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

Improving sequence similarity search for protein Homology Induction using Structural data and Machine Learning methods

Anuj Sharma

Επιβλέποντες: **Ηλίας Μανωλάκος,** Αναπληρωτής Καθηγητής ΕΚΠΑ
Ιωάννης Εμίρης, Καθηγητής ΕΚΠΑ
Ευαγγελία Χρυσίνα, Εθνικό Ίδρυμα Ερευνών

ΑΘΗΝΑ
ΝΟΕΜΒΡΙΟΣ 2011

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Improving sequence similarity search for protein Homology Induction using Structural data and Machine Learning methods

Anuj Sharma
A.M.: ΠΙΒ07013

ΕΠΙΒΛΕΠΟΝΤΕΣ: **Ηλίας Μανωλάκος,** Αναπληρωτής Καθηγητής ΕΚΠΑ
Ιωάννης Εμίρης, Καθηγητής ΕΚΠΑ
Ευαγγελία Χρυσίνα, Εθνικό Ίδρυμα Ερευνών

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ: **Ηλίας Μανωλάκος,** Αναπληρωτής Καθηγητής ΕΚΠΑ
Ιωάννης Εμίρης, Καθηγητής ΕΚΠΑ
Ευαγγελία Χρυσίνα, Εθνικό Ίδρυμα Ερευνών

Νοέμβριος 2011

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 οδηγήσει 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 ταξινόμητη 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 διαχωρισιμότητα των δύο τάξεων (ομόλογων ή μη ομόλογων πρωτεϊνών) και στη συνεχεία, 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.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Βιοπληροφορική, Σύγκριση δομών πρωτεϊνών
ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Πρωτεΐνη, Ομολογία, Μηχανική μάθηση, Σύγκριση αλληλουχιών, Σύγκριση δομών

Abstract

In this thesis, we present a work flow for improving Homology detection for Proteins. The proposed method combines protein sequence and structure comparison data for detecting Homology. Sequence similarity measures are the most commonly used tool for homology detection. However it is known that evolutionary divergence can lead to homologous proteins having very little sequence similarity. In the middle range of proteins found in the sequence similarity results this divergence leads to error in homology classification. This zone of proteins where maximum error occurs is referred to as the twilight zone of proteins. The work flow presented involves reclassifying twilight zone proteins, in the PSI-BLAST results, into 'true positives' and 'true negatives'. The reclassification is done using a Classifier built from the structural data.

Several parametric, non-parametric and committee classifiers were compared on standard metrics. Classifiers were built using structure comparison data and subsequently used for reclassifying the twilight zone proteins. We provide statistical data supporting the separability of the two classes (homologous and non-homologous proteins) and subsequently provide results of classification using various classifiers. Various combinations of structural features extracted were tried. Our tests show that the approach can be successfully used to improve a homology detection work flow by reducing errors that occur in the 'twilight zone' when plain sequence comparison is used as a metric.

SUBJECT AREA: Bioinformatics, Protein Structure Comparison
KEYWORDS: Protein, Homology, Machine learning, Sequence comparison, Structure comparison