**Improving Alignment techniques for DNA sequencing using machine learning**

Taonga E Kandiwo

May 4, 2022

Taonga E Kandiwo

May 4, 2022

**DNA sequencing using machine Learning**

Thesis Statement

1. Introduction
2. Method
3. Results
4. Conclusion
5. Refferences

**Introduction**

In organisms, DNA is a kind of biomacromolecule. It holds life's genetic information and directs biological growth as well as the proper functioning of life's functions. Machine learning is currently frequently utilized in sequence data analysis, and it has a wide range of applications in terms of enhancing data processing capacities and providing useful biological data. Sequential encoding, one-hot encoding, and k-mer encoding are the three most used methods for sequence encoding (Choong and Lee, 2017). Sequential encoding's performance is equivalent to one-hot encoding's, but the training period is substantially less. One-hot encoding is commonly employed in deep learning algorithms and is particularly well suited to CNN algorithms (convolutional neural networks). Furthermore, the performance of one-hot encoding is rather stable across data sets, although high performance requires an appropriate CNN. In some evaluation data sets, ordinal codes represented by matrices perform well. The success of CNN in detecting DNA motifs is determined on the design of sequence encoding and representation. The good performance of the ordinal coding method shows that there is still room for improvement in the single-point coding method.

**Method**

During DNA evolution, its sequence patterns are well conserved, which is of great significance for biological research. The DNA sequence pattern is usually a sequence fragment in the DNA sequence that has a specific function. In the process of DNA evolution, the more conserved regions in most sequences will form specific sequence patterns, and the structure and function of these sequences play an important role. Therefore, identifying these patterns is an important research content of DNA sequence data analysis. This helps to predict DNA sequence function and explain the evolutionary relationship between sequences. The purpose of DNA sequence pattern mining is to find such sequence patterns from DNA sequences and to identify genes and their functions.

There are two types of sequential pattern mining algorithms currently available: One form of sequential pattern mining method is mining alternative patterns, which looks for patterns in the sequence that surpass a given threshold. They can only mine in a single series of alternating patterns. However, in the field of biological sequence analysis, we frequently need to analyze sequence patterns in sequence sets at the same time, which this approach cannot do. As a result, a different kind of sequence pattern mining algorithm is required. This sort of sequence pattern mining method searches data sets for recurrent sequence patterns. When dealing with large amounts of biological data, such algorithms are typically very slow.

**Results**

DNA sequence analysis provides an opportunity to explore the genetic variation of organisms. The rapid growth of DNA sequence data has continuously expanded the demand for DNA sequence analysis. At present, there are still the following problems in DNA sequence data mining. Firstly, when processing large-scale DNA sequence data, there are still efficiency issues. Appropriate DNA sequence data mining methods should be built for various biological demands based on the relevant background information and sequence properties. Thirdly, it is critical to understand how to extract the sequence properties of DNA sequences and how to build an efficient similarity measure to assess sequence similarity. And lastly, due to the “black box” nature of machine learning, the output of machine learning is difficult to give a reasonable explanation from a biological perspective, which limits the application of the model to a certain extent.

On one side, DNA sequence analysis using machine learning illustrates the biological importance of DNA sequences. Currently, a large number of algorithms can analyze DNA sequences efficiently, but their mining results are highly sensitive and specific, resulting in a large deviation during use. As a result, figuring out how to include the biological importance of DNA sequences into the data mining process is an issue that deserves everyone's attention and investigation. Traditional analytic tools, on the other hand, are inefficient in terms of computation time as data volumes grow, and how to build fast calculating techniques is an important research topic. Mining efficiency will be considerably improved by combining distributed computing with parallel computing.

**Conclusion**

On one side, DNA sequence analysis using machine learning illustrates the biological importance of DNA sequences. Currently, a large number of algorithms can analyze DNA sequences efficiently, but their mining results are highly sensitive and specific, resulting in a large deviation during use. As a result, figuring out how to include the biological importance of DNA sequences into the data mining process is an issue that deserves everyone's attention and investigation. Traditional analytic tools, on the other hand, are inefficient in terms of computation time as data volumes grow, and how to build fast calculating techniques is an important research topic. Mining efficiency will be considerably improved by combining distributed computing with parallel computing.

**Refferences**

Naznin, F., Sarker, R., and Essam, D. (2011). Vertical decomposition with genetic algorithm for multiple sequence alignment. *BMC Bioinformatics* 12:353. doi: 10.1186/1471-2105-12-353

Nguyen, N. G., Tran, V. A., Ngo, D. L., Phan, D., Lumbanraja, F. R., Faisal, M. R., et al. (2016). DNA sequence classification by convolutional neural network. *J. Biomed. Sci. Eng.* 9:280. doi: 10.4236/jbise.2016.95021

Pearson, W. R. (2013). An introduction to sequence similarity (“homology”) searching. *Curr. Protoc. Bioinform.* 42, 1–8. doi: 10.1002/0471250953.bi0301s42

Pearson, W. R., and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U.S.A.* 85, 2444–2448. doi: 10.1073/pnas.85.8.2444

Ranawana, R., and Palade, V. (2005). A neural network based multi-classifier system for gene identification in DNA sequences. *Neural Comput. Appl.* 14, 122–131. doi: 10.1007/s00521-004-0447-7

Rogozin, I. B., Milanesi, L., and Kolchanov, N. A. (1996). Gene structure prediction using information on homologous protein sequence. *Comput. Appl. Biosci.* 12, 161–170. doi: 10.1093/bioinformatics/12.3.161