A computational perception of locating multiple longest common subsequence in DNA sequences

Abstract:

-Bioinformatics is a growing field that combines biological and computer science research.

-The longest common subsequence (LCS) is a problem in computational science that must be resolved. Finding LCS is a fundamental task in deoxyribonucleic acid (DNA) structure research and other molecular biology. for determining the LCS of two DNA sequences and their region. To do this, DNA sequences are stored in an array, and the matching algorithm is used to compare DNA sequences. After the matching process is completed, the longest common subsequence(s) is found. The maximum length of LCS obtained is 8. The calculation time depends on the length of the DNA.

Introduction:

-(LCS) method is used to find the longest common subsequence between two subsequences. It is important to detect the presence of disease-causing sequences in human DNA at an early stage in order to avoid disease's consequences. One of the goals of bioinformatics is to provide a way for the analysis of data.

- Every living thing is made up of one to trillions of cells, which are the basic unit of all living things. Cell performs various operations such as molecule transportation, energy conversion and reproduction. DNA is a macromolecule composed of a series of nucleotides, each of which contains a nitrogen base, as well as a deoxyribose sugar and phosphate. There are four nucleotide bases in a DNA molecule. Cytosine (C), thymine (T), adenine (A), and guanine (G). The instructions for making proteins are encoded in DNA. A gene is a DNA sequence that plays a role in the functional part of protein or RNA molecule development, and inherited characteristics from generation to other.

- Human genome contains three billion nucleotide bases and about 20,000 genes. There are 46 chromosomes in the human genome. DNA is a chemical substance that is used for a variety of purposes in industry.

Related works:

- **For locating LCS, Tripathi and Pandey(2016) discussed two similarity algorithms: maximum common sub stream (MCS) and Rabin-Karp (R-K). Rabin-Karp is better than (MCS).** International Conference System Modeling & Advancement in Research Trends (SMART), Moradabad, pp.334–338.
- **Dheenadayalan (2013) Using hash-based common substring with suffix tree (HCSST) and HCS with separate chaining (HCSSC) calculations, suggested memory-efficient solutions for discovering standard substrings in various arrangements.** International Conference on Technology, Informatics, Management, Engineering and Environment, Bandung, pp.140–145

- **Yang (2013) Pro-MLCS with prevailing point method to deal with locate the fundamental substrings was expected. Genius MLCS can quickly find an expected arrangement and then continue to deliver better results. until getting the ideal one.** 'a new progressive algorithm for a multiple longest common subsequences problem and its efficient parallelization', IEEE Transactions on Parallel and Distributed Systems, Vol. 24, No. 5, pp.862–870.
- **Rubi and Arockiam (2012) suggested Positional LCS as a way to reduce time complexity. This an algorithm based on position to find (LCS) in a sequence database (SDB). And in (2016) proposed a method called Decode-HMM-MLCS.** Journal of Chemical and Pharmaceutical Sciences (JCPS), Vol. 9, No. 1, pp.59–64. IEEE International Conference on Computational Intelligence and Computing Research, Coimbatore, pp.1–4
- **Alsmadi and Nuser (2012) proposed the two algorithms called longest common substring and longest common subsequences (LCS, LCSS) to compare the DNA sequences.** International Journal of Advanced Science and Technology, Vol. 47, pp.13–32.
- **Wang (2011) suggested another calculation for finding a LCS of any number of strings. This algorithm is focused a concept called dominant point approach. Its calculation depends on divide and conquer technique.** IEEE Trans. Knowledge and Data Eng., Vol. 23, No. 3, pp.321–334.
- **Shukla and Agarwal (2010) proposed a simple and time-efficient parallel calculation. That computes the relative places of characters, which is used to determine the LCS of DNA, RNA, protein.** International Conference on Computer and Communication Technology (ICCCT), Allahabad, Uttar Pradesh, pp.496–502.
- **Rizvi and Agarwal (2007) suggested a calculation for finding LCS from two given groupings of DNA and Proteins. and in (2006) suggested a technique which analyses the database grouping of DNA representing genome of some living.** (2007) IEEE 33rd Annual Northeast Bioengineering Conference, Long Island, NY, pp.302–306. (2016) Third International Conference on Information Technology: New Generations (ITNG'06), Las Vegas, NV, pp.560–561.
- **Beal (2015) proposed a compression method for solving LCS problem in genome resequencing data.** IEEE International Conference on Bioinformatics and Biomedicine 2015, doi: 10.1186/s12864-016-2793-0.
  Ozkan and Turksen (2015), suggested a fuzzy C-means (FCM) calculation for coordinating the LCS. That computes in two steps. https://www.researchgate.net/publication/281084370 (accessed 29 May 2018).
- **Lavanya and Murugan (2013), suggested two techniques for finding (MLCS) and (MSCS). In first technique uses support vector. In second technique uses Positional Weight Matrix.** International Journal of Engineering and Technology(IJET), Apr–May, Vol. 5, No. 2, pp.1153–1161.
- **Peng and Wang (2017) to recognize MLCS, proposed a graph-based model called levelled-DAG.** Frontiers in Genetics, Vol. 8, Article 104, pp.1–13.
- **Chen (2017) proposed an algorithm for automatic international disease classification based on similarities using LCS.** PLoS ONE, Vol. 12, No. 3, e0173410 [online] http://sci-hub.tw/https://doi.org/10.1371/journal.pone.0173410.

Methodology:

There are two DNA groupings and find the LCS among them. DNA sequences are stored in an array and the matching process is performed. the longest subsequence is identified and the time complexity is computed for various length of DNA sequences.by using **algorithm MATLAB R2012b.**

DNA sequence length is considered and used for process is 500 for comparing and locating the LCS of two DNA sequences. These two sequences are represented in X and Y

array separately. Another exhibit is made and named as Z where each array element contains the comparable subsequence of two successions. at least two matches in various areas in string Y share a similar area in string X. This kind of comparison is repeatedly performed until reach the last character present in X and Y array. Length and location of the matched subsequences are stored. From which, the greater length of common sub sequences are obtained. **DNA data is collected from the NCBI database.**

To compare between two DNA:

1- computed and stored Lengths of two DNA sequences.**2**-by using Algorithm 1 each character of DNA seq1 and DNA seq2 are compared. If the character is matched then it is stored and next character is checked. **3**- the subsequence is stored and the length and location of the matched subsequence is computed for each match.**4**- the process will be repeated. Until all the matches are found.

Finding longest common subsequences:

1- For finding the LCS The matching algorithm returns several number of subsequence.by using Algorithm3. **2**- Area and length data of each match between two DNA successions are stored.**3**-The calculation starts perusing all array elements. Every element holds the area and length of coordinating subsequence between given DNA arrangements.**4**- The maximum length can be identified from the obtained lengths of the sequences. From that the information will be retrieved and the LCS can be located.

Results:

1-DNA sequence comparison. The length of the DNA arrangements is calculated when the dataset accumulation of DNA successions is completed. Each character of two DNA sequences is used. are compared, with the length and location of matched subsequences being stored in the array. The locations and lengths of the subsequences are stored.

2- Finding longest common subsequence. After identifying all the subsequences, the array is traversed for finding the subsequence with maximum length. Then the location of that resultant subsequence is retrieved from the array and displayed.

3- Analyzed performance the program has been connected to DNA arrangements with five unique lengths and determines the LCS in each situation with their area. In each iteration, the length of the DNA sequences is varied, and the results are computed. The computation time of the process is calculated. Varying length of DNA sequence are used for performance analysis.