ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
"ΤΕΧΝΟΛΟΓΙΕΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΣΤΗΝ ΙΑΤΡΙΚΗ ΚΑΙ ΤΗ ΒΙΟΛΟΓΙΑ"

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αξιολόγηση και Σύγκριση Εργαλείων Βιοπληροφορικής που Μελετούν την Δομή των Πρωτεϊνών στον Χώρο

Θεοδώρα Π. Μανουσίδου

Επιβλέποντες: **Ιωάννης Εμίρης,** Καθηγητής Ε.Κ.Π.Α

Ευαγγελία Χρυσίνα, Διδάκτωρ Ε.Ι.Ε.

ΑΘΗΝΑ

ΑΠΡΙΛΙΟΣ 2012

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αξιολόγηση και Σύγκριση Εργαλείων Βιοπληροφορικής που Μελετούν την Δομή των Πρωτεϊνών στον Χώρο

Θεοδώρα Π. Μανουσίδου

Α.Μ.: ΠΙΒ 033

ΕΠΙΒΛΕΠΟΝΤΕΣ: **Ιωάννης Εμίρης,** Καθηγητής Ε.Κ.Π.Α.

Ευαγγελία Χρυσίνα, Διδάκτωρ Ε.Ι.Ε.

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ: **Ιωάννης Εμίρης,** Καθηγητής Ε.Κ.Π.Α.

Ευαγγελία Χρυσίνα, Διδάκτωρ Ε.Ι.Ε.

Ηλίας Μανωλάκος, Καθηγητής Ε.Κ.Π.Α.

Απρίλιος 2012

ΠΕΡΙΛΗΨΗ

Στην παρούσα διπλωματική εργασία διερευνάται η γεωμετρική αναπαράσταση της τρισδιάστατης δομής μακρομορίων όπως αυτή πραγματοποιείται από σύγχρονα υπολογιστικά εργαλεία βιοπληροφορικής. Αντικείμενο της εργασίας αποτελεί η συγκριτική μελέτη των εργαλείων αυτών αναφορικά με την ικανότητα πρόβλεψης και της αποτελεσματικής χαρτογράφησης των επί μέρους κέντρων σύνδεσης των μακρομοριακών στόχων.

Ειδικότερα, το σύνολο των εργαλείων που εξετάζονται έχουν ως στόχο την αναζήτηση και τον προσδιορισμό επιμέρους περιοχών στην πρωτεϊνική δομή, με πιθανό λειτουργικό ρόλο. Ειδικότερα, τα υπολογιστικά εργαλεία που εξετάζονται είναι η εφαρμογή ανοιχτού κώδικα *Fpocket* και ο αλγόριθμος *Cast*, τα οποία αναζητούν περιοχές στο μόριο των εξεταζόμενων πρωτεϊνών ικανών να προσδέσουν άλλα μόρια που χαρακτηρίζονται ως τροποποιητές, αναστολείς ή ενεργοποιητές. Επιπλέον, εξετάζεται και η εφαρμογή *Caver,* με την βοήθεια της οποίας πραγματοποιείται ο καθορισμός μοριακών μονοπατιών με κατεύθυνση από το εσωτερικό περιβάλλον προς το εξωτερικό περίβλημα των πρωτεϊνών.

Η εφαρμογή των επιλεγόμενων αλγορίθμων και μαθηματικών προσεγγίσεων πραγματοποιήθηκε σε ένα πλήρως χαρακτηρισμένο σύνολο δεδομένων. που περιλαμβάνει ένα σύνολο 260 πρωτεϊνικών δομών, η επιλογή του οποίου κατέστησε δυνατή την πραγματοποίηση της μελέτης αυτής.

Επιπλέον, επιλέχθηκε ένα δεύτερο σύνολο δεδομένων το οποίο αποτέλεσε το σύνολο ελέγχου κατά την διαδικασία αξιολόγησης των αποτελεσμάτων καθώς επίσης και της ικανότητας πρόβλεψης περιοχών ικανών να λειτουργήσουν ως ενεργές περιοχές. Με τον τρόπο αυτό προσδιορίζεται το σύνολο των χαρακτηριστικών που είναι ικανά να διακρίνουν τις επιμέρους περιοχές ενδιαφέροντος.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Γεωμετρική αναπαράσταση μακρομορίων

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: φωσφορυλάση του γλυκογόνου, κατευθυνόμενος από την δομή σχεδιασμός φαρμάκων, διάγραμμα voronoi, τριγωνοποίηση Delaunay, άλφα σχήμα

ABSTRACT

The subject of this thesis project is related to the geometrical representation of molecular structures in the three-dimensional space, as it is being carried out by different Bioinformatics tools. The main addressed goal is the comparative study of these tools, regarding the accuracy of their predictions and their ability to map the various binding sites that can be found on the targeted molecules.

The structure of Glycogen Phoshorylase, that has been extensively used as a molecular target in the process of drug design, for the study against the disease of Diabetes 2, was used as a control.

More specifically, the aim of the above mentioned tools is the geometrical description of the regions that have functional role for the proteins that are being studied. The open source application Fpocket and the algorithm Caver, process the input molecular structure, that the user specifies, for potential binding sites. These regions are characterized by their ability to attract other molecules that their function is to work as inhibitors or activators, related to protein functionality. Additionally, the application Caver performs the search in the protein structure for pathways that have their starting point inside the protein molecule and they are directed to the bulk solvent.

Also, in this project the identification of those features that can be used for the prediction of the various regions of interest and have the ability to describe their properties, is being studied. This is performed by the use of machine learning approaches and related feature selection methods.

For the testing of the selected algorithms, a manually curated data set of 260 protein complexes in total was used. A second data set, composed of native protein structures was used in order to test the prediction accuracy of the selected algorithms.

SUBJECT AREA: Geometrical representation of macromolecules

KEYWORDS: glycogen phosphorylase, drug design, voronoi diagrams, Delaunay triangulation, a-shape