NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCE
DEPARTMENT OF INFORMATICS AND TELLECOMUNICATIONS

POSTGRADUATE PROGRAM
"INFORMATION TECHNOLOGIES IN MEDICINE AND BIOLOGY"

MASTER THESIS

Development of data mining tools for identifying structural determinants that dictate protein-ligand interactions

Anaxagoras Apostolos Fotopoulos
Athanasios Vasileios Papathanasiou

SUPERVISORS: **Ioannis Z. Emiris** Professor.
Evangelia D. Chrysina Assoc. Professor.

ATHENS

JUNE 2015

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

Ανάπτυξη εργαλείων εξόρυξης δεδομένων για τον εντοπισμό καθοριστικών δομικών παραγόντων που υποδεικνύουν αλληλεπιδράσεις πρωτεΐνης-συνδέτη.

Αναξαγόρας Αποστόλου Φωτόπουλος,
Αθανάσιος Βασιλείου Παπαθανασίου

ΕΠΙΒΛΕΠΟΝΤΕΣ: **Γιάννης Ζ. Εμίρης** Καθηγητής.
Χρυσίνα Ευαγγελία Assoc. Professor.

ΑΘΗΝΑ

ΙΟΥΝΙΟΣ 2015

MASTER THESIS

Development of data mining tools for identifying structural determinants that dictate protein-ligand interactions.

Anaxagoras A. Fotopoulos
A.M.: PIV0113
Athanasios V. Papathanasiou
A.M.: PIV0125

SUPERVISORS	**Ioannis Z. Emiris**	Professor.
	Evangelia D. Chrysina	Assoc. Professor.

EXAMINATION COMMITTEE :	**Ioannis Z. Emiris**	Professor.
	Evangelia D. Chrysina	Assoc. Professor.
	Frédéric Cazals	Professor.

JUNE 2015

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ανάπτυξη εργαλείων εξόρυξης δεδομένων για τον εντοπισμό καθοριστικών δομικών παραγόντων που υποδεικνύουν αλληλεπιδράσεις πρωτεΐνης-συνδέτη.

Αναξαγόρας Α. Φωτόπουλος
Α.Μ.: ΠΙΒ0113
Αθανάσιος Β. Παπαθανασίου
Α.Μ.: ΠΙΒ0125

ΕΠΙΒΛΕΠΟΝΤΕΣ:	**Γιάννης Ζ. Εμίρης**	Καθηγητής.
	Χρυσίνα Ευαγγελία	Assoc. Professor.

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ:	**Γιάννης Ζ. Εμίρης**	Καθηγητής.
	Χρυσίνα Ευαγγελία	Assoc. Professor.
	Frédéric Cazals	Professor.

ΙΟΥΝΙΟΣ 2015

ABSTRACT

Modelling binding sites of enzymes is a fundamental but rather demanding task, of increased complexity since the residues forming these sites are not rigid. Similarly, binding studies of a ligand at such a site and complex formation raises difficulties mainly because most of the structural determinants that control binding are not known. Using a combination of sampling algorithms and statistical analysis techniques, we shall contribute towards developing much more accurate binding affinity predictions for macro-molecular docking. To this end, our ultimate aim is to study benchmark protein families with known 3D structure with the aim to identify specific geometric parameters for modeling their binding cavities. This is possible by studying the boundaries within which every residue in those cavities can move, in 3D Euclidean or conformational space. Key methods employed include structural alignment of secondary structure elements, RMSD heat-maps, sampling (e.g. in the space of rotamers), standard scoring functions, and (generously) allowed regions as defined in Ramachandran plot. Our algorithmic tools involve powerful methods, such as nearest-neighbor search and clustering, which shall be adapted to the specific context. The developed methods were tested on a subset of protein kinases with known 3D structure, which offer a number of target sites for one or several ligands. Different conformers were first produced based on the simulation of chi angles rotations and then clustered in a 2-level hierarchical process. For each conformer cluster, representative polygonal shapes were produced which can thereafter be exploited in ligand screening approaches. To further aid protein analysis a series of bioinformatics tools were developed and their application is also discussed.

SUBJECT AREA: Bioinformatics

KEYWORDS: Conformers, hierarchical clustering, structural determinants, conformation space, data mining

ΠΕΡΙΛΗΨΗ

Η μοντελοποίηση των περιοχών πρόσδεσης των ενζύμων είναι ένα θεμελιώδες, αλλά απαιτητικό έργο αυξημένης πολυπλοκότητας, δεδομένου ότι τα αμινοξέα που συνιστούν αυτές τις περιοχές δεν είναι άκαμπτα. Παρομοίως, οι μελέτες πρόσδεσης σε μία τέτοια περιοχή και ο σχηματισμός συμπλέγματος παρουσιάζουν δυσκολίες, κυρίως επειδή οι περισσότεροι από τους δομικούς παράγοντες που εμπλέκονται στην διαδικασία πρόσδεσης δεν είναι γνωστοί. Χρησιμοποιώντας έναν συνδυασμό αλγορίθμων δειγματοληψίας και μεθόδων στατιστικής ανάλυσης, θα συμβάλουμε στην ανάπτυξη πιο ακριβών προβλέψεων στην πρόσδεση μακρο-μορίων. Απώτερος στόχος είναι η μελέτη οικογενειών πρωτεϊνών με γνωστή τρισδιάστατη δομή με σκοπό να προσδιοριστούν συγκεκριμένες γεωμετρικές παράμετροι για τη μοντελοποίηση των κοιλοτήτων πρόσδεσής. Αυτό θα καθίσταται δυνατό με τη μελέτη των ορίων, εντός των οποίων κάθε αμινοξύ μπορεί να κινηθεί σε αυτές τις κοιλότητες, στον τρισδιάστατο Ευκλείδειο χώρο ή στο χώρο διαμόρφωσης. Βασικές μέθοδοι που χρησιμοποιούνται είναι η δομική στοίχιση της δευτεροταγούς δομής, τα διαγράμματα RMSD, δειγματοληψία (π.χ. στον χώρο των ροταμερών), συναρτήσεις βαθμολόγησης και επιτρεπόμενες περιοχές όπως ορίζονται από διάγραμμα Ramachandran. Τα αλγοριθμικά εργαλεία μας περιλαμβάνουν ισχυρές μεθόδους, όπως η αναζήτηση πλησιέστερων γειτόνων και η συσταδοποίηση, οι οποίες εντάσσονται στο συγκεκριμένο πλαίσιο. Οι μέθοδοι που αναπτύχθηκαν, δοκιμάστηκαν σε ένα υποσύνολο των πρωτεϊνικών κινασών με γνωστή τρισδιάστατη δομή, που προσφέρουν μια σειρά από κέντρων σύνδεσης για έναν ή περισσότερους προσδέτες. Διαφορετικά διαμορφομερή δημιουργήθηκαν αρχικά με βάση την προσομοίωση των χ γωνιών στρέψης των αμινοξέων και στη συνέχεια έγινε η ομαδοποίησή τους με δι-επίπεδη ιεραρχική συσταδοποίηση. Για κάθε συστάδα διαμορφομερών παρήχθησαν αντιπροσωπευτικά πολυγωνικά σχήματα, τα οποία μπορούν στη συνέχεια να αξιοποιηθούν στην διαδικασία επιλογής προσδέτη. Για την περαιτέρω υποβοήθηση της ανάλυσης των πρωτεϊνών, μια σειρά από εργαλεία Βιοπληροφορικής αναπτύχθηκαν η χρήση των οποίων περιγράφεται.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Βιοπληροφορική

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Διαμορφομερή, Ιεραρχική ομαδοποίηση, Δομικοί Παράγοντες, Χώρος Διαμόρφωσης, Εξόρυξη Δεδομένων