THE JOURNAL OF COMPUTER SCIENCE AND ITS APPLICATIONS

ncs nigeria computer society www.ncs.org.ng

Vol. 23, No 2, December 2016 ISSN: 2006-5523

AN INTERNATIONAL JOURNAL OF THE NIGERIA COMPUTER SOCIETY (NCS)

Title and Author(s)	Pages
CAPSULE ENDOSCOPY: AN APPLICATION OF PERVASIVE COMPUTING IN E- HEALTH A. S. Olaniyan, A. A. Araromi and F. M. Babalola	1-8
EVALUATION AND IMPLEMENTATION OF WEB-BASED OPEN SOURCE STUDENT INFORMATION SYSTEM (SIS) SOFTWARE Aishatu Ibrahim, Akin Fapohunda	9-16
INTERNET OF THINGS: AN INNOVATION, TECHNOLOGICAL BACKBONES, APPLICATIONS AND CHALLENGES. E.G. Onyedinma and H.C. Inyiama	17-27
NEUTROSOPHIC-BASED DECISION SUPPORT SYSTEM FOR THE DIAGNOSIS OF CONFUSABLE DISEASE A.E Okpako and P.O. Asagba	28-41
A MODEL FOR EVALUATION OF DUE PROCESS TENDERS IN PUBLIC PROCUREMENT U. D. George, O. C. Akinyokun, O. Olabode and O. U. Obot	42-52
DEVELOPING A NEURO-FUZZY MODEL FOR DETERMINING SHORTEST ROUTING PATH IN A COMPUTER NETWORK A.O. Adebare O. Folorunso, and O.L. Usman	53-63
HYBRID INTELLIGENT PREGNANCY RISK LEVEL PREDICTION U. A. Umoh1, P. U. Usip and E. Nyoho	64-75
VLSI DESIGN AND PHYSICAL LAYOUT OF A DECADE COUNTER DIGITAL INTEGRATED CIRCUIT USING AN OPEN SOURCE CAD TOOL O.O. Ilori and O. T. Ayodele	76-83
INVESTIGATING THE RELATIONSHIP BETWEEN THE SIZE OF TRAINING DATA AND ERROR-RATE A. O. Olamiti and A. O. Osofisan	84-89
PERFORMANCE ANALYSIS OF REAL-TIME DATA IN MOBILE NETWORKS B. C. Onyemaobi and E. Okafor	90-102
AN INTEGRATED INTER AGENCY INFORMATION SYSTEM FOR MANAGING MOTOR VEHICLES BUSINESS IN NIGERIA O. Aweh and R. Okpu	103-118
DEVELOPMENT OF A BIO-INFORMATIC KIT O.C.Abikoye, A.N. Babatunde, M.A. Mabayoje, O.I. Oladunni	119-126
A NEW RSA ENCRYPTION SCHEME WITH ERROR DETECTION AND CORRECTION Salifu abdul-Mumin and K. A. Gbolagade	127-137

TETFUND SPONSORED PROJECT, 2012

tetfund TERTIARY EDUCATION TRUST FUND

EDITOR-IN-CHIEF: DR. V.E. EJIOFOR

nigeria computer society
www.ncs.org.ng

THE JOURNAL OF COMPUTER SCIENCE AND ITS APPLICATIONS
Vol. 23, No 2 December 2016

DEVELOPMENT OF A BIO-INFORMATIC KIT

[1]Abikoye O.C., [2]Babatunde A.N., [3]Mabayoje M.A., [4]Oladunni O.I.
[1,3,4] Department of Computer Science, University of Ilorin, Ilorin, Nigeria.
[2]Department of Computer Science, Kwara State University, Malete, Nigeria
[1]kemi_adeoye@yahoo.com [2]akinbowale.babatunde@kwasu.edu.ng
[3]mmabayoje@gmail(yahoo).com

ABSTRACT
Due to the proliferation of biological data and the need to manage and analyse it's data, there is need for a computer technology to extract information stored in certain types of biological data. Several applications have been developed to assist people gain more knowledge and improve the users in this area of interest. In this work, a Bioinformatics Application Kit (BIOKIT) is developed for sequence analysis and manipulation. The work is made up of three (3) parts; the creation of alignment sequences which can either be done globally with the use of the Needleman Wunsch Algorithm or locally using the Smith Waterman Algorithm. Transcription which is performed using the Watson-crick base pairing algorithm was carried out and then followed by translation for the second phase. Pattern matching forms the third part of the work with the use of Boyer-Moore string matching algorithm. The designed system was then implemented using the .NetBio open source library on visual studio as the front end in c# programming language. The developed system having being analyzed and tested was found to be efficient, easy to use and a reliable means for the intended purpose of proposition.

KEY WORDS: Alignment, Bioinformatics, De-oxy Ribonucleic Acid, Pattern Matching, Ribonucleic Acid, Transcription, Translation.

1.0 INTRODUCTION

Bioinformatics is the application of computer technology to the management of biological information. Computers are used to gather, store, analyze and integrate biological and genetic information which can then be applied to gene-based drug discovery and development. The need for Bioinformatics capabilities has been precipitated by the explosion of publicly available genomic information resulting from the Human Genome Project [2].

The science of Bioinformatics, which is the melding of molecular biology with computer science, is essential to the use of genomic information in understanding human diseases and in the identification of new molecular targets for drug discovery. In recognition of this, many universities, government institutions and pharmaceutical firms have formed bioinformatics groups, consisting of computational biologists and bioinformatics computer scientists. Such groups will be keys to unravelling the mass of information generated by large scale sequencing efforts underway in laboratories around the world [7].

Through bioinformatics technology, biological data such as a string of molecule can be extracted from a database and displayed on a computer screen as a series of letters. Those series can be compared and contrasted among human and animal samples as a method of finding out how similar or different humans are to other animals and how that might help develop new drugs to fight life-threatening diseases [8].

This biological data is proliferating rapidly. Public databases such as GenBank and the protein data bank have been growing exponentially for some time now. With the advent of the world wide web and fast internet connection, the data contained in these databases and many special programs

can be accessed quickly, easily and cheaply from any location in the world. As a consequence, computer based tools now play an increasingly critical role in the advancement of biological research. Computational approaches are used to answer biological questions and these questions require that investigators or researchers take advantage of large, complex data set (both public and private) in a rigorous fashion to reach valid biological conclusions [1].

Given the need of laboratory scientists to conduct the analysis of large numbers of genes and sequences, this project is aimed at providing a user-friendly bioinformatics system for biological analysis.

2.0 RELATED WORKS

2.1 Molecular Evolutionary Genetics Analysis (MEGA3) software.

[9] proposed the MEGA3; The MEGA software which was developed in the early1990's was an integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment which aimed at making available a wide variety of statistical and computational methods for comparative sequence analysis in a user-friendly environment. The tool was tested and results were obtained.

2.2 DIALIGN (DIagonal ALIGNment).

DIALIGN is a widely used software tool for multiple DNA and protein sequence alignment. The program combines local and global alignment features and can therefore be applied to sequence data that cannot be correctly aligned by more traditional approaches [6]. Dialign is a local algorithm which aligns whole segments rather than single residues. The name 'Dialign' comes from these regions as they would appear as diagonals on a dotplot. A consistent set of diagonals is determined and iteratively added to the alignment.

Exon's discovery by genomic sequence alignment [9]. The aim of the research was to describe the implementation of the DIALIGN sequence alignment program that has been developed for the alignment of large genomic sequences. This implementation was compared with other alignment programs such as PipMaker, WABA and BLAST. It was discovered that among the long-ranged alignment programs tested, PipMaker is the most sensitive method while DIALIGN is the most specific and considerably slower than the other programs in the study.

2.3 BIOEDIT

[5] presented a report in the his work "BioEdit. BioEdit, a biological sequence editor is an important software for molecular biology". His work which is aimed at providing basic functions for protein and nucleic sequence editing, alignment, manipulation and analysis runs in Windows 95/98/NT/2000/XP. Although BioEdit is not a powerful sequence analysis program, it offers many quick and easy functions for sequence editing, annotation and manipulation, as well as a few links to external sequence analysis programs. It is a C++ program written in Borland's C++ Builder.

3.0 METHODOLOGY

Materials for this work were gotten majorly online: eBook, journals and DNA sequences were downloaded online for the alignment and manipulation. The research work was developed using the .NET bio platform and on the windows form platform. The .NET Bio Framework is an open source, reusable .NET library and application programming interface (API) for bioinformatics research. .NET Bio Framework is essentially a bioinformatics toolkit built on top of the Microsoft .NET Framework 4.0 providing a foundation upon which other tools can be built.

The proposed system contains the three (3) modules, they are:

i. Creating alignments from sequences.
ii. Transcription and translation
iii. Pattern matching against sequences

3.1 Creating Alignment from Sequences

DNA sequence is the precise ordering of the bases adenine, thymine, guanine and cytosine (A, T, G, and C) from which DNA is composed.

Sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. It could be pairwise or multiple alignment.

Pairwise sequence alignment is used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences (protein or nucleic acid).

Multiple sequence alignment (MSA) is a sequence alignment of three or more biological sequences, generally protein, DNA, or RNA. Sequence alignment can either be global or local:

Global sequence alignment: A global alignment contains all letters from both the query and target sequences. The alignment takes place in a two-dimensional matrix in which each cell corresponds to a pairing of one letter from each sequence. The alignment starts at the upper left and follows a mostly diagonal path down and to the right. When two letters are aligned, the path follows a diagonal trajectory. The global alignment was done using the Needleman Wunsch algorithm: There are three major phases here: Initialization, fill and trace back as shown in figures 3.1, 3.2 and 3.3 respectively.

Initialization phase
In the initialization phase, you assign values for the first row and column. The score of each cell is set to the gap score multiplied by the distance from the origin. The arrows all point back to the origin, which ensures that alignments go-all the way back to the origin (a requirement for global alignment).

Fill
In the fill phase, the entire matrix is filled with scores and pointers using a simple operation that requires the scores from the diagonal, vertical, and horizontal neighbouring cells. Three scores are computed: a match score, a vertical gap score, and a horizontal gap score and the maximum is assigned to the cell with the arrow pointed in the direction of the maximum score. The operation is continued until the entire matrix is filled.

Trace back
The trace back recovers the alignment from the matrix. It transverses from the bottom right corner and follows the arrow until it gets to the beginning.
Example: perform the alignment of the sequences GAAT and GGA. With matching score 1, mismatch score 0 and gap penalty 0;

Step1: Initialization.

Figure 3.1: Initialization matrix.

		G	A	A	T
	0	0	0	0	0
G	0				
G	0				
A	0				

Step 2: Calculation of scores and filling the matrix.

Match score = sum of diagonal cell and score for a match.
Horizontal gap score = sum of the cell to the left and the gap score.
Vertical gap score = sum of the top cell and the gap score.

Figure 3.2: filled in matrix.

		G	A	A	T
	0	0 ←	0 ←	0 ←	0 ←
G	0 ↑	↖ 1	1 ←	1 ←	1 ←
G	0 ↑	↖ 1	↖ 1	↖ 1	↖ 1
A	0 ↑	1 ↑	↖ 2	↖ 2	2 ←

Step 3: Trace back

Figure 3.3: Trace back.

		G	A	A	T
	0	0	0	0	0
G	0	↖ 1			
G	0		↖ 1		
A	0			↖ 2	2 ←

The best alignment using the Needleman Wunsch algorithm is GAAT GGA-.

Local sequence alignment: Local alignment aligns a substring of the query sequence to a substring of the target sequence. Local alignment was done using the **Smith Waterman algorithm**: There are also three major phases here:

1. Initialization of matrix as shown in figure 3.4

2. Matrix Filling with the appropriate scores as shown in figure 3.5

3. Trace back the sequences for a suitable alignment as shown in figure 3.6

The two sequences are arranged in a matrix form with A+1columns and B+1rows. The values in the first row and first column are set to zero.

The Smith-Waterman algorithm is a very simple modification of Needleman-Wunsch. There are only three changes:

1) The edges of the matrix are initialized to 0 instead of increasing gap penalties.
2) The maximum score is never less than 0, and no pointer is recorded unless the score is greater than 0.
3) The trace-back starts from the highest score in the matrix (rather than at the end of the matrix) and ends at a score of 0 (rather than the start of the matrix).

Example: perform the alignment of the sequences CAATG and GAATC. With matching score 1, mismatch score 0 and gap penalty 0.

Step1: Initialization

Figure 3.4: Initialization matrix.

		C	A	A	T	G
	0	0	0	0	0	0
G	0					
A	0					
A	0					
T	0					
C	0					

Step2: Calculation of scores and filling the matrix.

Match score = sum of diagonal cell and score for a match.
Horizontal gap score = sum of the cell to the left and the gap score.
Vertical gap score = sum of the top cell and the gap score.

Table 3.5: Filled in matrix

		C	A	A	T	G
	0	0	0	0	0	0
G	0	0	0	0	0	↖
A	0	0	↖	↖	1←	1←
A	0	0	1↖	↖	2←	2←
T	0	0	1↑	2↑	↖	3←
C	0	↖	1	2↑	3↑	3↑

Step3: Trace back.

Table 3.6: Trace back

		C	A	A	T	G
	0	0	0	0	0	0
G	0	0				
A	0		↖			
A	0			↖		
T	0				↖	←
C	0					↑

The local alignment using the Smith Waterman algorithm results are:
Sequence 1 = CAAT-
Sequence 2 = -AATC

3.2 Transcription and translation

3.2.1 Transcription: Transcription is the process by which DNA is copied (*transcribed*) to mRNA, which carries the information needed for protein synthesis. Transcription takes place in two broad steps. First, pre-messenger RNA is formed, with the involvement of RNA polymerase enzymes. The process relies on Watson-Crick base pairing, and the resultant single strand of RNA is the reverse-complement of the original DNA sequence. The pre-messenger RNA is then "edited" to produce the desired messenger RNA (mRNA) molecule in a process called *RNA splicing*.

3.2.2 Translation: The mRNA formed in transcription is transported out of the nucleus, into the cytoplasm, to the ribosome (the cell's protein synthesis factory). Here, it directs protein synthesis. Messenger RNA is not directly involved in protein synthesis, transfer RNA (tRNA) is required for this. The process by which mRNA directs protein synthesis with the assistance of tRNA is called *translation*.

3.3 Pattern matching against sequences

The pattern matching was done using the **Boyer- Moore string matching algorithm.** The Boyer-Moore algorithm is considered as the most efficient string-matching algorithm in usual applications. Unlike other algorithms, Boyer Moore compares the letters of the pattern from left to right. The main ideas of Boyer-Moore in order to improve the performance are some observations of the pattern. In the terminology of this algorithm they are called good-suffix and bad-character shifts.

Boyer-Moore string matching algorithm.
Given a text string T and a pattern string P, find all occurrences of P in T.
Step 1: Scan characters of the pattern from right to left beginning with the rightmost character.
Step 2: During the testing of the possible placement of the pattern P against text T, a mismatch of text character T[i] = c with the corresponding pattern character P[j] is handled as follows:
Step 2(a): If c is not contained anywhere in P, then shift the pattern P completely past T[i].
Step 2(b): Otherwise, shift P until an occurrence of character c in P gets aligned with T[i]. Adopted from (Greg, 2005).

4.0 RESULTS AND DISCUSSION

This section presents the results and discussions using extracts from the system testing.

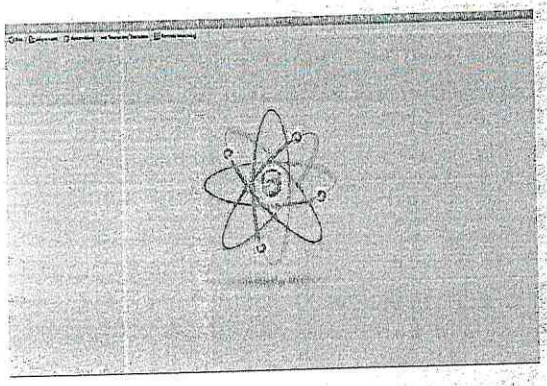


Figure 4.1: Main Menu Interface.

Figure 4.1, the main menu interface displays the link to select file and the links to the alignment, assembling, transcribe/translate and pattern matching interface.

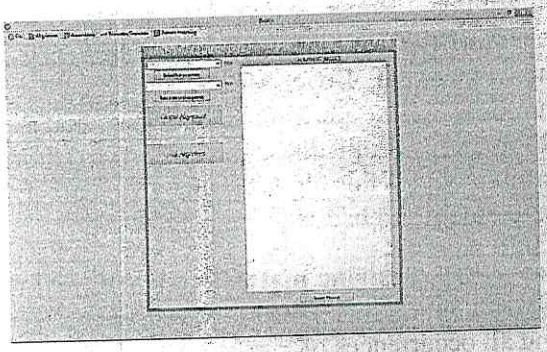


Figure 4.2(a): Alignment interface.

Figure 4.2 (a), the alignment interface is where the user selects the sequences to be aligned and either performs a global or local alignment.

In figure 4.2 (b) the interface where the user selects the first sequence to be aligned is shown.

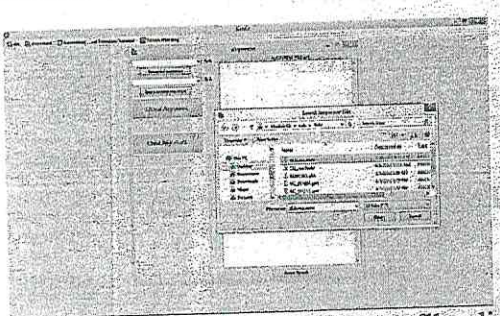


Figure 4.2(b): Alignment open file dialog box.

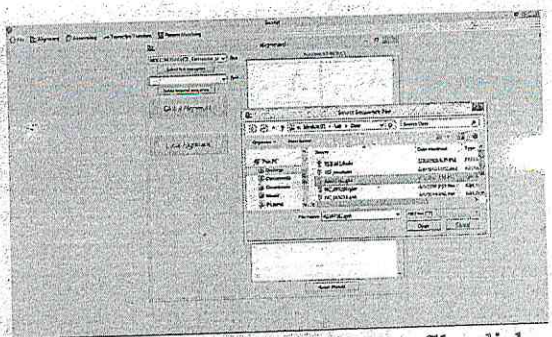


Figure 4.2(c): Alignment open file dialog box.

This is the interface where the users will be able to select the second sequence to be aligned.

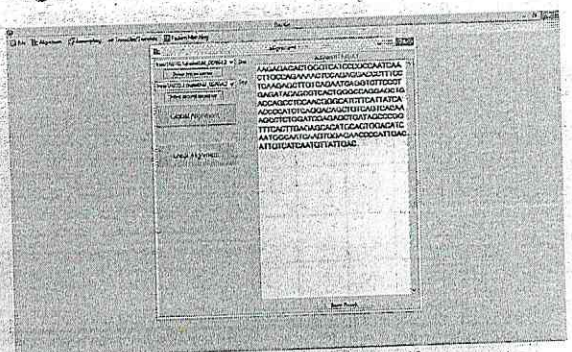


Figure 4.2(d): Result of Global alignment.

Figure 4.2 (d) is the interface displaying the result of a global alignment and allows the user save by clicking on the save button.

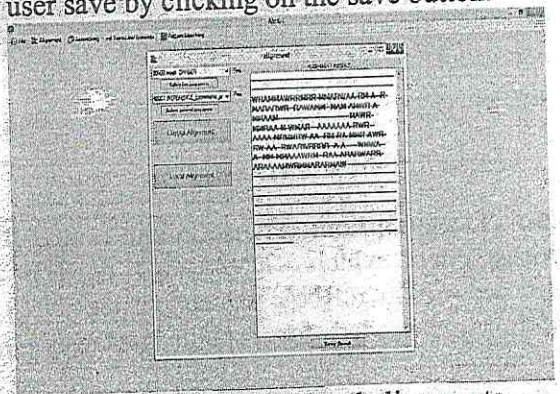


Figure 4.2(e): Result of local alignment.

This interface displays the result of a local alignment and allows the user save the aligned sequence by clicking on the save button.

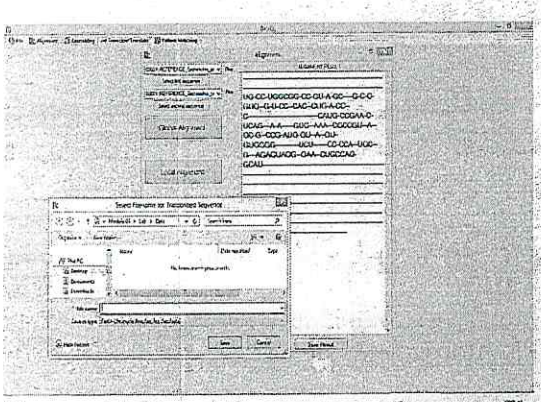


Figure 4.2(f): Alignment result save file dialog box.

This interface allows the user save the result of an alignment.

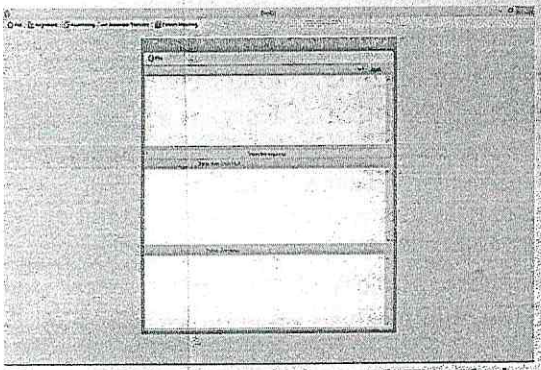


Figure 4.3(a): Transcription and Translation interface.
This is the transcription and translation interface where the user selects the sequence to be transcribed and translated.

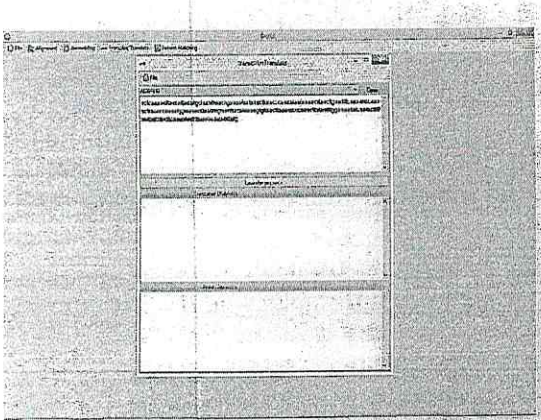


Figure 4.3(b): Transcription and Translation Interface with loaded sequence.

Figure 4.3 (b) shows the interface displaying the selected sequence to be transcribed and translated.

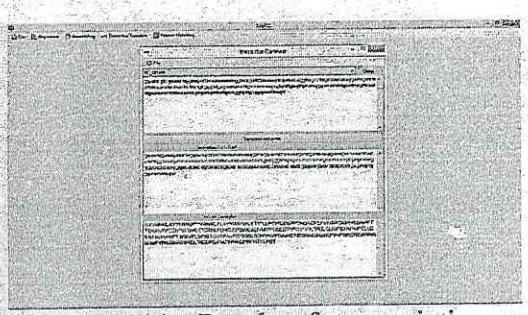


Figure 4.3(c): Result of transcription and translation.

This interface (figure 4.3(c) displays the transcribed and translated sequences.

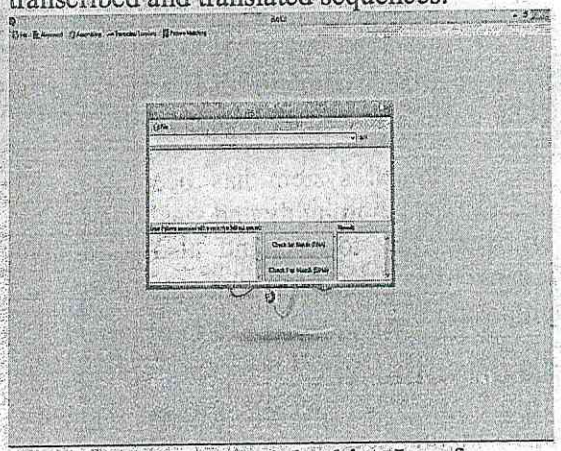


Figure 4.4(a): Pattern Matching Interface.

The interface is the pattern matching interface where the user either selects a DNA or RNA sequence and checks for a pattern in it.

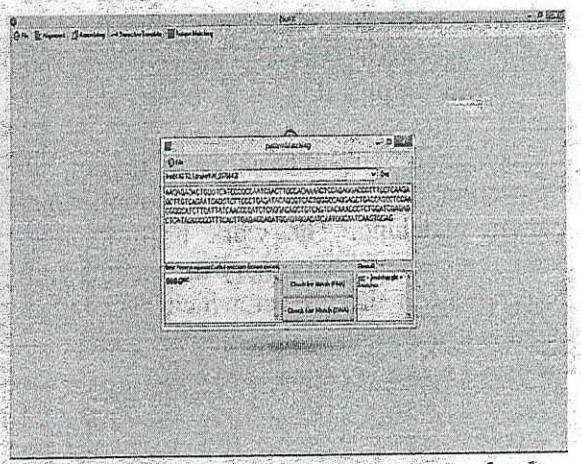


Figure 4.4(b): Result for DNA pattern check.

Figure 4.4 (b) displays the result for a pattern check in a DNA sequence.

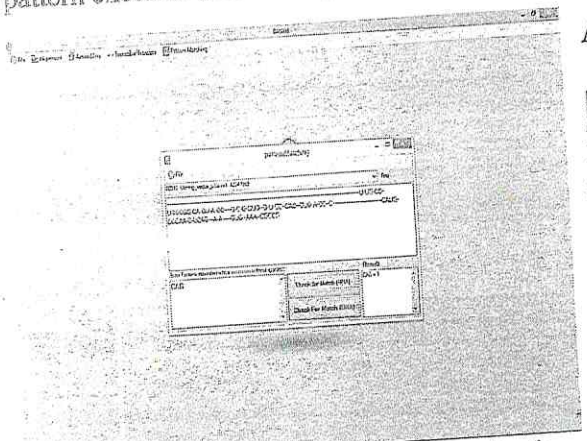


Figure 4.4(c): Result for RNA pattern check.

This interface displays the result for a pattern check in a RNA sequence.

The developed system has been analyzed and assessed thoroughly and was found to be efficient, easy to use and also a reliable means for the intended purpose of proposition.

5.0 CONCLUSION

This work is an electronic sequence manipulation system which attempts to help students and researchers in the area of sequence manipulation with reference to alignment, translation, transcription and pattern matching on a computerized system using c# as the programming language. The developed system was analyzed and evaluated and at the end of this work the aim and objectives were achieved. The developed application has been able to handle issues posed on sequence manipulation with reference to alignment, translation, assembling, transcription and pattern matching. The work itself can be seen as a personal laboratory to students and researchers in this aspect which also helps them gain speed in their work.

6.0 REFERENCE

[1]. Andreas, D.B. and Francis, B.F. (2001). "Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, 2nd edn.

[2]. Bioplanet 2013. Retrieved may 12, 2016, from Bioplanet.com: http://www.bioplanet.com/what-is-bioinformatics/.

[3]. Burkhard M. (2004). DIALIGN: Multiple DNA and Protein Sequence Alignment at BiBiSery. University of Gottingen, Institute of Microbiology and Genetics, Goldschmidtstr. 1, 37077 Gottingen, Germany.

[4]. Greg P. (2005). String Matching: Boyer-Moore Algorithm. Theory in Programming Practice. Department of Computer Science, University of Texas at Austin

[5]. Hall, T.A. (1999). BioEdit: A User-Friendly Biological Sequence Alignment Editor and Analysis Program for Windows.

[6]. Morgenstern, B., Rinner, O., Abdeddaim, S., Haase, D., Mayer, K., Dress, A. and Mewes, H. 2001. Exon discovery by Genomic Sequence Alignment.

[7]. .Net Bio. (2014). Retrieved May 13, 2016, from bio.codeplex.com: https://bio.codeplex.com.

[8]. Smith, S. & Steve S. (2011). What is bioinformatics? Retrieved September 10, 2016, from http://www.rpi.edu/academics/interdisciplinary/what_is_bioinformatics.html

[9]. Sudhir Kumar, Koichiro Tamura and Masatoshi Nei, (2004). MEGA (Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment). URL: http://www.megasoftware.net.