Abstract

Emergent interest in genomic research has resulted in the creation of huge biological sequence databases, however search and retrieval of relevant information from these databases takes a lot of processing time, when performed conventionally as size of databases containing DNA sequences is huge. Hence, providing an efficient searching mechanism is mandatory. In this paper we present an efficient search mechanism using Hashing techniques. Initially, the data is hashed and indexed according to different window sizes. During this process, we eliminate redundancies and only record patterns with distinct elements and provide them with corresponding hash values. During the search phase, the search string is checked for the size of the window and if it exceeds the maximum limit of 4, then it is divided. The first part is considered as the search string and the search is made. After the confirmation of the index, the strings that follow the current indexed string are matched with the search string and finally the confirmation is made. The simulation results show that the current methodology provides faster results, while occupying lesser memory.
Fast and Efficient Hashing for Sequence Similarity Search using Substring Extraction in DNA Sequence Databases

References

- Petri Kalsi, Hannu Peltola, and Jorma Tarhio, &quot;Comparison of exact string matching algorithms for biological sequences&quot;, BIRD, 417-426, 2008.
- David Nassimi, Milind Joshi, Andrew Sohn, &quot;H-PBS: A Hash-Based Scalable Technique for Parallel Bidirectional Search&quot;, 1063-6374/95 1995 IEEE
- Zhou Bai Stefan C. Kremer, &quot;Sequence Learning: Analysis and Solutions for Sparse Data in High Dimensional Spaces&quot;, 978-1-4673-1191-5/12/$31.00 ©2012 IEEE.
- David Dittman, Taghi Khoshgoftaar, Randall Wald, and Amri Napolitano, &quot;Similarity Analysis of Feature Ranking Techniques on Imbalanced DNA Microarray Datasets&quot;, 2012 IEEE International Conference on Bioinformatics and Biomedicine.
- V. Hari Prasad, Dr. P. Y. Kumar, Dr. D. Vasmuthi, &quot;Adaptive segmentation of DNA sequences using SBC Technique: A novel Algorithm&quot;, ICCCNT'12, July 2012, Coimbatore, India.
- Maryam Nuser, Izzat Alsmadi, &quot;Evaluating Graphical and Statistical Techniques for Measuring Similarity in DNA Sequences&quot;, 978-1-4673-1550-0/12/$31.00 ©2012 IEEE.

Index Terms

Computer Science Applied Sciences
Keywords
Hashing; Sequence Similarity by Hashing; Substring Extraction; DNA Sequence