

Μεταπτυχιακός φοιτητής: **Mr. Sharma Anuj**

Τίτλος διπλωματικής εργασίας:

## **Βελτίωση της αναζήτησης ομοιότητας αλληλουχιών για την ανίχνευση ομολόγων πρωτεϊνών με την χρήση δομικών δεδομένων και μεθόδων μηχανικής μάθησης**

Περίληψη:

Στην πτυχιακή αυτή, παρουσιάζουμε μια μέθοδο βελτίωσης της ανίχνευσης ομολογίας των πρωτεϊνών. Η προτεινόμενη μέθοδος συνδυάζει δεδομένα σύγκρισης αλληλουχίας και δομών πρωτεϊνών για την ανίχνευση ομολογίας. Η σύγκριση ομοιότητας αλληλουχιών είναι το πιο συχνά χρησιμοποιούμενο μέτρο για την ανίχνευση ομολογίας. Ωστόσο, είναι γνωστό ότι, σε κάποιες περιπτώσεις, ομόλογες πρωτεΐνες μπορεί να παρουσιάζουν ελάχιστη (< 35% ) ομοιότητα αλληλουχιών. Στην μέση του φάσματος των αποτελεσμάτων, που επιστρέφονται από την σύγκριση ομοιότητας αλληλουχιών, η απόκλιση αυτή οδηγεί σε λάθος στην ταξινόμηση ομολογίας. Αυτή η ζώνη των πρωτεϊνών, στην οποία παρουσιάζεται το μέγιστο σφάλμα, ονομάζεται διφορούμενη ζώνη (twilight zone) των πρωτεϊνών. Η μέθοδος που παρουσιάζουμε περιλαμβάνει την αναταξινόμηση των αποτελεσμάτων αυτών, από το PSI-BLAST, σε «αληθώς θετικά» και «αληθώς αρνητικά». Η αναταξινόμηση γίνεται με τη χρήση ενός ταξινομητή που χρησιμοποιεί πληροφορίες δομής των πρωτεϊνών.

Διάφοροι παραμετρικοί και μη παραμετρικοί ταξινομητές, καθώς και συνδυασμοί ταξινομητών, συγκρίθηκαν με τυπικά μέτρα αξιολόγησης. Αναπτύχθηκαν ταξινομητές, με τη χρήση δεδομένων σύγκρισης δομής, και στη συνέχεια χρησιμοποιήθηκαν για την αναταξινόμηση των πρωτεϊνών της διφορούμενης ζώνης. Παρέχουμε στατιστικά στοιχεία που υποστηρίζουν την διαχωριστικότητα των δύο τάξεων (ομόλογων ή μη ομόλογων πρωτεϊνών) και στη συνέχεια, παρουσιάζουμε τα αποτελέσματα των διαφόρων ταξινομητών. Δοκιμάστηκαν ταξινομητές για διαφορετικούς συνδυασμούς δομικών χαρακτηριστικών που ανακτήθηκαν από την σύγκριση των δομών των πρωτεϊνών. Τα αποτελέσματά μας επιβεβαιώνουν ότι η προσέγγιση αυτή μπορεί να χρησιμοποιηθεί επιτυχώς για την σημαντική βελτίωση της ανίχνευσης ομολογίας, μειώνοντας τα σφάλματα που συμβαίνουν στη διφορούμενη ζώνη κατά την σύγκριση πρωτεϊνών βάση της αλληλουχίας τους.

Εξεταστική Επιτροπή

Ηλίας Μανωλάκος, Αν. Καθηγητής, Τμήμα Πληροφορικής και Τηλεπικοινωνιών, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών (επιβλέπων)

Ιωάννης Εμίρης, Καθηγητής, Τμήμα Πληροφορικής και Τηλεπικοινωνιών, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

Δρ. Ευαγγελία Δ. Χρυσίνα Ομάδα Δομικής Βιολογίας & Χημείας Ινστιτούτο Οργανικής & Φαρμακευτικής Χημείας, Εθνικό Ίδρυμα Ερευνών