ΠΕΡΙΛΗΨΗ

Τα προς αποθήκευση δεδομένα αυξάνονται ραγδαία τα τελευταία χρόνια, τα οποία μπορεί να είναι είτε δεδομένα του παγκόσμιου ιστού (κείμενα, εικόνες κ.λπ.) είτε βιολογικά δεδομένα. Στην δεύτερη κατηγορία δεδομένων λόγω της ταχύτατης ανάπτυξης της τεχνολογίας απαιτείται η κατάλληλη αποθήκευση, απαιτώντας όσο το δυνατόν λιγότερο χώρο, παρέχοντας ταυτόχρονα την ταχύτατη ανάκτηση αυτής.

Μας ενδιαφέρει ο τρόπος με τον οποίο θα γίνει η αποθήκευση μιας αλληλουχίας DNA με τη χρήση ανεστραμμένων ευρετηρίων και δέντρων επιθεμάτων. Η δομή των δέντρων επιθεμάτων μας δίνει την δυνατότητα να βρούμε αποδοτικά το πλήθος εμφάνισης μίας επιθυμητής υποσυμβολοσειράς σε μία μεγαλύτερη συμβολοσειρά. Αυτό είναι κάτι που απαιτείται για την κατάλληλη αποθήκευση στην δομή των ανεστραμμένων ευρετηρίων είτε αυτές είναι ενός επιπέδου, είτε είναι δύο επιπέδων. Ο βέλτιστος τρόπος για αυτό το διαχωρισμό κατά την αποθήκευση της αλληλουχίας είναι το πρόβλημα το οποίο επιλύουμε.

Στα δεδομένα βιολογικών ακολουθιών πέρα από τις κανονικές συμβολοσειρές πάνω στις οποίες βασίστηκε η προπτυχιακή μου διπλωματική και για τις οποίες θα δώσουμε μία γρήγορη ανάλυση, θα ασχοληθούμε κυρίως με τον τρόπο διαχείρισης και κατάλληλης αποθήκευσης των βεβαρυμμένων αλληλουχιών. Βεβαρυμμένες ονομάζονται οι συμβολοσειρές οι οποίες σε κάποια σημεία αντί να εμφανίζεται ένα συγκεκριμένο γράμμα από το αλφάβητο, έχουν τη δυνατότητα να εμφανιστούν όλα τα γράμματα του αλφαβήτου με βάση κάποια πιθανότητα το καθένα.

Στην μεταπτυχιακή διπλωματική εργασία αναλύουμε τις βεβαρυμμένες αλληλουχίες και τον τρόπο που θα εφαρμοστεί η μετρική για την αποθήκευση των υποσυμβολοσειρών σε κάθε ευρετήριο. Πέρα από αυτό ένα από τα καίρια θέματα που εφαρμόζονται σε ανεστραμμένα ευρετήρια είναι και η συμπίεση. Υπάρχουν διάφοροι αλγόριθμοι συμπίεσης που εφαρμόζονται σε ανεστραμμένα ευρετήρια με διαφορετικά πλεονεκτήματα και μειονεκτήματα ο κάθε ένας. Θα παρουσιάσουμε, μια συγκριτική ανάλυση των καλύτερων αλγορίθμων συμπίεσης και πως αυτοί μπορούν να εφαρμοστούν στη μονοκειμενική συλλογή βιολογικής αλληλουχίας που χρησιμοποιούμε στα πειράματα μας. Τέλος, θα παραθέσουμε και τα αποτελέσματα της συμπίεσης που δημιουργήθηκαν από την εφαρμογή ενός από τους καλύτερους αλγορίθμους συμπίεσης επάνω στο ευρετήριο μας καθώς και κάποιες παραλλαγές που έγιναν πάνω στην μετρική για την προσπάθεια να πάρουμε καλύτερα αποτελέσματα συμπίεσης.

SUMMARY

The data that must be stored are increasing fast over the past years, which can be either web data (text, images etc.) or biological sequences. On the other hand, the data due to the rapid development demand better storage, inside the least space, providing in the same time a fast retrieval of this information.

In this diploma thesis, the great interest is the way that a DNA sequence will be stored with the use of inverted indexes and suffix trees. Suffix trees as a data structure makes possible to find the number of occurrences of a subsequence inside a larger sequence. The above data structure is something that is needed for the best storage, either in one level inverted index or a two-level. The most efficient way for this split during the storage process is the main problem that we solve in the current dissertation.

In biological sequences there are two types of strings, the first type is classic strings where my undergraduate diploma thesis was base, and we will give a quick analysis. The other type is weighted strings in which we will analyze the way of handling the best storage. Weighted are called the sequences where in some certain positions over the sequence instead of occurring a certain letter from the alphabet, there is a possibility to occur more than one letters of the alphabet based in a certain probability for each one.

In the current postgraduate diploma thesis, we analyze the weighted sequences and the way we will apply the mathematic formula to store the subsequences in each inverted index. Beyond that, one of the most relevant topics to inverted indexes are the compression techniques. There are several compression algorithms that are applied to the inverted indexes with different advantages and disadvantages each one. We will introduce a comparative analysis of the best algorithms and in which way the could be implemented in the biological sequence that is used in our experiments. Finally, we will take the results that were created by compressing our inverted index with one of the best compression algorithms, as well as some changes that we made to our formula to take better compression rate results.