Prof. Jayant Harista

Title: Indexing Techniques for Biological Data

Abstract: The indexing technique commonly used for long strings, such as genomes, is the suffix tree, which is based on a vertical (intra-path) compaction of the underlying trie structure. Tin this talk, we investigate an alternative approach to index building, based on horizontal (inter-path) compaction of the trie. In particular, we present SPINE, a carefully engineered horizontally-compacted trie index. SPINE consists of a backbone formed by a linear chain of nodes representing the underlying string, with the nodes connected by a rich set of edges for facilitating fast forward and backward traversals over the backbone during index construction and query search. A special feature of SPINE is that it collapses the trie into a linear structure, representing the logical extreme of horizontal compaction. We describe algorithms for SPINE construction and for searching this index to find the occurrences of query patterns. Our experimental results on a variety of real genomic and proteomic strings show that SPINE requires significantly less space than standard implementations of suffix trees. Further, SPINE takes lesser time for both construction and search as compared to suffix trees, especially when the index is disk-resident. Finally, the linearity of its structure makes it more amenable for integration with database engines.