PERFORMANCE EVALUATION OF DNA PATTERN MATCHING ALGORITHMS

I. Aziz[1], S. Shoaib[2], K. S. Khurshid[3], T. Ahmad[4] and M. Awais[5]

Department of Computer Science University of Engineering and Technology Lahore, Pakistan
Corresponding Author email: iram.aziz@uet.edu.pk[1] , khaldoon@uet.edu.pk[2]

ABSTRACT: Bioinformatics is a new area of research in which DNA, RNA and proteins sequences are dealt. To store, retrieve, analyze, match, align, search, and process these sequences different techniques are existed. Currently a lot of advancements in sequence analysis cause the drastic increase in the DNA database sizes that require more efficient approaches encompass accuracy. Search and analysis of DNA patterns can be performed by using various pattern matching algorithms in the computational biology. The aim of present study is to present taxonomy and performance evaluation of these pattern matching algorithms. The objective of this SLR is to set a research trend and to find a mathematical model to estimate execution search time before scanning whole DNA sequence by following a search strategy.

INDEX TERMS: DNA, DNA pattern matching, DNA string matching, DNA pattern matching algorithm, time complexity, and space complexity.

(Received 09.02.2022 Accepted 09.05.2022)

INTRODUCTION

DNA is a genetic sequence and a biological molecule considered as a basic blueprint of life [1][4]. The basic function of this molecule is to store and transfer the genetic instructions. This molecule can be viewed as a long string of four alphabets A, C, G and T. These characters are actually nucleotides of four types, Adenine (A), Cytosine (C), Guanine (G), and Thymine (T) [3]. A large quantity of genes sequences from different microorganisms, various species of plants and animals illustrate complex and similar genetic patterns with one another that's why many biologists from all over the world are sure that understanding of biological information like DNA, RNA and proteins sequences is the first step to understand the process of evolution.

Now a days, sequence analysis not only used to find out the origin of evolution but also in DNA profiling, forensic science, crime investigations, disease prediction, gene therapies, paternity issues, detecting muted genes and mutation rate evaluation, protein-protein interaction in cellular activities and for many other applications [1, 2]. The storage of extracted DNA sequences of different species of plants and animals are increasing exponentially day by day. With the drastic increase in the size of DNA databases it becomes more difficult to store and retrieve, classify, alignment, clustering, and matching necessary patterns efficiently in computational biology [14]. Consequently, for fast retrieval and pattern matching techniques more efficient and robust algorithms and methods are required. By acquiring these methods, the goal of bioinformatics, finding possible defected (mutated) genes position, identification of diseases and diagnosing some treatment for that particular mutated region can be achieved. The mutation is changes in a tiny part of DNA [3].

The pattern matching is a task of identifying all occurrences of subsequences or patterns p and their location within large DNA sequence/text T [4]. It can be defined as following.

Given that we have two strings text: 'T' and pattern: 'P' over the alphabet '∑'.
Where ∑ is a set of character A, T, C, G such as ∑ = {A, T, C, G}

- 'i' are the indexes or location of pattern 'P' in the text 'T'
- 'n' is the size of text T
- 'm' is the size of pattern P
- Here we must find all occurrences of 'P' in 'T' with locations 'i'0

Example:

	Pattern P												
Text T	A	T	C	G	G	T	A	A	A	C	G	T	T
indexes	0	1	2	3	4	5	6	7	8	9	10	11	12

This study precisely presents a comparative survey on different string-matching algorithms which are based on character searching. The novelty of this study is that this is a first systematic literature review in the field of DNA pattern matching algorithm to the best of author's knowledge The existing surveys mostly provides

a detail examination of performance of different pattern matching algorithm on the bases of time and space complexity only. Two main factors mutation rate (relevancy ratio) and positions of muted patterns are not included for evaluation which are one of the main challenges of bioinformatics. No standard database like GenBank, NCBI, DDBJ, NDNAD is used for taking DNA test samples for empirical results. Table 1 gives differences among existing reviews based on eight major perspectives time complexity, space complexity, number of identified positions of a pattern, mutation rate, and prediction of possible genes position, quality assessment and targeted digital repositories.

This systematic literature reiew only included comparison for those reviews published in quality journals, conferences and workshops are almost excluded in this study, because more established research gets published in journals. This comparison helps us to build the need of this survey.

Most of comparative studies just focusing on different methodologies, taxonomy, time, and space complexity of pattern matching algorithms and not considering other aspects.

A study conducted in 2019 by [5] provides a detail discussion on issues and taxonomy of at least 50 pattern matching algorithms. Taxonomy is briefly explained bellow. String matching further divided into two categories

Exact string matching: This approach is involved to find all occurrences of given pattern P from text T [7,15]. In this type of research two windows pattern window and text windows are given for matching. Windows should be of same length during comparison phase. If mismatch found, then shifting of window is necessary to improve searching efficiency [5].

Approximate string matching: In this approach all occurrences of substring that are close to given pattern P must found. This approach is like exact string-matching algorithm excepts a full match. The degree of closeness is an interesting and important factor in this case.

There are many other approaches for string or pattern matching which are based on matrix, prefix postfix, suffix, factoring, Streaming filter, maximum shift, sliding windows, data compression, character based, automata based, bit parallel based, hashing based and hybrid based [5]. This paper discussed 50 algorithms their searching methodology and performance in term of time and space complexity only and didn't discuss other issues like identification of muted part on long sequence of DNA & relevancy ratio.

Another research analyzed performance and characteristics of different algorithms on natural genome and tested randomly by different pattern sizes. Those algorithms are included Knuth-Morris-Pratt (KMP), Naïve exact matching and Boyer Moore algorithm. It is found that for DNA sequences **Boyer Moore algorithm is faster** and efficient. It has no needless comparison. It works sub linearly on best case. Time & space complexity of this algorithm is O(m+n) [6]. This research doesn't focus on identification of locations of mutated region. Any quality assessment methods for literature review are not specified. No comprehensive and precise approach is followed to conduct survey.

Another approach [4] explores a method called 2-jump, which avoids unnecessary comparisons in the DNA sequences, while comparing with other algorithms. This algorithm doesn't work for single character (though this is not much important). This research also focused on time and space complexity for performance comparison, only. Although this is not the focus of current research, but this might be possible to predict the maximum possible locations of given pattern without scanning by using purposed approach. These researches also not consider other factors like relevancy ratio and locations of mutated region. Standard database samples are also not taken for proving novelty of given study.

Currently a lot of advancements in sequence analysis cause the significant increase in the DNA database sizes. Another powerful technique, machine learning is also in used for analyzing large scale data with sequence analysis to obtain a lot of research achievements. Basic process of data mining and machine learning algorithms are elaborated in this paper. This research also emphasizes on four machine learning applications. DNA sequence alignment, DNA sequence clustering, DNA pattern mining and DNA sequence classification [1]. The performance of all discussed approaches is evaluated in term of time and space complexity only and not identify relevancy ratio, locations of similar region and disease prediction issues.

The novelty of our study is that it is a first SLR in the field of DNA pattern matching algorithm to the best of author's knowledge. Current review also distinguished itself from above discusses reviews on the bases of challenges considers in bioinformatics as well as string matching algorithm. Our review also focused on publication channels. All relevant study selected based on strict inclusion and exclusion criteria. The parameters we are considering are quality assessment, time complexity, space complexity, similarity ratio or relevancy ratio, mutation rate, prediction of possible muted regions and targeted digital repositories.

Table 1: Comparison of existing reviews.

Paper	Comprehensive search strategy	Survey Approach	Quality assessment Method	Time complexity	Space Complexity	Standard database specification	Identified regions	Relevance Ratio/ mutation	Prediction of mutated region	Total digital repository
[6] 2018	×	Informal	×	✓	✓	✓	×	✓	×	NA
[13] 2017	×	Informal	×	×	×	×	×	✓	×	NA
[5] 2019	×	Informal	×	✓	✓	×	×	×	×	NA
[1] 2020	×	Informal	×	✓	✓	×	×	×	×	NA
[4] 2011	×	Informal	×	✓	✓	×	×	×	×	NA
Current Survey 2021	✓	Formal	✓	✓	✓	✓	✓	✓	✓	5

MATERIAL AND METHOD

We followed more comprehensive and precise approach to conduct our research than all the above-mentioned reviews. Our research methodology includes formulating research questions for specific objective, making search query and executing that query on multiple repositories and then extracting relevant information on the basis of strict inclusion and exclusion criteria.

I. RESEARCH QUESTIONS AND OBJECTIVES

Table 2: Highlights research questions with motivations and objective.

	Research question statement	Motivation / objective
RQ1	Which are related Publication channels for DNA pattern matching algorithms? Which channel types of target bioinformatics research?	To find 1. Quality assessed publication venues for bioinformatics research 2. Research on DNA pattern matching algorithms published during last ten years.
RQ2	Which factors or parameters are used to analyze and to evaluate the performance of algorithms?	To determine different assessment parameters described so far to evaluate performance of pattern matching algorithm. And, to find correlation & tradeoff among these parameters
RQ3	How many approaches or methods have been adopted to solve the problem of DNA pattern matching& what is their significance?	To identify the various approaches for pattern matching algorithms.
RQ4	What are main challenges faced in pattern matching algorithms and ultimate goal of bioinformatics?	To discover different technique to reduce the number of comparisons during search so that searching time can be reduced and to find a way to store biological data to reduce memory consumptions.

II. **Formation of Search Query:** Multiple digital libraries are accessed systematically to filter out only relevant and appropriate information. For research of DNA pattern matching algorithm, primary keywords are selected as key identifiers then primary keywords are combined with any of secondary or additional keywords. Primary keywords are identified based on formulated research questions. Secondary keywords and synonyms are also identified for additional keywords. Boolean operators like 'AND' and 'OR' have been incorporated with keywords to develop a search string. A sample of query string has been shown in Listing.

Listing 1: Search Query.

```
[DNA] OR [DNA string] OR [DNA Patterns] OR [DNA Pattern matching Algorithm ] OR [DNA string matching Algorithm] AND [Performance] OR [Evaluation] AND [Time Complexity] OR [Space Complexity] OR [Performance Evaluation of DNA string]
```

III. Selection Based on Inclusion/Exclusion Criteria

Inclusion criteria: Papers are selected for the review that are in the domain of computational biology, according to the research questions and published in journals or conferences. Papers discussing DNA pattern matching and focusing on performance evaluation parameters, discussing tools and methods for DNA pattern matching, disease predictions and genome are also included

Exclusion criteria: Papers are excluded written in non-English, published before 2010 and not emphasizing on character comparison approach. Some papers from snow balling and seems to be very important are included which are published before 2010.

Search query executed on multiple repositories as shown in the Table 5. The automated count for search result is a very big number however that is reduced to 45 after inclusion / exclusion criteria. Quality assessment (QA) plays the most important role in the selection of relevant study for conducting any review. Each of selected study based on keywords like DNA pattern matching algorithms, performance, time complexity and space complexity of DNA pattern matching algorithms. After selection, filtration of required information applied according to the research question.

Table3: Query and Keyword based search.

Selection criteria	Google scholar	ACM	IEEE	PLOSS ONE
Query base	26,400	573,485	4,930,000	14,152
Keyword base	133,000	458,290	17,00	37,044

DISCUSSION AND RESULTS OF RESEARCH QUESTIONS

A. ASSESSMENT OF RQ1: Which are related publication channels for DNA pattern matching algorithms? Which channel types target bioinformatics research?

Listening 2.

No. of paper from Conferences = 5/50 *100 = 10%
No. of paper from Reports = 1 /50 *100 = 2%
No. of paper from Journals = 44 / 50 * 100 = 88

The analysis and development of efficient DNA sequencing tools, DNA pattern matching methods and algorithms are key challenges for the researchers in the field of bioinformatics. For this purpose, high quality publication venues and scientometric analysis based on meta-information in the domain of computational biology is required. In this section publication source, year, research type, methodology and channel of publication of selected studies is presented. After analysis phase few of studies selected as shown in Table 3, proving that publication sources as result world largest professional societies publications more than 40 scholarly in computational biology.

According to years, total number of publications has been shown in figure 1. From 2017 to 2021, maximum publications have be selected which are showing importance of research in computational biology. It is found from figure 2 that most of the studies have been selected from recognized journals and second highest numbers of studies have been chosen from good rank conferences, whereas very few relevant studies have been selected as reports. Figure 2 and Listening 2 proving it percentage wise.

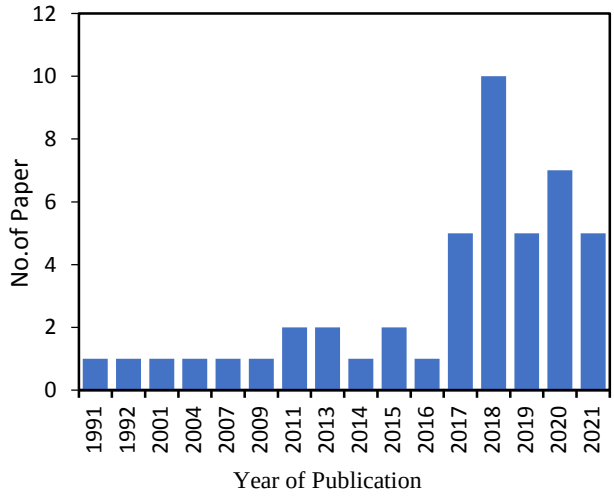

Figure 1: Published paper count identified by our search

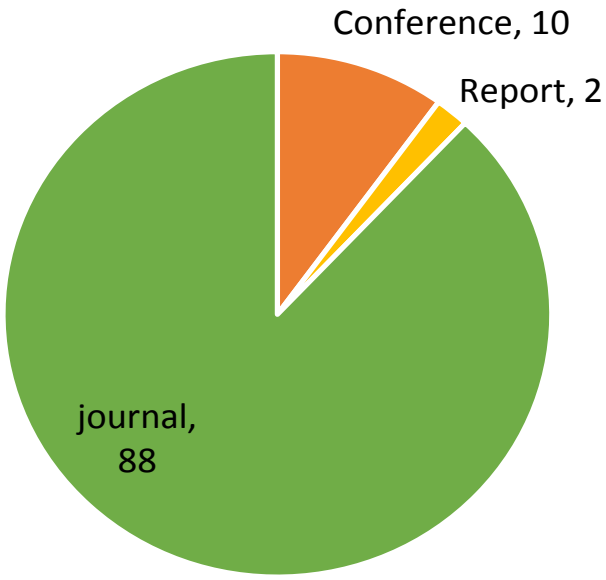

Figure 2: Percentage of publications type

B. ASSESSMENT OF RQ2: Which factors or parameters are used to analyze and to evaluate the performance of algorithms?

The aim of this question is to find main factors that affect performance of algorithms. Time complexity and space complexity are considered as basic parameters for evaluation of performance of any algorithm. Studies like [6] [7][9][10][11][12] assessed the impact of these parameters as shown in the TABLE 1.Time complexity of algorithm deals with number of iterations required to find a specific pattern in a large sequence of DNA while space complexity deals with the space require to an algorithm to acquire for temporary preparation of pre-search phase. It depends on design methodology of an algorithm which technique is followed to use main memory either suffix trees or arrays or stacks or tables. A study [6] considered relevancy as another parameter to check performance of discussed algorithms in survey.

C. ASSESSMENT OF RQ3: How many approaches or methods have been adopted to solve the problem of DNA pattern matching? Multiple tools like PSI-BLAST, FASTA, Roche/454 Life Science BLAST and Illumina Genome Analyzer [16] are used for sequence analysis. Behind tools algorithms are implemented. The taxonomy of these algorithms is presented by [5]. The taxonomy is based on matrix; prefix postfix, suffix, factoring, Streaming filter, maximum shift, sliding windows, data compression, character based, automata based, bit parallel based, hashing based and hybrid-based approaches.

D. ASSESSMENT OF QUESTION 4: What are main challenges faced in pattern matching algorithms and goal of bioinformatics?

The purpose of this question is to identify main challenges faced in pattern matching algorithms and goal of bioinformatics in term of minimizing comparisons for pattern matching to reduce time complexity, disease prediction, disease type, diseases stages, mutation rate for a disease, identifying mutated region on a gene and to find basics of evolution.[2][1] gets all the places and number of duplications of the specified pattern inside a DNA sequence. These two parameters are tremendously essential in determining the type and intensity of any disease.

CONCLUSION AND RECOMMENDATION: An understanding of search trend is built in the field of bioinformatics by conducting this study and following a systematic literature review. The search is performed using as many terms as known to be associated with performance evaluation of pattern matching algorithm. Most of the research we have included are published in general. Four globally available digital repositories and around fifty articles are explored for this research. This is identified that studies which is appeared in recognized journals is mostly selected and only a few had published in workshops or as reports. Most of the selected studies about DNA pattern matching algorithms is statistically proved by implementations and using graphs. Time complexity and space complexity are the frequently addressed aspects of existing studies whereas relevancy, mutation rate and location of mutated regions are less addressed aspects.

Short comings of any review are generally related to improper search strategy, unstandardised data set and not focusing on challenges faced by bioinformatics. However, our search strategy based on all relevant keywords from multiple repositories, proper data extraction based on scoring, strict inclusion/ exclusion criteria.

For future research on DNA pattern matching algorithms, more attention should be paid on giving a mathematical model for prediction of time taken to find a pattern in a large sequence before scanning. Maximum possible location for mutation can be predicted by extracting indexes of minimum occurred character in given DNA pattern in pre-search phase of algorithms. An idea for DNA string storage can be built once the indexes are extracted and can be stored permanently in the database in the form of digits instead of character. Another algorithm can be proposed by improving comparison style of 2-jump string matching algorithm, like matching could be started first, from extreme left then extreme right recursively, of given pattern once the pivot is decided.

REFERENCES

1. Yang, A., Zhang, W., Wang, J., Yang, K., Han, Y., & Zhang, L. (2020). Review on the Application of Machine Learning Algorithms in the Sequence Data Mining of DNA. *Frontiers in Bioengineering and Biotechnology*, *8(1032),* 1-13.
2. Mukherjee A., Dutta G., Chowdhury C. R. (2017). DNA Pattern Matching A Comprehensive Study of Three Pattern Matching Algorithms. International Journal of Computer Application. *International Journal Of Computer Engineering & Application, 13(4)*, 1-5.
3. Loewe, L., & Hill, W. G. (2010). The population genetics of mutations: good, bad and indifferent. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *365*(1544), 1153–1167. https://doi.org/10.1098/rstb.2009.0317
4. Bhukya, R. &Somayajulu, D. (2011) 2-Jump DNA Search Multiple Pattern Matching Algorithm. *International Journal of Computer Science Issues,8(3),* 1694-0814.
5. Hakak, S. I., Kamsin, A., Shivakumara, P., Gilkar, G. A., Khan, W. Z., & Imran, M. (2019). Exact String Matching Algorithms: Survey, Issues, and Future Research Directions. *IEEE Access*, 7, 69614–69637.
6. Hossen, M.H., Azam, M., R. &Rana, H.,K. (2018). Performance Evaluation of Various DNA Pattern Matching Algorithms Using Different Genome Datasets. *Pabna University of Science and Technology Studies.* 3, 14-18.
7. Charras, C. and Lecroq, T. (2004). Handbook of Exact String Matching Algorithms. King's College. *Journal of Computer and Communications*, 4(13)
8. Sun Wu and Udi Manber. 1992. Fast text searching: allowing errors. Commun. *ACM* 35, 83–91. DOI:https://doi.org/10.1145/135239.135244
9. NYOME TUN. (2017). DNA Pattern Matching – A comparison of three pattern matching algorithms. *ISSN*, 3(35), 6916-6920
10. CORE Conference Portal. (2018). http://portal.core.edu.au/conf-ranks/. [Online; accessed 06-Jan-2020].
11. Scimago Journal Country Rank. (2018). https://www.scimagojr.com/. [Online; accessed 06-Jan-2020].
12. Jolanta Kawulok. *(2013)*. Approximate String Matching for Searching DNA Sequences. *International Journal of Bioscience, Biochemistry and Bioinformatics, 3(2).*

13. Citation: Marki´c, I.; Štula, M.; Zori´c, M.; Stipani˘cev, D. (2021). Entropy-Based Approach in Selection Exact String-Matching Algorithms. *Entropy*, 23, 31. https://doi.org/ 10.3390/e23010031
14. P. Brereton, B. A. Kitchenham, D. Budgen, M. Turner, and M. Khalil. (2007). Lessons from applying the systematic literature review process within the software engineering domain. *Journal of systems and software*, 80(4),571–583.
15. P.D. Michailidis& K.G. Margaritis. (2001). On-line string matching algorithms: survey and experimental results. *International Journal of Computer Mathematics*, 76(4)s, 411-434, DOI: 10.1080/00207160108805036
16. Gasperskaja, E., &Kučinskas, V. (2017). The most common technologies and tools for functional genome analysis. *Acta medica Lituanica*, *24*(1), 1–11. https://doi.org/10.6001/actamedica.v24i1.3457.