx.doi.org/10.1093/nar/28.1.235.

[5] David M. Webster. *Protein Structure Prediction: Methods and Protocols*. English. Humana Press, 2000. ISBN: 0-89603-637-5.

[6] Yachdav G. et al. "PredictProtein—an open resource for online prediction of protein structural and functional features". English. In: *Nucleic acids research* (2014), gku366.

[7] J. A. Cuff et al. "JPred: a consensus secondary structure prediction server". English. In: *Bioinformatics* 14.10 (1998), pp. 892–893.

[8] D. T. Jones. "Protein secondary structure prediction based on position-specific scoring matrices". English. In: *J. Mol. Biol.* 292.2 (1999), pp. 195–202.

[9] S. F. Altschul et al. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs". English. In: *Nucleic Acids Res.* 25.17 (1997), pp. 3389–3402.

[10] "PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth". English. In: *Proceedings of the 17th International Conference on Data Engineering*. ICDE '01. Washington, DC, USA: IEEE Computer Society, 2001, pp. 215–. URL: http://dl.acm.org/citation.cfm?id=876881.879716.

[11] Ramakrishnan Srikant and Rakesh Agrawal. "Mining sequential patterns: Generalizations and performance improvements". English. In: *Advances in Database Technology — EDBT '96: 5th International Conference on Extending Database Technology Avignon, France, March 25–29, 1996 Proceedings*. Ed. by Peter Apers, Mokrane Bouzeghoub, and Georges Gardarin. Berlin, Heidelberg: Springer Berlin Heidelberg, 1996, pp. 1–17. ISBN: 978-3-540-49943-5. DOI: 10.1007/BFb0014140. URL: http://dx.doi.org/10.1007/BFb0014140.

[12] Mohammed J. Zaki. "SPADE: An Efficient Algorithm for Mining Frequent Sequences". English. In: *Machine Learning* 42.1 (2001), pp. 31–60. ISSN: 1573-0565. DOI: 10.1023/A:1007652502315. URL: http://dx.doi.org/10.1023/A:1007652502315.

[13] J. Ayres et al. "Sequential Pattern Mining Using Bitmaps". English. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2002.

[14] Z. Yang & Y. Wang & M. Kitsuregawa. *LAPIN: Effective Sequential Pattern Mining Algorithms*. English. Tech. rep. 2005.

[15] Fournier-Viger P. & Gomariz A. & Campos M. & Thomas R. "Fast Vertical Mining of Sequential Patterns Using Co-occurrence Information". English. In: Proc. 18th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2014). 2014.

[16] J. Wang & J. Han. "BIDE: Efficient Mining of Frequent Closed Sequences". English. In: ICDE 2004. 2004, pp. 79–90.

[17] Fournier-Viger P. & Wu C.-W. & Tseng V.-S. "Mining Maximal Sequential Patterns without Candidate Maintenance". English. In: Proc. 9th International Conference on Advanced Data Mining and Applications (ADMA 2013). Springer LNAI 8346, 2013, pp. 169–180.

[18] Jiawei Han. *GSP: Apriori-Based Sequential Pattern Mining*. English. University of Illinois at Urbana-Champaign. URL: http://www.coursera.org/learn/data-patterns/lecture/qtiPk/5-2-gsp-apriori-based-sequential-pattern-mining.

[19] Jiawei Han. *SPADE—Sequential Pattern Mining in Vertical Data Format*. English. University of Illinois at Urbana-Champaign. URL: http://www.coursera.org/learn/data-patterns/lecture/sOm9A/5-3-spade-sequential-pattern-mining-in-vertical-data-format.

[20] Jiawei Han. *PrefixSpan—Sequential Pattern Mining by Pattern-Growth*. English. University of Illinois at Urbana-Champaign. URL: http://www.coursera.org/learn/data-patterns/lecture/ZJHTA/5-4-prefixspan-sequential-pattern-mining-by-pattern-growth.

[21] Themis P. Exarchos et al. "A two-stage methodology for sequence classification based on sequential pattern mining and optimization". English. In: *Data & Knowledge Engineering* 66.3 (2008), pp. 467 –487. ISSN: 0169-023X. DOI: http://dx.doi.org/10.1016/j.datak.2008.05.007. URL: http://www.sciencedirect.com/science/article/pii/S0169023X08000748.

[22] Themis P. Exarchos et al. "An optimized sequential pattern matching methodology for sequence classification". English. In: *Knowledge and Information Systems* 19.2 (2009), pp. 249–264. ISSN: 0219-3116. DOI: 10.1007/s10115-008-0146-2. URL: http://dx.doi.org/10.1007/s10115-008-0146-2.

[23] Maral Azizi and Mohammad Saniee Abade. "Protein structure prediction by means of sequential pattern mining". English. In: *International Journal of Artificial Intelligence & Applications (IJAIA)* 6.4 (2015). DOI: 10.5121/ijaia.2015.6403. URL: http://dx.doi.org/10.5121/ijaia.2015.6403.

[24] Chieh-Yuan Tsai and Chih-Jung Chen. "A PSO-AB Classifier for Solving Sequence Classification Problems". English. In: *Applied Soft Computing* 27.C (2015), pp. 11–27. ISSN: 1568-4946. DOI: 10.1016/j.asoc.2014.10.029. URL: http://dx.doi.org/10.1016/j.asoc.2014.10.029.

[25] Kusum Sharma and Asha Ambhaikar. "A New Classification Approach using Gapped Subsequences". English. In: *International Journal of Science and Research (IJSR)* 2.5 (2013), pp. 368–373. ISSN: 2319-7064.

[26] Naomi K. Fox, Steven E. Brenner, and John-Marc Chandonia. "SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures". English. In: *Nucleic Acids Research* 42.D1 (2013), p. D304. DOI: 10.1093/nar/gkt1240. URL: +http://dx.doi.org/10.1093/nar/gkt1240.

[27] R.R. Picard and R.D. Cook. "Cross-validation of regression models". English. In: *Journal of the American Statistical Association* 79.387 (1984), pp. 575–583. DOI: 10.1080/01621459.1984.10478083. URL: https://www.scopus.com/inward/record.uri?eid=2s2.0-84950445313&doi=10.1080%2f01621459.1984.10478083&partnerID=40&md5=2012efbb19b16b256463851d88008ec9.

[28] Λύρας Π. Δημήτριος. 'Παραμετροποίηση στοχαστικών μεθόδων εξόρυξης γνώσης από δεδομένα, μετασχηματισμού συμβολοσειρών και τεχνικών συμπερασματικού λογικού προγραμματισμού'. Διδακτορική διατρ. Πανεπιστήμιο Πατρών, Τμήμα Ηλεκτρολόγων Μηχανικών και Τεχνολογίας Υπολογιστών, 2010.