B

Biological Sequences

Amarnath Gupta
San Diego Supercomputer Center, University of
California San Diego, La Jolla, CA, USA

Synonyms

[DNA sequences](#); [Protein sequence](#)

Definition

A biological sequence is a sequence with a small fixed alphabet, and represents a naturally occurring or experimental generated fragment of genetic or protein material or any intermediate product (like the messenger RNA).

Example: A DNA fragment has the 4 character alphabet ‘A’, ‘C’, ‘T’, ‘G’. Chromosomes are long strings over this alphabet.

Key Points

Biological sequences can be long. A full chromosome may have millions of characters. Therefore development of proper storing and indexing strategies is very important for fast retrieval. Suffix tree based indexes have been used successfully for long biological sequences. Further,

approximate string matching techniques with potential deletions and insertions are important for biological sequences. BLAST is a well known algorithm used for approximate matching and ranking of biological sequences.

Cross-References

- ▶ [Index Structures for Biological Sequences](#)
- ▶ [Query Languages and Evaluation Techniques for Biological Sequence Data](#)
- ▶ [Query Languages for the Life Sciences](#)

Recommended Reading

1. Brown AL. Constructing genome scale suffix trees. In: Proceedings of 2nd Asia-Pacific Bioinformatics Conference; 2004.
2. Hunt E, Atkinson MP, Irving RW. A database index to large biological sequences. In: Proceedings of 27th International Conference on Very Large Data Bases; 2001. p. 139–48.
3. Phoophakdee B, Zaki MJ. TRELLIS +: an effective approach for indexing genome-scale sequences using suffix trees. In: Proceedings of the Pacific Symposium on Biocomputing (online proceedings); 2008. p. 90–101.
4. Tian Y, Tata S, Hankins RA, Patel JM. Practical methods for constructing suffix trees. VLDB J. 2005;14(3):281–99.