Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:12:2022
DOI10.17605/OSF.IO/WK9Y2

IMPLEMENTING ARTIFICIAL IMMUNE SYSTEM FOR ANALYZING AND PREDICTING WALDENSTROM MACROGLOBULINEMIA FROM MYD88 AND CXCR4

NICHENAMETLA RAJESH

Asst. Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Andra Pradesh. Email.id: nichenametlarajeshklef@gmail.com, nrajeshcse@kluniversity.in

NARESH VURUKONDA

Associate Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India. Email.id: naresh.vurukonda@kluniversity.in

Abstract

One rare type of cancer that slowly grows and affects the human blood cells is termed as Waldenstrom Macroglobulinemia (WM). With the formation of excess WBCs in the bone marrow region, WM occurs. Healthcare industries can provide better treatments to eliminate the symptoms that cannot be cured. Everyone in the healthcare industry knows that genetic mutations trigger WM but do not know what causes the mutations. The risk factors that lead to the multiplication of WBCs in the bone marrow regions causing WM have been identified. Healthcare industries are trying to provide better treatment to save patients. When detected at an earlier stage, the possibilities of curing the disease are bright. Several earlier research works have proposed conventional algorithms and software models for analyzing healthcare data related to WM. But the accuracy is poor and not efficient in terms of cost and time. This paper proposed an Artificial Immune System (AIS) algorithm for analyzing the genomic dataset and identifying Waldenstrom's Macroglobulinemia or its symptoms. The experiment is carried out in Python software for verifying the results. By comparing the experimental results with the other methods, the performance is evaluated. The comparison shows that the proposed AIS algorithm outperforms the others.

Keywords: Artificial Immune System, Waldenstrom Macroglobulinemia, Blood Cancer, Data Analysis.

INTRODUCTION

Human Immune System (HIS), which is complex in nature has the inherent ability to learn about pathogens that enter it and fight infection. This complex nature of HIS has been a source of inspiration for problem-solving in various domains, computer science, and engineering in particular. A majority of research works carried out these days using AI and molecular biology, revolve around the concept of genome mapping to cater to the needs of microbiologists. As these applications generate a large amount of DNA/RNA/mRNA sequences there is an urgent need for automatic computational models for increasing the speed and accuracy of the sequence in analysis. The task of identifying abnormal B-cells (WM affected) is a complex one for the reasons that follow, such as irregular coding, difficulty in recognizing and translating the sequences, redundancy, and complexity in understanding the patterns.

WM is a type of cancer that affects the B-Cells in the human body. It can be identified either using medical image diagnosis or genomic data analysis methods. This researcher in the previous research works has dealt with image processing methods for identifying

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:12:2022
DOI10.17605/OSF.IO/WK9Y2

the presence of WM but the accuracy was not so accurate and satisfactory. Since the presence of WM is better identified biologically, the researcher has shifted to the genomic data analysis method for identifying WM cancer.

Most of the earlier research works published in the National Library of Medicine (NLB) have discussed clinical methods for identifying cancer that have made it difficult for candidates from non-clinical domains to comprehend. Though a major part of the research works proposed medical image processing methods for detecting and identifying WM cancer, very few had suggested biologically inspired algorithms for analyzing human genomic data. One of the popular and potentially effective biological-inspired algorithms is AIS. So, Artificial Immune System (AIS) algorithm has been used for analyzing the genomic data of WM patients. Consequently, the abnormal regions are identified in the human genome microarray data. The AIS algorithm mainly focuses on analyzing gene symbol, chromosomal location, GB-ACC, and cytoband for determining the presence of cancer. This paper contributes towards

- Preparing the data in a manner where the proposed algorithm can learn each entity.
- Implementing the AIS algorithm and experimenting with the prepared genomic data.
- Detecting WM cancer in terms of accuracy and comparing it with the other state-of-the-art methods to evaluate the performance.

Before implementing the AIS algorithms, a detailed literature survey is carried out to understand the issues and challenges present in WM detection.

LITERATURE REVIEW

This section presents a detailed literature survey on the methods adopted so far in detecting and classifying WM cancer using image processing and genomic data analysis methods. For example, the Author in [1] explained that the human immune system is the basis for data analysis methods. It helps to create an artificial immune system model to solve biological problems in terms of representing, comparing, and visualizing genomic datasets. The authors in [2] attempted to utilize the advanced merits of the deep learning algorithms for analyzing lymphoma and cancer of lymph nodes. The experimental results obtained 97.33% of accuracy in analyzing small sized lymphoma dataset. For a large size dataset, the author in [3] proposed an R-CNN algorithm for diagnosing lymph nodes. The R-CNN algorithm is faster than other algorithms and analyses the data region-wise using convolution neural networks. The training accuracy is compared with the testing accuracy.

Three different immune-based algorithms, such as negative selection, clonal selection, and immune network are used for analyzing the DNA/RNA/mRNA sequences and recognizing the normal and abnormal patterns [11]. These algorithms are compared with the AIS, and ANN for pattern recognition. The core functionalities of the computation like mechanisms are used in the computation, basic components, etc. For example in CNN,

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:12:2022
DOI10.17605/OSF.IO/WK9Y2

MLP [12], associative memories [13], and SON [14] are used for pattern recognition. The neural networks are classified based on the basic units like artificial neurons, storage patterns, learning methods, and the architectures are used for weight calculation, etc, to increase the efficiency of the network [15].

From the statement, it has been identified that it is essential to choose the appropriate algorithm for pattern recognition according to the dataset. Thus, this paper is motivated to implement the AIS algorithm for genomic data analysis.

Immune system

The Human Immune System (HIS) is made up of a large number of cells and substances. It can be identified and tested to get rid of hazardous organisms (pathogens). HIS cells have receptors on their interfaces, and several of them chemically attach to pathogens. Others attach to certain HIS cells or substances to simplify the complicated structure of signaling that facilitates the immune reaction. As a result, these relationships are localized since they rely on chemical coupling. The majority of HIS cells travel throughout the body through the lymphatic cells and blood streams, providing a dynamic system of scattered observation and reaction in the absence of a systematic organization and centralized management. Pathogen identification and eradication are the results of functioning of billions of cells according to a small-scale, localized criteria. As a result, the HIS is immune to both the threats on the HIS and the malfunctioning of specific elements. The differences in identifying "self (elements of the body)" from "non-self (Pathogens)" are used to illustrate the parameters that help in detecting the diseases. Many pathogens are not hazardous, but are at the risk of being harmed by the HIS system, which could affect the metabolism of the human body. The HIS system should be capable of differentiating the normal pathogens from the hazardous ones which is represented here as the self and non-self pathogens. The "self" refers to non-toxic components like all physiologically healthy body cells and the "nonself" refers to any pathogen that damages the immune system.

Once infections are detected, the HIS automatically removes the hazardous pathogen from the body. The units of the HIS that carry out this elimination of various pathogens are referred to as effectors. The HIS 's challenge with elimination is to select the appropriate effectors for the specific pathogen to be muted. Appropriate effectors need to be chosen for eliminating a particular type of pathogen and IS usually faces certain issues while choosing the correct effectors. In the absence of a healthy immune system, an artificial immune system must be built to protect against numerous diseases and pathogens. Artificial immune systems (AIS) are dynamic systems that use empirical immune processes, theories, and modeling as well as conceptual immunology to treat the deficiency.

Artificial Immune System (AIS)

Artificial immune system (AIS) is a family of rule-based, smart machine learning systems that draw inspiration from the functions and theories of the mammalian immune system. For application in problem-solving, the techniques are frequently based on the training

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:12:2022
DOI10.17605/OSF.IO/WK9Y2

and storage capabilities of the immune system. This section outlines the design of our Artificial Immune System (AIS). Since AIS is closely patterned after the biological immune system, we will also detail the corresponding biological systems that serve as the inspiration of the model. The immunologic data is short and incomplete; see [1] for a thorough introduction of immunity to the layman.

Problem Definition

In the IS, covalent linkages that develop among protein molecules serve as the foundation for all distinctions into self and non-self. To maintain flexibility, protein links are modeled as binary threads of a specified size(l). The IS must use proteins to differentiate self from non-self; AIS solves a related issue, which we characterize as below. The totality,U, which is made up of all threads of length$l$ is divided into two disjoint subgroups that we refer to as self,S, and nonself,N. It can be represented as $U = S \cup N$and$S \cap N$**.** A detection or classification operation lays AIS and it needs to determine whether a random thread from U is normal (related to self) or abnormal (related to nonself).

A false positive happens if a self-thread is identified as abnormal, and a false negative happens if a non-self-thread is labelled as normal. These are two different types of classification errors that AIS could generate. Comparable mistakes are made by the HIS as well. A false negative error happens if the HIS is unable to identify and eliminate infections, and a false positive error happens if the HIS starts an assault to the body (a process called an auto-immune reaction). Both types of errors are hazardous to the body, thus the HIS has developed to reduce them; likewise, AIS seeks to reduce both types of errors. Figure 1 is a 2D illustration of a thread environment. Figure-1 shows how the immune identification mechanism classifies threads as either normal (related to self) or abnormal (related to non-self) to capture the border within the two groups.

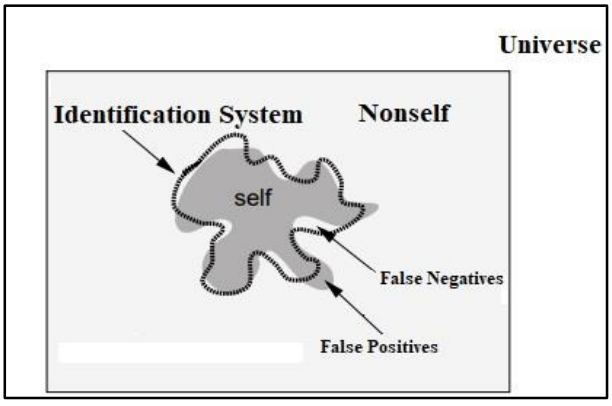

Figure-1: Thread Environment

Self and non-self could not be distinct if real-world issues are transferred to this concept since some threads may characterize into both self and non-self. The classification of such threads as either of the type may cause unavoidable errors under this circumstance. This case is not the only one we take into account. It highlights the need of selecting the

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:12:2022
DOI10.17605/OSF.IO/WK9Y2

appropriate feature for the application area. It is crucial to select proteins that could be utilized to accurately distinguish the self and non-self.

DETECTORS

Numerous types of cells and substances that have been found and scientifically examined make up of innate immune systems. By adopting a single fundamental kind of detector that is based on the lymphocyte type of immunological cells, we can optimize our approach. This detector incorporates characteristics of antibodies, T-cells, and B-cells. In that, it is made up of a large number of portable detectors that move throughout a dispersed area and AIS is equivalent to the HIS. We use a graph $G = (V; E)$ to represent the dispersed surroundings; every vertex ($v \in V$has a small cluster of detectors (referred to as a detection unit), and detectors move from one node to the subsequent node along the edges. The graph design also introduces the concept of location. At the identical node, detectors can communicate with each other. This location concept is helpful.

Since lymphocytes have millions of similar receptors on their membrane, they are referred to as monoclonal. Such receptors attach to areas on pathogens termed epitopes. Because adhesion is based on chemical composition, receptors are expected to attach to a small number of epitopes of a similar type. The sensitivity among the receptors or epitopes increases with the probability of binding formation. In AIS, chemical coupling among epitopes or receptors is treated as a roughly threaded match. All epitopes and receptors are modelled as binary threads of the same size$l$. Every detector has a binary thread that correlates to its receptors.

Hamming length or edit length are two viable rough matching criteria, however, we had chosen the more immunologically appealing r-contiguous bits criterion [2]. If two threads share r-contiguous bits, they are considered to be equal and are shown in Figure-2. Indicating the dimension of the subgroup of threads that a unique detector may match, the number r serves as a cut-off and establishes the specificity of the detector. For example, if the value of r equals$l$, the match is entirely accurate and the detector will exactly match one string (itself). When r equals 0, the matching is completely generic and the detector matches with every single thread with length$l$.

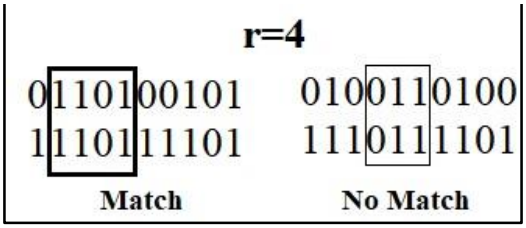

Figure-2: Matching Using Contiguous Match Rule

In the above figure, the matching is found for the value of r equal to 3 instead of 4. Due to the trade-off between the number of detectors utilized and the specificity, the incomplete matching rules with cutoff, such as r-contiguous bits, have this effect. The number of detectors that are needed to get a specific degree of detection. Through this

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:12:2022
DOI10.17605/OSF.IO/WK9Y2

measure the specificity of the detectors can be found. The ideal r is the one that provides better classification while requiring the fewest possible detectors.

If lymphocytic receptors attach to epitopes, a lymphocyte is stimulated. The process of activation alters the lymphocytic state and stimulates the sequence of responses. These responses may create the removal of pathogens. The stimulation of a lymphocyte will be achieved if the quantity of the receptors attached to the epitopes is more than the cutoff. A lymphocyte must attach enough receptors in a short time frame since the chemical interactions among receptors as well as epitopes are not long-lasting. This could be designed with activation cutoffs where at least T strings should match over a specific period. The detector is allowed to consolidate matches and the time delay should match over a short duration of time. This design provides the probability of attachment between the receptor and the epitope. If the detector is stimulated, the count of the match is reset to zero.

The B cell model

1. In the human DNA pattern specification, a variety of intermediate expressions (such as the nuclear RNA and the mRNA), a genome library, and the antibody are all available in the B cell. It keeps track of the B cell's degree of stimulation and its connections with both sister cells and daughter cells. The parent B cells may also be present.

There are two approaches for making a novel antibody. The initial technique aims to replicate the gene transcription, folding, and translation processes that take place in the B cells of the body's immune system. The B cell is given access to a genome library, and the gene selecting procedures are launched. From matured mRNA, the antibody's paratope is produced. To accomplish the above step, the mRNA thread is replicated in a similar manner (for instance **G→C and C→G, T→A and A→T**) (We are aware that this is not a true representation of mRNA translating, which requires identifying the appropriate amino acids to integrate into proteins with the nucleotide triplet sequence).

The second technique employs an antigen's string definition as the model for the mRNA thread. Then, symmetrically, this is replicated as earlier. If the immune system component produces fresh B cells after each cycle, the first approach is applied. When no B cell would attach the present antigen, the second approach is employed to initialize the B cell population.

The Immune network

The present set of B cells that have been created as well as the connections between such B cells are actively maintained in the immune system memories, which are organized like a network. Figure-3 demonstrates a B cell network memory case. A freshly generated B cell is put next to such B cells for which it has an affinity in immune-based storage. The two B cells with which the fresh B cell exhibits the strongest affinity are then identified, which is how the B cell gets incorporated into the network. It is then connected to such two B cells as well as any additional B cells connected to them. This eventually

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:12:2022
DOI10.17605/OSF.IO/WK9Y2

allows the formation of B cell-containing areas inside the network that can handle related issues. Some bridges connect those areas, indicating shared features between various issue areas.

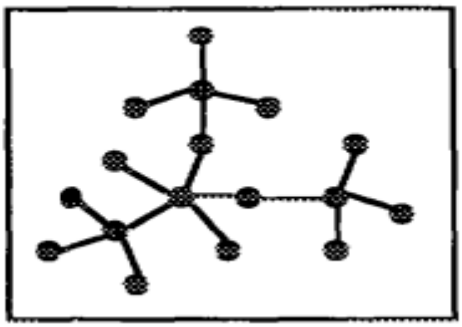

Figure-3: Interpretation Of An Immune Network

An Antibody Model

The antibody has a paratope in the AIS that serves as a design for the layout it will employ to bind to the antigen. The depiction we have selected makes use of the letters A, T, G, C, and the wildcard X. It was discovered that this wild card, which complements all nucleotides, is essential. For the below antigens, for instance, we would seek to discover a typical antibody:

TATAATGCCGTATA
TATAATCGGCTATA
TATAATGATCTATA

We can create a B cell containing the wild card as the antibody (Given below):

TATAATXXXXTATA

We could not create a B cell that could attach all three in the absence of it. This is crucial because our ability to distinguish whether new DNA segments are promoter positives or negatives depends on our ability to find shared characteristics across the antigens.

An Antigen Model

We employ the DNA segments to instruct the AIS (Towell, Shavlik & Noordewier 1990). Those variants have a length of 57 nucleotides. A few are promoter-negative, whereas others have recognized Escherichia coli (procaryotic) gene promoters. The nucleotides in the promoter, which span from 50 nucleotides forward of the protein-coding area to 7 nucleotides within the protein-coding area, are coordinated so that the translation beginning location is at the end of the segment. Figure-4 in the section below offers a few examples.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:12:2022
DOI10.17605/OSF.IO/WK9Y2

```
+,S10,      tactagcaatacgcttgcgttcggtggttaagtatgtataatgcgcgggcttgtcgt
+,AMPC,     tgctatcctgacagttgtcacgctgattggtgtcgttacaatctaacgcatcgccaa
+,AROH,     gtactagagaactagtgcattagcttatttttttgttatcatgctaaccacccggcg
+,DEOP2,    aattgtgatgtgtatcgaagtgtgttgcggagtagatgttagaatactaacaaactc

-,1218,     ttcgtctccgcgactacgatgagatgcctgagtgcttccgttactggattgtcacca
-, 668,     catgtcagcctcgacaacttgcataaatgctttcttgtagacgtgccctacgcgctt
-, 413,     aggaggaactacgcaaggttggaacatcggagagatgccagccagcgcacctgcacg
-, 991,     tctcaacaagattaaccgacagattcaatctcgtggatggacgttcaacattgagga
```

Figure-4: Some Examples Of Promoters With Positive And Negative DNA Segments

52 of the 53 positive samples comprised the antigen population. A, G, C, and T are thus the components of every antigen (Wild card is absent). Our AIS uses a very basic antigen model. An antigen component stands for every potential antigen. This component has a unique epitope, or antigen-representing thread, that antibodies will try to bind. The antigen population component loads the antigens onto the AIS after they have been specified in an independent ASCII file. This component creates antigen instances from a collection of lists that it receives from a data file.

The Behavior of the Artificial Immune System

Antibody/Antigen Association

A match score is generated by determining how strongly the antibody and antigen coincide during the immune reaction. The antibody could attach the antigen when the match score is greater than a predetermined threshold. The B cell will compute its level of stimulation after calculating the binding intensity. Here, every action is discussed in detail. Every component between the antigen and the antibody that correlates in a complementing way is counted by the matching algorithm. Additionally, continuous match locations are given more importance in the matching algorithm. If there are 4 matching components in a continuous area, the particular location would receive a score of 2 to the power of 4. (i.e. 2^4).

Figure-5 shows what happens when an antibody and an antigen are "matched." The figure illustrates that there are 12 matching components. Moreover, every match location's value (for instance, the six components that match at the pattern's top) should include this number. As a result, the example's ultimate match score equals 98.

Antigen	c	g	c	t	t	g	c	g	t	t	c	g	g	t	g	
Antibody	g	c	x	x	a	c	a	c	a	c	g	c	t	a	c	
Evaluation	2	2	1	1	2	2	0	2	2	0	2	2	0	2	2	=>22
Length						6			2			2			2	
Match value						$22+2^6+2^2+2^2+2^2$=> 98										

Figure-5: Computing A Match Value

A measure of the effective bonding between two molecules is called the binding value, which is generated using the match score. An antibody should attach to an antigen for

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:12:2022
DOI10.17605/OSF.IO/WK9Y2

the bonding to be durable, which means that the match score should be higher than a specific level before the initiation of bonding. The AIS itself establishes this threshold. It accomplishes it by changing the threshold's value in relation to the present mean binding value by a factor of 0.5, i.e., by either raising or lowering the threshold by the difference between the two thresholds. This strategy is a modification of the matching method employed by Hightower, Forrest, and Perelson (1993), where we have expanded to add a wild card "X" for justification. But it was crucial that we would not simply produce a batch of B cells where its antibodies were all wild cards (those antibodies may attach to whatever theoretically). As a result, we assigned a score of 1 to a "wild card match" as opposed to a score of 2 for a "complete match," which seemed to function effectively.

B-Cell Simulation Modeling

According to the immunological network hypothesis, a B cell's level of stimulation depends on how strongly its antibody attaches to both the antigen and adjacent B cells. In AIS, an antigen (Complementary to B cells) and B cells (Similar to it) stimulate the B cell. Also, B cells are suppressed by certain other B cells (Complementary to it). The stimulation intensity is determined using the method from (Farmer, Packard, & Perelson 1986). The resultant stimulation intensity must be greater than a specific threshold for the B cell to produce copies of itself. Those clones activate a hypermutation operator and so they are introduced to the immunological network.

Somatic Hypermutation Modeling

To apply somatic hypermutation to a newly formed B cell, we adopt the method described by Farmer, Packard, and Perelson (1986). The three kinds of mutation used by the AIS are simple substitution, substring regeneration, and multi-point mutation. The specific mutation that is used is selected at random.

Every component of the antibody is synthesized individually at a time during multi-point mutation. The component is mutated when an arbitrarily produced value exceeds the mutation threshold. This indicates that to substitute the actual component, a value of A, T, G, C, or X is created at random. Two locations in the paratope of the antibody are randomly chosen for substring regeneration. Subsequently, any of the constants A, T, G, C, or X, which are selected at random, is substituted for every component within those two locations. The simple substitution function chooses a different B cell to utilize as a resource of fresh components for the present B cell using the roulette wheel algorithm (Goldberg 1989). The operator then chooses a paratope from the initial antibody or an antibody from the recovered B cell. The novel "mutated" B cells belong to the immunological system when they could attach to the antigen available or when an affinity could be identified for them anywhere in the system, regardless of the mutation operator used.

Diversity Maintenance

Every time the primary algorithm iterates, the lowermost 5% of the B cell number is eliminated. Equivalent numbers of new B cells are taken by the immune system to

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:12:2022
DOI10.17605/OSF.IO/WK9Y2

replace them. As a result, a network is formed where the population of B cells is constantly changing in terms of both quantity and kind. This promotes the heterogeneity of the system's B cells. The overall functionalities of the AIS algorithm are given in the form of an algorithm and flowchart to understand in a better manner and will be implemented in any computer programming language to experiment and verify the performance.

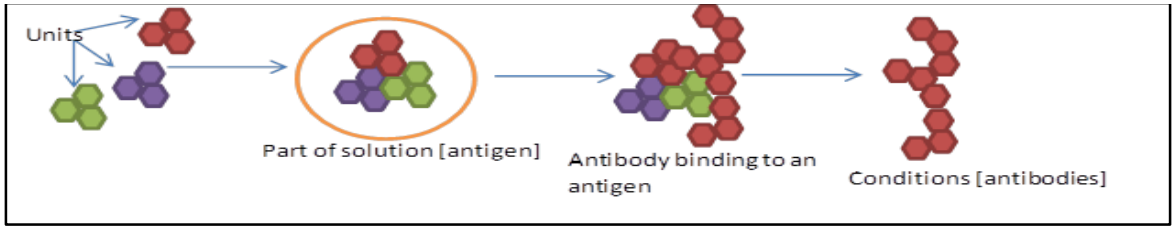

Figure-6: Building Blocks, Partial Solutions And Conditions Mapped To The AIS Metaphor

Artificial Immune System

If a population of responses is important, whether throughout the search or as a result, AISs are important. The issue must also involve some kind of "matching" concept. Because AISs are fundamentally evolving algorithms, they are more suited for recurring optimizations as compared to one-off issues that shift over time and have to be addressed repeatedly. To execute the AIS, four basic considerations must be made: the indoctrination design, the likeness measurement, the choice, and the modification. Once the model has been formed, locate the likeness measure, then select and modify the best option [10].

Algorithm of AIS

- The AIS operation is typically described as sequential procedures:
- Defining N as the total number of sequences and S denote the string.
- K is defined as the size of the sequence taken.
- The target pattern (TP) is designated as SK.
- SK is mapped over K using the pattern matching technique.
- If SK is found in the string Si, then i will be updated in the target array.
- If not SK is not found in the string Si, we need to clone the string using reverse mutation and inverse position mutation techniques.
- The above-mentioned procedures are repeated till the length of the string and length of K.

Pseudo Code for AIS

1 Let N be the total number of sequences for every individual

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:12:2022
DOI10.17605/OSF.IO/WK9Y2

2 Let S be the string

3 Designate the Target pattern (TP) as SK

4 Map SK over K (pattern matching technique)

5 If (SK= Si)

6 Update i in the target array

7 If (SK≠ Si)

8 Clone the string (Reverse Mutation and Inverse Position Mutation techniques)

9 Repeat 4

The algorithm provides a step-by-step overview of how to solve an issue. To create a computer program in any programming language and arrive at the result, pseudo-code is utilized. The schematic representation of the algorithm is shown in Figure-2. The flowchart shows the general operation of AIS.

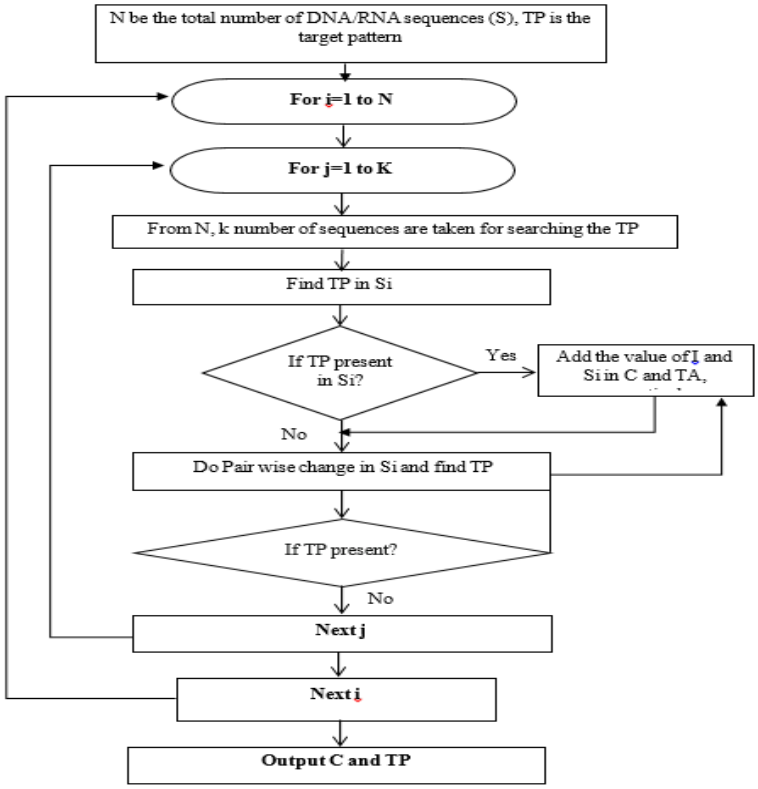

Figure-7: AIS Flowchart

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:12:2022
DOI10.17605/OSF.IO/WK9Y2

Experimental setup

The experimental setup consists of a laptop with an i7-7th generation processor, 8GB ram, 1TB storage, and 1650GPU. The experiment is run in the latest version of python with the required packages to run the MAS5.0 in the system. The open source dataset is loaded from the 1TB storage which is a solid-state drive. The following dataset is used to create an automatic immune system by analyzing the medical data of the patients. The Genetic data of the patients are analyzed and real-time monitoring is done to check the variations in the gene to come up with remedies automatically. For comparing the proposed method the obtained dataset consists of the results calculated with the MAS5.0 method. It is considered a benchmark and the proposed method is compared with it. The sample taken for comparison is MC1336_GEP from the GSM318148. The details of the patient corresponding to this sample data are age 89, gender male, bone marrow involvement is 90%, Hemoglobin level is 11.4 g/dL and B2M is 4.49 mg/L. The RNA is extracted from the bone marrow tissue through Trizol extraction. Under these conditions, the various places of the bone marrow region are scanned. The obtained readings are used in the MAS5.0 method through which the signal intensity, detection, and detection p-value are calculated. The same is done through the proposed method and the obtained results are compared with each other.

Dataset

The dataset taken for the WM pattern recognition is the Series GSE12668 which is available publicly since Aug 24, 2009. This dataset is provided by the NCBI(National Center for Biotechnology Information) as an open source. (Waldenstrom's Macroglobulinemia is very difficult to identify and it is distinct from other clinical observations. It can be defined as the B-cell neoplasm caused by the infiltration of the lymphoplasmacytic in the bone marrow region. Due to this, there is the generation of immunoglobulin M paraprotein. The tumor metaphases are difficult to be obtained through cytogenetic analysis. Also, the genetic reason for the development of this disease is not clearly defined. The dataset consists of data from 42 WM patients which are analyzed through the high-resolution array-based comparative genomic hybridization. The microarray taken for the comparative study is the Human Genome 244A. The dataset consists of 64 samples and each 64 sample has 22863 rows of data. Each sample of the 64 consists of the details of the patients like the sample type, name, organism, characteristics, extracted molecule, protocol, label and label protocol, hybridization protocol, scan protocol, description, and data processing.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:12:2022
DOI10.17605/OSF.IO/WK9Y2

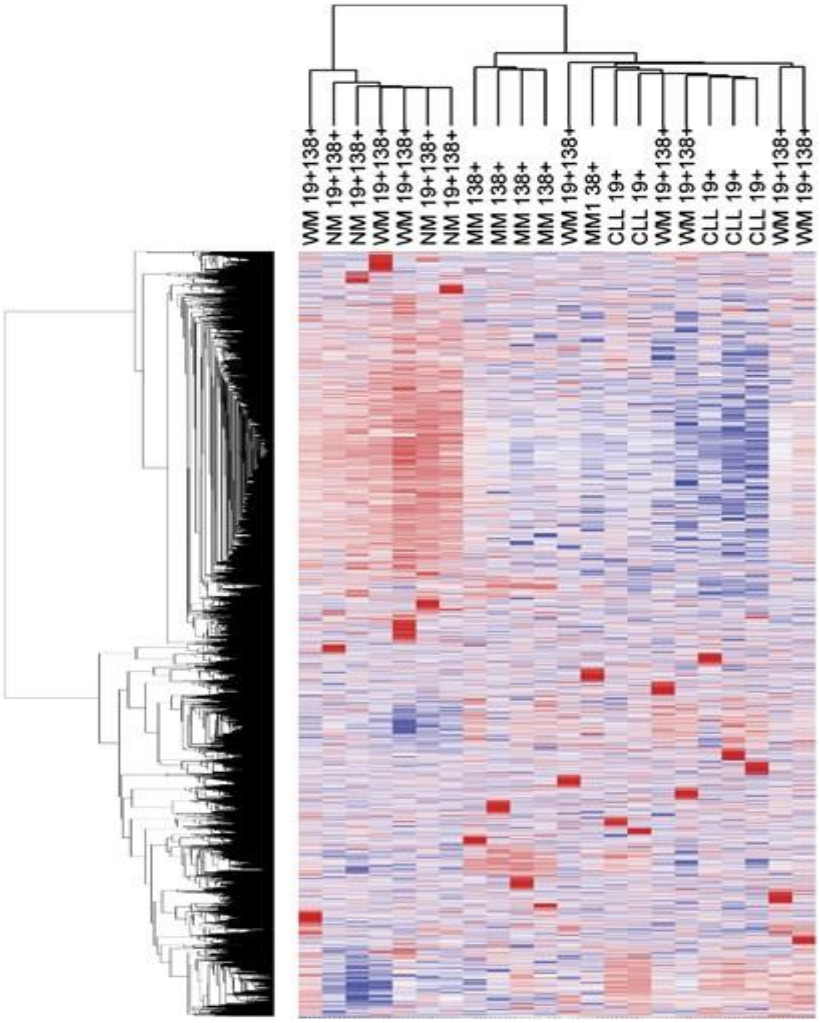

Figure-8: Classification of WM-Cells and Other Cells in the RNA Array

The characteristics of the samples say the health level of the patients like hemoglobin level, BMI, and bone marrow involvement. It also contains specific details about the source of the data obtained and the details about the clinicians who obtained these results. The obtained data is processed through the MAS5.0 method. This method processes the data and gives the overall value for each ID in the sample containing 22283 IDs. It also contains the status of the images obtained from the patients. It also provides the Detection P-value for each entry in the samples. Initially, the AIS algorithm is applied to check the target pattern availability in the DAN sequences. It uses cloning and mutation methods for improving accuracy by increasing the population. From the experiment, the heat map was obtained for classifying the WM cells with other cells like CLL, and MM according to their pattern availability. Waldenstroms' Macroglobulinemia (WM), chronic Lymphocytic Leukemia (CLL), Multiple Myeloma (MM), and Non-Malignant (NM) are the classified cells shown in Figure-8. According to the data availability matching, the given data sequence is labeled as WM-abnormal if the target pattern (TP) is present, or else

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:12:2022
DOI10.17605/OSF.IO/WK9Y2

normal. Table-1 shows the classified results obtained using AIS on the dataset. The TP is the pattern that indicates a corresponding class of the sequence (see Figure-8). For simple understanding, there are only two different classes are obtained from the entire dataset, such as normal and WM-abnormal.

Table-1: Normal vs. Abnormal Number of Data

# of sequences	Normal	WM-Abnormal
4000	2696	1211
8000	3775	4060
12000	5410	6366
16000	7086	8606
20000	8564	11031
22283	9170	12656

Table-1 shows the number of normal and WM predicted IDs from the sample. On the total 22283 IDs, the total WM affected IDs are 12656 and the normal IDs are 9170. This is the result obtained through the MAS5.0 method. It is provided by default in the dataset. We can see the number of IDs that are positive and negative with a consistent increase in the IDs.

Table-2: Accuracy of the AIS Pattern Matching Algorithm

No of ID	Normal Predicted	WM predicted	Accuracy
4000	2498	1122	92.6278
8000	3573	3843	94.6608
12000	5067	5962	93.6654
16000	6581	7992	92.8765
20000	8003	10309	93.4563
22283	8616	11892	93.9675

Table-2 shows the predictions made by the proposed algorithm and the accuracy of the prediction. The accuracy is calculated by comparing it with the predictions given in the algorithm. The results obtained to provide more than 90% accurate predictions. The provided results can be utilized to create an artificial immune system.

Table-3: Time Consumed By the Proposed Method

No of ID	Time (ms)
4000	3425
8000	7653
12000	10987
16000	14532
20000	17689
22283	20382

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:12:2022
DOI10.17605/OSF.IO/WK9Y2

Table-3 gives the time consumed by the proposed algorithm to process and give the results. The time is given in ms. when the algorithm is compared with similar methods like MAS5.0 and RCA the proposed algorithm is very much faster in computing the IDs.

Table-4: Confusion Matrix

Class	Positive	Negative
True	11892	8616
False	763	553

Table-4 shows the confusion matrix for the proposed algorithm. The wrong predictions are relatively lesser than the correct predictions. It is also comparable to the previous methods. Through this confusion matrix, the accuracy, sensitivity, and specificity of the algorithm are calculated.

Table-5: Performance Comparison

Methods	Accuracy	Sensitivity	Specificity
MAS5.0	99.05	0.98	0.94
Proposed AIS	93.96	0.95	0.91

Table-5 shows the performance evaluation by comparing the sensitivity, specificity, and accuracy obtained using various earlier methods and the proposed AIS method. The accuracy of the benchmark is 99.05%, whereas the proposed AIS obtained 93.96%. The obtained accuracy is closer to the benchmark and hence is considered a better method for small-size DNA data analysis.

CONCLUSION

This paper proposed a biologically inspired natural algorithm for analyzing genomic data for identifying and detecting WM cancer through B-cell immune system. A learning methodology is used for analyzing the taxonomy of the immune system for creating a classifier for DNA/RNA sequences. The AIS is one of the best learning models which can learn the HIS thoroughly and predict the sequences having WM cancer. It is done by comparing the cancer patterns in the DNA sequences with the normal patterns using AIS and giving the output DNA sequences having WM. Since the AIS can generate new antibodies, determining the emerging patterns is easy and fast. AIS performs like a neural network and it provides 93.96% accuracy in classifying the abnormal sequences from the dataset.

Future Work

The AIS algorithm provided better accuracy only with a limited amount of data. But the dataset that needs to be analyzed is crowd sourced. Thus, it is necessary to analyze the data using a deep learning model.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:12:2022
DOI10.17605/OSF.IO/WK9Y2

References

1. Jon Timmis, Mark Neal, John Hunt, (2000), "An artificial immune system for data analysis", BioSystems, Elsevier, Vol. 55, PP. 143–150, DOI: 10.1016/S0303-2647(99)00092-1.
2. Rucha Tambe, Sarang Mahajan, Urmil Shah, Mohit Agrawal, Bhushan Garware, (2019), "Towards Designing an Automated Classification of Lymphoma subtypes using Deep Neural Networks", ACM, PP. 143-149, DOI: https://doi.org/10.1145/3297001.3297019.
3. Lei Ding, Guangwei Liu, Xianxiang Zhang, Shanglong Liu, Shuai Li, Zhengdong Zhang, Yuting Guo, Yun Lu, (2020), "A deep learning nomogram kit for predicting metastatic lymph nodes in rectal cancer", Cancer Medicine.
4. Steven A. Hofmeyr. A Immunological Model of Distributed Detection and its Application to Computer Security. PhD thesis, Department of Computer Sciences, University of New Mexico, April 1999.
5. J. K. Percus, O. E. Percus, and A. S. Perelson. Predicting the size of the antibody-combining region from consideration of efficient self/nonself discrimination. In Procedings of the National Academy of Science 90, pages 1691–1695, 1993.
6. Towell, G. G., Shavlik, J. W., & Noordewier, M. O. (1990, July). Refinement of approximate domain theories by knowledge-based neural networks. In Proceedings of the eighth National conference on Artificial intelligence (Vol. 2, pp. 861-866).
7. Forrest, S., Hightower, R., & Perelson, A. S. (1996). The Baldwin effect in the immune system: Learning by somatic hypermutation. Individual Plasticity in Evolving Populations: Models and Algorithms.
8. Farmer, J. D., Packard, N. H., & Perelson, A. S. (1986). The immune system, adaptation, and machine learning. Physica D: Nonlinear Phenomena, 22(1-3), 187-204.
9. Golberg, D. E. (1989). Genetic algorithms in search, optimization, and machine learning. Addison Wesley, 1989(102), 36.
10. U. Aickelin and D. Dasgupta, Edmund K. Burke (Editor), Graham Kendall (Editor), "Search Methodologies: Introductory Tutorials in Optimization and Decision Support Techniques", University of Nottingham, Nottingham NG8 1BB, UK.
11. L. N. de Castro and J. Timmis, (2014), "Artificial immune system: A novel paradigm to pattern recognition", Vol. 8, ANNIS.
12. Rumelhart, D. E., McClelland, J. L. & The PDP Research Group, Eds. (1986), Parallel Distributed Processing, Cambridge MIT Press.
13. Hopfield, J. J. (1984), "Neurons with Graded Response Have Collective Computational Properties Like Those of Two-State Neurons", Proc. Natl. Acad. Sci. USA, 81, pp. 3088-3092.
14. Kohonen T. (1982), "Self-Organized Formation of Topologically Correct Feature Maps", Biological Cybernetics, 43, pp. 59-69.
15. Haykin S. (1999), Neural Networks – A Comprehensive Foundation, Prentice Hall, 2 nd Ed.