

Contents

Chapter 1: Introduction (cca 1,5 pages)

Motivation for data file comparison, difference with applicaiton in bioinformatics, text and data file analysis, plagiarism detection

Chapter 2: String comparison (cca 5 pages)

Definition of different types of string comparison. Simple alignment, maximal matches, edit distance, longest common subsequence.

Chapter 3. Text processing data structures and their implementation (cca 10 pages)

Enhanced suffix array, lcp-table, lcp-interval tree, implementation details

Chapter 4: Computing maximal matches/repeats (cca 10 pages)

Finding maximal matches/repeats using enhanced suffix array, implementation details
Match report tree - Linear space representation of maximal matches/repeats.

Chapter 5: Computing longest common subsequence using maximal matches (cca 5 pages)

Relation between LCS and maximal matches
Computation of LCS using maximal matches.

Appendix A: DiffEngine library programmers manual (cca 3 pages)

Appendix B: <application name> user manual (cca 4 pages)